AP Biology

Chi Square Analysis
M & M Statistics

Introduction

Consider a trait that exhibits the pattern of simple dominance. If we were to cross two heterozygous individuals (i.e., Aa x Aa), then we would expect a 3:1 ratio of dominant to recessive phenotypes in the offspring.

But what if we actually did this cross and did not get the expected 3:1 ratio? The difference from the expected ratio could be due to random chance or some type of sampling error. But it is also possible that the difference from the expected ratio is due to the fact that our original expectation was incorrect (i.e., the trait does not actually exhibit the pattern of simple dominance).

How can we determine which of these is the most likely cause of the difference between our expectation and our actual observation?

We can conduct a Chi-Square (χ^2) analysis!

Background Information

When starting a Chi-Square analysis, we must first identify the null hypothesis. A null hypothesis is a prediction that something is not present, that a treatment will have no effect, or that there is no difference between a treatment and a control. Another way of saying this is the hypothesis that an observed pattern of data and an expected pattern are effectively the same, differing only by chance, not because they are truly different.

The null hypothesis is for a Chi-Square analysis is ALWAYS the same:

Any difference between the observed and expected data is due to CHANCE.

The goal of the Chi-Square analysis is to confirm or refute this null hypothesis.

Once we have calculated a value for the Chi-Square, we will compare it to a table of critical values. If the calculated Chi-Square value is smaller than the critical value, we ACCEPT our null hypothesis because our data is consistent with what we would expect—any slight difference is due to chance. If the calculated Chi-Square is larger than the critical value, we REJECT our null hypothesis because our data is too different from what was expected to explain the differences by chance—there must be some other explanation.

This investigation will let you practice using the Chi-Square test with a "population" of familiar objects, M&M® candies. Later on, we will use this same method to analyze the outcome of our fruit fly crosses.

After completing the investigation you should be able to:

- write and test a null hypothesis that pertains to the investigation
- determine the degrees of freedom for an investigation
- calculate the χ^2 value from observed data
- determine if the Chi-Square value exceeds the critical value and if the null hypothesis is accepted or rejected

Let's get started!

A Candy-Coated Chi-Square?

Have you ever wondered why the package of M&Ms you just bought never seems to have enough of your favorite color? Why do you always seem to get the package of mostly brown M&Ms? What's going on at the Mars Company? Is the number of the different colors of M&Ms in a package really different from one package to the next? Or, does the Mars Company do something to insure that each package gets the correct number of each color?

Here is some information from the M&M website:

% color	Milk Chocolate	Peanut	Crispy	Minis	Peanut Butter	Almond
Brown	13%	12%	17%	13%	10%	10%
Blue	24%	23%	17%	25%	20%	20%
Orange	20%	23%	16%	25%	20%	20%
Green	16%	15%	16%	12%	20%	20%
Red	13%	12%	17%	12%	10%	10%
Yellow	14%	15%	17%	13%	20%	20%

One way that we could determine if the Mars Company is true to its word is to sample a package of M&Ms and do a type of statistical test known as a "goodness of fit" test. This type of statistical test allows us to determine if any differences between our observed measurements (counts of colors from our M&M sample) and our expected (what the M&M website claims) are simply due to chance or some other reason (i.e. the Mars Company's sorters are not putting the correct number of M&M's in each package). The goodness of fit test we will be using is called a Chi-Square (χ^2) Analysis.

We begin by stating the null hypothesis. Remember, the null hypothesis for a Chi-Square analysis is always the same. What is our null hypothesis for this experiment?

Null Hypothesis: _____

To test this hypothesis, we will need to determine the χ^2 value, which is calculated in the following way:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

"O" is the observed number (actual count) and "E" is the expected number (based on the information in the table above) for each color category. The "Σ" symbol means that we find the sum of the results of $(O-E)^2/E$ for each of the six color categories. The main thing to note about this formula is that, when all else is equal, the value of χ^2 increases as the difference between the observed and expected values increase.

Materials (per group of 2): 1 bag of M&Ms, 1 paper plate, clean hands

Procedure

1. Wash your hands with soap and water. You will be handling food that you may want to eat at the end of this activity.
2. Gather the materials listed above.
3. Open a bag of M&Ms and pour them out onto the paper plate. DO NOT EAT ANY OF THE M&M'S YET!
4. Separate the M&M's into color categories and count the number of each color you have.
5. Record your counts in the first row of Data Table 1 on the next page.
6. Use the table on the first page of this handout to calculate the expected number of each color. Record these numbers in the second row of Data Table 1.
7. Complete the calculations indicated in the remaining rows of Data Table 1 to determine the Chi-Square value for your data.

Data Table 1	Color Categories						
	Brown	Blue	Orange	Green	Red	Yellow	Total
Observed - (O)							
Expected - (E)							
Difference - (O-E)							
Difference Squared - $(O-E)^2$							
$(O-E)^2/E$							
$\chi^2 = \Sigma\ (O-E^2)/E$							

Once you have finished the table above, you may eat the M&Ms! ☺

Now, let's determine the probability that the difference between the observed and expected values occurred simply by *chance*. In order to do so, we must to compare the calculated value of the Chi-Square to the appropriate value in the table below.

