11.1.1 Comparing Observed and Expected Counts: The Chi-Square Test Statistic

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

*On average, 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.*

Jerome did an M&M’S Activity above recorded the different colors of M&M’s chocolate candies in a bag. The *one-way table* below summarizes the data from Jerome’s 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.

<table>
<thead>
<tr>
<th>Color:</th>
<th>Blue</th>
<th>Orange</th>
<th>Green</th>
<th>Yellow</th>
<th>Red</th>
<th>Brown</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count:</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>60</td>
</tr>
</tbody>
</table>

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, Jerome might believe that something fishy is going on. We could use the one-sample z test for a proportion from *Chapter 9* to test the hypotheses

\[ H_0: p = 0.24 \]
\[ H_a: p \neq 0.24 \]

where \( p \) is the true population proportion of blue M&M’S. We could then perform additional significance tests for each of the remaining colors. Not only would this method be pretty inefficient, but it would also lead to the problem of multiple comparisons, (to be discussed in further detail later). More important, this approach wouldn’t tell us how likely it is to get a random sample of 60 candies with a color distribution that differs as much from the one claimed by the company as Jerome’s 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**.

As with any test, we begin by stating hypotheses. 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 the case of the M&M’S Milk Chocolate Candies, the categorical variable we’re measuring is color. The appropriate null hypothesis is:

\[ H_0: \text{The company’s stated color distribution for M&M’S Milk Chocolate Candies is correct.} \]

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

\[ H_a: \text{The company’s stated color distribution for M&M’S Milk Chocolate Candies is not correct.} \]
We can also write the hypotheses in symbols as:

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

where \( p_{\text{color}} \) = the true population proportion of M&M’S Milk Chocolate Candies of that color.

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

\[ H_a: P_{\text{blue}} \neq 0.24, P_{\text{orange}} \neq 0.20, P_{\text{green}} \neq 0.16, \]
\[ P_{\text{yellow}} \neq 0.14, P_{\text{red}} \neq 0.13, P_{\text{brown}} \neq 0.13, \]

The idea of the chi-square 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. (Remember: we always assume that \( H_0 \) is true when performing a significance test.) The more the observed counts differ from the expected counts, the more evidence we have against the null hypothesis. In general, the expected counts can be obtained by multiplying the proportion of the population distribution in each category by the sample size. Here’s an example that illustrates the process.

**Observed Counts** - The actual numbers of individuals in the sample that fall in each cell of the one-way or two-way table.

**Expected Counts** - The expected numbers of individuals in the sample that would fall in each cell of the one-way or two-way table if \( H_0 \) were true.

Example – Return of M&M’s

Computing Expected Counts

Jerome’s bag of M&M’S Milk Chocolate Candies contained 60 candies. Calculate the expected counts for each color. Show your work.

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 find the expected counts for the other color categories:

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>9</td>
<td>14.40</td>
</tr>
<tr>
<td>Orange</td>
<td>8</td>
<td>12.00</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
<td>9.60</td>
</tr>
<tr>
<td>Yellow</td>
<td>15</td>
<td>8.40</td>
</tr>
<tr>
<td>Red</td>
<td>10</td>
<td>7.80</td>
</tr>
<tr>
<td>Brown</td>
<td>6</td>
<td>7.80</td>
</tr>
</tbody>
</table>
**Chi-square statistic:** A measure of how far the observed counts are from the expected counts. The formula is:

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

Example – Return of the M&M’s
Calculating the chi-square statistic

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

<table>
<thead>
<tr>
<th>Color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>9</td>
<td>14.40</td>
</tr>
<tr>
<td>Orange</td>
<td>8</td>
<td>12.00</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
<td>9.60</td>
</tr>
<tr>
<td>Yellow</td>
<td>15</td>
<td>8.40</td>
</tr>
<tr>
<td>Red</td>
<td>10</td>
<td>7.80</td>
</tr>
<tr>
<td>Brown</td>
<td>6</td>
<td>7.80</td>
</tr>
</tbody>
</table>

Think of \( \chi^2 \) as a measure of the distance of the observed counts from the expected counts. Like any distance, it is always zero or positive, and it is zero only when the observed counts are exactly equal to the expected counts. Large values of \( \chi^2 \) 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. Is \( \chi^2 = 10.180 \) a large value? You know the drill: compare the observed value 10.180 against the sampling distribution that shows how \( \chi^2 \) would vary in repeated random sampling if the null hypothesis were true.
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 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 peanut M&M’S.

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.
11.1.2 The Chi-Square Distributions and P-Values

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

Recall that larger values of $\chi^2$ give more convincing evidence against $H_0$. According to Fathom, 37 of the 500 simulated samples resulted in a chi-square statistic of 10.180 or higher. Our estimated $P$-value is $37/500 = 0.074$. Since the $P$-value exceeds the default $\alpha = 0.05$ significance level, we do not have enough evidence to refute the company’s claim about the color distribution.

