PART VII

Inference When Variables Are Related

CHAPTER

26

Comparing Counts

WHO	Executives of Fortune 400 companies
WHAT	Zodiac birth sign
WHY	Maybe the researcher was a Gemini and naturally curious?

A S ***Activity:*** **Children at Risk.** See how a contingency table helps us understand the different risks to which an incident exposed children.

Does your zodiac sign predict how successful you will be later in life? *Fortune* magazine collected the zodiac signs of 256 heads of the largest 400 companies. The table shows the number of births for each sign.

We can see some variation in the number of births per sign, and there *are* more Pisces, but is that enough to claim that successful people are more likely to be born under some signs than others?

Births	Sign
23	Aries
20	Taurus
18	Gemini
23	Cancer
20	Leo
19	Virgo
18	Libra
21	Scorpio
19	Sagittarius
22	Capricorn
24	Aquarius
29	Pisces

Birth totals by sign for 256 Fortune 400 executives.

Goodness-of-Fit

"All creatures have their determined time for giving birth and carrying fetus, only a man is born all year long, not in determined time, one in the seventh month, the other in the eighth, and so on till the beginning of the eleventh month."

—Aristotle

If births were distributed uniformly across the year, we would expect about 1/12 of them to occur under each sign of the zodiac. That suggests 256/12, or about 21.3 births per sign. How closely do the observed numbers of births per sign fit this simple "null" model?

A hypothesis test to address this question is called a test of **"goodness-of-fit."** The name suggests a certain badness-of-grammar, but it is quite standard. After all, we are asking whether the model that births are uniformly distributed over the signs fits the data good, . . . er, well. Goodness-of-fit involves testing a hypothesis. We have specified a model for the distribution and want to know whether it fits. There is no single parameter to estimate, so a confidence interval wouldn't make much sense.

If the question were about only one astrological sign (for example, "Are executives more likely to be Pisces?"[1]), we could use a one-proportion z-test and ask if

[1] A question actually asked us by someone who was undoubtedly a Pisces.

the true proportion of executives with that sign is equal to 1/12. However, here we have 12 hypothesized proportions, one for each sign. We need a test that considers all of them together and gives an overall idea of whether the observed distribution differs from the hypothesized one.

FOR EXAMPLE Finding expected counts

Birth month may not be related to success as a CEO, but what about on the ball field? It has been proposed by some researchers that children who are the older ones in their class at school naturally perform better in sports and that these children then get more coaching and encouragement. Could that make a difference in who makes it to the professional level in sports?

Baseball is a remarkable sport, in part because so much data are available. We have the birth dates of every one of the 16,804 players who ever played in a major league game. Since the effect we're suspecting may be due to relatively recent policies (and to keep the sample size moderate), we'll consider the birth months of the 1478 major league players born since 1975 and who have played through 2006. We can also look up the national demographic statistics to find what percentage of people were born in each month. Let's test whether the observed distribution of ballplayers' birth months shows just random fluctuations or whether it represents a real deviation from the national pattern.

Month	Ballplayer count	National birth %	Month	Ballplayer count	National birth %
1	137	8%	7	102	9%
2	121	7%	8	165	9%
3	116	8%	9	134	9%
4	121	8%	10	115	9%
5	126	8%	11	105	8%
6	114	8%	12	122	9%
			Total	1478	100%

Question: How can we find the expected counts?

There are 1478 players in this set of data. I'd expect 8% of them to have been born in January, and 1478(0.08) = 118.24. I won't round off, because expected "counts" needn't be integers. Multiplying 1478 by each of the birth percentages gives the expected counts shown in the table.

Month	Expected	Month	Expected
1	118.24	7	133.02
2	103.46	8	133.02
3	118.24	9	133.02
4	118.24	10	133.02
5	118.24	11	118.24
6	118.24	12	133.02

Assumptions and Conditions

These data are organized in tables as we saw in Chapter 3, and the assumptions and conditions reflect that. Rather than having an observation for each individual, we typically work with summary counts in categories. In our example, we don't see the birth signs of each of the 256 executives, only the totals for each sign.

Counted Data Condition: The data must be *counts* for the categories of a categorical variable. This might seem a simplistic, even silly condition. But many kinds of values can be assigned to categories, and it is unfortunately common to find the methods of this chapter applied incorrectly to proportions, percentages, or measurements just because they happen to be organized in a table. So check to be sure the values in each **cell** really are counts.

Independence Assumption

Independence Assumption: The counts in the cells should be independent of each other. The easiest case is when the individuals who are counted in the cells are sampled independently from some population. That's what we'd like to have if we want to draw conclusions about that population. Randomness can arise in

other ways, though. For example, these Fortune 400 executives are not a random sample, but we might still think that their birth dates are randomly distributed throughout the year. If we want to generalize to a large population, we should check the Randomization Condition.

Randomization Condition: The individuals who have been counted should be a random sample from the population of interest.

Sample Size Assumption

We must have enough data for the methods to work. We usually check the following:

Expected Cell Frequency Condition: We should expect to see at least 5 individuals in each cell.

The Expected Cell Frequency Condition sounds like—and is, in fact, quite similar to—the condition that np and nq be at least 10 when we tested proportions. In our astrology example, assuming equal births in each month leads us to expect 21.3 births per month, so the condition is easily met here.

FOR EXAMPLE Checking assumptions and conditions

Recap: Are professional baseball players more likely to be born in some months than in others? We have observed and expected counts for the 1478 players born since 1975.

Question: Are the assumptions and conditions met for performing a goodness-of-fit test?

- ✓ **Counted Data Condition:** I have month-by-month counts of ballplayer births.
- ✓ **Independence Assumption:** These births were independent.
- ✓ **Randomization Condition:** Although they are not a random sample, we can take these players to be representative of players past and future.
- ✓ **Expected Cell Frequency Condition:** The expected counts range from 103.46 to 133.02, all much greater than 5.
- ✓ **10% Condition:** These 1478 players are less than 10% of the population of 16,804 players who have ever played (or will play) major league baseball.

It's okay to use these data for a goodness-of-fit test.

Calculations

We have observed a count in each category from the data, and have an expected count for each category from the hypothesized proportions. Are the differences just natural sampling variability, or are they so large that they indicate something important? It's natural to look at the *differences* between these observed and expected counts, denoted ($Obs - Exp$). We'd like to think about the total of the differences, but just adding them won't work because some differences are positive, others negative. We've been in this predicament before—once when we looked at deviations from the mean and again when we dealt with residuals. In fact, these *are* residuals. They're just the differences between the observed data and the counts given by the (null) model. We handle these residuals in essentially the same way we did in regression: We square them. That gives us positive values and focuses attention on any cells with large differences from what we expected. Because the differences between observed and expected counts generally get larger the more data we have, we also need to get an idea of the *relative* sizes of the differences. To do that, we divide each squared difference by the expected count for that cell.

NOTATION ALERT:

We compare the counts *observed* in each cell with the counts we *expect* to find. The usual notation uses O's and E's or abbreviations such as those we've used here. The method for finding the expected counts depends on the model.

NOTATION ALERT:

The only use of the Greek letter χ in Statistics is to represent this statistic and the associated sampling distribution. This is another violation of our "rule" that Greek letters represent population parameters. Here we are using a Greek letter simply to name a family of distribution models and a statistic.

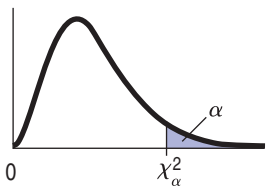
The test statistic, called the **chi-square** (or chi-squared) **statistic,** is found by adding up the sum of the squares of the deviations between the observed and expected counts divided by the expected counts:

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}.$$

The chi-square statistic is denoted χ^2, where χ is the Greek letter chi (pronounced "ky" as in "sky"). It refers to a family of sampling distribution models we have not seen before called (remarkably enough) the **chi-square models.**

This family of models, like the Student's t-models, differ only in the number of degrees of freedom. The number of degrees of freedom for a goodness-of-fit test is $n - 1$. Here, however, n is *not* the sample size, but instead is the number of categories. For the zodiac example, we have 12 signs, so our χ^2 statistic has 11 degrees of freedom.

