Exploring the χ^2 (Chi Square) Distribution – Goodness of Fit Test

The χ^2 Distribution is used for inference for categorical variables. It can answer questions such as, "Are the colors of M&Ms distributed as the company says?", or "Do the phenotypes for a particular gene of the cross of two different types of roses follow the theoretical percentages dictated by the punnet square? In other words, how well does the observed data "fit" what is expected?

Let's look at the M&M distribution and follow the **4 steps for an inference test**. According to the Mars company website, the distribution of plain M&Ms is: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown. Therefore, for the χ^2 test:

Step 1: The population is all of the M&Ms and the parameter of interest is the proportions of the colors of M&Ms

The **hypotheses** are:

H_0: the true proportion of blue = .24, the true proportion of orange = .2, …, the true proportion of red = .13.

H_a: the proportions are different

NOTE: The alternative hypothesis is a general statement that the proportions are different. This does not necessarily mean ALL of the proportions are different. Just one or two being different (significantly) could yield a small p-value.

Step 2: State the inference procedure: **Chi Square Goodness of Fit Test** and verify the conditions have been met.

NOTE: To use inference for the χ^2 Distribution, you MUST verify that all expected counts are > 1, and 80% or more of the expected counts are > 5. We will need to begin our procedure on the handheld.

Using a King size bag of plain M&Ms, sort the M&Ms into their colors.

1. Open a new document on your handheld and add a Lists & Spreadsheet.

2. Name the first column **mmcolor**, and enter the colors of the M&Ms. Remember to use quotes when entering categorical information.

3. Enter the counts of your M&Ms in column B, naming it **obscounts** (observed counts).

4. In column C, name it **mms** and create a frequency list for the colors of M&Ms. In the formula bar type freqTable(**mmcolor**, **obscounts**).

5. Graph **mms** on a separate Data & Statistics page.

List the percentages of the colors in your bag shown in the bar chart or pie chart. Based on this, do you think your bag of M&Ms follow the distribution stated by the company. Support your answer.

6. In your spreadsheet, name the next free column **expercent** (expected percent) and type the percent distribution for the colors of M&Ms.

For the χ^2 test, we need to calculate the **expected counts**. This is accomplished by finding the product of the total number of M&Ms in the bag by the expected percent. NOTE: Your number may not be a whole number.

7. In the next column, named **expcounts**, use a product of the **sum**(**obscount**) and **expercent** to calculate the expected counts.

Check the expected counts and verify that the conditions for inference for Chi Square are satisfied. Use supporting statements.

Write your counts on the board as we will be performing a χ^2 test on the class counts.

Step 3: Calculate the test statistic and find the p-value.

The χ^2 test uses the following formula: $\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

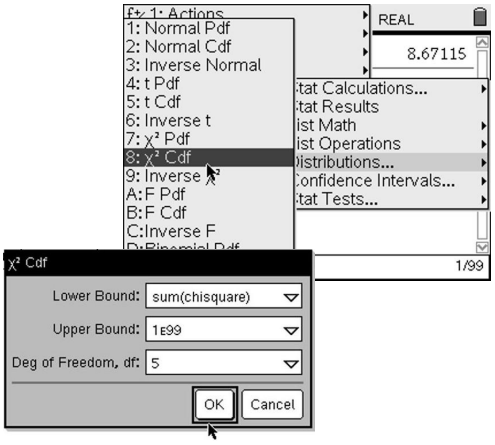
8. In the next column, named **chisquare**, use **obscounts** and **expcounts** to calculate the χ^2 statistic.

State the critical list of values you just calculated. Which colors seem to have a particularly high value compared to the others? Do you think this will be significant? Support you answer.

9. On a calculator page, find the **sum(chisquare)**.

Now, let's calculate the p-value of this statistic. The **degrees of freedom** for the χ^2 Goodness-of-Fit test is the number of items in the column – 1. For this example, it would be 6 colors – 1, so 5 **degrees of freedom**.

10. On the calculator page, press (menu) , *Statistics*, *Distributions*, χ^2 *cdf*.



11. In the dialogue box, you will be asked for the Lower Bound (**sum**(**chisquare**)), Upper Bound (1E99), and degrees of freedom. Fill in the appropriate numbers as listed on the screen shot here.

List the p-value you calculated.

Step 4: Interpret your results in the context of the problem.

What can you conclude based on the p-value? What does this mean about the M&Ms?

12. Once the entire class has written their counts on the board, add the counts for each color and enter this data into a separate column **classcounts**.

