ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 6, Issue 8, August 2018

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Application of chi-square test in social science research

Dr. Dilip M Patil
Associate Professor,
Department of Mathematics & Statistics,
Jashbhai Maganbhai Patel College of Commerce,
Mumbai – India.

Abstract: ***In social science research, when data is collected on two categorical variables the interest lies in studying the interrelationship between these variables. The relationship may be cause-effect relationship (like smoking habit & lung cancer). The data collected on the count can be summarized and expressed in the form of a mxn table called as 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 observed frequencies and expected frequencies. The test is referred as 'Test of Independence' or 'Test of goodness of fit'.

Chi-square test applied in this paper on 3 different sets of data revealed the significance of Chi-square test in research. The test can also be used to verify the validity of data supplied by the research.

Keywords: Chi-square test, level of significance, p-value, degrees of freedom, contingency table.

I. INTRODUCTION

The chi-square test is mostly used to study the significance of relationship between two categorical variables or attributes. Since this is a distribution free test it is also referred as non-parametric test & often find applications in social sciences two study relationship pattern between two social characteristics like 'parents income' & 'student's progress in examination' or 'husband's education level' & 'freedom of movement to wife' or 'smoking habit' & 'lung cancer'.

If the data sets achieve only nominal or ordinal levels of measurement, then non parametric method of analysis is employed. The statistical analysis method often involved experimental work where the data consist in frequencies or 'counts' e.g. 'How many students feel that intensive guidance program was effective?' or 'How many employees feel that, office work environment is stress-free?'. The most significant feature of Chi-square test is that, it doesn't require the assumption of normality to hold by the data. This test is invented by Pearson in 1900.

This paper studies the applications of Chi-square test to the results of social science research with the help of practical examples based on data collected.

II. CHI-SQUARE ($\chi 2$) DISTRIBUTION & $\chi 2$ STATISTIC

If $X_1, X_2, X_3 \ldots\ldots X_n$ is a set of n i.i.d. standard normal variables then statistic $\chi 2 = \sum Xi^2$ **follows Chi-square distribution.**

f($\chi 2$) = Chi-square statistic is defined as a square of standard normal variable.it is mostly used to study the significance of difference between observed values (experimental frequencies) & expected values (theoretical frequencies) under stated null

© IJARCSMS (www.ijarcsms.com), All Rights Reserved

hypothesis. For a set of n i.i.d. (s.n.v.) standard normal variables $\chi 2$ statistic is the sum of square of these n i.i.d. normal variable with n d.f.(degrees of freedom).

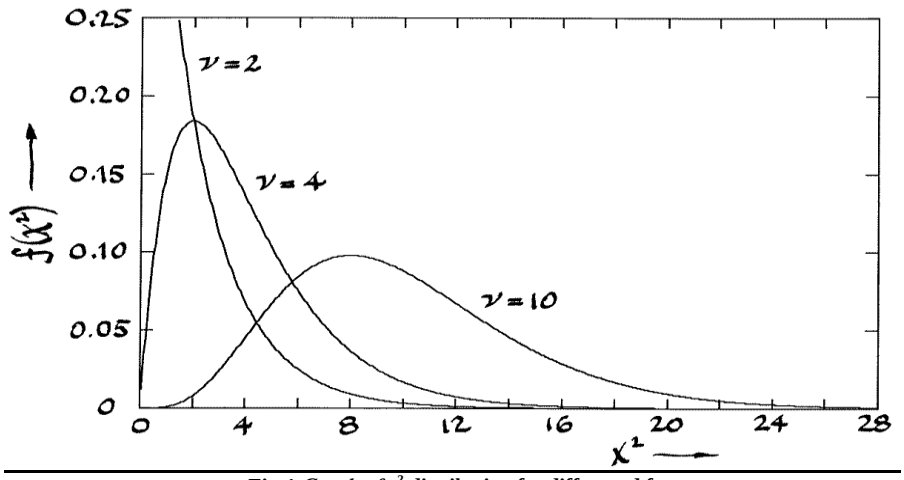

Fig-1 ***Graph of χ^2 distribution for different d.f.***

2.1 Application of Chi-square test;

- To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables) contingency table; 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) consistency of data recorded with the theoretical values.

2.2 Underlying Assumptions for 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.

2.3 Calculation of Chi-square static:

Chi-square statistic:

It is calculated by the formula, $\chi^2 = \sum \frac{(O-E)^2}{E}$, where,

$\chi 2$ … the test statistic that asymptotically approaches a chi-square distribution,

O … the observed frequency or recorded observation

E … the expected (theoretical) frequency calculated by the formula, $E=\frac{axb}{N}$,

a= Row total, b= Column total & N= Gross total

© 2018, IJARCSMS All Rights Reserved

r … the number of rows in the contingency table,

c … the number of columns in the contingency table.

