Inference for Distributions of Categorical Data

lesson 1
Chi-Square Goodness-of-Fit Tests

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*

- Here’s what the company says about the color distribution of its M&M’S Milk Chocolate Candies: On average, the M&M’S Milk Chocolate Candies will contain 13% of each of browns and reds, 14% yellows, 16% greens, 20% oranges and 24% blues.

- 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**

- Since the company claims that 24% of all M&M’S Milk Chocolate Candies are blue, we could use the one-sample z test for a proportion.

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

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.

× pretty inefficient ×

■ Comparing Observed and Expected Counts

We need a new kind of significance test, called

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:

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

H_a: **At least one** of the p_i's is incorrect

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

The idea of the chi-square goodness-of-fit test:

- Assume H_0 is true, compare the **observed counts** (from sample) with the counts that would be expected.
- The more the observed counts differ from the **expected counts**, the more evidence we have against the null hypothesis.

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 the company is true.

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

Ho: 13% of each of browns and reds, 14% yellows,16% greens, 20% oranges and 24% blues.

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 variable.

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

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

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

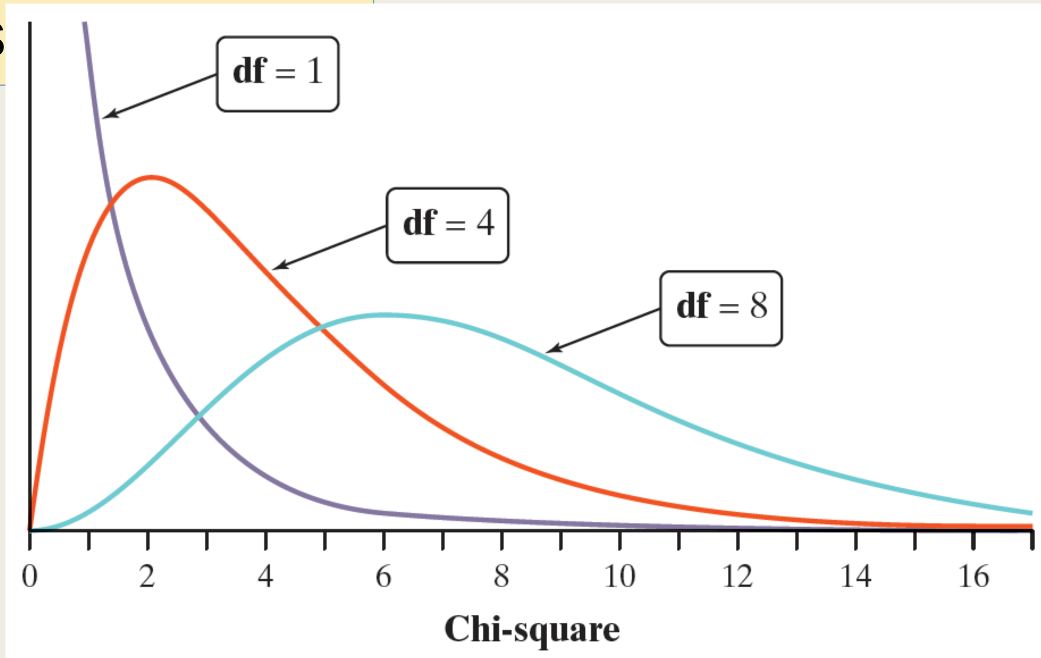
Chi-Square Goodness-of-Fit Tests

■ The Chi-Square Distributions

The Chi-Square Distributions

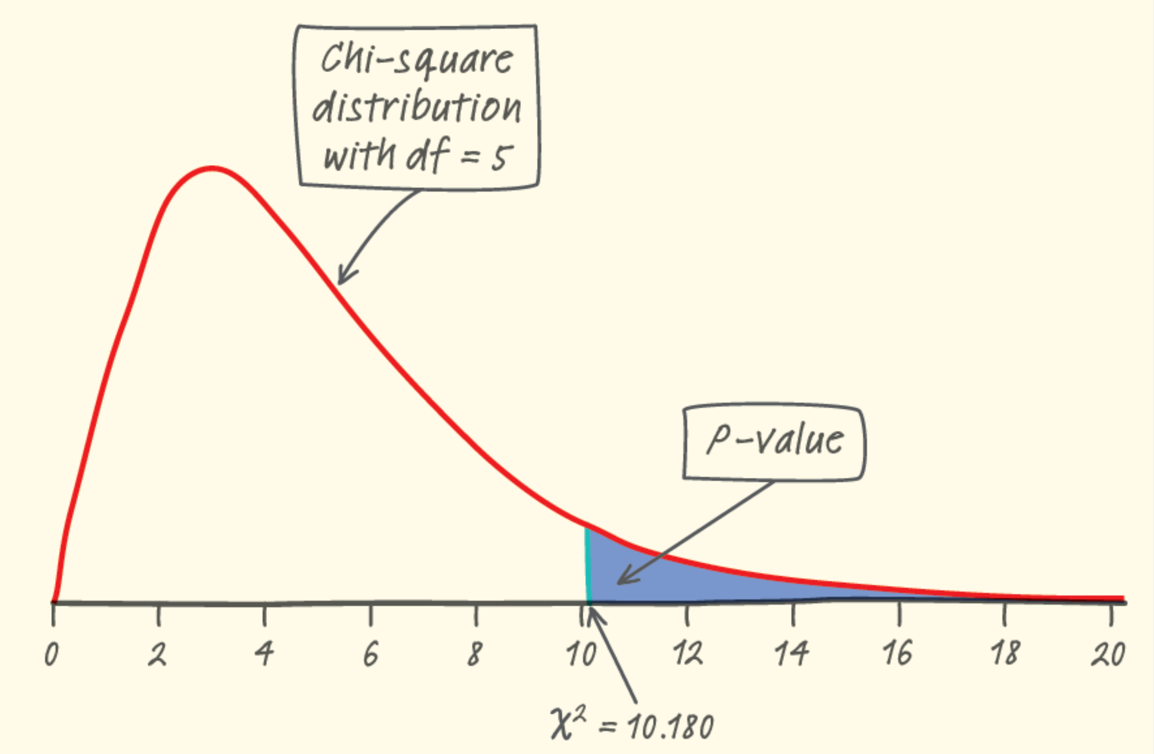
The chi-square distributions are a family of distributions that take only positive values and are skewed to the right.

