# CHAPTER 11
# Inference for Distributions of Categorical Data

## 11.1
## Chi-Square Tests for Goodness of Fit

The Practice of Statistics, 5th Edition

Starnes, Tabor, Yates, Moore

Bedford Freeman Worth Publishers

# Chi-Square Tests for Goodness of Fit

## Learning Objectives

After this section, you should be able to:

- ✓ STATE appropriate hypotheses and COMPUTE expected counts for a chi-square test for goodness of fit.
- ✓ CALCULATE the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test for goodness of fit.
- ✓ PERFORM a chi-square test for goodness of fit.
- ✓ CONDUCT a follow-up analysis when the results of a chi-square test are statistically significant.

# Introduction

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**.

We will often organize our data in a two-way table.

It is also possible to use the information in a two-way table to study the relationship between two categorical variables. The **chi-square test for independence** allows us to determine if there is convincing evidence of an association between the variables in the population at large.

# The 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.

The **one-way table** 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 $\hat{p} = \dfrac{9}{60} = 0.15.$

# The Candy Man Can

| **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 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.

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

# The Chi-Square Statistic

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.

$H_0$: 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.

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

# The Chi-Square Statistic

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.

# The Chi-Square Statistic

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®.

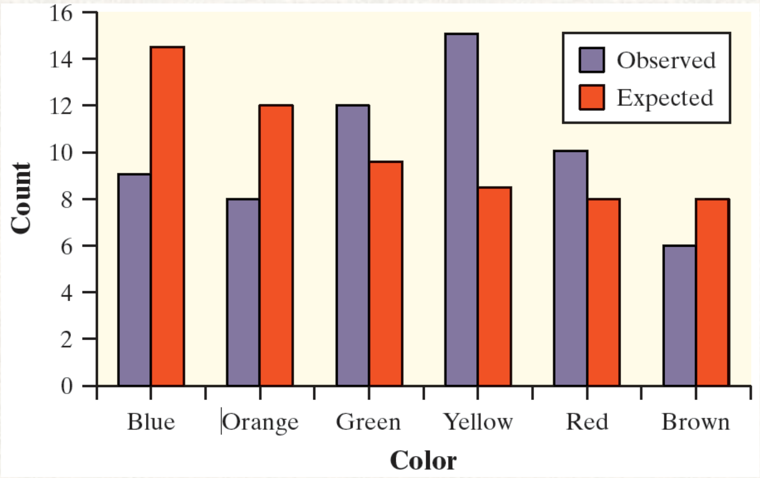
Using this same method, we can find the expected counts for the other color categories:

| | |
|---|---|
| **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.**

# The Chi-Square Statistic

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

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

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

# The Chi-Square Statistic

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

| 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 |

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

$$\begin{aligned}\chi^2 &= 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 \\ &= 10.180\end{aligned}$$

Think of $\chi^2$ as a measure of the distance of the observed counts from the expected counts. Large values of $\chi^2$ 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 $\chi^2$ suggest that the data are consistent with the null hypothesis.

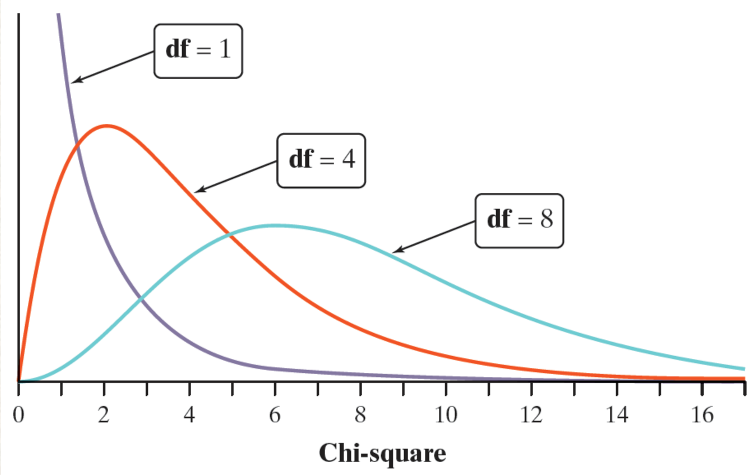
# The Chi-Square Distributions and *P*-Values

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 $c^2$ can never be negative.

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

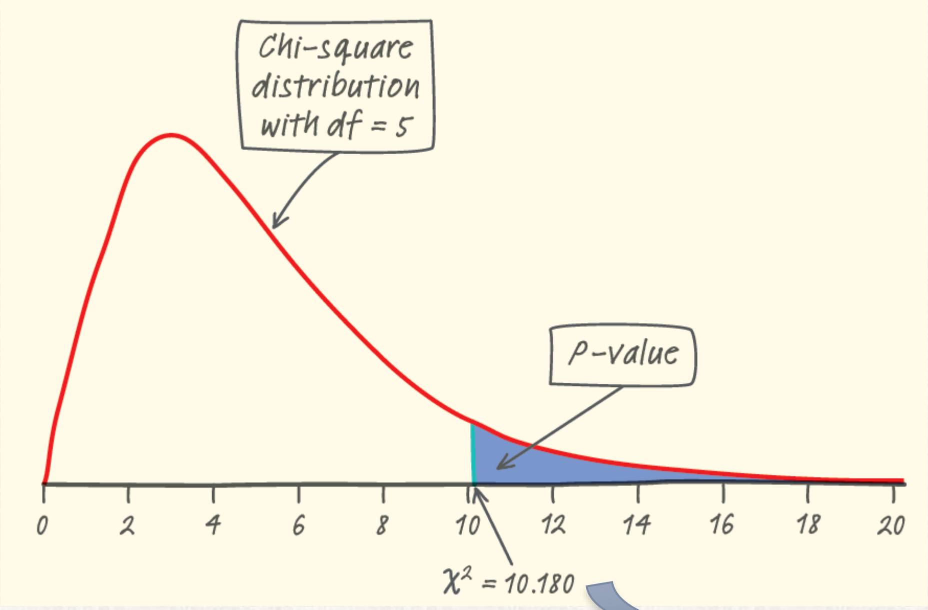
**The Chi-Square Distributions**

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.



