Chi Squared Test

The Chi Squared Test is a statistical method used to find the strength of association between two sets of categorical data. Categorical data is simply that which has been placed into sets such as age ranges or income level. The test relies on comparing the observed frequency of an occurrence (i.e. that which happens in the field, collected through primary data) with expected frequency of an occurrence (i.e. that which theoretically should be true, or that which would be true if no variables were acting upon the outcome of the data).

As the Chi Squared Test uses **categorical data**, one must ensure that both sets of data fit neatly into categories. As such Chi Squared relies on the researcher using discrete data or continuous data grouped into discrete fields. For example **distance** is a form of continuous data but can be put into discrete fields such as 0 to 5cm, 6 to 10cm, 11 to 15cm and so on.

The advantage of using the Chi Squared Test is that one is not restricted to comparing simply one set of data with another as might happen with a traditional correlation test. Instead, the number of categories which can be used for each data set is in theory infinite, though for practical ease and to make the results meaningful, it may be wise to have no more than four categories in either data set. Equally the Chi Squared Test is meaningless if fewer than two categories are made.

How to carry out a Chi Squared Test

This example will look at how frequently people of different age groups visited a site that had undergone regeneration in a coastal town. The data for such a test came from a questionnaire that was asked on the street of the town using predefined categories to assess how frequently people visited a certain area. After the completion of the questionnaire, respondents were then also categorised into discrete age bands.

The researcher had a hunch that the regeneration scheme was targeting the older generation who were viewed anecdotally as those that held the highest levels of wealth in the town. This created the following hypothesis in the mind of the researcher:

"There will be an association between age range and the frequency with which people use the regeneration site in Market Square."

The researcher collected the following observed data:

	Daily	Weekly	Monthly	Less than monthly
0 -18 years	0	2	3	2
19 - 40 years	4	7	4	4
41 - 60 years	5	12	8	1
61 years +	9	14	6	4

The totals for each age category and frequency category were then calculated.

	Daily	Weekly	Monthly	Less than monthly	Total
0 -18 years	0	2	3	2	7
19 - 40 years	4	7	4	4	19
41 - 60 years	5	12	8	1	26
61 years +	9	14	6	4	33
Total	18	35	21	11	85

The expected data was then calculated. The expected data assumes that all age groups use the regenerated areas frequently as each other and that there is no association between age and frequency of use. To calculate the expected data the researcher used the following equation:

$$\text{Expected} = \frac{\text{column total x row total}}{\text{overall total}}$$

So the expected number of 0 to 18 year olds who visited the regenerated area daily is

$$\text{Expected} = \frac{18 \times 7}{85} = 1.48$$

Therefore, a table of expected data can also be drawn:

	Daily	Weekly	Monthly	Less than monthly	Total
0 -18 years	1.48	2.88	1.73	0.91	7
19 - 40 years	4.02	7.82	4.70	2.46	19
41 - 60 years	5.51	10.71	6.42	3.36	26
61 years +	6.99	13.59	8.15	4.27	33
Total	18	35	21	11	85

A Chi Squared (χ^2) value was then be calculated using the following equation, where O is the observed value and E is the expected value:

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

This would then produce the following table...

	Daily	Weekly	Monthly	Less than monthly	Total
0 -18 years	1.48	0.269	0.932	1.306	3.987
19 - 40 years	0.000	0.086	0.104	0.964	1.154
41 - 60 years	0.047	0.155	0.389	1.658	2.249
61 years +	0.578	0.012	0.567	0.017	1.175
Total	2.105	0.522	1.992	3.945	8.565

... with this as the Chi Squared Value (χ^2)

The larger the Chi Squared value, the more likely there is an association between the two sets of categorical data. However, as a stand alone figure, the Chi Squared value has very little meaning. To understand whether the researcher can accept their hypothesis, they have to look at a Chi Squared significance table (also available to download from the island geographer) to see if the amount of data they have collected makes the result significant or not.

To be deemed significant, the critical value read from the significance table must be lower than the Chi Squared value. In this case, to a 95% significance level, the critical value is 16.919. Therefore, the researcher cannot accept the hypothesis and one must instead accept that there is no association between a visitor's age and how frequently they visit the regenerated site.