A NON- PARAMETRIC AND GOODNESS OF FIT TEST: CHI-SQUARE

Sanjay Kumar Gupta

Department of Mathematics Master GSSS Sangwari (2543) Rewari

---***---

Abstract - There are various tests which is based on assumption that the samples were drawn from normal distributed population. Since the testing procedure requires assumption about the type of population or parameters such test is known as parametric test. But there are many situations in which it is not possible to make any assumption about the distribution of the population from which samples are being drawn. This limitation has led an alternative technique known as non-parametric tests. The chi square test is one of the simplest and most widely used non parametric tests in statistical work. This test is commonly used for testing relationships between categorical variables. It is used to evaluate tests of Independence when using a cross tabulation. Cross tabulation presents the distributions of two categorical variables simultaneously, it is used to analyze categorical data (e.g. male or female students, smokers and non-smokers, etc.), it is not meant to analyze parametric or continuous data (e.g., height measured in centimeters or weight measured in kg, etc.). with the intersections of the categories of the variables appearing in the cells of the table.

Key Words: Assumption, Non parametric, Chi-square test, hypothesis testing, Goodness of Fit test

INTRODUCTION *(Size 11, Times New roman)*

The **chi-squared test** is **used** to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. A **chi-squared test** can be **used** to attempt rejection of the null hypothesis that the data are independent. The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.

An example research question that could be answered using a Chi-Square analysis would be:

In research, there are studies which often collect data on categorical variables that can be summarized as a series of counts. These counts are commonly arranged in a tabular format known as a contingency table. The chi-square test statistic can be used to evaluate whether there is an association between the rows and columns in a contingency table. More specifically, this statistic can be used to determine whether there is any difference between the study groups in the proportions of the risk factor of interest. Chi-square test and the logic of hypothesis testing were developed by Karl Pearson (1857-1936) was born in London England his interest in analytical statistics was kindled only in late 1880.s after he has become a professor of applied Mathematics. This article describes in detail what is a chi-square test, on which type of data it is used, the assumptions associated with its application, how to manually calculate it and how to make use of an online calculator for calculating the Chi-square statistics and its associated *P*-value.

Assumptions Underlying a Chi-square Test

The data are randomly drawn from a population

The values in the cells are considered adequate when expected counts are not <5 and there are no cells with zero count

The sample size is sufficiently large. The application of the Chi-square test to a smaller sample could lead to type II error (i.e. accepting the null hypothesis when it is actually false). There is no expected cut-off for the sample size; however, the minimum sample size varies from 20 to 50

The variables under consideration must be mutually exclusive. It means that each variable must only be counted once in a particular category and should not be allowed to appear in other category. In other, words no item shall be counted twice.

How to Calculate a Chi-square Statistics?

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where,

O stands for the observed frequency,

E stands for the expected frequency.

Expected count is subtracted from the observed count to find the difference between the two. Then the square of the difference is calculated to get rid of the negative vales (as the squares of 2 and −2 are, of course, both 4). Then the square of the difference is divided by the expected count to normalize bigger and smaller values (because we don't want to get bigger Chi-square values just because we are working on large data sets). The sigma sign in front of them denotes that we have, to sum up, these values calculated for each cell.

As an example, suppose we want to find out that whether there is an association between Economic Condition and IQ level of School Students.

The null and alternative hypothesis will be:

H_0: There is no association between Economic Condition and IQ level of Students

H_1: There is an association between Economic Condition and IQ level of Students.

The general formula for calculating the expected counts from observed count for a particular cell is [(corresponding row total * corresponding column total) /Total no. of students. Before we proceed further, we need to know how many degrees of freedom (df) we have. When a comparison is made between one sample and another, a simple rule is that the df equals (number of columns − 1) × (number of rows − 1) excluding the rows and column containing the total. Hence, in our example df = (2−1) × (3−1) = 2.

Chi-square test is a nonparametric test used for two specific purpose: (a) To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables); (b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). It is used to analyze categorical data (e.g. male or female patients, smokers and non-smokers, etc.), it is not meant to analyze parametric or continuous data (e.g., height measured in centimeters or weight measured in kg, etc.). For example if we want to test the level of IQ of school students according their economic condition and graded accordingly. Use chi square test to find out whether any association between economic condition and level of IQ 400 Stratified random samples taken of school students. A 2x3 contingency table also known as cross tables can be constructed for calculating a Chi-square statistic. Stratified random samples of 400 student's classification according level of IQ and their economic condition as follow

Level of IQ/Economic Condition	**High**	**Medium**	**low**	**Total**
Poor	102	66	44	212
Rich	88	64	36	188
Total	190	130	80	400

For this formulate a suitable hypothesis Apply chi square test

Let us take a hypothesis that the samples are drawn from the same population or Economic condition influence the most IQ level of the students for this null hypothesis is set

Null hypothesis, H0:- There in no relationship between the IQ level and economic condition of the school students or Economic Condition do not influence to IQ level of students. Difference is due to fluctuation of samples

Alternative Hypothesis H1: - There is relationship between the IQ level and Economic condition of the school students or Economic Condition influences to IQ level of students. Samples are insignificant at 5% level of significance.