The chi-square goodness-of-fit test uses the chi-square distribution with df = n - 1. n is the number of categories



Chi-Square Goodness-of-

Example: Return of the M&M's



- Since our P-value(0.07) is greater than α(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.

■ Carrying Out a Test

Conditions: Use the chi-square goodness-of-fit test when

✓ **Random**

The data come from a random sample or a randomized experiment.

✓ **Large Sample Size**

All expected counts are at least 5.

✓ **Independent**

Individual observations are independent.

When sampling without replacement, check that the population is at least 10 times as large as the sample (the 10% condition).

The Ch uar lnes Fit Test

Suppose the ndtions are met. To ribution, express tegory, pe

tru

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 sum is ove alue is the area to the right of χ^2 under the density curve the chi-square distribution 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.

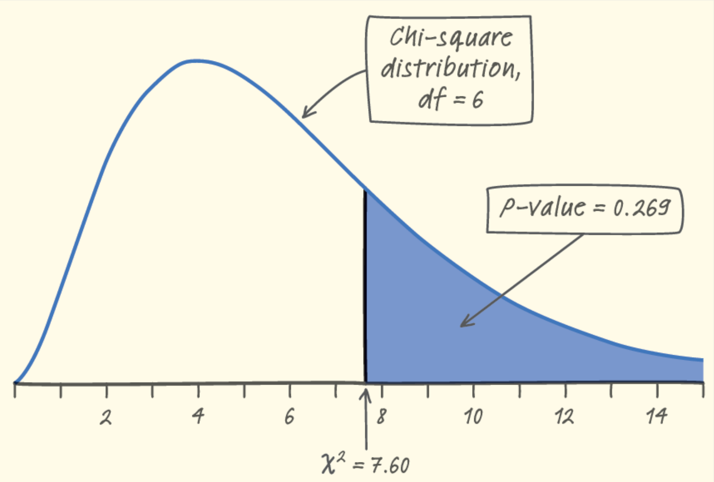
Test statistic:

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

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

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

$$= 7.60$$



***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.

Inference for Distributions of Categorical Data

lesson 2
Inference for Relationships

Inference for Relationships

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 test for homogeneity to determine whether the distribution of a categorical variable differs for several populations or treatments
- ✓ PERFORM a chi-square test for association/independence to determine whether there is convincing evidence of an association between two categorical variables
- ✓ EXAMINE individual components of the chi-square statistic as part of a follow-up analysis
- ✓ INTERPRET computer output for a chi-square test based on a two-way table

Example: Does Music Influence Purchases?

