Chi-Square Statistic

by Sophia

WHAT'S COVERED

This tutorial will cover the chi-square statistic, discussing how to calculate the observed frequency and expected frequency of a data set. Our discussion breaks down as follows:

1. The Chi-Square Statistic
2. Finding Observed and Expected Frequencies

1. The Chi-Square Statistic

A **chi-square statistic** is a particular test used for categorical data. It measures how expected frequency differs from observed frequency.

The **observed frequency** is the number of observations we actually see for a value, or what actually happened. The **expected frequency** is what we would expect to happen. It is the number of observations we would see for a value if the null hypothesis was true.

HINT

In this tutorial, you will not run any significance tests because the chi-square tests have many different versions, and each of them will have their own tutorial. This tutorial is going to focus on how the statistic is calculated, as it's calculated the same regardless of the test you're running.

To measure the discrepancy between what you observed and what you expected, we need to calculate the chi-square statistic, which is calculated this way:

STEP BY STEP

1. Take the observed values.
2. Subtract the expected values.
3. Square that difference.
4. Divide by the expected values.
5. Add up all of those fractions.

FORMULA

Chi-Square Statistic

$$X^2 = \sum \frac{(O-E)^2}{E}$$

↗ EXAMPLE Suppose you have a tin of colored beads, and you claim that the tin contains the colored beads in these proportions: 35% blue, 35% green, 15% yellow, and 15% red. These will be used to find the expected frequencies.

You draw 10 beads from the tin: 4 red, 3 blue, 1 green, and 2 yellow. These will be your observed frequencies.

Is what you drew consistent with the percentages you claimed or not? Why or why not?

If the claim were true, we would have expected that out of 10 beads, 3 1/2 of them would be blue, 3 1/2 green, 1 1/2 yellow, and 1 1/2 red. This is called the expected frequencies and can be calculated by multiplying the sample size by the hypothesized proportion.

Color	Percentage	Expected out of 10
Blue	35%	3.5
Green	35%	3.5
Yellow	15%	1.5
Red	15%	1.5

You can't actually pull 3 1/2 blue beads, because you can't have half of a bead. Therefore, this is an idealized scenario, representative of what you might expect in the long-term in samples of 10.
In your one sample of 10 beads, what you actually got was: 3 blue, 1 green, 2 yellow, and 4 red. The two yellow beads drawn seems fairly close with the 15% claim. However, the four red beds that were drawn does not seem consistent with the 15% claim for red.
How can you measure that discrepancy? We can calculate the chi-square statistic using the above formula. First, subtract the each expected frequency from the observed frequency, square that value, and divide by the expected frequency. Finally, add up all those calculations.

Color	Expected	Observed	$\frac{(O-E)^2}{E}$
Blue	3.5	3	0.0714
Green	3.5	1	1.7857
Yellow	1.5	2	0.1667
Red	1.5	4	4.1667
		Sum	6.1905

The 3 1/2, 3 1/2, 1 1/2, and 1 1/2 were the expected frequencies and the observed frequencies were the 3, 1, 2, and 4. Using the formula, we get a chi-square statistic value of 6.1905.
So what do we do with this chi-squared statistic? We can find this value, along with the degrees of freedom, in a chi-squared distribution table to determine if we reject or fail to reject the null hypothesis by

comparing it to the pre-determined significance level.

HINT

You can use a table to calculate the chi-square statistic or you can use technology.
Now, it's worth noting that in this case, the conditions for inference with a chi-square test are not met. This is only meant to illustrate how a chi-square statistic would be calculated, although you can't do any real chi-square inference on this because the sample size isn't large enough.

TERMS TO KNOW

Chi-Square Statistic
The sum of the ratios of the squared differences between the expected and observed counts to the expected counts.

Observed Frequencies
The number of occurrences that were observed within each of the categories in a qualitative distribution.

Expected Frequencies
The number of occurrences we would have expected within each of the categories in a qualitative distribution if the null hypothesis were true.

2. Finding Observed and Expected Frequencies

IN CONTEXT

Suppose there are four flavors of candy in a bag: cherry, lemon, orange, and strawberry. The company claims the flavors are equally distributed in each bag.

After opening a bag of candy and sorting the flavors, the following counts were produced:

Flavor	Observed
Cherry	11
Lemon	15
Orange	12
Strawberry	12
Total	**50**

In equal distribution, it is helpful to think of the proportions of each flavor and then make a hypothesis based on those proportions. For the null hypothesis, we can assume that the proportions for the four flavors are the same. The alternate hypothesis would state that is is not true; that the proportions are not the same.

H_0: $p_C = p_L = p_O = p_S$

H_a: The proportions of the flavors are not the same.

Next, we need to compare the observed frequency with the expected frequency. The observed frequencies are the same as the above counts.

To find the expected frequency, we need to find the number of occurrences if the null hypothesis is true, which in this case, was that the flavor proportions are equal, or if the four flavor categories were all evenly distributed. Counting up all the flavors in that bag of candy gives us a total of 50 candies. If the flavor categories were evenly distributed among the 50 candies, we would need to divide the total candies evenly between the four flavors, so 50 divided by 4, or 12.5 candies. This means we would expect 12.5 candies in each flavor.

Flavor	Observed	Expected
Cherry	11	12.5
Lemon	15	12.5
Orange	12	12.5
Strawberry	12	12.5

We can then use the chi-squared formula to calculate the chi-square statistic to compare the discrepency between the expected and observed frequencies.

A middle school is gathering information on its after-school clubs because it was assumed that the distribution of students in each grade was evenly distributed across the clubs, meaning there were the same amount of 6th graders in each club, the same amount of 7th graders in each club, and the same amount of 8th graders in each club.

This table lists the number of students from each grade participating in each club.

	6th graders	**7th graders**	**8th graders**
Coding Club	12	14	8
Photography Club	7	11	15
Debate Club	9	5	13

Suppose we want to find the observed frequency for 7th graders participating in the photography club. Using the chart, we can directly see the observed frequency for 7th graders participating in the photography club is 11.

To find the expected frequency for 7th graders participating in the photography club, we need to find the number of occurrences if the null hypothesis is true, which in this case, was that the three options are equally likely, or if the students in each grade were all evenly distributed across the

clubs.

First, add up all the students in the 7th-grade column:

$$14 + 11 + 5 = 30$$

If each of these three clubs were evenly distributed among the 30 7th graders, we would need to divide the total evenly between the three options:

$$30 \div 3 = 10$$

This means we would expect 10 7th graders to participate in the coding club, 10 7th graders to participate in the photography club, and 10 7th graders to participate in the debate club.

In summary, the observed and expected frequencies for 7th graders participating in photography club is:

- Observed: 11
- Expected: 10

SUMMARY

The chi-square statistic is a measure of discrepancy across categories from what you would have expected in categorical data. You can only use it for data that appear in categories or qualitative data. The expected values may not be whole numbers since the expected values are long-term average values.

Good luck!

Source: Adapted from Sophia tutorial by Jonathan Osters.

TERMS TO KNOW

Chi-Square Statistic
The sum of the ratios of the squared differences between the expected and observed counts to the expected counts.

Expected Frequencies
The number of occurrences we would have expected within each of the categories in a qualitative distribution if the null hypothesis were true.

Observed Frequencies
The number of occurrences that were observed within each of the categories in a qualitative distribution.

FORMULAS TO KNOW

Chi-Square Statistic

$$X^2 = \sum \frac{(O - E)^2}{E}$$