AP Statistics (1.69oz)

"The Candy Man Can"

Name: MA10 (S-2014)

Background: Mars, Incorporated, makes milk chocolate candies including M&M'S Milk Chocolate Candies. They claim that, *"On average, the new mix of M&M'S Milk Chocolate Candies will contain 13% of each of browns and reds, 14% yellows, 16% greens, 20% oranges, and 24% blues."*

Activity: The purpose of this activity is to compare the color distribution of M&M'S in your individual bag with the advertised distribution. We want to see if there is enough evidence to dispute the company's claim. For the time being, assume that your bag is a random sample of M&M'S Milk Chocolate Candies from the population of candies produced in a particular batch.

1. Open your bag and carefully count the number of M&M'S of each color as well as the total number of M&M'S in the bag. Complete the observed column in the table below. (From p. 678)

	Color	**Observed**	**Expected**	**Observed - Expected**	**(Observed - Expected)^2**	$\frac{(Observed - Expected)^2}{Expected}$
0.24	Blue	9	(0.24)(60) = 14.40	-5.4	29.16	2.025
0.2	Orange	8	(0.2)(60) = 12.00	-4.0	16.00	1.333
0.16	Green	12	(0.16)(60) = 9.60	2.4	5.76	0.600
0.14	Yellow	15	(0.14)(60) = 8.40	6.6	43.56	5.186
0.13	Red	10	(0.13)(60) = 7.80	2.2	4.84	0.621
0.13	Brown	6	(0.13)(60) = 7.80	-1.8	3.24	0.415
	Total	60	60.00 ✓	0		10.180 = χ^2

2. Assuming that the company's claim is true, how many of each color would you expect in *your* bag? These are called the *expected counts*. Compute your expected counts (to two decimal places) and record your results in the expected column. Check that the sum is equal to the sum of M&M'S in your bag.

3. How close are your observed counts to the expected counts? To answer this, compute Observed - Expected for each color. Record these differences in your table. Find the sum of the values in this column. What do you notice about this sum? (= 0)

4. In step 3, the sum of the differences should be 0. You get a total difference of 0 because the positive and negative values cancel each other out. We can fix this by squaring the differences. Compute the values of (Observed - Expected)^2 and fill in this column in the table. Compare your results with other teams.

5. In the last column of the table, divide the values of (Observed - Expected)^2 by the corresponding Expected values then find the sum. This final value is called the *chi-square statistic* and is denoted by χ^2.

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:

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

H_A: The company's stated color distribution for M&M'S is not correct.

We can also state the hypotheses in symbols:

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

Let us assume the following counts were found:

Color:	Blue	Orange	Green	Yellow	Red	Brown	Total
Count:	9	8	12	15	10	6	60

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

Blue: $(0.24)(60) = 14.40$

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: $(0.13)(60) = 7.80$

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

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

FORMULA SHEET.

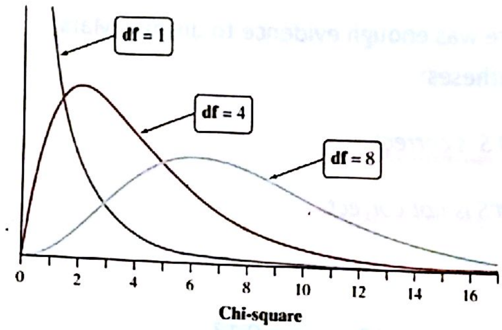
For our example:

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

$$= 2.025 + 1.333 + 0.600 + 5.186 + 0.621 + 0.415 = \boxed{10.180}$$

As you probably suspect, 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.

Chi-Square Distributions and *P*-Values



df = c - 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.

When looking at the distribution of $\frac{(Observed-Expected)^2}{Expected}$, 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.

*

The *mean* of a particular chi-square distribution is equal to its degrees of freedom.