Researchers suspect that background music may lead to different buying behaviors. One study in a supermarket compared three randomly assigned treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. The table below summarizes the data.

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

H_0: There is no difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.

H_a: There is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.

- Finding Expected Counts

Finding Expected Counts

The expected count in any cell of a two-way table **when H_0 is true** is

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

Expected Counts

	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

Calculating the Chi-Square Statistic

In order to calculate a chi-square statistic for the wine example, we must check to make sure the conditions are met:

- ✓All the **expected** counts in the music and wine study are at least **5**.
- ✓The treatments were assigned at random.

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

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

■ Calculating The Chi-Square Statistic

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

Expected Counts				
	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	34.22	30.56	34.22	**99**
Italian	10.72	9.57	10.72	**31**
Other	39.06	34.88	39.06	**113**
Total	**84**	**75**	**84**	**243**

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

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

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

Example: Does Music Influence Purchases?

P-value = P(χ^2>18.28) = ….. , where $\chi^2 \sim {\chi^2}_{df} = {\chi^2}_4$.

The small *P*-value gives us convincing evidence to reject H_0 and conclude that there is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played. Furthermore, the random assignment allows us to say that the difference is caused by the music that's played.

The appropriate hypotheses for a chi-square test for homogeneity are:

H_0: There is no difference in distributions of a categorical variable across populations or treatments.

H_a: There is a difference in distributions of a categorical variable across populations or treatments.

VAR-8.I.2

The appropriate hypotheses for a chi-square test for independence are:

H_0: There is no association between two categorical variables in a given population or the two categorical variables are independent.

H_a: Two categorical variables in a population are associated or dependent.

Example: Cell-Only Telephone Users

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

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

State: We want to perform a test of

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

We will use $\alpha = 0.05$.

■ Example: Cell-Only Telephone Users

Plan: If the conditions are met, we should conduct a chi-square test for homogeneity.

• *Random* The data came from separate random samples of 96 cell-only and 104 landline users.

• *Large Sample Size* We followed the steps in the Technology Corner (page 705) to get the expected counts. The calculator screenshot confirms all expected counts ≥ 5.

```
[B]
 [[46.08 49.92]
  [20.16 21.84]
  [29.76 32.24]]
```

• *Independent R*esearchers took independent samples of cell-only and landline phone users. Sampling without replacement was used, so there need to be at least 10(96) = 960 cell-only users under age 30 and at least 10(104) = 1040 landline users under age 30. This is safe to assume.

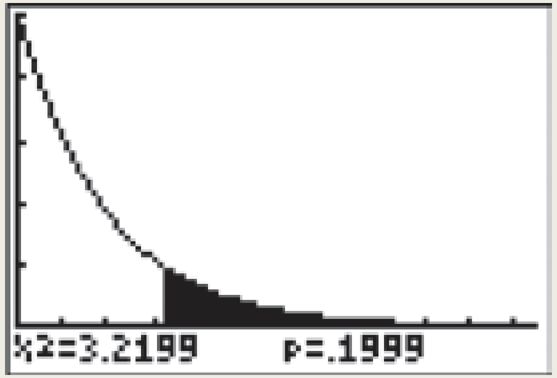
■ Example: Cell-Only Telephone Users

Do: Since the conditions are satisfied, we can a perform chi-test for homogeneity. We begin by calculating the test statistic.

Test statistic:

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

$$= \frac{(49 - 46.08)^2}{46.08} + \frac{(47 - 49.92)^2}{49.92} + \ldots + \frac{(30 - 32.24)^2}{32.24} = 3.22$$



***P*-Value**:
Using df = (3 – 1)(2 – 1) = 2, the *P*-value is 0.20.

Conclude: Because the *P*-value, 0.20, is greater than $\alpha = 0.05$, we fail to reject H_0. There is not enough evidence to conclude that the distribution of party affiliation differs in the cell-only and landline user populations.