

The Chi-Square (χ^2) Statistical Test

Paul K. Strode, Ph.D., Fairview High School, Boulder, Colorado

Scenario #1

Say that you are in charge of preparing a bracket for a tournament that involves 12 teams. You need to randomly decide which team in each pairing gets to be the home team. You hypothesize that the coin you are using is a fair coin, but you want to test that hypothesis before proceeding. You predict that if you flip the coin 50 times, based on a 50/50 probability, you will get 25 heads and 25 tails. These are your **expected** (predicted) values. Of course, you would rarely get exactly 25 and 25 but how close must you be to 25 and 25 to still provide support for your hypothesis that the coin is fair? How far off these numbers can you be without the results being significantly different from what you expected (which might indicate that there is maybe something wrong with the coin.)

After you conduct your experiment your **observed** result is 21 heads and 29 tails. Is the difference between **observed** and **expected** results purely due to chance, or are the differences true differences and the coin is unfair? The Chi-square test (pronounced with a “K” and an “eye” sound, like in “kite”) can help you answer that question.

Equation

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

Where: O = observed values

E = expected values

χ^2 = Chi-squared

Σ = summation

Explanation and steps of equation:

Step 1: Calculate the Chi-square Value

Table 1 below outlines the steps required to calculate the Chi-square value and test the null hypothesis based on the coin-flipping example discussed above. The equations for calculating a Chi-square value are provided in each column heading.

Table 1. Coin toss results.

Side of Coin	Observed (O)	Expected (E)	(O-E)	(O-E) ²	(O-E) ² /E
Heads	21	25	-4	16	0.64
Tails	29	25	4	16	0.64
					$\chi^2 = \sum(O-E)^2/E$
					$\chi^2 = 1.28$

Step 2: Determine the Degrees of Freedom

The degrees of freedom value is calculated as follows:

$$df = \text{number of categories minus } 1$$

In the example above, there are two categories (heads and tails).

$$df = (2-1) = 1$$

Step 3: Use the Critical Values Table to Determine the Probability (p) Value of getting the sum Chi-square value by chance.

The critical values table below (Table 2) shows the probability (or p -value) of obtaining a X^2 value as large as the listed value if the null hypothesis is correct. A p -value of 0.05 means there is only a 5% chance that differences between the observed values and expected values have occurred simply by accident.

For example, for $df = 1$, there is a 5% probability (p -value = 0.05) of obtaining a X^2 value of 3.841 or larger. In statistics, a probability of 5% is considered so rare that it is unlikely that the null hypothesis is correct. 5% probability is called a *statistically significant* result. If the X^2 value obtained was 4.5, then the null hypothesis can be rejected. If the X^2 value was 3.1, then the null hypothesis cannot be rejected.

(In science, you also see p -values of 0.01 being used. Some studies use a more stringent 1% as the criterion for rejecting the null hypothesis. 1% probability is called a highly significant result.)

To use the critical values table, locate the calculated X^2 value in the row corresponding to the appropriate number of degrees of freedom. For the coin flipping example, locate the calculated X^2 value in the $df = 1$ row. The X^2 value is 1.28, which falls between 1.10 and 1.32 and which is smaller than 3.84 (the X^2 value at the $p=0.05$ cutoff). Therefore, the null hypothesis, that the results have likely occurred simply by chance, is not rejected. Based on a $p=0.05$ cutoff value, we can conclude that the differences between the observed and expected values are due to random chance alone. The coin is fair!

Table 2. Chi-square Critical Values Table

		Chi-Square Table of Critical Values											
		Probability of Distribution Occurring by Chance											
		0.99	0.95	0.90	0.75	0.70	0.50	0.30	0.25	0.10	0.05	0.03	0.01
df	# of Categ	Critical Values											
1	2	0.00	0.00	0.02	0.10	0.15	0.45	1.10	1.32	2.71	3.84	5.02	6.63
2	3	0.02	0.10	0.21	0.58	0.71	1.39	2.40	2.77	4.61	5.99	7.38	9.21
3	4	0.11	0.35	0.58	1.21	1.42	2.37	3.70	4.11	6.25	7.81	9.35	11.34
4	5	0.30	0.71	1.06	1.92	2.02	3.36	4.90	5.39	7.78	9.49	11.14	13.28
5	6	0.55	1.15	1.61	2.67	3.00	4.35	6.10	6.63	9.24	11.07	12.83	15.09
6	7	0.87	1.64	2.20	3.45	3.83	5.35	7.20	7.84	10.64	12.59	14.45	16.81
7	8	1.24	2.17	2.83	4.25	4.67	6.35	8.40	9.04	12.02	14.07	16.01	18.48
8	9	1.65	2.73	3.49	5.07	5.53	7.34	9.50	10.22	13.36	15.51	17.53	20.09
9	10	2.09	3.33	4.17	5.90	6.39	8.34	10.60	11.39	14.68	16.92	19.02	21.67

Note that this test must be used on **RAW** data; values cannot be transformed to frequencies or percentages because using percentages artificially sets the sample size to 100 even if the sample is small or very large. The size of the sample is an important aspect of the Chi-square test—it is generally more difficult to detect a statistically significant difference between experimental and observed results in a small sample than in a large sample.

