Chapter 11: Inference for Distributions of Categorical Data

Section 11.1
Chi-Square Goodness-of-Fit Tests

The Practice of Statistics, 4th edition – For AP*
STARNES, YATES, MOORE

Chapter 11
Inference for Distributions of Categorical Data

- **11.1 Chi-Square Goodness-of-Fit Tests**
- 11.2 Inference for Relationships

Section 11.1
Chi-Square Goodness-of-Fit Tests

Learning Objectives

After this section, you should be able to…

- ✓ COMPUTE expected counts, conditional distributions, and contributions to the chi-square statistic
- ✓ CHECK the Random, Large sample size, and Independent conditions before performing a chi-square test
- ✓ PERFORM a chi-square goodness-of-fit test to determine whether sample data are consistent with a specified distribution of a categorical variable
- ✓ EXAMINE individual components of the chi-square statistic as part of a follow-up analysis

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 goodness-of-fit test** allows us to determine whether a hypothesized distribution seems valid.

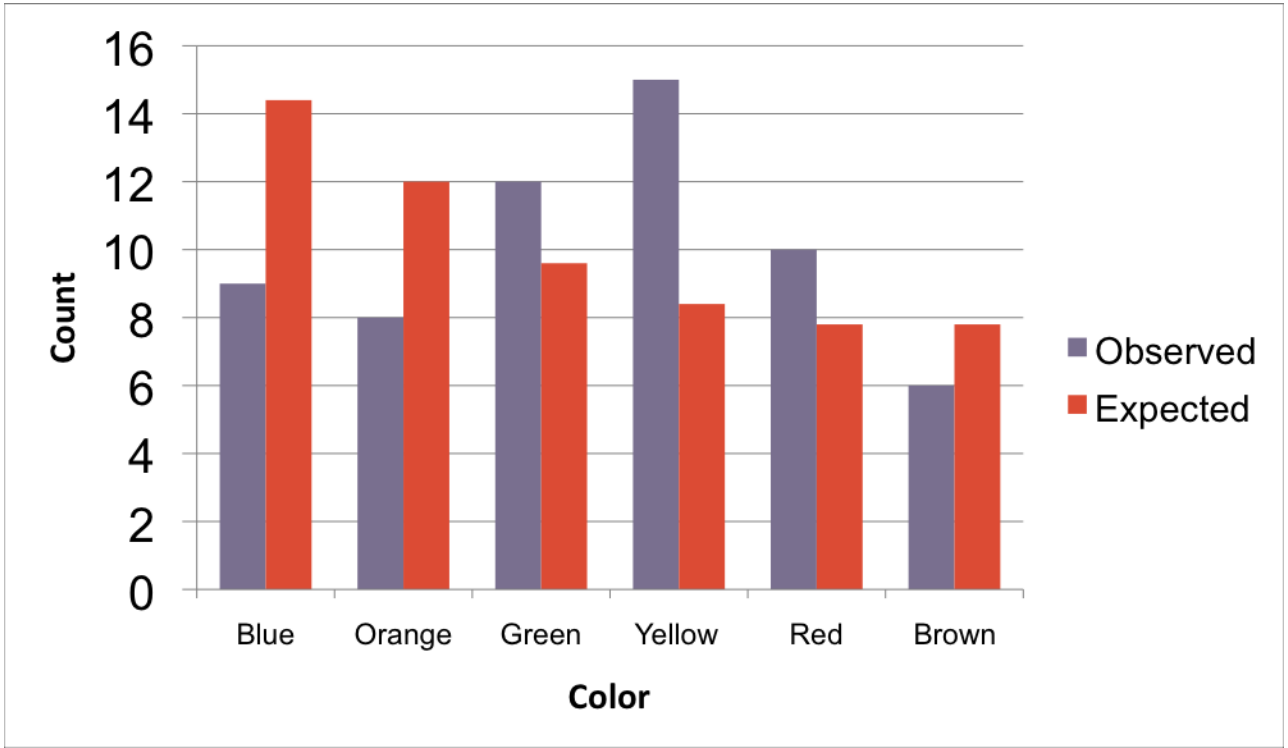
We can decide whether the distribution of a categorical variable differs for two or more populations or treatments using a **chi-square test for homogeneity**. In doing so, we will often organize our data in a two-way table.

