4f – A Guide to
Chi-Squared Testing

Royal Geographical Society with IBG
Advancing geography and geographical learning

The Chi-Squared test (also known as the Index of Association) looks at how closely data the researcher has collected fits with a theoretical or expected set of data. It is sometimes referred to as a 'goodness of fit' test as it is used to see how closely observed findings match those that are known to be true on a wider scale.

Why would we use the Chi-Squared test?

A researcher would use the Chi-Squared test when their data is showing the frequency of an occurrence of something. This test only checks to see if there is an association between two sets of data, not what the nature of the relationship might be between those sets, nor the strength of any relationship.

Worked Example:

The following example will look at the age structure of visitors to a tourist attraction in a particular town. A Chi-Squared test will be carried out to see if the observed age structure at the attraction is indicative of the age structure of the town in general.

The researcher has found secondary data which represents the annual number of visitors to the attraction in different age groups. Census data was also found to show the age structure of the town in which the attraction was situated.

Age grp	Number of visitors	Population of Town
0 – 19	280,580	41,370
20 – 39	240,500	48,830
40 – 59	121,250	44,730
60 – 79	80,160	30,500
80+	2,080	7,650
Total	724,570	173,080

Nuffield Foundation

This project was funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily those of the Foundation.

1. The researcher should firstly work out the expected data (in this case the expected number of visitors). To do this, convert the known data of the town's population into percentages and use these same percentages to predict the number of visitors there will be to the attraction in each of the age categories.

Age grp	Population of Town	Population as a %
0 – 19	41,370	23.9
20 – 39	48,830	28.2
40 – 59	44,730	25.8
60 – 79	30,500	17.6
80+	7,650	4.5
Total	173,080	100

Total number of visitors = 724,570

Therefore, the number of predicted visitors aged 0 – 19 is

$$\frac{23.9 \times 724570}{100} = 173{,}172$$

This calculation should then be done for each age category for the number of visitors to create a set of predicted (or expected) values:

Age grp	Expected visitor numbers	As a %
0 – 19	173,172	23.9
20 – 39	204,329	28.2
40 – 59	186,939	25.8
60 – 79	127,524	17.6
80+	32,603	4.5
Total	724,570	100

Where no real data is available on which to base an expected set of values, the researcher should divide the total value (724,570) equally between the number of categories in order to get expected values. This would suggest that the category or variable in question would have no influence over the observed data as the values would all be the same (in this case 144,914 predicted visitors in each age category). It is then the role of the researcher, through carrying out the Chi-Squared test, to see if the categories or variable have a role to play in the geography being observed.

The expected frequencies or values do not have to be whole numbers to further work within the Chi-Squared calculations.

2. The researcher should then table the observed and expected values as follows:

	0 – 19	20 - 39	40 - 59	60 - 79	80+
Observed value or frequency (O)	280,580	240,500	121,250	80,160	2,080
Expected value or frequency (E)	173,172	204,329	186,939	127,524	32,603

3. For each category, the researcher should then calculate the value of $\frac{(O - E)^2}{E}$

For example, 0 – 19 age group

$$\frac{(280580 - 173172)^2}{173172} = \frac{107408^2}{173172} = \frac{11536478464}{173172} = 66{,}618$$

	0 – 19	20 - 39	40 - 59	60 - 79	80+
$\frac{(O - E)^2}{E}$	66,618	6,403	23,082	17,592	28,575

4. Add all these values together:

$\sum \frac{(O - E)^2}{E}$ = 142,270 This is the Chi-Squared value (shown as the symbol χ^2).

5. The researcher can then use a data table (part of which printed here or otherwise found through online searches) to work out if this value is significant.

   The 'degree of freedom' is the number of categories you have in your test minus one (in this case, we have 4 degrees of freedom).

> For the purpose of most geography research, an exact understanding of what 'degrees of freedom' means is not necessary and the reader should not worry at this stage if this term remains unexplained except for how it should be used in a significance table such as that below.

Significance Table (also known as the Critical Values of Chi-Squared Table):

Degree of freedom (df)	Significance Level				
	10%	5%	2.5%	1%	0.5%
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	**9.49**	11.14	13.23	14.86
5	9.24	11.07	12.83	15.09	20.52

A significance table tells us how reliable the result is and whether the observed data is as a result of a geographical factor, or due to complete chance. Using a 5% (sometimes written as 0.05) significance level tells us that there is only a 5% possibility of the results being as they are by chance. A 5% significance level is commonly used by geographical researchers.

If the Chi-Squared value is greater than the appropriate value in the table, the observed data is significantly different from the expected data. Therefore, there is more likely to be an association between the collected data and the variable in question (in this example, the tourist attraction). The data is said to be dependent on something, and not just a result of random numbers coinciding.

If the Chi-Squared value is less than the appropriate value in the table, then the observed data is not significantly different from the expected data. This does not mean there is no association between the observed data and the variable in question, just that there is not enough evidence to show this.

In this example, the size of the sample is huge so it is not unexpected that the Chi-Squared value is also very large, suggesting an association between the attraction and age of visitor. In many student-based studies the researcher will be working with primary rather than secondary data so it is highly likely that a smaller Chi-Squared value will be used when looking up the result in a significance table.