First, examine the table. Note the term "degrees of freedom." For this statistical test, the degrees of freedom equal the number of classes (i.e. color categories) minus one:

degrees of freedom = number of categories –1

In your M&M experiment, what is the number of degrees of freedom? _____

The reason 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 classes. Therefore, the natural increase in the value of Chi-Square with an increase in classes must be taken into account.

Scan across the row corresponding to your degrees of freedom. Values of the Chi-Square are given for several different probabilities, ranging from 0.90 (90%) on the left to 0.01 (1%) on the right.

Remember that the Chi-Square value is a measure of the difference between the observed and expected numbers. We are using it to test whether the observed and expected numbers are close enough to accept the null hypothesis (that chance alone can explain the difference) or so far apart that the null hypothesis must be rejected.

Accept the null hypothesis ▮ Reject the null hypothesis

Degrees of Freedom	Probability					
	0.90	0.50	0.25	0.10	0.05	0.01
1	0.016	0.46	1.32	2.71	3.84	6.64
2	0.21	1.39	2.77	4.61	5.99	9.21
3	0.58	2.37	4.11	6.25	7.82	11.35
4	1.06	3.36	5.39	7.78	9.49	13.28
5	1.61	4.35	6.63	9.24	11.07	15.09

Please note that the probability decreases as the Chi-Square value increases. Therefore, the lower the Chi-Square value, the higher the probability that the difference between the observed results and the expected results is due to chance alone. Usually, a scientist is hoping to find a low Chi-Square value because it means there is a high probability that the deviation from the expected results is due to chance alone. This tells the scientist that the proposed explanation is likely to be correct. If, however, the Chi-Square value is high, it means that there is a low probability that the deviation is due to chance alone. This tells the scientist that the explanation is probably incorrect and that the true reason for the deviation is something other than chance alone. At that point, it's back to the drawing board!

Notice that dark black line separating the 0.10 and 0.05 probability columns? Here's why that is important.

If the probability of getting the observed deviation from the expected results by chance is greater than 5%, then a scientist will usually accept the null hypothesis. In other words, when the amount of deviation represented by the Chi-Square value is expected by chance more than 5% of the time, scientists DO NOT have a significant reason to reject the null hypothesis. Five percent may seem like a low probability, but it is enough for scientists to accept that the deviation is likely due to chance alone.

If, however, the probability of getting the observed deviation from the expected results by chance is less than 5%, then a scientist will usually reject the null hypothesis. In other words, when the amount of deviation represented by the Chi-Square value would be expected by chance less than 5% of the time, we DO have significant reason to reject the null hypothesis. There is more deviation than would be expected due to chance alone, and something else must be going on.

So, if your calculated Chi-Square value is less than the value listed on the appropriate degrees of freedom row in the table under 0.05, then you can ACCEPT your null hypothesis. This means that any differences between what the Mars Company claims and what is actually in a bag of M&Ms can be attributed to chance sampling errors, such as the fact that there are only around 50 M&Ms in a bag.

However, if your calculated Chi-Square value is greater than the value listed, then you must REJECT your null hypothesis. Any differences you observed between what the Mars Company claims and what is actually in a bag of M&Ms did not occur due to chance only. There must be some other explanation for the difference.

With all of this in mind, based on your individual sample, should you ACCEPT or REJECT the null hypothesis? Why?

If you rejected your null hypothesis, what might be some explanations for your outcome?

Now that you have completed this Chi-Square analysis for your data, let's do it for the entire class, as if we had one huge bag of M&Ms. Using the information reported on the board, complete Data Table 2.

Data Table 2	Color Categories						
	Brown	Blue	Orange	Green	Red	Yellow	Total
Observed (O)							
Expected (E)							
Difference (O-E)							
Difference Squared $(O-E)^2$							
$(O-E)^2/E$							
$\chi^2 = \Sigma\,(O-E^2)/E$							

Based on the class data, should you ACCEPT or REJECT the null hypothesis? Why ?

If you rejected the null hypothesis based on the class data, what might be some of the explanations for your outcome?

If you accepted the null hypothesis, how do you explain it—particularly if you rejected the null based on your individual group's data?

What is the purpose of collecting data from the entire group?

Practice Problem

In pea plants, green color (G) is dominant to albino (g). If we cross two pea plants that are heterozygous for color, what would be the expected phenotype ratio of the offspring? Do the Punnett square and write the phenotypic ratio in the space below:

Let's say that we actually did cross two heterozygous pea plants and obtained the following data:

Phenotype	# Offspring Observed
Green	72
Albino	12
Total	84

Now we will calculate a Chi-Square value for this data and find out if any difference between what we observe and what we expect can or cannot be explained by chance.

First, write out our standard null hypothesis:

Null Hypothesis: _____

Next, we will follow the same procedure that we did for the M&Ms.

Fill out the table below by figuring out the number of expected offspring among these phenotypes and calculating the Chi-Square value. Then, determine the degrees of freedom.

Data Table 3	Phenotypes		
	Green	Albino	Total
Observed (O)			
Expected (E)			
Difference (O-E)			
Difference Squared (O-E)^2			
(O-E)^2/E			
$\chi^2 = \Sigma$ (O-E^2)/E			

What is the number of degrees of freedom for this problem? _____

(Remember: degrees of freedom = number of categories –1)

Compare your Chi-Square value to the same table that you used for the M&Ms, this time using the new number of degrees of freedom.

Based on the data, should you ACCEPT or REJECT the null hypothesis? Why?