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 association/independence** allows us to determine if there is convincing evidence of an association between the variables in the population at large.

Activity: The Candy Man Can

- Mars, Incorporated 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, the new mix of colors of 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.*

- Follow the instructions on page 676. Teacher: Right-click (control-click) on the graph to edit the observed counts.



Chi-Square Goodness-of-Fit Tests

The **one-way table** below summarizes the data from a sample bag of M&M'S Milk Chocolate Candies. In general, one-way tables display the distribution of a 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 $\hat{p} = \frac{9}{60} = 0.15$.

Since the company claims that 24% of all M&M'S Milk Chocolate Candies are blue, we 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. We could then perform additional significance tests for each of the remaining colors.

However, performing a one-sample *z* test for each proportion would be pretty inefficient and would lead to the problem of multiple comparisons.

Alternate Example – A fair die?

Jenny made a six-sided die in her ceramics class and rolled it 60 times to test if each side was equally likely to show up on top.

Problem: Assuming that her die is fair, calculate the expected counts for each side.

Solution: If the die is fair, each of the six sides has a 1/6 probability of ending up on top. This means that each expected count is 1/6(60) = 10.

Comparing Observed and Expected Counts

More important, performing one-sample z tests for each color 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 this bag does (taking *all* the colors into consideration at one time).

For that, we need a new kind of significance test, called a **chi-square goodness-of-fit test.**

The null hypothesis in a chi-square goodness-of-fit test should state a claim about the distribution of a single categorical variable in the population of interest. In our example, the appropriate null hypothesis is

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

The alternative hypothesis in a chi-square goodness-of-fit test is that the categorical variable does *not* have the specified distribution. In our example, the alternative hypothesis is

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

Comparing Observed and Expected Counts

We can also write the hypotheses in symbols as

$$H_0: p_{blue} = 0.24,\ p_{orange} = 0.20,\ p_{green} = 0.16,$$
$$p_{yellow} = 0.14,\ p_{red} = 0.13,\ p_{brown} = 0.13,$$
$$H_a: \text{At least one of the } p_i\text{'s is incorrect}$$

where p_{color} = the true population proportion of M&M'S Milk Chocolate Candies of that color.

The idea of the chi-square goodness-of-fit test is this: we compare the **observed counts** from our sample with the counts that would be expected if H_0 is true. 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 proportion of the population distribution in each category by the sample size.

Example: Computing Expected Counts

A sample bag of M&M's milk Chocolate Candies contained 60 candies. Calculate the expected counts for each color.

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 (0.24)(60) = 14.40. This is our expected count of blue M&M's.

Using this same method, we can find the expected counts for the other color categories:

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

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

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

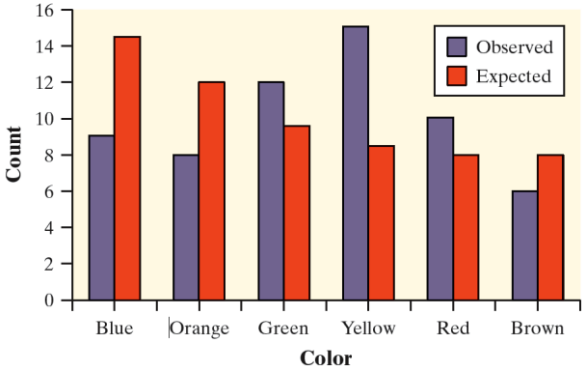
Red: **(0.13)(60) = 7.80**

Brown: **(0.13)(60) = 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

The Chi-Square Statistic

To see if the data give convincing evidence against the null hypothesis, we compare the observed counts from our sample with the expected counts assuming H_0 is true. If the observed counts are far from the expected counts, that's the evidence we were seeking.