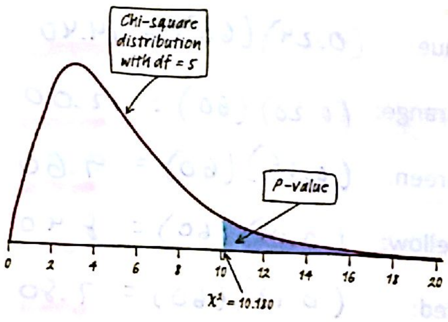
For df > 2, the *mode* (peak) of the chi-square density curve is at df - 2.

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 $\chi^2 = 10.180$.
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.
CRITICAL VALUE.

To use technology, use χ^2cdf with entry arguments *lower bound, upper bound, df*



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.

SAME EXPECTED COUNTS SINCE Σ = 60 BLUE = 14.40, ORANGE = 12.00, GREEN = 9.60, YELLOW = 8.40, RED = 7.80, BROWN = 7.80 $\boxed{>5 \checkmark}$

b) Compute the chi-square statistic.

$$\chi^2 = \frac{(10-14.40)^2}{14.40} + \frac{(7-12)^2}{12} + \frac{(12-9.6)^2}{9.6} + \frac{(14-8.4)^2}{8.4} + \frac{(11-7.8)^2}{7.8} + \frac{(6-7.8)^2}{7.8} = 9.49$$

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

DISCUSS LISTS.

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

df = 6 - 1 = 5 ~~α = 0.05~~

CALC: 0.0910

$0.05 < p < 0.10$

χ^2CDF LOWER UPPER DF

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

H_0: p_1 = ___, p_2 = ___, . . . , p_n = ___.   IS SAME AS CLAIMED DISTR.

H_a: *At least one of the p_i's is incorrect.*   IS DIFF THAN CLAIMED DISTR.

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

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

where the sum is over the *k* different categories. The *P*-value is the area to the right of χ^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 (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?

n=80

STATE: WE WANT TO TEST THE FOLLOWING HYPOTHESES AT $\alpha = 0.05$.

H_0: BIRTHDAYS OF NHL PLAYERS ARE EQUALLY LIKELY TO OCCUR IN EACH QTR OF THE YR.

H_a: BIRTHDAYS OF NHL PLAYERS ARE NOT EQUALLY LIKELY TO OCCUR IN EACH QTR OF THE YR.

PLAN: IF COND'S ARE MET, WE WILL PERFORM A χ^2 GOODNESS OF FIT TEST.

- RANDOM: RANDOM SAMPLE. ✓
- LARGE SAMPLE SIZE: EXPECTED COUNTS = $\frac{1}{4}(80) = 20 > 5$ ✓
- INDEPENDENT: IT IS SAFE TO ASSUME MORE THAN $10(80) = 800$ NHL PLAYERS ✓   ∴ COND'S ARE MET.

DO: TEST STATISTIC – $\chi^2 = \frac{(32-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(12-20)^2}{20}$

$= \boxed{11.2}$

P-VALUE: USING 4-1 = 3 df P-VALUE = $\boxed{0.011}$ (TECHNOLOGY)

CONCLUDE: BECAUSE $P = 0.011 < 0.05 = \alpha$, REJECT H_0. WE HAVE CONVINCING EVIDENCE THAT THE BIRTHDAYS OF NHL PLAYERS ARE NOT UNIFORMLY DISTRIBUTED THROUGHOUT THE YR.

$(\chi^2 = 7.2 + 0 + 0.8 + 3.2 = 11.2)$

Technology - To perform the chi-square goodness-of-fit test on the calculator, put the observed counts in L_1 and the expected counts in L_2. Choose [STAT] <TESTS> D: χ^2GOF-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 $\frac{(Observed-Expected)^2}{Expected}$. These *components* show which terms contribute most to the test statistic.

FROM THE NHL EXAMPLE THE FIRST QTR HAD THE LARGEST CONTRIBUTION TO THE χ^2 STATISTIC WHICH IS IN CONCERT WITH MR. GLADWELL'S THEORY.

CALCULATOR GIVES CONTRIBUTIONS.

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

1-9 ODD, 17, 19-22, 25, 26