Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Chi Square (X2) Goodness of Fit Modeling Using Candy**

The Chi Square test is often used in science to determine if data you observe from an experiment is close enough to the predicted data. In genetics, for instance, you might expect to get a 75% to 25% ratio if you crossed two heterozygous tall plants (Tt x Tt). *Calculating the X2 values help you determine whether the results follow the prediction and if the variations from the exact ratio are due to random chance.* It's the question of "how close is close enough?" If the numbers differ greatly from your expected results, then it's possible that other factors may be influencing your results.

The Chi-square test is intended to test how likely it is that an observed distribution is due to chance. It is also called a **"goodness of fit"** statistic, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

Another way to describe the Chi-square test is that it tests the **null hypothesis** that the variables are independent, that there is no relationship between the two things being tested. Wherever the observed data do not fit the model, the likelihood that the variables are dependent becomes stronger.

*According to the manufacturer of M&M candy, the color distribution for plain chocolate M&Ms in 2016 was 13% brown, 13% red, 14% yellow, 24% blue, 20% orange, and 16% green.*

**ScientificQuestion: Are the M&Ms in our sample distributed in the same frequencies in 2019 as they were in 2016?**

**Null Hypothesis H0**: There is no statistical difference between observed and expected numbers.(The distribution in our 2019 sample is the same as the manufacturer states for 2016).
**Alternate/Predicting Hypothesis HA**: One (or more) colors does not match the expected values.

**Procedure:**

1) Sort the candy onto a paper towel and write down the number of each color into Table 1 under **"Number Observed"**

2) Complete the table by determining the **total** number of candies and **"number expected"** columns by multiplying the TOTAL number of candies by the % Expected.

*For example: if my total number of candies is 50 and my % expected is 16%, 50 X 0.16 is 8 candies as the number expected.*

Data Table 1:

|  |  |  |  |
| --- | --- | --- | --- |
| Color of Candy | Percentage Expected | Number Observed | Number Expected *(total # of candy x percentage expected)* |
|  brown |  13% |   |   |
|  red |  13% |   |   |
|  yellow |  14% |   |   |
|  blue |  24% |   |   |
|  orange |  20% |   |   |
|  green | 16% |  |   |
|  |  |  Total # of candies = |  |

As you look at the data above, consider the two comparable numbers. The number you would expect to count if your percentage estimate was correct, and then the number you actually counted (number observed). For example, if you initially thought that you'd see 14% yellow candies, and you counted 200 pieces, you would then expect to see 28 yellow candies. You may have only counted 26 yellows. Is this “close enough” and can we say this data is accurate?

The Chi Square (*X2*) Equation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Classes (colors) | Expected (e) | Observed (o) | formula |
| 1 |  brown |   |   |   |
| 2 |  red |   |   |   |
| 3 |  yellow |   |   |   |
| 4 |  blue |   |   |   |
| 5 |  orange |   |   |   |
| 6 | green |  |  |  |
|  | Sum (add the values from row 1-5); this is your *X2* value |  |

Use the chi square chart below to determine if your *X2* rejects or fails to reject (accepts) your hypothesis. The **degrees of freedom** is determined by subtracting 1 from the number of colors you analyzed. (For example, if you had 4 colors to count, the degrees of freedom is 3) **Circle the number in the chart below that is closest to your *X2*.**

Fail to Reject Reject

Null Hypothesis Null Hypothesis



**Summary and Analysis**

1) What was your initial null hypothesis?

 2) Would you reject or fail to reject the null hypothesis? Why?

3) Explain what is meant by a "good fit" of your data?

4) Enter your data into the google doc for class totals. Repeat the *X2* and *p* value calculations using the class data.

Data Table 2: Class Totals

|  |  |  |  |
| --- | --- | --- | --- |
| Color of Candy | Percentage Expected | Number Observed | Number Expected *(total # of candy x percentage expected)* |
|  brown |  13% |   |   |
|  red |  13% |   |   |
|  yellow |  14% |   |   |
|  blue |  24% |   |   |
|  orange |  20% |   |   |
|  green | 16% |  |   |
|  |  |  Total # of candies = |  |

*X2* Data Table Class Totals:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Classes (colors) | Expected (e) | Observed (o) | formula |
| 1 |  brown |   |   |   |
| 2 |  red |   |   |   |
| 3 |  yellow |   |   |   |
| 4 |  blue |   |   |   |
| 5 |  orange |   |   |   |
| 6 | green |  |  |  |
|  | Sum (add the values from row 1-5); this is your ***X2* value** |  |
|  | Use the probability chart to determine your ***p value*** |  |

5) Explain how the *X2* and *p* value of the class data compare to your group’s data.