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.**

Definition:

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 variab

Example: Return of the M&M's

The table shows the observed and expected counts for our sample of 60 M&M's Milk Chocolate Candies. Calculate the chi-square statistic.

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

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

$$\chi^2 = \frac{(9-14.40)^2}{14.40} + \frac{(8-12.00)^2}{12.00} + \frac{(12-9.60)^2}{9.60} + \frac{(15-8.40)^2}{8.40} + \frac{(10-7.80)^2}{7.80} + \frac{(6-7.80)^2}{7.80}$$

$$\chi^2 = 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 = 10.180$$

Think of χ^2 as a measure of the distance of the observed counts from the expected counts. Large values of χ^2 are stronger evidence against H_0 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.

Alternate Example: A fair die?

Below are the results of Jenny's 60 rolls of her ceramic die and the expected counts. Calculate the value of the chi-square statistic.

Outcome	Observed	Expected
1	13	10
2	11	10
3	6	10
4	12	10
5	10	10
6	8	10
Total	60	60

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

$$\chi^2 = \frac{(13-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(8-10)^2}{10}$$

$$\chi^2 = 0.9 + 0.1 + 1.6 + 0.4 + 0 + 0.4 = 3.4$$

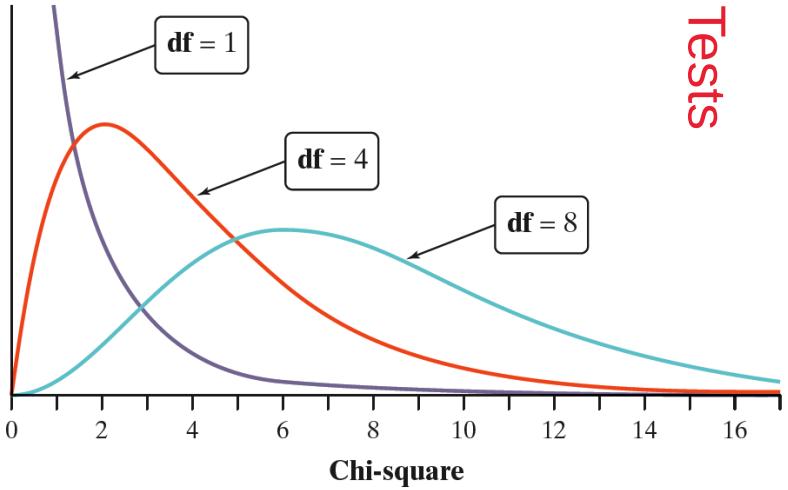
The Chi-Square Distributions and *P*-Values

The sampling distribution of the chi-square statistic is not a Normal distribution. It is a right-skewed distribution that allows only positive values because χ^2 can never be negative.

When the expected counts are all at least 5, the sampling distribution of the χ^2 statistic is close to a **chi-square distribution** with degrees of freedom (df) equal to the number of categories minus 1.