Degrees of freedom (*d.f.*): In general d.f. is defined as the no of values researcher can assumed independently subject to given condition. It can be explained by example given below:

If percentage contribution of 5 groups is known as 23% 40%, 10% & 17% then the contribution of fifth group must be 10% as the total must be 100. Hence here the researcher has a d.f.4(=5-1) as (s)he can choose only 4 values randomly and the fifth one is determined by the condition on total must be 100.

Level of significance (α)- It is the maximum probability of rejecting Ho when it is true. This has to be fixed by the researcher before conducting the experiment or collecting the data observations. It is generally 5% or 1% and denoted by 'α', 1% l.o.s. means researcher is more confident in not rejecting Ho when it is true, i.e. committing an error .

Contingency table: It is a rxc table reading the observed frequencies on two characteristics when divided into r-rows & c-columns. Here degrees of freedom is calculate as (r-1)x(c-1).

2.4 Steps in applying the chi-square test:

1. State the null and alternate hypothesis
2. State the level of significance (α)
3. Select the random sample and record the observed frequencies (O) in each cell of contingency table and also calculate the expected (theoretical) frequencies (E)
4. Check whether any expected frequency is <5, if so club it to the nearest cell frequency which is more than 5.
5. Calculate the degrees of freedom (d.f.)= (r-1)x(c-1)- no of cell frequencies clubbed
6. Calculate for $\chi 2$ test statistic, read the p-value for the d.f. & stated l.o.s.
7. Compare p-value with l.o.s. smaller p-value (<5% or 1% selected) Ho must be rejected.
8. Accept Ho otherwise i.e. when p-value is more than l.o.s.
9. Report & interpret the results

III. Work Examples

3.1 Example-1: Data on 348 women is collected with respect to their **Education level** & **Religion** they follow. Apply the **Chi-square** test to test the significance of 'Religion' in the 'Education' of women in the study area.

Table 1: Working women by Religion & Education level (Observed frequency (O))

Religion	EDUCATION LEVEL					
	NONE	SSC	HSC	GRADUATE	PG	TOTAL
Hindu	25	35	68	65	21	214
Muslim	8	8	12	11	8	47
Buddhist	6	13	16	17	8	60
Christian	1	1	2	8	15	27
TOTAL	40	57	98	101	52	348

© 2018, IJARCSMS All Rights Reserved

Solution: Here, two characteristics are 'Religion' divided into 4 groups and 'Education level' in 5 categories. Hence the contingency table is 4x5 as shown above.

The null & alternate hypotheses are stated as below.

Ho: Education level of women has no significant relationship with the religion they follow

H_1: Education level of women is significantly related to the religion they follow

Table 2: Working women by Religion & Education level(Expected frequency)

Religion	EDUCATION LEVEL					
	NONE	SSC	HSC	GRADUATE	PG	TOTAL
Hindu	24.5977	35.05172	60.26437	62.1092	31.97701	214
Muslim	5.402299	7.698276	13.23563	13.6408	7.022989	47
Buddhist	6.896552	9.827586	16.89655	17.41379	8.965517	60
Christian	**3.103448**	**4.422414**	7.603448	7.836207	**4.034483**	27
TOTAL	40	57	98	101	52	348

In the calculation table we can note that, 3 cell frequencies are less than 5 hence are clubbed with the adjacent cells. This will reduce the d.f. by 3,

d.f. = (r-1)x(c-1) - no of cell frequencies clubbed = (4-1)x(5-1)-3=15

Chi-square table value with d.f. 15 = 24.99 at 5% l.o.s

Chi-square value= 17.88 and p-value = 0.2686 > 0.05,

Decision: Accept Ho

Conclusion: The data shows that, '**Religion**' women follow has no significant impact on their '**Education level**' attended. Hence, we can say that, in any religion education of women mainly depends upon financial condition of their family. This finding is for the sample selected, the result may be different for another sample drawn from the same or other study area.

3.2 Example-2

In a small school running classes 5th to 10th standard, 120 students participated in the 'Annual day' cultural programme.

Class (Standard):	V	VI	VII	VIII	IX	X
No of students participated:	15	25	25	22	18	15

Does the data support the claim that, all classes have equal representation?

Solution: In this example first we should understand that, why the question of uniformity occurred? The answer is very simple; when the representation is equal/uniform in all classes no of students participated should be equal i.e. all observed frequencies should be equal, which are not. Hence, the researcher question is, 'Is this difference significant'?

To test this claim we apply the χ^2 test of goodness of fit. The null & alternate hypotheses are stated as below.