Scenario #2

A teacher had been teaching Biology for 4 years. Over those 4 years, she had kept track of the letter grade distributions in all of her classes—a total of 570 students. The teacher wondered if her observed letter grade distribution among her 570 students was rare when compared to what would be expected in a population of students that fit a bell curve. In other words, is her **observed** distribution significantly different from what would be **expected** in a bell curve distribution, so much that it probably did not occur by chance?

The teacher may hypothesize at first that the students in her Biology course are “normal” students. She may then predict that, if her hypothesis is valid, her students will have a letter grade distribution statistically indistinguishable from a bell curve distribution and that the probability of her student grade distribution occurring by chance is large (greater than 1/20, or 0.05). In other words, are the deviations (differences between **observed** and **expected**) the result of chance, or are they due to other factors (i.e. nonrandom choice, math skills, previous science knowledge)? How much deviation can occur before she must conclude that something other than chance is at work, causing the observed distribution to differ from the expected? The Chi-square test, like other statistical tests, is always testing what scientists call the null statistical hypothesis (**H₀**, the no difference or nothing happened hypothesis). In this case, H₀ is that there is no significant difference between the expected and observed result, and any differences are likely to have occurred by chance (H₀: O = E).

If the probability that her distribution occurred by chance is equal to or less than 0.05 (1/20), she can conclude that the differences between her observed distribution and what she would expect in a bell curve distribution are true differences and did not occur by chance alone. In other words, her students are not “normal” and her grade distribution may require a different explanation.

The teacher’s data are in Table 3.

Table 3. Student grade distributions from 2008-2012 in Biology and the expected distributions for a bell curve population.

Grade	Observed Distribution of Students in each Letter Grade Category	Expected % Distribution of Letter grades in a “Normal” Population	Expected Distribution of Students in each Letter Grade Category out of 570	(O-E)²	(O-E)²/E
A	74	10	57	289	5.07
B	148	20	114	1,156	10.14
C	194	40	228	1,156	5.07
D	103	20	114	121	1.06
F	51	10	57	36	0.632
Totals	570	100	570	$\sum X^2 =$	21.97

Practice:

Write a conclusion using statistical language: What was the teacher’s experimental hypothesis? What was her prediction? What can the teacher conclude about her experimental hypothesis given the results of the Chi-square test? What is the probability (p) that

- 1) Her distribution occurred by chance.
- 2) If she rejects the H_0 , she will be in error.

1 and 2 above are different ways we can report the results of a statistical test (p).

How does the Chi-square test react to sample size?

Scenario #1: Consider an average high school biology classroom. If the students in the class are representative of the school’s entire population, which we assume to be of a normal gender distribution, then we would expect to find 15 boys and 15 girls in a class of 30 students. What if a teacher’s second hour class has a distribution of boys and girls of 12 and 18, respectively? Could the distribution in this small sample have occurred by chance and still be representative of a normal population of half boys and half girls? Use the Chi-square test to test the null statistical hypothesis that the distribution of 12 boys and 18 girls is not significantly different than an even distribution of 15 and 15, and that the observed distribution could have occurred by chance.

Category	Observed	Expected	$(O - E)^2/E$
Boys			
Girls			
		$\Sigma \chi^2 =$	

What can you conclude? What is the probability that the null statistical hypothesis is correct?

Scenario #2: Now consider the same teacher. She counts all of her boys and girls in all of her classes and her colleague’s classes. Together they have 120 boys and 180 girls. Is this larger sample still representative of the school’s entire population? Or, is something else going on to explain the observation that there are more girls taking biology than boys? Use the Chi-square test to test the null statistical hypothesis that the distribution of 120 boys and 180 girls is not significantly different than an even distribution of 150 and 150, and that the observed distribution could have occurred by chance.

Category	Observed	Expected	$(O - E)^2/E$
Boys			
Girls			
		$\Sigma \chi^2 =$	

What can you conclude? What is the probability now that the null statistical hypothesis is correct?

If you found that the probability that the null statistical hypothesis is correct to be less than 5% ($p = 0.05$), you must reject the null statistical hypothesis—the null statistical hypothesis is not likely to be true. There must be some other hypothesis to explain this distribution of boys and girls. What hypothesis can you think of to explain this distribution?