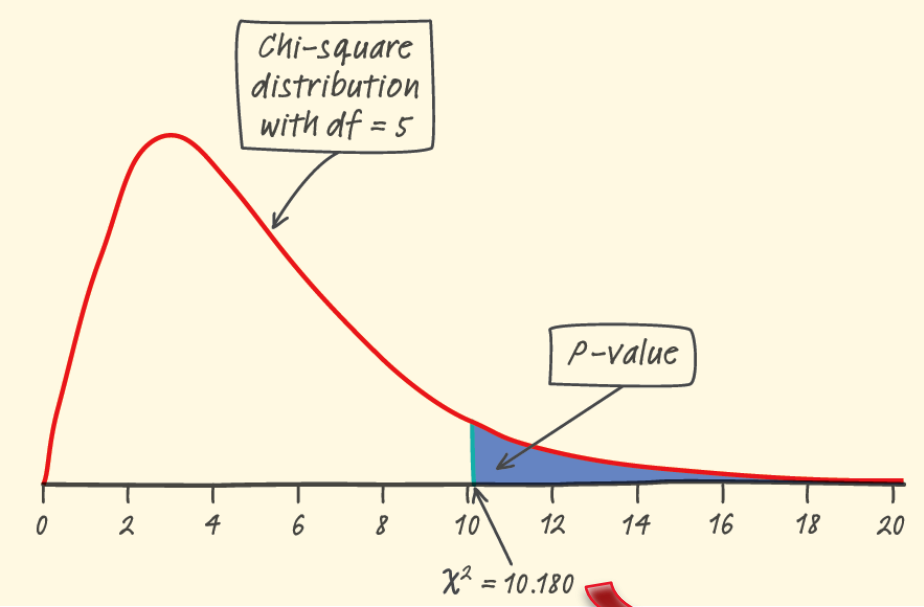
The Chi-Square Distributions

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



Example: Return of the M&M's

We computed the chi-square statistic for our sample of 60 M&M's to be $\chi^2 = 10.180$. Because all of the expected counts are at least 5, the χ^2 statistic will follow a chi-square distribution with df = 6 - 1 = 5 reasonably well when H_0 is true.



To find the *P*-value, use Table C and look in the df = 5 row.

	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

Since our *P*-value is between 0.05 and 0.10, it is greater than $\alpha = 0.05$. Therefore, we fail to reject H_0. We don't have sufficient evidence to conclude that the company's claimed color distribution is incorrect.

Alternate Example: A fair die?

When Jenny rolled her ceramic die 60 times and calculated the chi-square statistic, she got χ^2 =3.4.

Problem: Using the appropriate degrees of freedom, calculate the P-value. What conclusion can you make about Jenny's die?

Solution: Since there are six possible outcomes when rolling her die, the degrees of freedom = 6 – 1 = 5. Using Table C, the P-value is greater than 0.25 since the χ^2 statistic is smaller than the lowest critical value in the df = 5 row.
Using technology, $\chi^2\text{cdf}(3.4,\ 1000,\ 5) = 0.64$. Since the P-value is quite large, we do not have convincing evidence that her die is unfair. However, that doesn't prove that her die is fair.

Carrying Out a Test

The Chi-Square Goodness-of-Fit Test

Suppose the Random, Large ... independent conditions are met. To determine wh... ecified distribution, ex... each possible ...

Before we start using the chi-square goodness-of-fit test, 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 Sample Size condition, be sure to examine the *expected* counts, not the observed counts.

where the su... e area to the right of χ^2 under ... istribution with $k-1$ degrees of freedom.

Example: When Were You Born?

Are births evenly distributed across the days of the week? The one-way table below shows the distribution of births across the days of the week in a random sample of 140 births from local records in a large city. Do these data give significant evidence that local births are not equally likely on all days of the week?

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Births	13	23	24	20	27	18	15

State: We want to perform a test of

H_0: Birth days in this local area are evenly distributed across the days of the week.
H_a: Birth days in this local area are not evenly distributed across the days of the week.

The null hypothesis says that the proportions of births are the same on all days. In that case, all 7 proportions must be 1/7. So we could also write the hypotheses as

H_0: $p_{Sun} = p_{Mon} = p_{Tues} = \ldots = p_{Sat} = 1/7$.
H_a: At least one of the proportions is not 1/7.

We will use $\alpha = 0.05$.

Plan: If the conditions are met, we should conduct a chi-square goodness-of-fit test.

• *Random* The data came from a random sample of local births.

• *Large Sample Size* Assuming H_0 is true, we would expect one-seventh of the births to occur on each day of the week. For the sample of 140 births, the expected count for all 7 days would be 1/7(140) = 20 births. Since 20 ≥ 5, this condition is met.