# The Chi-Square Distributions and *P*-Values

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 $\chi^2$ 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 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 for Performing a Chi-Square Test for Goodness of Fit**

- **Random**: The data come a well-designed random sample or from a randomized experiment.
  - **10%**: When sampling without replacement, check that

    $n \leq (1/10)N.$
- **Large Counts**: All *expected* counts are greater than 5

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

•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.

•When checking the Large Sample Size condition, be sure to examine the *expected* counts, not the observed counts.

# Carrying Out a Test

## The Chi-Square Test for Goodness of Fit

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

$H_0$: The stated distribution of the categorical variable in the population of interest is correct.
$H_a$: The stated distribution of the categorical variable in the population of interest is not correct.

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

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

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

## CHECK YOUR UNDERSTANDING

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

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

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

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

# Example: A test for equal proportions

**Problem**: 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 cut-off date for youth leagues in Canada (where many National Hockey League (NHL) players come from), players born in January will be competing against players up to 12 months younger. The older players tend to be bigger, stronger, and more coordinated and hence get more playing time, more coaching, and have a better chance of being successful.

To see if birth date is related to success (judged by whether a player makes it into the NHL), a random sample of 80 National Hockey League players from a recent season was selected and their birthdays were recorded.

# Example: A test for equal proportions

**Problem**: The one-way table below summarizes the data on birthdays for these 80 players:

| Birthday | Jan – Mar | Apr – Jun | Jul – Sep | Oct – Dec |
|---|---|---|---|---|
| **Number of Players** | 32 | 20 | 16 | 12 |

Do these data provide convincing evidence that the birthdays of all NHL players are evenly distributed among the four quarters of the year?

**State**: We want to perform a test of
$H_0$: The birthdays of all NHL players are evenly distributed among the four quarters of the year.
$H_a$: The birthdays of all NHL players are not evenly distributed among the four quarters of the year.

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

# Example: A test for equal proportions

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

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

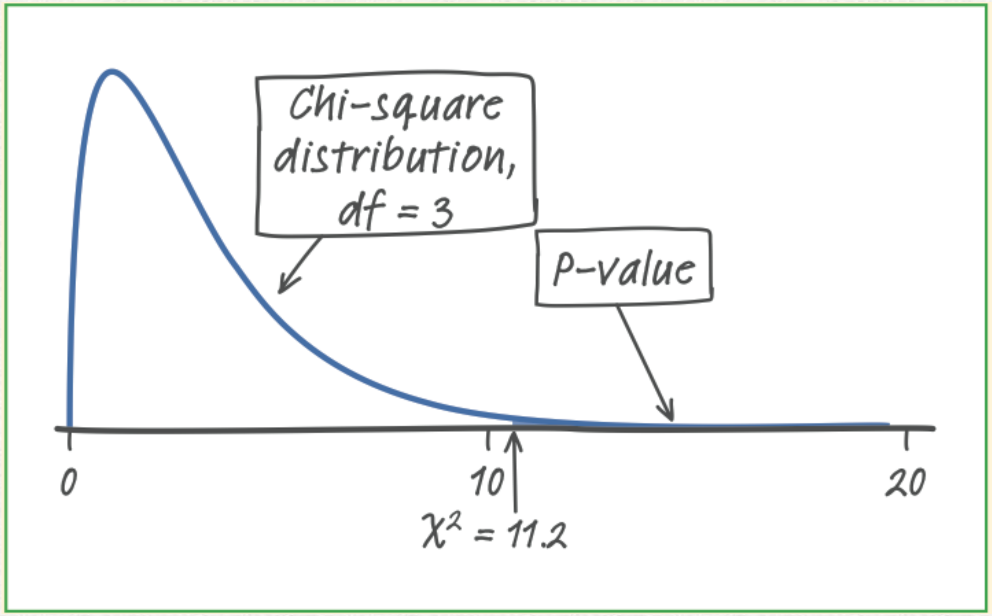
# Example: A test for equal proportions

**Do**: Test statistic

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

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

$$= 11.2$$




As the excerpt shows, $\chi^2$ corresponds to a *P*-value between 0.01 and 0.02.

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

# Example: A test for equal proportions

**Conclude**:

Because the $P$-value, 0.011, is less than $\alpha = 0.05$, we reject $H_0$.

We have convincing evidence that the birthdays of NHL players are not evenly distributed across the four quarters of the year.

# Chi-Square Test for Goodness of Fit

## Section Summary

In this section, we learned how to…

- ✓ STATE appropriate hypotheses and COMPUTE expected counts for a chi-square test for goodness of fit.
- ✓ CALCULATE the chi-square statistic, degrees of freedom, and *P*-value for a chi-square test for goodness of fit.
- ✓ PERFORM a chi-square test for goodness of fit.
- ✓ CONDUCT a follow-up analysis when the results of a chi-square test are statistically significant.