As the figure suggests, 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 $\chi^2$ can never be negative. The sampling distribution of $\chi^2$ differs depending on the number of possible values for the categorical variable (that is, on the number of categories).

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

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

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

When $df = 8$, for example, the chi-square distribution has a mean of 8 and a mode of 6.
Example – Return of the M&M’s
Finding the P-value

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

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

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

<table>
<thead>
<tr>
<th>$df$</th>
<th>.15</th>
<th>.10</th>
<th>.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6.74</td>
<td>7.78</td>
<td>9.49</td>
</tr>
<tr>
<td>5</td>
<td>8.12</td>
<td>9.24</td>
<td>11.07</td>
</tr>
<tr>
<td>6</td>
<td>9.45</td>
<td>10.64</td>
<td>12.59</td>
</tr>
</tbody>
</table>

Learn Finding $P$-values for chi-square tests on the calculator
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 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. Confirm that the expected counts are large enough to use a chi-square distribution. Which distribution (specify the degrees of freedom) should we use?

2. Sketch a graph that shows the $P$-value.

3. Use the chi-square table to find the $P$-value. Then use your calculator’s $\chi^2$cdf command.

11.1.3 Carrying Out a Test

Like our test for a population proportion, the chi-square goodness-of-fit test uses some approximations that become more accurate as we take more observations. Our rule of thumb is that all expected counts must be at least 5. This **Large Sample Size condition** takes the place of the Normal condition for \( z \) and \( t \) procedures. To use the chi-square goodness-of-fit test, we must also check that the Random and Independent conditions are met.

### The Chi-Square Goodness-of-Fit Test

Suppose the Random, Large Sample Size, and Independent conditions are met. To determine whether a categorical variable has a specified distribution, expressed as the proportion of individuals falling into each possible category, perform a test of

- **\( H_0 \):** The specified distribution of the categorical variable is correct.
- **\( H_a \):** The specified distribution of the categorical variable is not correct.

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

\[
H_0: \ p_1 = \_\, \_\, \_\, \_\, \_\, \_\, \_, \ p_2 = \_\, \_\, \_\, \_\, \_\, \_\, \_, \ \ldots, \ p_k = \_\, \_\, \_\, \_\, \_\, \_\, \_.
H_a: \ \text{At least one of the } p_i \text{'s is incorrect.}
\]

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.

**Conditions:** Use this 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).

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.
Example – When Were You Born?
A test for equal proportions

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:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Births:</td>
<td>13</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>27</td>
<td>18</td>
<td>15</td>
</tr>
</tbody>
</table>

Do these data give significant evidence that local births are not equally likely on all days of the week?
Learn Chi Square goodness-of-fit test on the calculator

Example – Inherited Traits
Testing a genetic model

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

<table>
<thead>
<tr>
<th>Parent 1 passes on:</th>
<th>G</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent 2 passes on:</td>
<td>GG</td>
<td>Gg</td>
</tr>
<tr>
<td>G</td>
<td>Gg</td>
<td>gg</td>
</tr>
</tbody>
</table>

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.
CHECK YOUR UNDERSTANDING

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

<table>
<thead>
<tr>
<th>Parent 1 passes on:</th>
<th>Parent 2 passes on:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RC</td>
</tr>
<tr>
<td>RC</td>
<td>RRCC (x)</td>
</tr>
<tr>
<td>Rc</td>
<td>RRCC (x)</td>
</tr>
<tr>
<td>rC</td>
<td>RrCC (x)</td>
</tr>
<tr>
<td>rc</td>
<td>RrCc (x)</td>
</tr>
</tbody>
</table>

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

To test their hypothesis about the distribution of offspring, the biologists mate a random sample of pairs of fruit flies. Of 200 offspring, 99 had red eyes and straight wings, 42 had red eyes and curly wings, 49 had white eyes and straight wings, and 10 had white eyes and curly wings. Do these data differ significantly from what the biologists have predicted? Carry out a test at the $\alpha = 0.01$ significance level.
11.1.4 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

\[
\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}
\]

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

Let’s return to the tobacco plant example. The table of observed and expected counts for the 84 offspring is repeated below. Looking at these counts, we see that there were many more yellow-green tobacco offspring than expected and many fewer albino plants than expected.

<table>
<thead>
<tr>
<th>Offspring color</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>Yellow-green</td>
<td>50</td>
<td>42</td>
</tr>
<tr>
<td>Albino</td>
<td>11</td>
<td>21</td>
</tr>
</tbody>
</table>

The chi-square statistic is

\[
\chi^2 = \frac{(23 - 21)^2}{21} + \frac{(50 - 42)^2}{42} + \frac{(11 - 21)^2}{21} = 0.190 + 1.524 + 4.762 = 6.476
\]

We see that the component for the albino offspring, 4.762, made the largest contribution to the chi-square statistic.

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