Volume 2, Issue 1 January 2016 ISSN NO: 2454 - 7514

# North Asian International Research Journal Consortium

*North Asian International Research Journal*

*Of*

*Science, Engineering and Information Technology*

NAIRJC JOURNAL PUBLICATION

North Asian International Research Journal Consortium

# Welcome to NAIRJC

**ISSN NO: 2454 -7514**

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi, Urdu all research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

## Editorial Board



**Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,**
**Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com**

# CHI SQUARE TEST FOR INDEPENDENT VARIABLES

**A.JEYANTHI***
*Faculty in Mathematics, Anna University, Regional Campus, Madurai, Tamilnadu, India.

***ABSTRACT***

***The Test of Independence assesses whether an association exists between the two variables by carefully examining the pattern of responses in the cells, calculating the Chi-Square statistic and comparing it against a critical value from the Chi-Square distribution allows the researcher to assess whether the association seen between the variables in a particular sample is likely to represent an actual relationship between those variables in the population. The Chi Square statistic is commonly used for testing relationships on categorical variables.***

***KEYWORD:*** ***Cross tabulation, Chi-Square statistic,***

## INTRODUCTION

Chi Square analysis would be useful is to see if there are differences between male and female college students on choice of major field of study- Engineering or English. The null hypothesis would be: There is no difference between male and female college students on their choice between taking quantitative versus qualitative elective courses. Both variables GENDER and COURSE are categorical. Perfect for a Chi Square. Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if, according to Mendel's laws, you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the "goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. How much deviation can occur before you, the investigator, must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what scientists call the null hypothesis, which states that there is no significant difference between the expected and observed result.

## CHI-SQUARE STATISTIC

The null hypothesis is that no relationship exists on these categorical variables in the population; they are independent. The Chi-Square statistic is most commonly used to evaluate Tests of Independence when using a cross tabulation. Cross tabulation presents the distributions of two categorical variables simultaneously, with the intersections of the categories of the variables appearing in the cells of the table. The calculation of the Chi-Square statistic is quite straight-forward and intuitive,

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Where $f_o$ = the observed frequency (the observed counts in the cells) and $f_e$ = the expected frequency if no relationship existed between the variables. As depicted in the formula, the Chi-Square statistic is based on the difference between what is actually observed in the table and what would be expected if there was truly no relationship between the variables. The Chi-Square statistic appears as an option when requesting a cross-tabulation in SPSS. The output is labeled Chi-Square Tests; the Chi-Square statistic used in the Test of Independence is in the first row labeled Pearson Chi-Square. This statistic can be evaluated by comparing the actual value against a critical value found in a Chi-Square distribution where degrees of freedom is calculated as # of rows – 1 x # of columns – 1, but it is easier to simply examine the *p*-value provided by SPSS. To make a conclusion about the hypothesis with 95% confidence. (Which is the *p*-value of the Chi-Square statistic) should be less than .05 (which is the alpha level associated with a 95% confidence level).

Is the *p*-value < .05. If so, conclude that the variables are dependent in the population and that there is a statistical relationship between the categorical variables. There are a number of important considerations when using the Chi-Square statistic to evaluate a cross-tabulation. Because of how the Chi-Square value is calculated, it is extremely sensitive to sample size, when the sample size is sufficiently large (~500), almost any small difference appears significant. It is also sensitive to the distribution within the cells, and SPSS gives a warning message if cells have fewer than 5 cases. This can be addressed by always using categorical variables with a limited number of categories by combining categories if necessary to produce a smaller table. Statistics Solutions can assist with your quantitative analysis by assisting you to develop your methodology and results chapters. In a chi-square test, tell us is if there is a large difference between collected numbers and expected numbers. If the difference is large, it tells us that there may be something causing a significant change. A significantly large difference will allow us to reject the null hypothesis, which is defined as the prediction that there is no interaction between variables. Basically, if there is a big enough difference between the scores, then we can say something significant happened. If the scores are too close, then we have to conclude that they are basically the same. The actual formula for running a chi-square is actually very simple:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \quad \text{or} \quad \chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$\chi 2 \rightarrow$ the test statistic that asymptotically approaches a chi-square distribution,
$f_0 = O_i \rightarrow$ the observed frequency of the ith row and jth column,
$f_e = E_i \rightarrow$ the expected (theoretical) frequency of the ith row and jth column,
$r \rightarrow$ the number of rows in the contingency table,
$k \rightarrow$ the number of columns in the contingency table.
we take your observed data (*o*) and subtract what we expected (*e*). You square the results, and then divide by the expected data in all the categories.
To use the number we find, we refer to the degrees of freedom, usually labeled as *df* for short, and is defined for the chi-square as the number of categories minus 1. Due to the nature of the chi-square test, you will always use

the number of categories minus 1 to find the degrees of freedom. The reason this is done is because there is an assumption that your sample data is biased, and this helps shift your scores to allow for error.

