**Section 11.1 - Chi-Square Goodness-of-Fit Tests** (pp. 678-695)

**M&M’S Activity** - In the activity, we were determining if there was enough evidence to dispute Mars, Inc.’s claimed color distribution. We could write this as hypotheses:

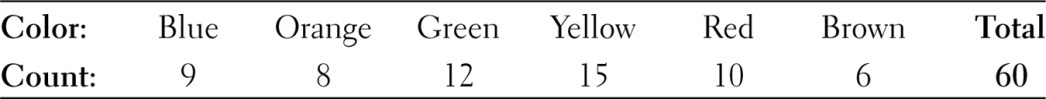
*H0: The company’s stated color distribution for M&M’S is correct.*

*HA: The company’s stated color distribution for M&M’S is not correct.*

We can also state the hypotheses in symbols:

*H0: pblue = 0.24, porange = 0.20, pgreen = 0.16, pyellow = 0.14, pred = 0.13, pbrown = 0.13*

*Ha:; At least one of the pi’s is incorrect*

Let us assume the following counts were found:   
  


To calculate what we would expect to see (*expected counts*), we use the claimed proportions and the total:

Blue:

Orange:

Green:

Yellow:

Red:

Brown:

To measure how far the observed counts are from the expected counts, we are going to use the **chi-square statistic**:

For our example:

As you probably suspect, large values of are stronger evidence against *H0* because they say that the observed counts are far from what we would expect if *H0* were true.

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

|  |  |
| --- | --- |
| F11.03.jpg | **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.  When looking at the distribution of , if the *expected* counts are all at least 5, the distribution is chi-square. The chi-square goodness-of-fit test uses the chi-square distribution with degrees of freedom equal to the number of categories minus 1. |

Then *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.

To compute ***P*-values**, we can either used Table C or technology.

|  |  |
| --- | --- |
| **Example** - In the M&M’S example, we had a chi-square statistic of = 10.180.  There are 6 color categories, making df = 6 - 1 = 5.  To find the *P*-value from Table C, enter with df = 5 and “tail probability” of 0.05. The *P*-value is 11.07.  To use technology, use cdf with entry arguments *lower bound, upper bound, df* | F11.04.jpg |

**Application** - Suppose a different sample of M&M’S gave: 10 blue, 7 orange, 12 green, 14 yellow, 11 red and 6 brown.

a) Find the expected counts and confirm that they are large enough to use a chi-square distribution.

b) Compute the chi-square statistic.

c) Sketch a graph that shows the *P*-value.

d) Use Table C to find the *P*-value and then confirm it with your calculator.

**Carrying Out a 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 specific distribution, expressed as the proportion of individuals falling into each possible category, perform a test of

*H0: p1 = \_\_\_, p2 = \_\_\_, . . . , pn = \_\_\_ .*

*Ha: At least one of the pi’s is incorrect.*

Start by finding the expected count for each category assuming that *H0* is true. Then calculate the chi-square statistic

where the sum is over the *k* different categories. The *P*-value is the area to the right of 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 (10% condition).

**Example** - in the 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 to the NHL), 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?

**Technology** - To perform the chi-square goodness-of-fit test on the calculator, put the observed counts in L1 and the expected counts in L2. Choose [STAT} <TESTS> D:GOF-Test

**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 . These *components* show which terms contribute most to the test statistic.

HW: Read Sec 11.1; do problems 1-9 odd, 17, 19-22, 25, 26 on pp. 692-695.