• *Independent* Individual births in the random sample should occur independently (assuming no twins). Because we are sampling without replacement, there need to be at least 10(140) = 1400 births in the local area. This should be the case in a large city.

Example: When Were You Born?

Do: Since the conditions are satisfied, we can perform a chi-square goodness-of-fit test. We begin by calculating the test statistic.

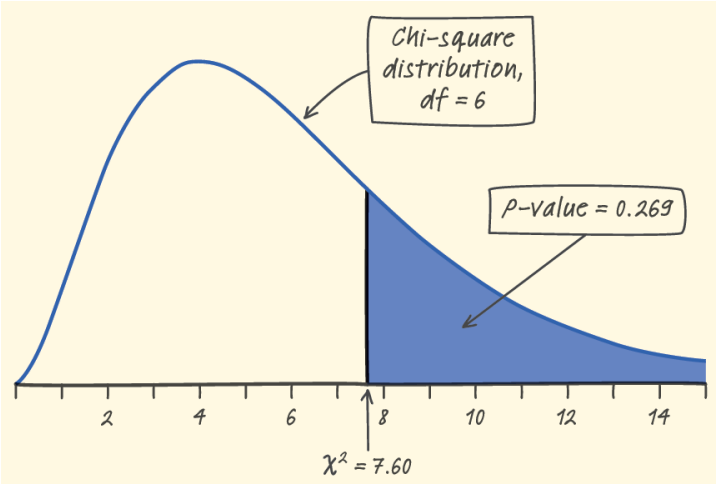
Test statistic

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

$$= \frac{(13-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(24-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(15-20)^2}{20}$$

$$= 2.45 + 0.45 + 0.80 + 0.00 + 2.45 + 0.20 + 1.25$$

$$= 7.60$$



***P*-Value**:
Using Table C: $\chi^2 = 7.60$ is less than the smallest entry in the df = 6 row, which corresponds to tail area 0.25. The *P*-value is therefore greater than 0.25.
Using technology: We can find the exact *P*-value with a calculator: $\chi^2\text{cdf}(7.60,1000,6) = 0.269$.

Conclude: Because the *P*-value, 0.269, is greater than $\alpha = 0.05$, we fail to reject H_0. These 140 births don't provide enough evidence to say that all local births in this area are not evenly distributed across the days of the week.

Alternate Example: Landline Surveys

According to the 2000 census, of all U.S. residents aged 20 and older, 19.1% are in their 20s, 21.5% are in their 30s, 21.1% are in their 40s, 15.5% are in their 50s, and 22.8% are 60 or older. The table below shows the age distribution for a sample of U.S. residents aged 20 and older. Members of the sample were chosen by randomly dialing landline telephone numbers. Do these date provide convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents?

Category	20-29	30-39	40-49	50-59	60+	Total
Count	141	186	224	211	286	1048

State: We want to perform a test of

H_0: The age distribution of people who answer landline telephone surveys is the same as the age distribution of all U.S. residents,
H_a: The age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents.

We will use $\alpha = 0.05$.

Plan: If the conditions are met, we should conduct a chi-square goodness-of-fit test.

• *Random* The data came from a random sample of U.S. residents who answer landline telephone surveys.

• *Large Sample Size* The expected counts are 1048(0.191)=200.2, 1048(0.215)=225.3, 1048(0.211)=221.1, 1048(0.155)=162.4, 1048(0.228)=238.9. All expected counts are at least 5.

• *Independent* Because we are sampling without replacement, there needs to be at least 10(1048) = 10,480 U.S. residents who answer telephone landline surveys. This is reasonable to assume.

Alternate Example: Landline Surveys

Do: Since the conditions are satisfied, we can perform a chi-square goodness-of-fit test. We begin by calculating the test statistic.