Ho: All the classes have equal representation, i.e. the difference in the no of students participated is insignificant

H_1: All the classes do not have equal representation, i.e. the difference in the no of students participated is significant

© 2018, IJARCSMS All Rights Reserved

Table 3: Calculation on χ^2 statistic

Class	O	E= $\frac{120}{6}$	$\frac{(O-E)^2}{E}$
V	15	20	1.25
VI	25	20	1.25
VII	25	20	1.25
VIII	22	20	0.2
IX	18	20	0.2
X	15	20	1.25
Total	120=N	120	χ^2 =5.4

d.f. = n-1 6-1=5

Chi-square table value with d.f. 5 = 11.07 at 5% l.o.s

Whereas,

Chi-square value = 5.4 with p-value = 0.369 > 0.05

Decision: Accept Ho

Conclusion: The data shows that, Class representation is uniform. i.e. the difference in students participation count is insignificant.

3.3 Let us study the application of chi-square test to the same data used in another situation by the researcher.

Example: We know that, on a cubic die the no of dots are all equally likely, i.e. probability of any score from 1, 2, 3, 4, 5 & 6 is $\frac{1}{6}$. To verify this data on count on no of dots when the cubic die is rolled 210 times is given below. Apply Chi-square test to check the consistency of data with the theory.

No of dots: 1	2	3	4	5	6
Frequency: 25	45	55	32	28	25

Does this data support the claim that, the die is fair?

Solution: In this example first we should understand that, why the question of uniformity occurred? The answer is very simple; when the die is uniform all no of dots should occur equal no of times i.e. all observed frequencies should be equal, which are not. Hence the researcher question is, 'Is this difference significant'?

To test this claim we apply the χ^2 test of goodness of fit. The null & alternate hypotheses are stated as below.

Ho: The cubic die is fair, i.e. the difference in the count is insignificant

H_1: The cubic die is not fair, i.e. the difference in the count is significant

Table 4: Calculation on χ^2 statistic

No of dots	O	E= $\frac{210}{6}$	$\frac{(O-E)^2}{E}$
1	25	35	2.857143

2	45	35	2.857143
3	55	35	11.42857
4	32	35	0.257143
5	28	35	1.4
6=n	25	35	2.857143
Total	210=N	210	χ^2=21.65714

d.f. = n-1 6-1=5

Chi-square table value = 11.07 at 5% l.o.s and p-value = 0.0.0006 < 0.05

Whereas, Chi-square value = 21.65,

Decision: Reject Ho

Conclusion: The data shows that, 'The cubic die is not uniform (fair), i.e. the difference in the count is significant. Now we should note that, when a **cubic die** is used the differences should be insignificant.

Hence, we can conclude here that, the data is manipulated data and not an experimental data.

3.4 Let us apply the test in another data set in the same experiment.

No of dots: 1 2 3 4 5 6

Frequency: 30 40 45 32 33 30

Table 5: Calculation on χ^2 statistic

No of dots	O	E=$\frac{210}{6}$	$\frac{(O-E)^2}{E}$
1	30	35	0.714286
2	40	35	0.714286
3	45	35	2.857143
4	32	35	0.257143
5	33	35	0.114286
6=n	30	35	0.714286
Total	210=N	210	χ^2=5.371429

Here, the Chi-square value is 5.37 with

p-value: 0.37 for 4 d.f. which is greater than 0.05.

Hence, our null hypothesis Ho is accepted and it support the assumption/claim that the data is real.

Therefore, we can note that, Chi-square test also help us in testing the consistency of the data supplied.

IV. CONCLUSION

From the above discussion we can note that, Chi-square test helps us to,

- Study the significance of relationship or independence between 2 ordinal variables or attributes.

- Test the goodness of fit or consistency of the data with the theory assumption.

However the Chi-square test does not reveal the direction of relationship when exists neither the numerical measure of the relationship. At the same time when too menay cell frequencies are less than 5 pooling the frequencies reduces the d.f. to great extent. This may produce some misleading results.

References

1. Alena Košťálová:' APPLICATION OF CHI-SQUARE TEST OF INDEPENDENCE IN THE UTILIZATION OF POSTAL AND TELECOMMUNICATION SERVICES', The 10th International Conference "RELIABILITY and STATISTICS in TRANSPORTATION and COMMUNICATION - 2010
2. 'Chi-square: Testing for goodness of fit', http://maxwell.ucsc.edu/~drip/133/ch4.pdf
3. Isiwele A. Joseph, Aikpehae A. Moses, Adamolekun M. Olusegun. Application of Chi-Square and T-Test in Architectural Research Methods. Open Science Journal of Mathematics and Application. Vol. 4, No. 5, 2016, pp. 28-32.
4. Rana R, Singhal R. Chi-square test and its application in hypothesis testing. J Pract Cardiovasc Sci 2015;1:69-71

© 2018, IJARCSMS All Rights Reserved