One-Sided or Two-Sided?



TI-*nspire*

The χ^2 Models. See what a χ^2 model looks like, and watch it change as you change the degrees as freedom.

The chi-square statistic is used only for testing hypotheses, not for constructing confidence intervals. If the observed counts don't match the expected, the statistic will be large. It can't be "too small." That would just mean that our model *really* fit the data well. So the chi-square test is always one-sided. If the calculated statistic value is large enough, we'll reject the null hypothesis. What could be simpler?

Even though its mechanics work like a one-sided test, the interpretation of a chi-square test is in some sense *many*-sided. With more than two proportions, there are many ways the null hypothesis can be wrong. By squaring the differences, we made all the deviations positive, whether our observed counts were higher or lower than expected. There's no direction to the rejection of the null model. All we know is that it doesn't fit.

FOR EXAMPLE — Doing a goodness-of-fit test

Recap: We're looking at data on the birth months of major league baseball players. We've checked the assumptions and conditions for performing a χ^2 test.

Questions: What are the hypotheses, and what does the test show?

H_0: The distribution of birth months for major league ballplayers is the same as that for the general population.

H_A: The distribution of birth months for major league ballplayers differs from that of the rest of the population.

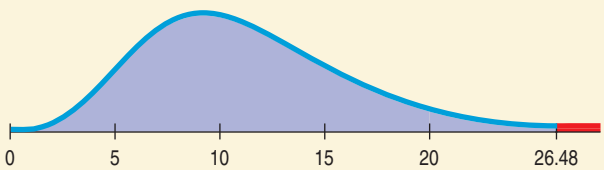
$$df = 12 - 1 = 11$$

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

$$= \frac{(137 - 118.24)^2}{118.24} + \frac{(121 - 103.46)^2}{103.46} + \cdots$$

$$= 26.48 \text{ (by technology)}$$

$$P\text{-value} = P(\chi^2_{11} \geq 26.48) = 0.0055 \text{ (by technology)}$$



Because of the small P-value, I reject H_0; there's evidence that birth months of major league ballplayers have a different distribution from the rest of us.

STEP-BY-STEP EXAMPLE — A Chi-Square Test for Goodness-of-Fit

We have counts of 256 executives in 12 zodiac sign categories. The natural null hypothesis is that birth dates of executives are divided equally among all the zodiac signs. The test statistic looks at how closely the observed data match this idealized situation.

Question: Are zodiac signs of CEOs distributed uniformly?

Plan State what you want to know.

Identify the variables and check the W's.

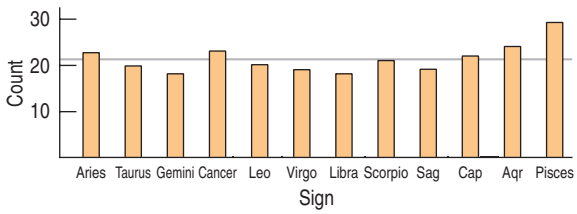
I want to know whether births of successful people are uniformly distributed across the signs of the zodiac. I have counts of 256 Fortune 400 executives, categorized by their birth sign.

Hypotheses State the null and alternative hypotheses. For χ^2 tests, it's usually easier to do that in words than in symbols.

H_O: Births are uniformly distributed over zodiac signs.[2]

H_A: Births are not uniformly distributed over zodiac signs.

Model Make a picture. The null hypothesis is that the frequencies are equal, so a bar chart (with a line at the hypothesized "equal" value) is a good display.



The bar chart shows some variation from sign to sign, and Pisces is the most frequent. But it is hard to tell whether the variation is more than I'd expect from random variation.

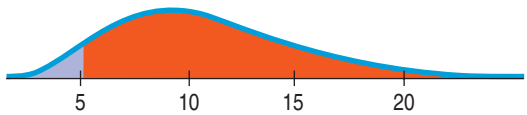
Think about the assumptions and check the conditions.

- ✓ **Counted Data Condition:** I have counts of the number of executives in 12 categories.
- ✓ **Independence Assumption:** The birth dates of executives should be independent of each other.
- ✓ **Randomization Condition:** This is a convenience sample of executives, but there's no reason to suspect bias.
- ✓ **Expected Cell Frequency Condition:** The null hypothesis expects that 1/12 of the 256 births, or 21.333, should occur in each sign. These expected values are all at least 5, so the condition is satisfied.

[2] It may seem that we have broken our rule of thumb that null hypotheses should specify parameter values. If you want to get formal about it, the null hypothesis is that

$$p_{\text{Aries}} = p_{\text{Taurus}} = \cdots = p_{\text{Pisces}}.$$

That is, we hypothesize that the true proportions of births of CEOs under each sign are equal. The role of the null hypothesis is to specify the model so that we can compute the test statistic. That's what this one does.

Specify the sampling distribution model. Name the test you will use.	*The conditions are satisfied, so I'll use a* χ^2 *model with* $12 - 1 = 11$ *degrees of freedom and do a* ***chi-square goodness-of-fit test.***
Mechanics Each cell contributes an $\frac{(Obs - Exp)^2}{Exp}$ value to the chi-square sum. We add up these components for each zodiac sign. If you do it by hand, it can be helpful to arrange the calculation in a table. We show that after this Step-By-Step. The P-value is the area in the upper tail of the χ^2 model above the computed χ^2 value. The χ^2 models are skewed to the high end, and change shape depending on the degrees of freedom. The P-value considers only the right tail. Large χ^2 statistic values correspond to small P-values, which lead us to reject the null hypothesis.	*The expected value for each zodiac sign is 21.333.* $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(23 - 21.333)^2}{21.333} + \frac{(20 - 21.333)^2}{21.333} + \cdots$ $= 5.094$ *for all 12 signs.*  $P\text{-value} = P(\chi^2 > 5.094) = 0.926$
Conclusion Link the P-value to your decision. Remember to state your conclusion in terms of what the data mean, rather than just making a statement about the distribution of counts.	*The P-value of 0.926 says that if the zodiac signs of executives were in fact distributed uniformly, an observed chi-square value of 5.09 or higher would occur about 93% of the time. This certainly isn't unusual, so I fail to reject the null hypothesis, and conclude that these data show virtually no evidence of nonuniform distribution of zodiac signs among executives.*

The Chi-Square Calculation

***Activity:* Calculating Standardized Residuals.** Women were at risk, too. Standardized residuals help us understand the relative risks.

Let's make the chi-square procedure very clear. Here are the steps:

1. **Find the expected values.** These come from the null hypothesis model. Every model gives a hypothesized proportion for each cell. The expected value is the product of the total number of observations times this proportion.

   For our example, the null model hypothesizes *equal* proportions. With 12 signs, 1/12 of the 256 executives should be in each category. The expected number for each sign is 21.333.
2. **Compute the residuals.** Once you have expected values for each cell, find the residuals, *Observed* − *Expected*.
3. **Square the residuals.**
4. **Compute the components.** Now find the component, $\frac{(Observed - Expected)^2}{Expected}$, for each cell.

A S ***Activity:*** **The Chi-Square Test.** This animation completes the calculation of the chi-square statistic and the hypothesis test based on it.

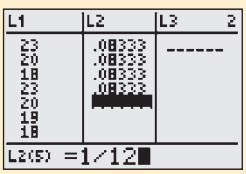
5. **Find the sum of the components.** That's the chi-square statistic.
6. **Find the degrees of freedom.** It's equal to the number of cells minus one. For the zodiac signs, that's 12 − 1 = 11 degrees of freedom.
7. **Test the hypothesis.** Large chi-square values mean lots of deviation from the hypothesized model, so they give small P-values. Look up the critical value from a table of chi-square values, or use technology to find the P-value directly.