Test statistic:

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

$$= \frac{(141 - 200.2)^2}{200.2} + \ldots = 48.2$$

***P*-Value:**

Using technology: We can find the exact *P*-value with a calculator:

χ^2cdf(48.2,1000,4) = 0.

Conclude: Because the *P*-value is approximately is less than $\alpha = 0.05$, we reject H_0. We have convincing evidence that the age distribution of people who answer landline telephone surveys is not the same as the age distribution of all U.S. residents.

Example: Inherited Traits

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.

		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 these data differ significantly from what the biologists have predicted? Carry out an appropriate test at the $\alpha = 0.05$ level to help answer this question.

Example: Inherited Traits

State: We want to perform a test of

H_0: The biologists' predicted color distribution for tobacco plant offspring is correct. That is, $p_{green} = 0.25$, $p_{yellow\text{-}green} = 0.5$, $p_{albino} = 0.25$
H_a: The biologists' predicted color distribution isn't correct. That is, at least one of the stated proportions is incorrect.

We will use $\alpha = 0.05$.

Plan: If the conditions are met, we should conduct a chi-square goodness-of-fit test.

• *Random* The data came from a random sample of local births.

• *Large Sample Size* We check that all expected counts are at least 5. Assuming H_0 is true, the expected counts for the different colors of offspring are
green: (0.25)(84) = 21; yellow-green: (0.50)(84) = 42; albino: (0.25)(84) = 21
The complete table of observed and expected counts is shown below.

• *Independent* Individual offspring inherit their traits independently from one another. Since we are sampling without replacement, there would need to be at least 10(84) = 840 tobacco plants in the population. This seems reasonable to believe.

Offspring color	Observed	Expected
Green	23	21
Yellow-green	50	42
Albino	11	21

Example: Inherited Traits

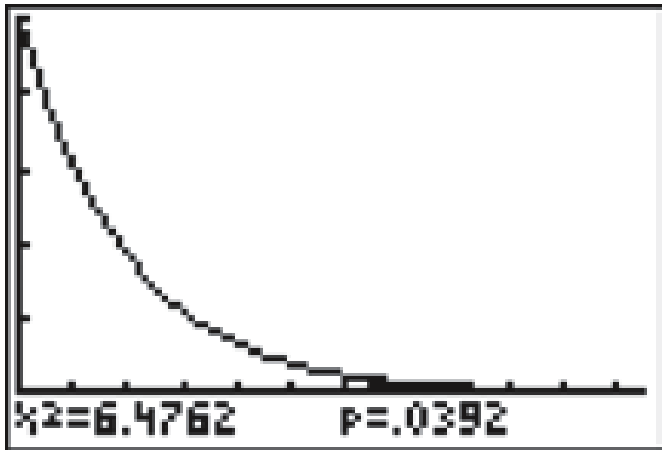
Do: Since the conditions are satisfied, we can perform a chi-square goodness-of-fit test. We begin by calculating the test statistic.

Test statistic

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

$$= \frac{(23-21)^2}{21} + \frac{(50-42)^2}{50} + \frac{(11-21)^2}{21}$$

$$= 6.476$$



***P*-Value**:
Note that df = number of categories - 1 = 3 - 1 = 2. Using df = 2, the *P*-value from the calculator is 0.0392

Conclude: Because the *P*-value, 0.0392, is less than $\alpha = 0.05$, we will reject H_0. We have convincing evidence that the biologists' hypothesized distribution for the color of tobacco plant offspring is incorrect.

Alternate Example: Birthdays and Hockey

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, since January 1 is the cutoff date for youth leagues in Canada (where many 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 NFL), a random sample of 80 NHL players from the 2009-2010 season was selected and their birthdays were recorded. Overall, 32 were born in the first quarter of the year, 20 in the second quarter, 16 in the third quarter, and 12 in the fourth quarter. Do these data provide convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the year.

