Topic 5: Genetics – 5a. Chi-Square Analysis of Data

Resources:

Triola, M. (1992). *Elementary Statistics,* 5th edition. New York, NY: Addison-Wesley Publishing Co.

Williams, J. Probability, Chi Square, and Pop Beads [Internet]. Southwest Tennessee Community College. Cited 28 July 2009.
Available from: http://faculty.southwest.tn.edu/jiwilliams/probability.htm

Eck, D., Ryan, J. The Chi Square Statistic [Internet]. Hobart and Williams Colleges: Mathbeans Project, National Science Foundation. Cited 28 July 2009. Available from: http://math.hws.edu/javamath/ryan/ChiSquare.html

Building on:

The study of *genetics* involves data and *data analysis.* In the introduction of *Mendelian genetics,* phenotypic and *genotypic* ratios are used to explain *dominant* and *recessive* traits. When students start to do actual *genetic crosses* with fruit flies or with albino tobacco seeds, etc. the data is usually far from perfect. The question arises, “How close to perfect does the data need to be?” This brings up the topic of *quantitative* versus *qualitative* data as well the concept of *subjective* versus *objective* decisions. In science, quantitative data is more valuable than qualitative data and objective decisions are very important. Chi-square analysis allows the scientist to take quantitative data that is less than perfect and make an objective decision about its meaning.

Links to Chemistry and Physics:

Statistics
Data analysis
Error analysis

Stories:

Chi-square statistical analysis has lots of applications. There are three ways that chi-square is traditionally used in statistics:

1. Goodness of Fit: This is the chi-square application that is used in this activity. It is used to compare collected data to an expected distribution.

2. 2 x 2 Contingency Table: This is used to compare the numerical responses between two independent groups. As an example, a 2 x 2 contingency table would be used to see if the use a particular fertilizer increases the survival of a particular crop. The plants would be separated into two groups, one receiving the fertilizer and the other receiving no fertilizer. The number of plants dying and surviving would be measured and the chi-square analysis would indicate whether or not the fertilizer made a significant difference in the survival of the

plants. The formula for calculating chi-square for 2 x 2 contingency tables is different from the goodness-of-fit formula.

3. Chi-Square of Independence: This statistical measure determines if two categories show dependence on each other. An example would be to compare the number of people with breast cancer, lung cancer, and liver cancer in four different cities in the United States to see if the types of cancers occurring are dependent or independent of the cities where the people live. The formula for this chi-square analysis is similar to the formula for goodness of fit. Note that this analysis would tell you if there was a relationship between the city and the type of cancer, but it would not identify the actual relationship.

For a good discussion and examples of these types of chi-square, go to the following website: http://math.hws.edu/javamath/ryan/ChiSquare.html

Materials for the Lab:

- Plain M&M's – It is best to stock up on these around Halloween when you can buy big bags that contain the small bags usually given out at Trick-or-Treating; if that is not possible, buy a large bag of plain M&M's and using a scoop, scoop them into individual zip lock bags and hand them out.
- Calculators are very helpful. Some students may have statistical calculators or calculators that have statistical functions.

Chi-Square “Goodness-of-Fit” Analysis Lab

Introduction: Have you ever wondered why the package of M&M’s you just bought never seems to have enough of your favorite color? Or why does your package have mainly brown M&M’s? I’ll bet you’ve stayed up nights thinking about this. . . .

Mars Company estimates that each package of plain M&M’s they sell should have the following percentages of each color of M&M’s:

<u>M&M’s Plain</u>
13% Brown
14% Yellow
13% Red
20% Orange
16% Green
24% Blue

One way to determine if the Mars Company is accurate is to sample a package of M&Ms and do a type of statistical test known as a “goodness-of-fit” or chi-squared test. This allows us to determine if any differences between or observed measurements and the expected measurements are simply due to chance sample errors or are the differences due to some other factor (such as, the Mars Company sorters didn’t do a very good job). This test is generally used when we are dealing with discrete data (i.e., count data). This test will be using a table to determine a probability for getting this particular chi-square and will tell us what the chances are that the differences in our data are due to simply a chance sample error.

Start with the Null Hypothesis: If the Mars Company sorters are doing their job and the company statistic is accurate, then there should be no difference in the M&M color ratios between actual bags of M&M’s and the color ratios the company predicts.

$$\text{Chi-Square Formula: } \Sigma \frac{(obs - \exp)^2}{\exp}$$

Note that when expected and observed are equal, chi-square is zero and as the difference between observed and expected increases chi-square increases.

The **critical value** is the maximum value of chi-square that will allow you to accept the null hypothesis that there is no significant difference between the data observed and the data expected. Any small differences would be due to random chance.

Evidence Table

Colors	Observed	Expected	Deviation	Deviation Squared	D2/E
Brown					
Yellow					
Red					
Orange					
Green					
Blue					
Total					

Note the "degrees of freedom." The reason why it is important to consider degrees of freedom is that the value of the chi-square statistic is calculated as the sum of the squared deviations for all of the possible categories. The more categories considered increases the chance for random error; therefore, the critical value for chi-square must increase as the degrees of freedom increase.

Degrees of freedom = number of categories – 1.

To view a chi-square critical value table, go to this website:
http://faculty.southwest.tn.edu/jiwilliams/probab2.gif

Notice that for 5 degrees of freedom, a chi-square value as large as 1.61 would be expected by chance in 90% (.9) of the cases, whereas a critical value as large as 11.07 would only be expected by chance in 5% (.05) of the cases. The column that we need to concern ourselves with is the one under "0.05. Scientists, in general, are willing to say that if their probability of getting the observed deviation from the expected results by chance is greater than 0.05 (5%) then we can accept the null hypothesis. In other words, there is really no difference in the ratios between what was observed and what was expected.

Therefore, the critical value for chi-square of six different colors of M&M's is 11.07. If your chi-square value is below 11.07, then you can accept the null hypothesis and conclude that any deviation was due to random chance. In other words, the sorts are doing a good job!