You will then locate a chi-square distribution table, which is found in almost every statistical textbook printed. Using your degrees of freedom, you will locate the *p*-value you're interested in using the process below; typically the *p*-value is .05. If you can, see if your number is greater than .01, which means that your results could only happen by chance 1 in 100 times. Because of copyright restriction issues, we won't be able to provide a full image of the chi-square distribution table, but this is basically what they look like and how you find the digit you're looking for.

To find your *p*-value, you follow the left hand column of the degrees of freedom. If we have 10 categories, we have 9 degrees of freedom. We would move 9 places down on the left hand side. Next, we will follow the row of 9 degrees of freedom to the right until we reach the .05 level. If the number from your formula is greater than the one found in the chart, then you have a statistically significant finding. It's sort of like playing Battleship, except it's with degrees of freedom and the *p*-value. One of the most valuable statistics is a non-parametric procedure called Chi Square analysis. It is also called the test of "goodness of fit". Its symbol is "( $\chi^2$ ) squared" ( $\chi^2$ )

Unlike the t-test and ANOVA procedures, the Chi Square analysis is not as powerful to reject the null. It does not use the mean or standard deviation for computation; it does not rely on an interval or ratio scaling. Because the Chi Square relies on frequency data, its value lays in the statistic's ability to answer questions about data that are nominal. Variables in many settings are measured very often by their categories - and not exact intervals. Chi Square allows you to answer important questions with variables measured with nominal or ordinal scales.

The formula for calculating chi-square ( $\chi^2$ ) is:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e} or \qquad \chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

That is, chi-square is the sum of the squared difference between observed (*o*) and the expected (*e*) data (or the deviation, *d*), divided by the expected data in all possible categories.

For example, suppose that a cross between two pea plants yields a population of 880 plants, 639 with green seeds and 241 with yellow seeds. You are asked to propose the genotypes of the parents. Your *hypothesis* is that the allele for green is dominant to the allele for yellow and that the parent plants were both heterozygous for this trait. If your hypothesis is true, then the predicted ratio of offspring from this cross would be 3:1 (based on Mendel's laws) as predicted from the results of the Punnett square (Figure 1).

**Figure 1 - Punnett Square. Predicted offspring from cross between green and yellow-seeded plants. Green (G) is dominant (3/4 green; 1/4 yellow).**

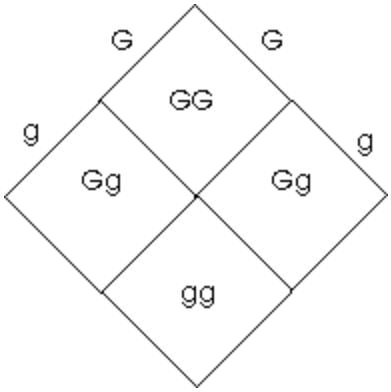


**Figure 1**

To calculate Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. For example, if, according to Mendel's laws, you expected 10 of 20 offspring from a cross to be male and the actual observed number was 8 males, then you might want to know about the "goodness to fit" between the observed and expected. Were the deviations (differences between observed and expected) the result of chance, or were they due to other factors. To know how much deviation can occur before you, the investigator must conclude that something other than chance is at work, causing the observed to differ from the expected. The chi-square test is always testing what scientists call the null hypothesis, which states that there is no significant difference between the expected and observed result.

## TESTS FOR DIFFERENT PURPOSES

1. Chi square test for testing goodness of fit is used to decide whether there is any difference between the observed (experimental) value and the expected (theoretical) value.
   For example given a sample, we may like to test if it has been drawn from a normal population. This can be tested using chi square goodness of fit procedure.

2. Chi square test for independence of two attributes. Suppose N observations are considered and classified according two characteristics say A and B. We may be interested to test whether the two characteristics are independent. In such a case, we can use Chi square test for independence of two attributes. The example considered above testing for independence of success in the English test vis a vis immigrant status is a case fit for analysis using this test.

3. Chi square test for single variance is used to test a hypothesis on a specific value of the population variance. Statistically speaking, we test the null hypothesis H0: $\sigma = \sigma_0$ against the research hypothesis H1: $\sigma \# \sigma_0$ where $\sigma$ is the population mean and $\sigma 0$ is a specific value of the population variance that we would like to test for acceptance. In other words, this test enables us to test if the given sample has been drawn from a population with specific variance $\sigma_0$. This is a small sample test to be used only if sample size is less than 30 in general.

