Chapter 10
Chi Square

Objectives

After completing this chapter, you will be able to:

- To learn Chi Square Statistics
- Does a Chi-Square Statistic Tell You
- Use of the Chi squared tests

What Is a Chi-Square Statistic?

A chi-square (χ^2) statistic is a test that measures how expectations compare to actual observed data (or model results). The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample. For example, the results of tossing a coin 100 times meet these criteria.
Chi-square tests are often used in hypothesis testing.

The Formula for Chi-Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:
c=Degrees of freedom *O*=Observed value(s) *E*=Expected value(s)

What Does a Chi-Square Statistic Tell You?

There are two main kinds of chi-square tests: the test of independence, which asks a question of relationship, such as, "Is there a relationship between gender and SAT scores?"; and the goodness-of-fit test, which asks something like "If a coin is tossed 100 times, will it come up heads 50 times and tails 50 times?"
For these tests, degrees of freedom are utilized to determine if a certain null hypothesis can be rejected based on the total number of variables and samples within the experiment.

For example, when considering students and course choice, a sample size of 30 or 40 students is likely not large enough to generate significant data. Getting the same or similar results from a study using a sample size of 400 or 500 students is more valid.
In another example, consider tossing a coin 100 times. The expected result of tossing a fair coin 100 times is that heads will come up 50 times and tails will come up 50 times. The actual result might be that heads will come up 45 times and tails will come up 55 times. The chi-square statistic shows any discrepancies between the expected results and the actual results.

Example of a Chi-Squared Test

Imagine a random poll was taken across 2,000 different voters, both male and female. The people who responded were classified by their gender and whether they were republican, democrat, or independent. Imagine a grid with the columns labeled republican, democrat, and independent, and two rows labeled male and female. Assume the data from the 2,000 respondents is as follows:

The first step to calculate the chi squared statistic is to find the expected frequencies. These are calculated for each "cell" in the grid. Since there are two categories of gender and three categories of political view, there are six total expected frequencies. The formula for the expected frequency is:

$$E(r,c) = \frac{n(r) \times c(r)}{n}$$

where:

r=Row in question *c*=Column in question *r*=Corresponding total

In this example, the expected frequencies are:

$$E(1,1) = \frac{900 \times 800}{2,000} = 360$$

$$E(1,2) = \frac{900 \times 800}{2,000} = 360$$

$$E(1,3) = \frac{200 \times 800}{2,000} = 80$$

$$E(2,1) = \frac{900 \times 1,200}{2,000} = 540$$

$$E(2,2) = \frac{900 \times 1,200}{2,000} = 540$$

$$E(2,3) = \frac{200 \times 1,200}{2,000} = 120$$

Next, these are used values to calculate the chi squared statistic using the following formula:

$$\text{Chi-squared} = \sum \frac{[O(r,c) - E(r,c)]^2}{E(r,c)}$$

where:

O(*r*,*c*)=Observed data for the given row and column

In this example, the expression for each observed value is:

$$O(1,1) = \frac{400 - 360^2}{360} = 4.44$$
$$O(1,2) = \frac{300 \times 360^2}{360} = 10$$
$$O(1,3) = \frac{100 - 80^2}{80} = 5$$
$$O(2,1) = \frac{500 - 540^2}{540} = 2.96$$
$$O(2,2) = \frac{600 - 540^2}{540} = 6.67$$
$$O(2,3) = \frac{100 - 120^2}{120} = 3.33$$

The chi-squared statistic then equals the sum of these value, or 32.41. We can then look at a chi-squared statistic table to see, given the degrees of freedom in our set-up, if the result is statistically significant or not.

What is a Chi Square Test?

There are **two types of chi-square tests**. Both use the chi-square statistic and distribution for different purposes:

- A **chi-square goodness of fit test** determines if a sample data matches a population. For more details on this type, see: *Goodness of Fit Test*.
- A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
  - A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
  - A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.

What is a Chi-Square Statistic?

The formula for the chi-square statistic used in the chi square test is:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The chi-square formula.