State: We want to perform a test of

H_0: The birthdays of NHL players are equally likely to occur in each quarter of the year,

H_a: The birthdays of NHL players are not equally likely to occur in each quarter of the year.

We will use $\alpha = 0.05$.

Alternate Example: Birthdays and Hockey

Plan: If the conditions are met, we should conduct a chi-square goodness-of-fit test.

• *Random* The data came from a random sample of NHL players.

• *Large Sample Size* If birthdays are equally likely to be in each quarter of the year, then the expected counts are all ¼(80) = 20. These counts are all at least 5.

• *Independent* Because we are sampling without replacement, there must be at least 10(80) = 800 players. In the 2009-2010 season, there are 879 NHL players, so this condition is met.

Do: **Test statistic**:

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

$$+ \frac{(16-20)^2}{20} + \frac{(12-20)^2}{20} = 11.2$$

P-value Using 4 – 1 = 3 degrees of freedom, P-value = χ^2cdf(11.2, 1000, 3) = 0.011

Conclude: Because the *P*-value, 0.011, is less than $\alpha = 0.05$, we will reject H_0. We have convincing evidence that the birthdays of NHL players are not uniformly distributed throughout the year.

Follow-up Analysis

In the chi-square goodness-of-fit test, we test the null hypothesis that a categorical variable has a specified distribution. If the sample data lead to a statistically significant result, we can conclude that our variable has a distribution different from the specified one.

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 that are added together to produce the test statistic χ^2. These **components** show which terms contribute most to the chi-square statistic.

In the tobacco plant example, we can see that the component for the albino offspring made the large contribution to the chi-square statitstic.

$$\chi^2 = \frac{(23-21)^2}{21} + \frac{(50-42)^2}{50} + \frac{(11-21)^2}{21}$$

$$= 0.190 + 1.524 + 4.762 = 6.476$$

Offspring color	Observed	Expected
Green	23	21
Yellow-green	50	42
Albino	11	21

Alternate Example: Birthdays and Hockey

In the previous Alternate Example, we conclude that the birthdays of NHL players were not uniformly distributed throughout the year. However, Gladwell's claim wasn't just that the distribution wasn't uniform – he specifically claimed that NHL players are more likely to be born early in the year. Comparing the observed and expected counts, it seems that he was correct. There were 12 more players born in the first quarter than expected, while there were 4 fewer than expected in the third quarter and 8 fewer than expected in the fourth quarter.

Section 11.1
Chi-Square Goodness-of-Fit Tests

Summary

In this section, we learned that…

- A **one-way table** is often used to display the distribution of a categorical variable for a sample of individuals.
- The **chi-square goodness-of-fit test** tests the null hypothesis that a categorical variable has a specified distribution.
- 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 specified proportion of the population distribution in that category by the sample size.
- The **chi-square 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

Section 11.1
Chi-Square Goodness-of-Fit Tests

Summary

- The test compares the value of the statistic χ^2 with critical values from the **chi-square distribution with** degrees of freedom df = number of categories - 1. Large values of χ^2 are evidence against H_0, so the *P*-value is the area under the chi-square density curve to the right of χ^2.

- The chi-square distribution is an approximation to the sampling distribution of the statistic χ^2. *You can safely use this approximation when all expected* cell counts are at least 5 (**Large Sample Size condition).**

- Be sure to check that the Random, Large Sample Size, and Independent conditions are met before performing a chi-square goodness-of-fit test.

- If the test finds a statistically significant result, do a follow-up analysis that compares the observed and expected counts and that looks for the largest **components** of the chi-square statistic**.**

Looking Ahead…

In the next Section…

We'll learn how to perform inference for relationships in distributions of categorical data.

We'll learn about

- ✓ **Comparing Distributions of a Categorical Variable**
- ✓ **The Chi-square Test for Homogeneity**
- ✓ **The Chi-square Test for Association/Independence**
- ✓ **Using Chi-square Tests Wisely**