## ASSUMPTIONS

The Chi square test for single variance has an assumption that the population from which the sample has been is normal. This normality assumption need not hold for chi square goodness of fit test and test for independence of attributes. However while implementing these two tests; one has to ensure that expected frequency in any cell is not less than 5. If it is so, then it has to be pooled with the preceding or succeeding cell so that expected frequency of the pooled cell is at least 5.

## NON PARAMETRIC AND DISTRIBUTION FREE VARIABLES

It has to be noted that the Chi square goodness of fit test and test for independence of attributes depend only on the set of observed and expected frequencies and degrees of freedom. These two tests do not need any assumption regarding distribution of the parent population from which the samples are taken. Since these tests do not involve any population parameters or characteristics, they are also termed as non parametric or distribution free tests. An additional important fact on these two tests is they are sample size independent and can be used for any sample size as long as the assumption on minimum expected cell frequency is met. Chi-square test for categorical variables determines whether there is a difference in the population proportions between two or more groups. In the medical literature, **the Chi-square is used most commonly to compare the incidence (or proportion) of a characteristic in one group to the incidence (or proportion) of a characteristic in other group(s).** Categorical data is also known as nominal data, meaning that one uses labels as opposed to numbers; for example, race and gender are categorical variables. The central tendency of categorical variables is given by its mode, since median and mean can only be computed on numerical data. Therefore, it does not follow a normal bell-curve distribution, and cannot be analyzed with tests that rely on a normal distribution such as the t-test or ANOVA.

The Chi Square test is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. A common usage of the Chi-square test is the Pearson's chi-square test, also known as the chi-square goodness-of-fit test or chi-square test for independence. The Chi square test is used to compare a group with a value, or to compare two or more groups, always using categorical data.

***Example: 1***

Suppose there is a city of 1 million residents with four neighbourhoods: A, B, C, and D. A random sample of 650 residents of the city is taken and their occupation is recorded as blue collar, white collar, or no collar. The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification. The data are tabulated as:

| | A | B | C | D | total |
|---|---|---|---|---|---|
| White collar | 90 | 60 | 104 | 95 | 349 |
| Blue collar | 30 | 50 | 51 | 20 | 151 |
| No collar | 30 | 40 | 45 | 35 | 150 |
| Total | 150 | 150 | 200 | 150 | 650 |

Let us take the sample living in neighborhood A, 150/650, to estimate what proportion of the whole 1 million people live in neighborhood A. Similarly we take 349/650 to estimate what proportion of the 1 million people are white-collar workers. By the assumption of independence under the hypothesis we should "expect" the number of white-collar workers in neighborhood A to be

$$\frac{150}{650} \times \frac{349}{650} \times 650 \approx 80.54$$

Then in that "cell" of the table, we have

$$\frac{(observed - \exp ected)^2}{\exp ected} = \frac{(90 - 80.54)^2}{80.54}$$

The sum of these quantities over all of the cells is the test statistic. Under the null hypothesis, it has approximately a chi-square distribution whose numbers of degrees of freedom are:

(Number of rows-1)(Number of columns-1) = (3-1) (4-1) = 6

If the test statistic is improbably large according to that chi-square distribution, then one rejects the null hypothesis of independence. A related issue is a test of homogeneity. Suppose that instead of giving every resident of each of the four neighborhoods an equal chance of inclusion in the sample, we decide in advance how many residents of each neighborhood to include. Then each resident has the same chance of being chosen as do all residents of the same neighborhood, but residents of different neighborhoods would have different probabilities of being chosen if the four sample sizes are not proportional to the populations of the four neighborhoods. In such a case, we would be testing "homogeneity" rather than "independence". The question is whether the proportions of blue-collar, white-collar, and no-collar workers in the four neighborhoods are the same. However, the test is done in the same way.

***Example: 2***

| Medication/Out come | Recovered | Did not Recover | Total |
|---|---|---|---|
| Prozan | 150 | 50 | |
| Drug X | 200 | 100 | |
| Total | | | |

Question: Was the proportion who recovered on Prozan the same as the proportion that recovered on Drug X. Some difference in proportions is not sufficient evidence – the difference could be due to sampling variability.

So we ask how probable is the observed difference if there is no actual difference between these populations distribution $\chi^2$ .That probability is governed by the r is the number of rows; k is the number of columns. The expression directs us to calculate the expected number for each row and column intersection, and then calculate the required squares. The expected number is the proportion in the total sample with characteristic R multiplied by the size of the subsample. The chi-square statistic has (r-1) × (k-1) degrees of freedom.

Chi-Square Calculation: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

How many would we expect to recover in the prozan group?