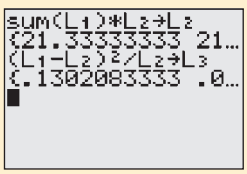
The steps of the chi-square calculations are often laid out in tables. Use one row for each category, and columns for observed counts, expected counts, residuals, squared residuals, and the contributions to the chi-square total like this:

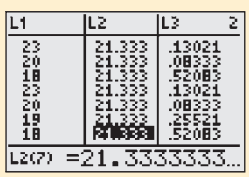
Sign	Observed	Expected	*Residual* = (*Obs* − *Exp*)	$(Obs - Exp)^2$	Component = $\frac{(Obs - Exp)^2}{Exp}$
Aries	23	21.333	1.667	2.778889	0.130262
Taurus	20	21.333	−1.333	1.776889	0.083293
Gemini	18	21.333	−3.333	11.108889	0.520737
Cancer	23	21.333	1.667	2.778889	0.130262
Leo	20	21.333	−1.333	1.776889	0.083293
Virgo	19	21.333	−2.333	5.442889	0.255139
Libra	18	21.333	−3.333	11.108889	0.520737
Scorpio	21	21.333	−0.333	0.110889	0.005198
Sagittarius	19	21.333	−2.333	5.442889	0.255139
Capricorn	22	21.333	0.667	0.444889	0.020854
Aquarius	24	21.333	2.667	7.112889	0.333422
Pisces	29	21.333	7.667	58.782889	2.755491
					Σ = 5.094

TI Tips

Testing goodness of fit



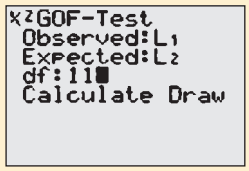


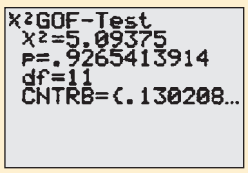


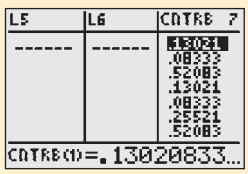
As always, the TI makes doing the mechanics of a goodness-of-fit test pretty easy, but it does take a little work to set it up. Let's use the zodiac data to run through the steps for a χ^2 GOF-Test.

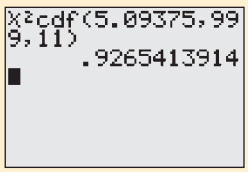
- Enter the counts of executives born under each star sign in L1.

  Those counts were: 23 20 18 23 20 19 18 21 19 22 24 29
- Enter the expected percentages (or fractions, here 1/12) in L2. In this example they are all the same value, but that's not always the case.
- Convert the expected percentages to expected counts by multiplying each of them by the total number of observations. We use the calculator's summation command in the LIST MATH menu to find the total count for the data summarized in L1 and then multiply that sum by the percentages stored in L2 to produce the expected counts. The command is sum(L1)*L2→L2. (We don't ever need the percentages again, so we can replace them by storing the expected counts in L2 instead.)
- Choose D: χ^2 GOF-Test from the STATS TESTS menu.
- Specify the lists where you stored the observed and expected counts, and enter the number of degrees of freedom, here 11.









- Ready, set, **Calculate**. . .
- . . . and there are the calculated value of χ^2 and your P-value.
- Notice, too, there's a list of values called **CNTRB**. You can scroll across them, or use **LIST NAMES** to display them as a data list (as seen on the next page). Those are the cell-by-cell components of the χ^2 calculation. We aren't very interested in them this time, because our data failed to provide evidence that the zodiac sign mattered. However, in a situation where we rejected the null hypothesis, we'd want to look at the components to see where the biggest effects occurred. You'll read more about doing that later in this chapter.

By hand?

If there are only a few cells, you may find that it's just as easy to write out the formula and then simply use the calculator to help you with the arithmetic. After you have found $\chi^2 = 5.09375$ you can use your TI to find the P-value, the probability of observing a χ^2 value at least as high as the one you calculated from your data. As you probably expect, that process is akin to **normalcdf** and **tcdf**. You'll find what you need in the **DISTR** menu at **8:χ^2cdf**. Just specify the left and right boundaries and the number of degrees of freedom.

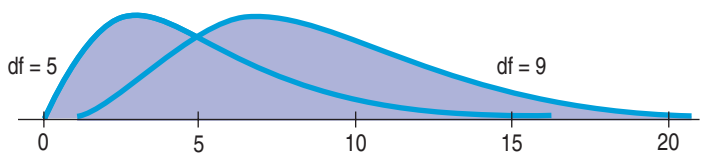
- Enter χ^2**cdf(5.09375,999,11)**, as shown. (Why 999? Unlike t and z, chi-square values can get pretty big, especially when there are many cells. You may need to go a long way to the right to get to where the curve's tail becomes essentially meaningless. You can see what we mean by looking at Table C, showing chi-square values.)

And there's the P-value, a whopping 0.93! There's nothing at all unusual about these data. (So much for the zodiac's predictive power.)

How big is big? When we calculated χ^2 for the zodiac sign example, we got 5.094. That value would have been big for z or t, leading us to reject the null hypothesis. Not here. Were you surprised that $\chi^2 = 5.094$ had a huge P-value of 0.926? What *is* big for a χ^2 statistic, anyway?

Think about how χ^2 is calculated. In every cell, any deviation from the expected count contributes to the sum. Large deviations generally contribute more, but if there are a lot of cells, even small deviations can add up, making the χ^2 value larger. So the more cells there are, the higher the value of χ^2 has to get before it becomes noteworthy. For χ^2, then, the decision about how big is big depends on the number of degrees of freedom.

Unlike the Normal and t families, χ^2 models are skewed. Curves in the χ^2 family change both shape and center as the number of degrees of freedom grows. Here, for example, are the χ^2 curves for 5 and 9 degrees of freedom.



Notice that the value $\chi^2 = 10$ might seem somewhat extreme when there are 5 degrees of freedom, but appears to be rather ordinary for 9 degrees of freedom. Here are two simple facts to help you think about χ^2 models:

A S ***Lesson:*** **The Chi-Square Family of Curves.** (Not an activity like the others, but there's no better way to see how χ^2 changes with more df.) Click on the Lesson Book's Resources tab and open the chi-square table. Watch the curve at the top as you click on a row and scroll down the degrees-of freedom column.

- The mode is at $\chi^2 = df - 2$. (Look back at the curves; their peaks are at 3 and 7, see?)
- The expected value (mean) of a χ^2 model is its number of degrees of freedom. That's a bit to the right of the mode—as we would expect for a skewed distribution.

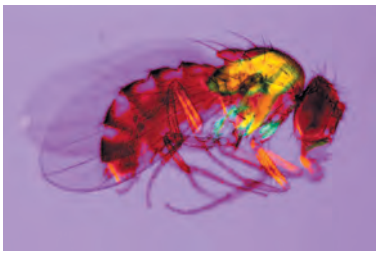
Our test for zodiac birthdays had 11 df, so the relevant χ^2 curve peaks at 9 and has a mean of 11. Knowing that, we might have easily guessed that the calculated χ^2 value of 5.094 wasn't going to be significant.

But I Believe the Model . . .

Goodness-of-fit tests are likely to be performed by people who have a theory of what the proportions *should* be in each category and who believe their theory to be true. Unfortunately, the only *null* hypothesis available for a goodness-of-fit test is that the theory is true. And as we know, the hypothesis-testing procedure allows us only to *reject* the null or *fail to reject* it. We can never confirm that a theory is in fact true, which is often what people want to do.

Unfortunately, they're stuck. At best, we can point out that the data are consistent with the proposed theory. But this doesn't *prove* the theory. The data *could* be consistent with the model even if the theory were wrong. In that case, we fail to reject the null hypothesis but can't conclude anything for sure about whether the theory is true.

And we can't fix the problem by turning things around. Suppose we try to make our favored hypothesis the alternative. Then it is impossible to pick a single null. For example, suppose, as a doubter of astrology, you want to prove that the distribution of executive births is uniform. If you choose uniform as the null hypothesis, you can only *fail* to reject it. So you'd like uniformity to be your alternative hypothesis. Which particular violation of equally distributed births would you choose as your null? The problem is that the model can be wrong in many, many ways. There's no way to frame a null hypothesis the other way around. There's just no way to prove that a favored model is true.