The subscript "c" are the degrees of freedom. "O" is your observed value and E is your expected value. It's very rare that you'll want to actually *use* this formula to find a critical chi-square value by hand. The summation symbol means that you'll have to perform a calculation for every single data item in your data set. As you can probably imagine, the calculations can get very, very, lengthy and tedious. Instead, you'll probably want to use technology:

- Chi Square Test in SPSS.
- Chi Square P-Value in Excel.

A chi-square statistic is one way to show a relationship between two categorical variables. In statistics, there are two types of variables: numerical (countable) variables and non-numerical (categorical) variables. The chi-squared statistic is a single number that tells you how much difference exists between your observed counts and the counts you would expect if there were no relationship at all in the population. There are a **few variations** on the chi-square statistic. Which one you use depends upon how you collected the data and which hypothesis is being tested. However, all of the variations use the same idea, which is that you are comparing your expected values with the values you actually collect. One of the most common forms can be used for contingency tables:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

Where O is the observed value, E is the expected value and "i" is the "ith" position in the contingency table.

A **low value** for chi-square means there is a high correlation between your two sets of data. In theory, if your observed and expected values were equal ("no difference") then chi-square would be zero — an event that is unlikely to happen in real life. Deciding whether a chi-square test statistic is large enough to indicate a statistically significant difference isn't as easy it seems. It would be nice if we could say a chi-square test statistic >10 means a difference, but unfortunately that isn't the case. You could take your calculated chi-square value and compare it to a critical value from a chi-square table. If the chi-square value is more than the critical value, then there is a significant difference.

You could also use a p-value. First state the null hypothesis and the alternate hypothesis. Then generate a chi-square curve for your results along with a p-value (See: Calculate a chi-square p-value Excel). Small p-values (under 5%) usually indicate that a difference is significant (or "small enough").

***Tip**: The Chi-square statistic can only be used on numbers. They can't be used for percentages, proportions, means or similar statistical value. For example, if you have 10 percent of 200 people, you would need to convert that to a number (20) before you can run a test statistic.*

Chi Square P-Values.

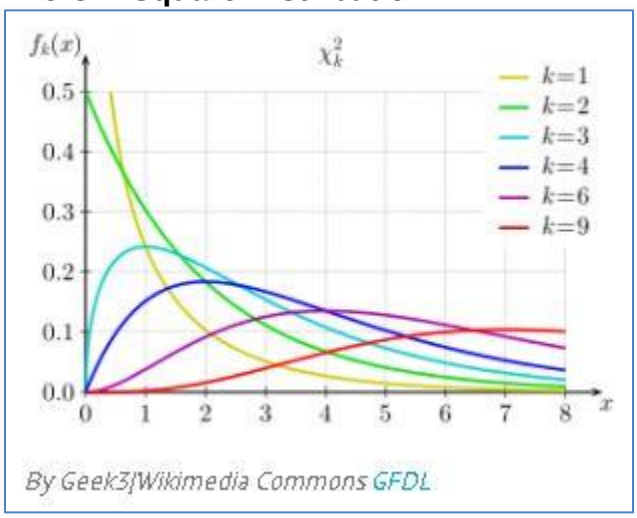
A chi square test will give you a p-value. The p-value will tell you if your test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:

1. Degrees of freedom. That's just the number of categories minus 1.
2. The alpha level(α). This is chosen by you, or the researcher. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

In elementary statistics or AP statistics, both the degrees of freedom(df) and the alpha level are usually given to you in a question. You don't normally have to figure out what they are. You *may* have to figure out the df yourself, but it's pretty simple: count the categories and subtract 1.

Degrees of freedom are placed as a subscript after the chi-square (X^2) symbol. For example, the following chi square shows 6 df:
X^2_6.
And this chi square shows 4 df:
X^2_4.

The Chi-Square Distribution



The chi-square distribution (also called the chi-squared distribution) is a special case of the gamma distribution; A chi square distribution with n degrees of freedom is equal to a gamma distribution with a = n / 2 and b = 0.5 (or β = 2).

Let's say you have a random sample taken from a normal distribution. The chi square distribution is the distribution of the sum of these random samples **squared** .