Of the total sample of 500,350 recovered so P(R) = 350/500 = 0.7

There are 200 in the prozan sample, so $E_{11}=0.7\times200=140$

Formula: $E_{ij} =[C_j/n)\times R_i$ where $C_j$ is the total in the j-th column and $R_i$ is the total in the i-row, while n is the size of the total sample.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{10^2}{140} + \frac{10^2}{210} + \frac{10^2}{60} + \frac{10^2}{90} = 3.9683$$

The null hypothesis is phrased as "Recovery is independent of drug taken." That is, knowing the patient took Prozan doesn't change the probability of recovery or, there's no difference in effectiveness, To reject $H_0$ means to conclude that there is a difference in effect Performing the hypothesis tests: ($\chi^2$) table

In these tests we are concerned only with right-tail area. At 5% significance with 1 degree of freedom, critical value = 3.841 < 3.9683

***Example: 3***

Calculate what numbers of "exposed" and "non-exposed" individuals would be expected in each disease group if the probability of disease were the same in both groups  If there was no association between exposure and disease, then the expected counts should nearly equal the observed counts, and the value of the chi-square statistic would be small  In this example, we can calculate: Overall proportion with exposure = 50/120 = 0.42 Overall proportion without exposure = 70/120 = 0.58 = 1– 0.42 

Under the assumption of no association between exposure and disease, the expected numbers or counts in the table are:

| Disease | | | |
|---|---|---|---|
| Exposure | Yes | No | Total |
| Yes | 50/120 x 54 = 22.5 | 50/120 x 66 = 27.5 50 | 50 |
| No | 70/120 x 54 = 31.5 | 70/120 x 66 = 38.5 | 70 |
| Total | 54 | 66 | 120 |

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} =$$

$$(37 - 22.5)^2/22.5 + (13 - 27.5)^2/27.5 + (17 - 31.5)^2/31.5 + (53 - 38.5)^2/38.5$$

The test statistic is: − χ2 = 29.1 with 1 degree of freedom  Assumption: no association between disease and exposure  A small value of the χ2 statistic supports this assumption (observed counts and expected counts would be similar)  A large value of the χ2 statistic would not support this assumption (observed counts and expected counts would differ). The probability of obtaining a statistic of this magnitude or larger when there is no association.

## CONCLUSION

The results of regular surveys can be used as groundwork for suggestions and recommendations concerning improving the quality of the postal and telecommunication services, to identify regional disparities in the utilization of postal and telecommunication services and also for market segmentation. Some of the assumed regional disparities were confirmed. Other regional disparities highlighted the need for further investigation of factors influencing customer behavior.

## REFERENCES

[1] Maxwell A. E. Analysing Qualitative Data. 4th Edition. Chapman and Hall Ltd., 1971. Library of Congress Catalog Card Number 75-10907.
[2] Žilinský kraj. (Zilina Region). [online]. [s.a.]. [Cited 2010-08-18]. Available on the internet: [3] Chajdiak, J., Komorník, J., Komorníková, M. Štatistické metódy. (Statistical methods). Bratislava: STATIS, 1999. 282 p. ISBN 80-85659-13-1.
[4] Linczényi A. Inžinierska štatistika. (Engineering Statistics). Bratislava: ALFA, 1974. 452 p. 63-025-74.
[5] Sojková, Z. Materiál na prednášky z predmetu Ekonomická štatistika. (Material for lectures on the subject Economic Statistics). Department of Statistics and Operation Research, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Slovak Republic. [online]. [Cited 2010-08-12]. Available on the internet: .
[6] Cochran, W. G. Some methods for strengthening the common χ2 tests, Biometrics, 1954. 10: pp. 417–451.
[7] Yates, D., Moore, D., McCabe, G. The Practice of Statistics. 1st Ed. New York: W.H. Freeman, 1999.
[8] Yates, F. Contingency table involving small numbers and the χ 2 test, Supplement to the Journal of the Royal Statistical Society, 1934, 1(2): 217-235. JSTOR Archive for the journal
[9] Nominal Association: Phi, Contingency Coefficient, Tschuprow's T, Cramer's V, Lambda, Uncertainty Coefficient. [online]. [s.a.]. [Cited 2010-08-18]. Available on the internet: .
[10] Cohen's scale for correlation coefficient. [online]. [s.a.]. [Cited 2010-08-16]. Available on the internet: .
[11] Zibran, M. F. CHI-Squared Test of Independence. Department of Computer Science, University of Calgary, Alberta, Canada. [online]. [Cited 2010-08-12]. Available on the internet.

# Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

**Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India**
**Cell: 09086405302, 09906662570,**
**Ph No: 01933212815**
**Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com**
**Website: www.nairjc.com**