The expected frequencies corresponding to each group and product can be obtained as follows.

Level of IQ/ Economic Condition	**High**	**Medium**	**low**	**Total**
Poor	101	69	42	212
Rich	89	61	38	188
Total	190	130	80	**400**

What conclusion can be drawn from the test result

Observed value	**Expected**	$(O-E)^2$	$(O-E)^2/E$
102	101	1	.009
88	89	1	.011
66	69	9	.130
64	61	9	.147
44	42	4	.095
36	38	4	.105
Total			.497

Since the calculated value of Chi square is less than the table value 5.99 so the null hypothesis is accepted hence Hypothetical data for calculating the Chi-square test for our example of testing an association between Economic Condition and Level of IQ is shown that there is no such association.

Chi-square test can be calculated manually by using the formula described above table for manual calculations. Chi-square value for our example as shown is 5.99, df = 2. If we want to test our hypothesis at 5% level of significance than our predetermined alpha level of significance is 0.05. Looking into the Chi-square distribution table with 2 degree of freedom and reading along the row we find our value of χ^2 *.497. That means that the* T value is above (it is actually 0.497). Since a *T* value of 5.99 is greater than the conventionally accepted significance level of 0.05 we fail to reject the null hypothesis or in other words we accept our null hypothesis and conclude that there is no association between Economic Condition and IQ Level of the students.

Approximate of P Value

Scientists and statisticians use large tables of values to calculate the *P* value for their experiment. These tables are generally set up with the vertical axis on the left corresponding to df and the horizontal axis on the top corresponding to *P* value. Use these tables by first finding our df, then reading that row across from the left to the right until we find the first value bigger than our Chi-square value. Look at the corresponding *P* value at the top of the column. Chi-square distribution tables are available from a variety of sources-they can easily be found online or in science and statistics textbooks.

Uses of Chi Square Test

The chi square test is one of the most popular statistical inference procedures today it applicable to a very large number of problems in practice which can be summed up under the following heads:

1. Chi Square test as a test of Independence: With the help of chi square test we can find out whether two or more attributes are associated or not. Suppose we have N observations classified according to some attributes we may ask whether the attributes are related or independent thus we can find out whether quinine is effective in controlling fever or not. Whether there is any association between marriage and failure, or eye color of husband and wife.
2. Chi square test as a test of goodness of fit. Chi square test is popularly known as test of goodness of fit for the reason that is enables us to ascertain how appropriately the theoretical distributions such us binomial, Poisson, normal etc. fit empirical distributions.
3. Chi test as test of homogeneity the chi square test of homogeneity is an extension of the chi square test of independence. Tests of homogeneity are designed to determine whether to or more independent random samples are drowning from same population or from different populations. Instead of one sample as we use with independence problem, we shall now have two or more samples. For example, we may be interested in finding whether or not university students of various levels, i.e., undergraduate. postgraduate, ph. D., feel the same in regard to the amount of work required by their professors i.e., too much work, right amount of work or too little work. we shall take the hypothesis that the three samples come from the same population; that is, the three classification are homogeneous in so far as the opinion of three different groups of students about the amount of work required by their professors is concerned. This also means there exists no difference in opinion among the three classes of people on the issue.

Limitation of chi square test

Chi square test is very widely use in practice however, in order to avoid the misapplication of test it following limitation should be kept in mind;

First frequencies of non-occurrence should not be omitted for binomial, muti nominal events for example, if five drugs were tried out on five separate groups of two hundred of patients each, the number of cures per drugs might be shown. It should not be applied until alternative outcomes are not present.

It may be clearly understood that Chi-square test only tells us the probability of independence of a distribution of data or in simple terms it will only test that whether two variables are associated with each other or not. It will not tell us that how closely they are associated. For instance, in the above example, the Chi-square test will only tell us that whether there is any relation between smoking and lung disease. It will not tell us that how likely it is, that smokers are prone to lung disease. However, once we got to know that there is a relation between these two variables, we can explore other methods to calculate the amount of association between them.

References and Bibliography

1. Magnello ME Karl Pearson and the origin of modern statistics: An elastic an becomes a statistician, Rutherford J, Vol. 1, 2005-2006.
2. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philon Mag Ser 1900; 50:157-75.
3. Plackett RL. Karl Pearson and the Chi-squared test. Int Stat Rev 1983; 51:59-72.
4. Yates F, Moore D, McCabe G. The Practice of Statistics 1st ed. New York: W. H. Freeman, 1999.
5. Yates F. Contingency table involving small numbers and the Chi-squared test. Suppl J R Stat Soc 1934; 1:217-35.
6. Sulthan chand & sons, S.P Gupta Statistical Method.