Why can't we prove the null? A biologist wanted to show that her inheritance theory about fruit flies is valid. It says that 10% of the flies should be type 1, 70% type 2, and 20% type 3. After her students collected data on 100 flies, she did a goodness-of-fit test and found a P-value of 0.07. She started celebrating, since her null hypothesis wasn't rejected—that is, until her students collected data on 100 more flies. With 200 flies, the P-value dropped to 0.02. Although she knew the answer was probably no, she asked the statistician somewhat hopefully if she could just ignore half the data and stick with the original 100. By this reasoning we could always "prove the null" just by not collecting much data. With only a little data, the chances are good that they'll be consistent with almost anything. But they also have little chance of disproving anything either. In this case, the test has no power. Don't let yourself be lured into this scientist's reasoning. With data, more is always better. But you can't ever prove that your null hypothesis is true.

Comparing Observed Distributions

Many colleges survey graduating classes to determine the plans of the graduates. We might wonder whether the plans of students are the same at different colleges. Here's a **two-way table** for Class of 2006 graduates from several colleges at one university. Each **cell** of the table shows how many students from a particular college made a certain choice.

WHO Graduates from 4 colleges at an upstate New York university
WHAT Post-graduation activities
WHEN 2006
WHY Survey for general information

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
Employed	379	305	243	125	**1052**
Grad School	186	238	202	96	**722**
Other	104	123	37	58	**322**
Total	**669**	**666**	**482**	**279**	**2096**

Table 26.1 Post-graduation activities of the class of 2006 for several colleges of a large university.

Because class sizes are so different, we see differences better by examining the proportions for each class rather than the counts:

A S ***Video:* The Incident.** You may have guessed which famous incident put women and children at risk. Here you can view the story complete with rare film footage.

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
Employed	56.7%	45.8%	50.4%	44.8%	**50.2**
Grad School	27.8	35.7	41.9	34.4	**34.4**
Other	15.5	18.5	7.7	20.8	**15.4**
Total	**100**	**100**	**100**	**100**	**100**

Table 26.2 Activities of graduates as a percentage of respondents from each college.

We already know how to test whether *two* proportions are the same. For example, we could use a two-proportion z-test to see whether the proportion of students choosing graduate school is the same for Agriculture students as for Engineering students. But now we have more than two groups. We want to test whether the students' choices are the same across all four colleges. The z-test for two proportions generalizes to a **chi-square test of homogeneity.**

Chi-square again? It turns out that the mechanics of this test are *identical* to the chi-square test for goodness-of-fit that we just saw. (How similar can you get?) Why a different name, then? The goodness-of-fit test compared counts with a theoretical model. But here we're asking whether choices are the same among different groups, so we find the expected counts for each category directly from the data. As a result, we count the degrees of freedom slightly differently as well.

The term "homogeneity" means that things are the same. Here, we ask whether the post-graduation choices made by students are the *same* for these four colleges. The homogeneity test comes with a built-in null hypothesis: We hypothesize that the distribution does not change from group to group. The test looks for differences large enough to step beyond what we might expect from random sample-to-sample variation. It can reveal a large deviation in a single category or small, but persistent, differences over all the categories—or anything in between.

Assumptions and Conditions

The assumptions and conditions are the same as for the chi-square test for goodness-of-fit. The **Counted Data Condition** says that these data must be counts. You can't do a test of homogeneity on proportions, so we have to work with the counts of graduates given in the first table. Also, you can't do a chi-square test on measurements. For example, if we had recorded GPAs for these same groups,

we wouldn't be able to determine whether the mean GPAs were different using this test.[3]

Often when we test for homogeneity, we aren't interested in some larger population, so we don't really need a random sample. (We would need one if we wanted to draw a more general conclusion—say, about the choices made by all members of the Class of '06.) Don't we need *some* randomness, though? Fortunately, the null hypothesis can be thought of as a model in which the counts in the table are distributed as if each student chose a plan randomly according to the overall proportions of the choices, regardless of the student's class. As long as we don't want to generalize, we don't have to check the **Randomization Condition** or the **10% Condition.**

We still must be sure we have enough data for this method to work. The **Expected Cell Frequency Condition** says that the expected count in each cell must be at least 5. We'll confirm that as we do the calculations.

Calculations

The null hypothesis says that the proportions of graduates choosing each alternative should be the same for all four colleges, so we can estimate those overall proportions by pooling our data from the four colleges together. Within each college, the expected proportion for each choice is just the overall proportion of all students making that choice. The expected counts are those proportions applied to the number of students in each graduating class.

For example, overall, 1052, or about 50.2%, of the 2096 students who responded to the survey were employed. If the distributions are homogeneous (as the null hypothesis asserts), then 50.2% of the 669 Agriculture school graduates (or about 335.8 students) should be employed. Similarly, 50.2% of the 482 Engineering grads (or about 241.96) should be employed.

Working in this way, we (or, more likely, the computer) can fill in expected values for each cell. Because these are theoretical values, they don't have to be integers. The expected values look like this:

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
Employed	**335.777**	334.271	241.920	140.032	**1052**
Grad School	230.448	229.414	166.032	96.106	**722**
Other	102.776	102.315	74.048	42.862	**322**
Total	**669**	**666**	**482**	**279**	**2096**

Table 26.3 Expected values for the '06 graduates.

Now check the **Expected Cell Frequency Condition.** Indeed, there are at least 5 individuals expected in each cell.

Following the pattern of the goodness-of-fit test, we compute the component for each cell of the table. For the highlighted cell, employed students graduating from the Ag school, that's

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(379 - 335.777)^2}{335.777} = 5.564$$

[3] To do that, you'd use a method called Analysis of Variance, discussed in a supplementary chapter on the DVD and in ActivStats.

Summing these components across all cells gives

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} = 54.51$$

How about the degrees of freedom? We don't really need to calculate all the expected values in the table. We know there is a total of 1052 employed students, so once we find the expected values for three of the colleges, we can determine the expected number for the fourth by just subtracting. Similarly, we know how many students graduated from each college, so after filling in three rows, we can find the expected values for the remaining row by subtracting. To fill out the table, we need to know the counts in only $R - 1$ rows and $C - 1$ columns. So the table has $(R - 1)(C - 1)$ degrees of freedom.

In our example, we need to calculate only 2 choices in each column and counts for 3 of the 4 colleges, for a total of $2 \times 3 = 6$ degrees of freedom. We'll need the degrees of freedom to find a P-value for the chi-square statistic.

NOTATION ALERT:

For a contingency table, R represents the number of rows and C the number of columns.

STEP-BY-STEP EXAMPLE — A Chi-Square Test for Homogeneity

We have reports from four colleges on the post-graduation activities of their 2006 graduating classes.

Question: Are students' choices of post-graduation activities the same across all the colleges?

THINK

Plan State what you want to know.

Identify the variables and check the W's.

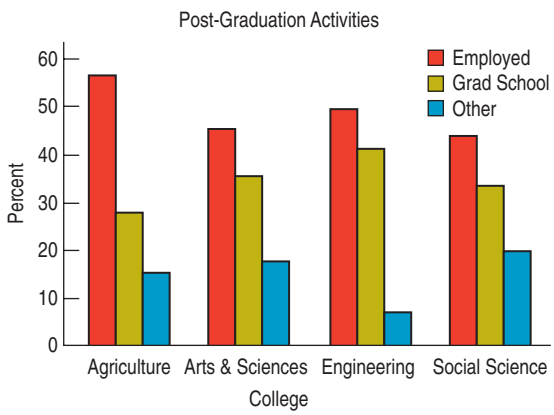
I want to know whether post-graduation choices are the same for students from each of four colleges. I have a table of counts classifying each college's Class of 2006 respondents according to their activities.

Hypotheses State the null and alternative hypotheses.

H_O: Students' post-graduation activities are distributed in the same way for all four colleges.

H_A: Students' plans do not have the same distribution.

Model Make a picture: A side-by-side bar chart shows the four distributions of post-graduation activities. Plot column percents to remove the effect of class size differences. A split bar chart would also be an appropriate choice.



A side-by-side bar chart shows how the distributions of choices differ across the four colleges.

Think about the assumptions and check the conditions.

- ✔ **Counted Data Condition:** I have counts of the number of students in categories.
- ✔ **Independence Assumption:** Student plans should be largely independent of each other. The occasional friends who decide to join Teach for America together or couples who make grad school decisions together are too rare to affect this analysis.
- ✔ **Randomization Condition:** I don't want to draw inferences to other colleges or other classes, so there is no need to check for a random sample.
- ✔ **Expected Cell Frequency Condition:** The expected values (shown below) are all at least 5.