13. Create a frequency list for **classcounts** and name this column **freqclass**. Graph this column in a separate Data & Statistics page.

List the percentages of the colors for the class in the bar chart or pie chart. Based on this, do you think the class's total M&Ms follow the distribution stated by the company. Support your answer.

14. Calculate the expected counts for the class, **classexp**, in the next column→ **sum(classcounts)*expercent**.

15. Using **classexp** and **classcounts**, calculate the **classchisq** critical values in the next column.

List the critical values. Are there any extremely large values? Does this lead you to believe you will find significant evidence that the distribution of M&Ms is different for the class?

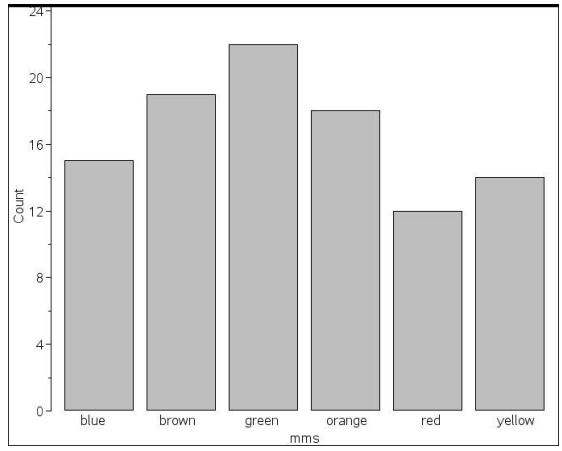
16. Calculate the χ^2 statistic on the calculator page and use the χ^2 cdf in Stat Distributions to find the p-value.

State the p-value and interpret it in context of the problem.

Here are the screen shots for a student sample, based on data a student of mine had collected in a previous AP Statistics class.

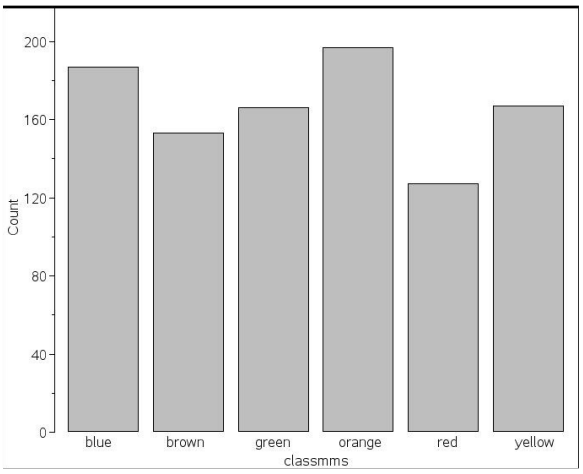
	A mmcol...	B obsc...	C mms	D expercent	E expcounts	F chisquare	G	H
◆			=freqtable		=sum(obsco	=('obscounts		
1	blue	15	blue	0.24	24.	3.375		
2	orange	18	blue	0.2	20.	0.2		
3	green	22	blue	0.16	16.	2.25		
4	yellow	14	blue	0.14	14.	0.		
5	red	12	blue	0.13	13.	0.076923		
6	brown	19	blue	0.13	13.	2.76923		
7			blue					
8			blue					
9			blue					
10			blue					
11			blue					
12			blue					
13			blue					
14			blue					
15			blue					
16			orange					

F $\mathbf{chisquare}:=\frac{(\mathbf{obscounts}-\text{'}\mathbf{expcounts})^2}{\text{'}\mathbf{expcounts}}$



sum(*chisquare*)	8.67115
χ^2Cdf(sum(*chisquare*),1.E99,5)	0.122922
χ^2Cdf(sum(*classchisq*),1.E99,5)	0.000688

3/99

	A classc...	B experc...	C classexp	D classchisq	E class...	F	G	H
◆			=sum(classe	=(classcount	=freqtable			
1	187	0.24	239.28	11.4226	blue			
2	197	0.2	199.4	0.028887	blue			
3	166	0.16	159.52	0.26323	blue			
4	167	0.14	139.58	5.38656	blue			
5	127	0.13	129.61	0.052558	blue			
6	153	0.13	129.61	4.22106	blue			
7					blue			
8					blue			
9					blue			
10					blue			
11					blue			
12					blue			
13					blue			
14					blue			
15					blue			
16					blue			

D $\mathbf{classchisq}:=\frac{(\mathbf{classcounts}-\mathbf{classexp})^2}{\mathbf{classexp}}$