The **degrees of freedom (k)** are equal to the number of samples being summed. For example, if you have taken 10 samples from the normal distribution, then df = 10. The degrees of freedom in a chi square distribution is also its **mean**. In this example, the mean of this particular distribution will be 10. Chi square distributions are always right skewed. However, the greater the degrees of freedom, the more the chi square distribution looks like a normal distribution.

Uses

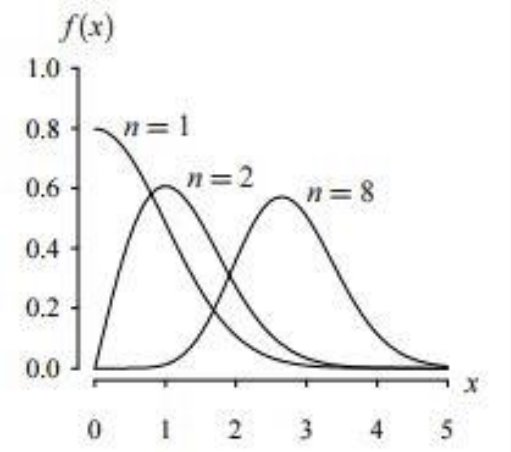
The chi-squared distribution has many uses in statistics, including:

- Confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
- Independence of two criteria of classification of qualitative variables.
- Relationships between categorical variables (contingency tables).
- Sample variance study when the underlying distribution is normal.
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The chi-square test (a goodness of fit test).

Chi Distribution

A similar distribution is the **chi distribution**. This distribution describes the **square root** of a variable distributed according to a chi-square distribution.; with df = n > 0 degrees of freedom has a probability density function of:

For values where x is positive.



The cdf for this function does not have a closed form, but it can be approximated with a series of integrals, using calculus.

The Chi squared tests

The χ^2tests

The distribution of a categorical variable in a sample often needs to be compared with the distribution of a categorical variable in another sample. For example, over a period of 2 years a psychiatrist has classified by socioeconomic class the women aged 20-64 admitted to her unit suffering from self poisoning sample A. At the same time she has likewise classified the women of similar age admitted to a gastroenterological unit in the same hospital sample B. She has employed the Registrar General's five socioeconomic classes, and generally classified the women by reference to their father's or husband's occupation. The results are set out in table 8.1.

Table 8.1 Distribution by socioeconomic class of patients admitted to self poisoning (sample A) and gastroenterological (sample B) units

Socioeconomic class	Samples		Total	Proportion in group A
	A	B		
	a	b	n = a + b	p = a/n
I	17	5	22	0.77
II	25	21	46	0.54
III	39	34	73	0.53
IV	42	49	91	0.46
V	32	25	57	0.56
Total	**155**	**134**	**289**	

The psychiatrist wants to investigate whether the distribution of the patients by social class differed in these two units. She therefore erects the null hypothesis that there is no difference between the two distributions. This is what is tested by the chi squared (χ^2) test (pronounced with a hard ch as in "sky").

By default, all χ^2 tests are two sided.

The psychiatrist wants to investigate whether the distribution of the patients by social class differed in these two units. She therefore erects the null hypothesis that there is no difference between the two distributions. This is what is tested by the chi squared (χ^2) test (pronounced with a hard ch as in "sky"). By default, all χ^2 tests are two sided.

It is important to emphasise here that χ^2 tests may be carried out for this purpose only on the *actual numbers* of occurrences, not on percentages, proportions, means of observations, or other derived statistics. Note, we distinguish here the Greek (χ^2) for the test and the distribution and the Roman (x^2) for the calculated statistic, which is what is obtained from the test.

The χ^2 test is carried out in the following steps:

For each observed number (0) in the table find an "expected" number (E); this procedure is discussed below.

Subtract each expected number from each observed number	$O - E$
Square the difference	$(O - E)^2$
Divide the squares so obtained for each cell of the table by the expected number for that cell	$(O - E)^2 / E$
x^2 is the sum of $(O - E)^2 / E$	

To calculate the expected number for each cell of the table consider the null hypothesis, which in this case is that the numbers in each cell are proportionately the same in sample A as they are in sample B. We therefore construct a parallel table in which the proportions are exactly the same for both samples. This has been done in columns (2) and (3) of table 8.2 . The proportions are obtained from the totals column in table 8.1 and are applied to the totals row. For instance, in table 8.2 , column (2), 11.80 = (22/289) x 155; 24.67 = (46/289) x 155; in column (3) 10.20 = (22/289) x 134; 21.33 = (46/289) x 134 and so on.