State the sampling distribution model and name the test you will use.

The conditions seem to be met, so I can use a χ^2 model with $(3 - 1) \times (4 - 1) = 6$ degrees of freedom and do a **chi-square test of homogeneity.**

SHOW

Mechanics Show the expected counts for each cell of the data table. You could make separate tables for the observed and expected counts, or put both counts in each cell as shown here. While observed counts must be whole numbers, expected counts rarely are—don't be tempted to round those off.

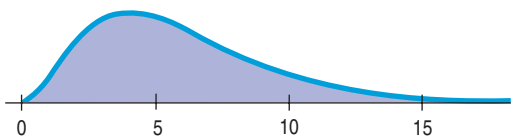
	Ag	A&S	Eng	Soc Sci
Empl.	379 / 335.777	305 / 334.271	243 / 241.920	125 / 140.032
Grad sch.	186 / 230.448	238 / 229.414	202 / 166.032	96 / 96.106
Other	104 / 102.776	123 / 102.315	37 / 74.048	58 / 42.862

Calculate χ^2.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}$$
$$= \frac{(379 - 335.777)^2}{335.777} + \cdots$$
$$= 54.52$$

The shape of a χ^2 model depends on the degrees of freedom. A χ^2 model with 6 df is skewed to the high end.

The P-value considers only the right tail. Here, the calculated value of the χ^2 statistic is off the scale, so the P-value is quite small.



$$P\text{-value} = P(\chi^2 > 54.52) < 0.0001$$

TELL

Conclusion State your conclusion in the context of the data. You should specifically talk about whether the distributions for the groups appear to be different.

The P-value is very small, so I reject the null hypothesis and conclude that there's evidence that the post-graduation activities of students from these four colleges don't have the same distribution.

If you find that simply rejecting the hypothesis of homogeneity is a bit unsatisfying, you're in good company. Ok, so the post-graduation plans are different. What we'd really like to know is what the differences are, where they're the greatest, and where they're smallest. The test for homogeneity doesn't answer these interesting questions, but it does provide some evidence that can help us.

Examining the Residuals

Whenever we reject the null hypothesis, it's a good idea to examine residuals. (We don't need to do that when we fail to reject because when the χ^2 value is small, all of its components must have been small.) For chi-square tests, we want to compare residuals for cells that may have very different counts. So we're better off standardizing the residuals. We know the mean residual is zero,[4] but we need to know each residual's standard deviation. When we tested proportions, we saw a link between the expected proportion and its standard deviation. For counts, there's a similar link. To standardize a cell's residual, we just divide by the square root of its expected value:

$$c = \frac{(Obs - Exp)}{\sqrt{Exp}}.$$

Notice that these **standardized residuals** are just the square roots of the **components** we calculated for each cell, and their sign indicates whether we observed more cases than we expected, or fewer.

The standardized residuals give us a chance to think about the underlying patterns and to consider the ways in which the distribution of post-graduation plans may differ from college to college. Now that we've subtracted the mean (zero) and divided by their standard deviations, these are z-scores. If the null hypothesis were true, we could even appeal to the Central Limit Theorem, think of the Normal model, and use the 68–95–99.7 Rule to judge how extraordinary the large ones are.

Here are the standardized residuals for the Class of '06 data:

	Ag	A&S	Eng	Soc Sci
Employed	2.359	−1.601	0.069	−1.270
Grad School	−2.928	0.567	2.791	−0.011
Other	0.121	2.045	−4.305	2.312

Table 26.4
Standardized residuals can help show how the table differs from the null hypothesis pattern.

The column for Engineering students immediately attracts our attention. It holds both the largest positive and the largest negative standardized residuals. It looks like Engineering college graduates are more likely to go on to graduate work and very unlikely to take time off for "volunteering and travel, among other activities" (as the "Other" category is explained). By contrast, Ag school graduates seem to be readily employed and less likely to pursue graduate work immediately after college.

[4] Residual = observed − expected. Because the total of the expected values is set to be the same as the observed total, the residuals must sum to zero.

FOR EXAMPLE Looking at χ^2 residuals

Recap: Some people suggest that school children who are the older ones in their class naturally perform better in sports and therefore get more coaching and encouragement. To see if there's any evidence for this, we looked at major league baseball players born since 1975. A goodness-of-fit test found their birth months to have a distribution that's significantly different from the rest of us. The table shows the standardized residuals.

Month	Residual	Month	Residual
1	1.73	7	−2.69
2	1.72	8	2.77
3	−0.21	9	0.08
4	0.25	10	−1.56
5	0.71	11	−1.22
6	−0.39	12	−0.96

Question: What's different about the distribution of birth months among major league ballplayers?

It appears that, compared to the general population, fewer ballplayers than expected were born in July and more than expected in August. Either month would make them the younger kids in their grades in school, so these data don't offer support for the conjecture that being older is an advantage in terms of a career as a pro athlete.

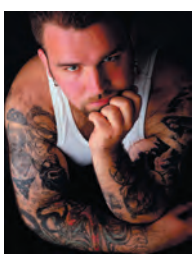
JUST CHECKING

Tiny black potato flea beetles can damage potato plants in a vegetable garden. These pests chew holes in the leaves, causing the plants to wither or die. They can be killed with an insecticide, but a canola oil spray has been suggested as a non-chemical "natural" method of controlling the beetles. To conduct an experiment to test the effectiveness of the natural spray, we gather 500 beetles and place them in three Plexiglas® containers. Two hundred beetles go in the first container, where we spray them with the canola oil mixture. Another 200 beetles go in the second container; we spray them with the insecticide. The remaining 100 beetles in the last container serve as a control group; we simply spray them with water. Then we wait 6 hours and count the number of surviving beetles in each container.

1. Why do we need the control group?
2. What would our null hypothesis be?
3. After the experiment is over, we could summarize the results in a table as shown. How many degrees of freedom does our χ^2 test have?
4. Suppose that, all together, 125 beetles survived. (That's the first-row total.) What's the expected count in the first cell—survivors among those sprayed with the natural spray?
5. If it turns out that only 40 of the beetles in the first container survived, what's the calculated component of χ^2 for that cell?
6. If the total calculated value of χ^2 for this table turns out to be around 10, would you expect the P-value of our test to be large or small? Explain.

	Natural spray	Insecticide	Water	Total
Survived				
Died				
Total	200	200	100	500

Independence



A study from the University of Texas Southwestern Medical Center examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. Hepatitis C causes about 10,000 deaths each year in the United States, but often lies undetected for years after infection.

The data from this study can be summarized in a two-way table, as follows:

WHO	Patients being treated for non–blood-related disorders
WHAT	Tattoo status and hepatitis C status
WHEN	1991, 1992
WHERE	Texas

	Hepatitis C	No Hepatitis C	Total
Tattoo, parlor	17	35	**52**
Tattoo, elsewhere	8	53	**61**
None	22	491	**513**
Total	**47**	**579**	**626**

Table 26.5
Counts of patients classified by their hepatitis C test status according to whether they had a tattoo from a tattoo parlor or from another source, or had no tattoo.

These data differ from the kinds of data we've considered before in this chapter because they categorize subjects from a single group on two categorical variables rather than on only one. The categorical variables here are *Hepatitis C Status* ("Hepatitis C" or "No Hepatitis C") and *Tattoo Status* ("Parlor," "Elsewhere," "None"). We've seen counts classified by two categorical variables displayed like this in Chapter 3, so we know such tables are called contingency tables. **Contingency tables** categorize counts on two (or more) variables so that we can see whether the distribution of counts on one variable is contingent on the other.

The natural question to ask of these data is whether the chance of having hepatitis C is *independent* of tattoo status. Recall that for events **A** and **B** to be independent $P(\mathbf{A})$ must equal $P(\mathbf{A} \mid \mathbf{B})$. Here, this means the probability that a randomly selected patient has hepatitis C should not change when we learn the patient's tattoo status. We examined the question of independence in just this way back in Chapter 15, but we lacked a way to test it. The rules for independent events are much too precise and absolute to work well with real data. A **chi-square test for independence** is called for here.

If *Hepatitis Status* is independent of tattoos, we'd expect the proportion of people testing positive for hepatitis to be the same for the three levels of *Tattoo Status*. This sounds a lot like the test of homogeneity. In fact, the mechanics of the calculation are identical.

The difference is that now we have two categorical variables measured on a single population. For the homogeneity test, we had a single categorical variable measured independently on two or more populations. But now we ask a different question: "Are the variables independent?" rather than "Are the groups homogeneous?" These are subtle differences, but they are important when we state hypotheses and draw conclusions.

A S ***Activity:* Independence and Chi-Square.** This unusual simulation shows how independence arises (and fails) in contingency tables.

The only difference between the test for homogeneity and the test for independence is in what you . . .

FOR EXAMPLE Which χ^2 test?

Many states and localities now collect data on traffic stops regarding the race of the driver. The initial concern was that Black drivers were being stopped more often (the "crime" ironically called "Driving While Black"). With more data in hand, attention has turned to other issues. For example, data from 2533 traffic stops in Cincinnati[5] report the race of the driver (Black, White, or Other) and whether the traffic stop resulted in a search of the vehicle.

Question: Which test would be appropriate to examine whether race is a factor in vehicle searches? What are the hypotheses?

		Race			
		Black	White	Other	Total
Search	No	787	594	27	**1408**
	Yes	813	293	19	**1125**
	Total	**1600**	**887**	**46**	**2533**

(*continued*)

[5] John E. Eck, Lin Liu, and Lisa Growette Bostaph, Police Vehicle Stops in Cincinnati, Oct. 1, 2003, available at http://www.cincinnati-oh.gov. Data for other localities can be found by searching from http://www.racialprofilinganalysis.neu.edu.

For Example (*continued*)

These data represent one group of traffic stops in Cincinnati, categorized on two variables, Race and Search. I'll do a chi-square test of independence.

H_O: Whether or not police search a vehicle is independent of the race of the driver.

H_A: Decisions to search vehicles are not independent of the driver's race.

Assumptions and Conditions

A S ***Activity:*** **Chi-Square Tables.** Work with *ActivStats*' interactive chi-square table to perform a hypothesis test.

Of course, we still need counts and enough data so that the expected values are at least 5 in each cell.

If we're interested in the independence of variables, we usually want to generalize from the data to some population. In that case, we'll need to check that the data are a representative random sample from, and fewer than 10% of, that population.

STEP-BY-STEP EXAMPLE A Chi-Square Test for Independence

We have counts of 626 individuals categorized according to their "tattoo status" and their "hepatitis status."

Question: Are tattoo status and hepatitis status independent?

Plan State what you want to know.

Identify the variables and check the W's.

I want to know whether the categorical variables *Tattoo Status* and *Hepatitis Status* are statistically independent. I have a contingency table of 626 Texas patients with an unrelated disease.

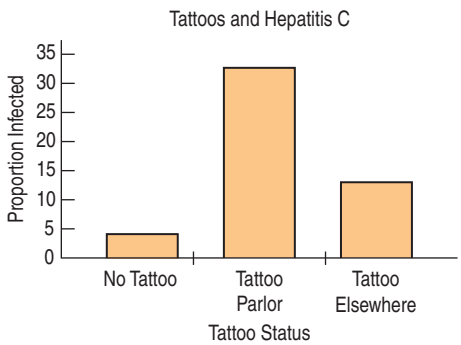
Hypotheses State the null and alternative hypotheses.

We perform a test of independence when we suspect the variables may not be independent. We are on the familiar ground of making a claim (in this case, that knowing *Tattoo Status* will change probabilities for *Hepatitis C Status*) and testing the null hypothesis that it is *not* true.

H_O: *Tattoo Status and Hepatitis Status are independent.*[6]

H_A: *Tattoo Status and Hepatitis Status are not independent.*

Model Make a picture. Because these are only two categories—Hepatitis C and No Hepatitis C—a simple bar chart of the distribution of tattoo sources for Hep C patients shows all the information.



The bar chart suggests strong differences in Hepatitis C risk based on tattoo status.

[6] Once again, parameters are hard to express. The hypothesis of independence itself tells us how to find expected values for each cell of the contingency table. That's all we need.

Think about the assumptions and check the conditions.

- ✓ **Counted Data Condition:** I have counts of individuals categorized on two variables.
- ✓ **Independence Assumption:** The people in this study are likely to be independent of each other.
- ✓ **Randomization Condition:** These data are from a retrospective study of patients being treated for something unrelated to hepatitis. Although they are not an SRS, they were selected to avoid biases.
- ✓ **10% Condition:** These 626 patients are far fewer than 10% of all those with tattoos or hepatitis.
- ✗ **Expected Cell Frequency Condition:** The expected values do not meet the condition that all are at least 5.

This table shows both the observed and expected counts for each cell. The expected counts are calculated exactly as they were for a test of homogeneity; in the first cell, for example, we expect $\frac{52}{626}$ (that's 8.3%) of 47.

	Hepatitis C	No Hepatitis C	Total
Tattoo, parlor	17 3.904	35 48.096	52
Tattoo, elsewhere	8 4.580	53 56.420	61
None	22 38.516	491 474.484	513
Total	47	579	626

Warning: Be wary of proceeding when there are small expected counts, If we see expected counts that fall far short of 5, or if many cells violate the condition, we should not use χ^2. (We will soon discuss ways you can fix the problem.) If you do continue, always check the residuals to be sure those cells did not have a major influence on your result.

Although the Expected Cell Frequency Condition is not satisfied, the values are close to 5. I'll go ahead, but I'll check the residuals carefully. I'll use a χ^2 model with $(3 - 1) \times (2 - 1) = 2$ df and do a **chi-square test of independence.**

Specify the model.

Name the test you will use.

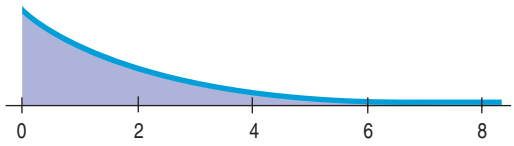
SHOW

Mechanics Calculate χ^2.

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}$$
$$= \frac{(17 - 3.094)^2}{3.094} + \cdots = 57.91$$

The shape of a chi-square model depends on its degrees of freedom. With 2 df, the model looks quite different, as you can

see here. We still care only about the right tail.



P-Value = $P(\chi^2 > 57.91) < 0.0001$

TELL

Conclusion Link the P-value to your decision. State your conclusion about the independence of the two variables.

(We should be wary of this conclusion because of the small expected counts. A complete solution must include the additional analysis, recalculation, and final conclusion discussed in the following section.)

The P-value is very small, so I reject the null hypothesis and conclude that Hepatitis Status is not independent of Tattoo Status. Because the Expected Cell Frequency Condition was violated, I need to check that the two cells with small expected counts did not influence this result too greatly.

FOR EXAMPLE Chi-square mechanics

Recap: We have data that allow us to investigate whether police searches of vehicles they stop are independent of the driver's race.

Questions: What are the degrees of freedom for this test? What is the expected frequency of searches for the Black drivers who were stopped? What's that cell's component in the χ^2 computation? And how is the standardized residual for that cell computed?

		Race			
		Black	**White**	**Other**	**Total**
Search	**No**	787	594	27	**1408**
	Yes	813	293	19	**1125**
	Total	**1600**	**887**	**46**	**2533**

This is a 2 × 3 contingency table, so $df = (2 - 1)(3 - 1) = 2$.

Overall, 1125 of 2533 vehicles were searched. If searches are conducted independent of race, then I'd expect $\frac{1125}{2533}$ *of the 1600 Black drivers to have been searched:* $\frac{1125}{2533} \times 1600 \approx 710.62$.

That cell's term in the χ^2 *calculation is* $\frac{(Obs - Exp)^2}{Exp} = \frac{(813 - 710.62)^2}{710.62} = 14.75$

The standardized residual for that cell is $\frac{Obs - Exp}{\sqrt{Exp}} = \frac{813 - 710.62}{\sqrt{710.62}} = 3.84$

Examine the Residuals

Each cell of the contingency table contributes a term to the chi-square sum. As we did earlier, we should examine the residuals because we have rejected the null hypothesis. In this instance, we have an additional concern that the cells with small expected frequencies not be the ones that make the chi-square statistic large.

Our interest in the data arises from the potential for improving public health. If patients with tattoos are more likely to test positive for hepatitis C, perhaps physicians should be advised to suggest blood tests for such patients.

The standardized residuals look like this:

	Hepatitis C	No Hepatitis C
Tattoo, parlor	6.628	−1.888
Tattoo, elsewhere	1.598	−0.455
None	−2.661	0.758

Table 26.6
Standardized residuals for the hepatitis and tattoos data. Are any of them particularly large in magnitude?

The chi-square value of 57.91 is the sum of the squares of these six values. The cell for people with tattoos obtained in a tattoo parlor who have hepatitis C is large and positive, indicating there are more people in that cell than the null hypothesis of independence would predict. Maybe tattoo parlors are a source of infection or maybe those who go to tattoo parlors also engage in risky behavior.

The second-largest component is a negative value for those with no tattoos who test positive for hepatitis C. A negative value says that there are fewer people in this cell than independence would expect. That is, those who have no tattoos are less likely to be infected with hepatitis C than we might expect if the two variables were independent.

What about the cells with small expected counts? The formula for the chi-square standardized residuals divides each residual by the square root of the expected frequency. Too small an expected frequency can arbitrarily inflate the residual and lead to an inflated chi-square statistic. Any expected count close to the arbitrary minimum of 5 calls for checking that cell's standardized residual to be sure it is not particularly large. In this case, the standardized residual for the "Hepatitis C and Tattoo, elsewhere" cell is not particularly large, but the standardized residual for the "Hepatitis C and Tattoo, parlor" cell is large.

We might choose not to report the results because of concern with the small expected frequency. Alternatively, we could include a warning along with our report of the results. Yet another approach is to combine categories to get a larger sample size and correspondingly larger expected frequencies, if there are some categories that can be appropriately combined. Here, we might naturally combine the two rows for tattoos, obtaining a 2 × 2 table:

	Hepatitis C	No Hepatitis C	Total
Tattoo	25	88	**113**
None	22	491	**513**
Total	**47**	**579**	**626**

Table 26.7
Combining the two tattoo categories gives a table with all expected counts greater than 5.

This table has expected values of at least 5 in every cell, and a chi-square value of 42.42 on 1 degree of freedom. The corresponding P-value is <0.0001.

We conclude that *Tattoo Status* and *Hepatitis C Status* are not independent. The data *suggest* that tattoo parlors may be a particular problem, but we haven't enough data to draw that conclusion.

FOR EXAMPLE — Writing conclusions for χ^2 tests

Recap: We're looking at Cincinnati traffic stop data to see if police decisions about searching cars show evidence of racial bias. With 2 df, technology calculates $\chi^2 = 73.25$, a P-value less than 0.0001, and these standardized residuals:

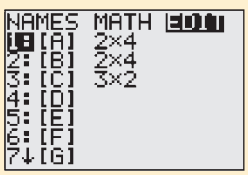
		Race		
		Black	White	Other
Search	No	–3.43	4.55	0.28
	Yes	3.84	–5.09	–0.31

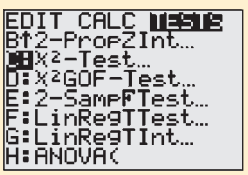
Question: What's your conclusion?

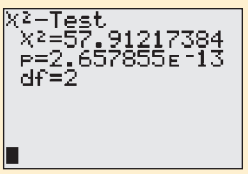
The very low P-value leads me to reject the null hypothesis. There's strong evidence that police decisions to search cars at traffic stops are associated with the driver's race.

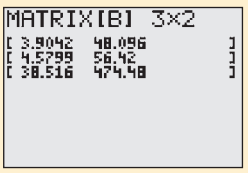
The largest residuals are for White drivers, who are searched less often than independence would predict. It appears that Black drivers' cars are searched more often.

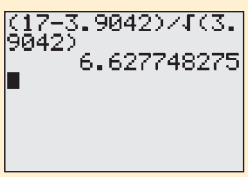
TI Tips — Testing homogeneity or independence











Yes, the TI will do chi-square tests of homogeneity and independence. Let's use the tattoo data. Here goes.

Test a hypothesis of homogeneity or independence

Stage 1: You need to enter the data as a matrix. A "matrix" is just a formal mathematical term for a table of numbers.

- Push the MATRIX button, and choose to EDIT matrix [A].
- First specify the dimensions of the table, rows × coloumns.
- Enter the appropriate counts, one cell at a time. The calculator automatically asks for them row by row.

Stage 2: Do the test.

- In the STAT TESTS menu choose C: χ^2-Test.
- The TI now confirms that you have placed the observed frequencies in [A]. It also tells you that when it finds the expected frequencies it will store those in [B] for you. Now Calculate the mechanics of the test.

The TI reports a calculated value of $\chi^2 = 57.91$ and an exceptionally small P-value.

Stage 3: Check the expected counts.

- Go back to MATRIX EDIT and choose [B].

Notice that two of the cells fail to meet the condition that expected counts be at least 5. This problem enters into our analysis and conclusions.

Stage 4: And now some bad news. There's no easy way to calculate the standardized residuals. Look at the two matrices, [A] and [B]. Large residuals will happen when the corresponding entries differ greatly, especially when the expected count in [B] is small (because you will divide by the square root of the entry in [B]). The first cell is a good candidate, so we show you the calculation of its standardized residual.

A residual of over 6 is pretty large—possibly an indication that you're more likely to get hepatitis in a tattoo parlor, but the expected count is smaller that 5. We're pretty sure that hepatitis status is not independent of having a tattoo, but we should be wary of saying anything more. Probably the best approach is to combine categories to get cells with expected counts above 5.

Chi-Square and Causation

Chi-square tests are common. Tests for independence are especially widespread. Unfortunately, many people interpret a small P-value as proof of causation. We know better. Just as correlation between quantitative variables does not demonstrate causation, a failure of independence between two categorical variables does not show a cause-and-effect relationship between them, nor should we say that one variable *depends* on the other.

The chi-square test for independence treats the two variables symmetrically. There is no way to differentiate the direction of any possible causation from one variable to the other. In our example, it is unlikely that having hepatitis causes one to crave a tattoo, but other examples are not so clear.

In this case it's easy to imagine that lurking variables are responsible for the observed lack of independence. Perhaps the lifestyles of some people include both tattoos and behaviors that put them at increased risk of hepatitis C, such as body piercings or even drug use. Even a small subpopulation of people with such a lifestyle among those with tattoos might be enough to create the observed result. After all, we observed only 25 patients with both tattoos and hepatitis.

In some sense, a failure of independence between two categorical variables is less impressive than a strong, consistent, linear association between quantitative variables. Two categorical variables can fail the test of independence in many ways, including ways that show no consistent pattern of failure. Examination of the chi-square standardized residuals can help you think about the underlying patterns.

JUST CHECKING

Which of the three chi-square tests—goodness-of-fit, homogeneity, or independence—would you use in each of the following situations?

7. A restaurant manager wonders whether customers who dine on Friday nights have the same preferences among the four "chef's special" entrées as those who dine on Saturday nights. One weekend he has the wait staff record which entrées were ordered each night. Assuming these customers to be typical of all weekend diners, he'll compare the distributions of meals chosen Friday and Saturday.

8. Company policy calls for parking spaces to be assigned to everyone at random, but you suspect that may not be so. There are three lots of equal size: lot A, next to the building; lot B, a bit farther away; and lot C, on the other side of the highway. You gather data about employees at middle management level and above to see how many were assigned parking in each lot.

9. Is a student's social life affected by where the student lives? A campus survey asked a random sample of students whether they lived in a dormitory, in off-campus housing, or at home, and whether they had been out on a date 0, 1–2, 3–4, or 5 or more times in the past two weeks.

WHAT CAN GO WRONG?

- **Don't use chi-square methods unless you have counts.** All three of the chi-square tests apply only to counts. Other kinds of data can be arrayed in two-way tables. Just because numbers are in a two-way table doesn't make them suitable for chi-square analysis. Data reported as proportions or percentages can be suitable for chi-square procedures, *but only after they are converted to counts*. If you try to do the calculations without first finding the counts, your results will be wrong.

(*continued*)

A S **Simulation: Sample Size and Chi-Square.** Chi-square statistics have a peculiar problem. They don't respond to increasing the sample size in quite the same way you might expect.

- **Beware large samples.** Beware *large* samples?! That's not the advice you're used to hearing. The chi-square tests, however, are unusual. You should be wary of chi-square tests performed on very large samples. No hypothesized distribution fits perfectly, no two groups are exactly homogeneous, and two variables are rarely perfectly independent. The degrees of freedom for chi-square tests don't grow with the sample size. With a sufficiently large sample size, a chi-square test can always reject the null hypothesis. But we have no measure of how far the data are from the null model. There are no confidence intervals to help us judge the effect size.
- **Don't say that one variable "depends" on the other just because they're not independent.** Dependence suggests a pattern and implies causation, but variables can fail to be independent in many different ways. When variables fail the test for independence, you might just say they are "associated."

CONNECTIONS

Chi-square methods relate naturally to inference methods for proportions. We can think of a test of homogeneity as stepping from a comparison of two proportions to a question of whether three or more proportions are equal. The standard deviations of the residuals in each cell are linked to the expected counts much like the standard deviations we found for proportions.

Independence is, of course, a fundamental concept in Statistics. But chi-square tests do not offer a general way to check on independence for all those times when we have had to assume it.

Stacked bar charts or side-by-side pie charts can help us think about patterns in two-way tables. A histogram or boxplot of the standardized residuals can help locate extraordinary values.

WHAT HAVE WE LEARNED?

We've learned how to test hypotheses about categorical variables. We use one of three related methods. All look at counts of data in categories, and all rely on chi-square models, a new family indexed by degrees of freedom.

- Goodness-of-fit tests compare the observed distribution of a single categorical variable to an expected distribution based on a theory or model.
- Tests of homogeneity compare the distribution of several groups for the same categorical variable.
- Tests of independence examine counts from a single group for evidence of an association between two categorical variables.

We've seen that, mechanically, these tests are almost identical. Although the tests appear to be one-sided, we've learned that conceptually they are many-sided, because there are many ways that a table of counts can deviate significantly from what we hypothesized. When that happens and we reject the null hypothesis, we've learned to examine standardized residuals in order to better understand patterns as in the table.

Terms

Chi-square model 621, 625. Chi-square models are skewed to the right. They are parameterized by their degrees of freedom and become less skewed with increasing degrees of freedom.

Cell 619, 626. A cell is one element of a table corresponding to a specific row and a specific column. Table cells can hold counts, percentages, or measurements on other variables. Or they can hold several values.

Chi-square statistic

621. The chi-square statistic can be used to test whether the observed counts in a frequency distribution or contingency table match the counts we would expect according to some model. It is calculated as

$$\chi^2 = \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp}.$$

Chi-square statistics differ in how expected counts are found, depending on the question asked.

Chi-square test of goodness-of-fit

618, 622. A test of whether the distribution of counts in one categorical variable matches the distribution predicted by a model is called a test of goodness-of-fit. In a chi-square goodness-of-fit test, the expected counts come from the predicting model. The test finds a P-value from a chi-square model with $n - 1$ degrees of freedom, where n is the number of categories in the categorical variable.

Chi-square test of homogeneity

627. A test comparing the distribution of counts for two or more groups on the same categorical variable is called a test of *homogeneity*. A chi-square test of homogeneity finds expected counts based on the overall frequencies, adjusted for the totals in each group under the (null hypothesis) assumption that the distributions are the same for each group. We find a P-value from a chi-square distribution with $(\#Rows - 1) \times (\#Cols - 1)$ degrees of freedom, where *#Rows* gives the number of categories and *#Cols* gives the number of independent groups.

Chi-square test of independence

633. A test of whether two categorical variables are independent examines the distribution of counts for one group of individuals classified according to both variables. A chi-square test of *independence* finds expected counts by assuming that knowing the marginal totals tells us the cell frequencies, assuming that there is no association between the variables. This turns out to be the same calculation as a test of homogeneity. We find a P-value from a chi-square distribution with $(\#Rows - 1) \times (\#Cols - 1)$ degrees of freedom, where *#Rows* gives the number of categories in one variable and *#Cols* gives the number of categories in the other.

Chi-square component

623, 628. The components of a chi-square calculation are

$$\frac{(Observed - Expected)^2}{Expected},$$

found for each cell of the table.

Standardized residual

631. In each cell of a two-way table, a standardized residual is the square root of the chi-square component for that cell with the sign of the $Observed - Expected$ difference:

$$\frac{(Obs - Exp)}{\sqrt{Exp}}.$$

When we reject a chi-square test, an examination of the standardized residuals can sometimes reveal more about how the data deviate from the null model.

Two-way table

626, 633. Each *cell* of a two-way table shows counts of individuals. One way classifies a sample according to a categorical variable. The other way can classify different groups of individuals according to the same variable or classify the same individuals according to a different categorical variable.

Contingency table

633. A two-way table that classifies individuals according to two categorical variables is called a *contingency table*.

Skills

- Be able to recognize when a test of goodness-of-fit, a test of homogeneity, or a test of independence would be appropriate for a table of counts.
- Understand that the degrees of freedom for a chi-square test depend on the dimensions of the table and not on the sample size. Understand that this means that increasing the sample size increases the ability of chi-square procedures to reject the null hypothesis.

- Be able to display and interpret counts in a two-way table.
- Know how to use the chi-square tables to perform chi-square tests.

- Know how to compute a chi-square test using your statistics software or calculator.
- Be able to examine the standardized residuals to explain the nature of the deviations from the null hypothesis.

- Know how to interpret chi-square as a test of goodness-of-fit in a few sentences.
- Know how to interpret chi-square as a test of homogeneity in a few sentences.
- Know how to interpret chi-square as a test of independence in a few sentences.

CHI-SQUARE ON THE COMPUTER

Most statistics packages associate chi-square tests with contingency tables. Often chi-square is available as an option only when you make a contingency table. This organization can make it hard to locate the chi-square test and may confuse the three different roles that the chi-square test can take. In particular, chi-square tests for goodness-of-fit may be hard to find or missing entirely. Chi-square tests for homogeneity are computationally the same as chi-square tests for independence, so you may have to perform the mechanics as if they were tests of independence and interpret them afterwards as tests of homogeneity.

Most statistics packages work with data on individuals rather than with the summary counts. If the only information you have is the table of counts, you may find it more difficult to get a statistics package to compute chi-square. Some packages offer a way to reconstruct the data from the summary counts so that they can then be passed back through the chi-square calculation, finding the cell counts again. Many packages offer chi-square standardized residuals (although they may be called something else).

EXERCISES

1. **Which test?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
   a) A brokerage firm wants to see whether the type of account a customer has (Silver, Gold, or Platinum) affects the type of trades that customer makes (in person, by phone, or on the Internet). It collects a random sample of trades made for its customers over the past year and performs a test.
   b) That brokerage firm also wants to know if the type of account affects the size of the account (in dollars). It performs a test to see if the mean size of the account is the same for the three account types.
   c) The academic research office at a large community college wants to see whether the distribution of courses chosen (Humanities, Social Science, or Science) is different for its residential and nonresidential students. It assembles last semester's data and performs a test.

2. **Which test again?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
   a) Is the quality of a car affected by what day it was built? A car manufacturer examines a random sample of the warranty claims filed over the past two years to test whether defects are randomly distributed across days of the work week.
   b) A medical researcher wants to know if blood cholesterol level is related to heart disease. She examines a database of 10,000 patients, testing whether the cholesterol level (in milligrams) is related to whether or not a person has heart disease.
   c) A student wants to find out whether political leaning (liberal, moderate, or conservative) is related to choice of major. He surveys 500 randomly chosen students and performs a test.

3. **Dice.** After getting trounced by your little brother in a children's game, you suspect the die he gave you to roll may be unfair. To check, you roll it 60 times, recording the number of times each face appears. Do these results cast doubt on the die's fairness?