Thus by simple proportions from the totals we find an expected number to match each observed number. The sum of the expected numbers for each sample must equal the sum of the observed numbers for each sample, which is a useful check. We now subtract each expected number from its corresponding observed number.

Table 8.2 Calculation of the χ^2 test on figures in table 8.1

Class (I)	Expected numbers		0 - E		$(0-E)^2/E$	
	A (2)	B (3)	A (4)	B (5)	A (6)	B (7)
I	11.80	10.20	5.20	-5.20	2.292	2.651
II	24.67	21.33	0.33	-0.33	0.004	0.005
III	39.15	33.85	-0.15	0.15	0.001	0.001
IV	48.81	42.19	-6.81	6.81	0.950	1.009
V	30.57	26.43	1.43	-1.43	0.067	0.077
Total	**30.57**	**134.00**	**0**	**0**	**3314**	**3.833**

$\chi^2 = 3.314 + 3.833 = 7.147$. d.f. = 4. $0.10<P<0.50$.

The results are given in columns (4) and (5) of table 8.2 . Here two points may be noted.

1. The sum of these differences always equals zero in each column.
2. Each difference for sample A is matched by the same figure, but with opposite sign, for sample B.

Again these are useful checks.

The figures in columns (4) and (5) are then each squared and divided by the corresponding expected numbers in columns (2) and (3). The results are given in columns (6) and (7). Finally these results, $(O-E)^2/E$ are added. The sum of them is x^2

A helpful technical procedure in calculating the expected numbers may be noted here. Most electronic calculators allow successive multiplication by a constant multiplier by a short cut of some kind. To calculate the expected numbers a constant multiplier for each sample is obtained by dividing the total of the sample by the grand total for both samples. In table 8.1 for sample A this is 155/289 = 0.5363. This fraction is then successively multiplied by 22, 46, 73, 91, and 57. For sample B the fraction is 134/289 = 0.4636. This too is successively multiplied by 22, 46, 73, 91, and 57.

The results are shown in table 8.2 , columns (2) and (3).

Having obtained a value for we look up in a table of χ^2 distribution the probability attached to it (Appendix Table C.pdf). Just as with the *t* table, we must enter this table at a certain number of degrees of freedom. To ascertain these requires some care. When a comparison is made between one sample and another, as in table 8.1 , a simple rule is that the degrees of freedom equal (number of columns minus one) x (number of rows minus one) (not counting the row and column containing the totals). For the data in table 8.1 this gives (2 - 1) x (5 - 1) = 4. Another way of looking at this is to ask for the minimum number of figures that must be supplied in table 8.1 , in addition to all the totals, to allow us to complete the whole table. Four numbers disposed anyhow in samples A and B provided they are in separate rows will suffice.

Entering Table C at four degrees of freedom and reading along the row we find that the value of $x^2(7.147)$ lies between 3.357 and 7.779. The corresponding probability is: $0.10<P<0.50$. This is well above the conventionally significant level of 0.05, or 5%, so

the null hypothesis is not disproved. It is therefore quite conceivable that in the distribution of the patients between socioeconomic classes the population from which sample A was drawn were the same as the population from which sample B was drawn.

Video Links:

Chi-squared Test

- https://www.youtube.com/watch?v=WXPBoFDqNVk

Chi Squared Test

- https://www.youtube.com/watch?v=qYOMO83Z1WU

Chi Square Test - with contingency table

- https://www.youtube.com/watch?v=misMgRRV3jQ

References

- https://www.statisticshowto.com/probability-and-statistics/chi-square/
- https://www.investopedia.com/terms/c/chi-square-statistic.asp#:~:text=A%20chi%2Dsquare%20(%CF%872,from%20a%20large%20enough%20sample.
- https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests