Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) DOI: 10.7176/MTM
Vol.9, No.2, 2019

Hypothesis Testing for the Association Between Categorical Variables: Empirical Application of Chi-square Test

Basil Msuha
Division of Sector Coordination, President's Office, Regional Administration and Local Government, P.O. Box 1923, Dodoma, Tanzania
E-mail: basil.msuha@tamisemi.go.tz

Tiberio Mdendemi
Institute of Rural Development Planning, P.O. Box 138 Dodoma, Tanzania.
E-mail: tmdendemi@irdp.ac.tz

Abstract

Chi-square test and the logic of hypothesis testing were developed by Karl Pearson. In this article we demonstrate theoretically and empirically the hypothesis testing for the association between categorical variables using Chi-square Test. 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. We show in this paper how the chi-square test statistic can be used to evaluate whether there is an association between the rows and columns in a contingency table. We describes in detail what is a chi-square test, on which type of data it is used and the assumptions associated with its application. We consider both theoretical and empirical cases. On empirical case we use the data from the study which was conducted between September 2017 and March, 2018 in two municipalities of Dodoma and Morogoro, Tanzania. We conclude in this article that the Chi-square test, only tells us the probability of independence of a distribution of data or in simple terms it does only test that whether two categorical variables are associated with each other or not. It does not tell us that how closely they are associated. Therefore, once we got to know that there is a relation between these two variables, we need to explore other methods to calculate the amount of association between them.

Key words: Contingency table**,** categorical data analysis, Chi-square test, hypothesis testing

DOI: 10.7176/MTM/9-2-02

1.0 Introduction

The logic of hypothesis testing was first invented by Karl Pearson (1857–1936), a renaissance scientist and famous statistician in Victorian London in 1900. Pearson's Chi-square distribution and the Chi-square test also known as test for goodness of fit and test of independence are his most important contribution to the modern theory of statistics. The importance of this distribution is that one should not depend much on only the normal distribution only for inferencing about the data and hypothesis. Just to iterate, it is a statistical methods that does not depend on the normal distribution to interpret the findings. Karl Pearson invented the Chi square distribution mainly to address the needs of economists, biologists and psychologists (Magnello, 2006). His paper in 1900 published in Philosophical

magazine elaborates the invention of Chi-square distribution and goodness of fit test (Pearson, 1992; Plackett, 1983).

The chi square test is mainly used for the categorical values or variables and it is a non-parametric test method. Non parametric test methods are not concerned with the aspects of shape of the distribution population that is why they are called as the distribution free tests. A chi-squared test is also written as χ^2 test.

2.0 Theoretical Considerations

2.1 Uses of Chi-square test

Chi-square test is 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 (such as male or female patients, smokers and non-smokers) it is not meant to analyze parametric or continuous data (such as height measured in centimeters or weight measured in kg).

A Chi-square test compares proportions actually observed in a study with the expected to establish if they are significantly different. The Chi-square value increases as the difference between observed and expected increase. Whether the calculated Chi-square value is significant is determined by comparing it with the value from table. If the calculated value exceeds the table value, the difference between the observed and expected frequencies is taken as significant otherwise it is considered insignificant.

2.1 Assumptions Underlying a Chi-square Test

i. The data are randomly drawn from a population
ii. The values in the cells are considered adequate when expected counts are not <5 and there are no cells with zero count
iii. 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
iv. 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 Manual Calculation of Chi-Square Statistic

First we have to calculate the expected value of the two nominal variables. We can calculate the expected value of the two nominal variables by using this formula:

$$E_{ij} = \frac{T_i \times T_j}{N} \quad \text{.....(1)}$$

E_{ij} = the expected frequency for the cell in the ith row and the jth column,
T_i = total in the ith row
T_j = total in the jth column
N = is the total number of subjects in the whole table
Note: You can think of this equation more simply as *(row total * column total)/grand total*.

After calculating the expected value, we then apply the following formula to calculate the value of the Chi-Square test of Independence:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad \text{.....(2)}$$

Where:
O is the observed frequency
E is the expected frequency
c is degrees of freedom (df)

In the above formula (2) the expected frequencies are subtracted from the observed frequency values and the resultant values are squared and then they are divided by the expected frequency finally to produce the result.

Degree of freedom is calculated by using the following formula:
DF = (r-1)(c-1) .....(3)
Where:
DF = Degree of freedom
r = number of rows
c = number of columns

2.4 Hypothesis

- **Null hypothesis:** Assumes that there is no association between the two categorical variables.
- **Alternative hypothesis:** Assumes that there is an association between the two categorical variables.

2.5 Hypothesis Testing

Hypothesis testing for the chi-square test of independence as it is for other tests, where a test statistic (calculated) is computed and compared to a critical value (tabulated). The critical value for the chi-square statistic is determined by the level of significance (typically 0.05) and the degrees of freedom. The degrees of freedom for the chi-square are calculated using the following formula: df = (r-1)(c-1) where r is the number of rows and c is the number of columns. If the observed chi-square test statistic is greater than the critical value, the null hypothesis can be rejected.

2.5.1 General notation for 2×2 contingency table

The general notation for a 2×2 contingency table is given in table 1 below. Note that we can also have a 3x2, a 2x3 or 3×3 tables.

Table 1: General notation for a 2×2 contingency table (observed values for the data)

	Column 1	**Column 2**	**Totals**
Row 1	a	b	**a+b**
Row 2	c	d	**c+d**
Totals	**a+c**	**b+d**	**a+b+c+d= total number of samples (N)**

Table 2: General notation for a 3×3 contingency table (observed values for the data)

	Column 1	Column 2	Column 3	**Totals**
Row 1	a	b	**c**	**a+b+c**
Row 2	d	e	f	**d+e+f**
Raw 3	g	h	i	**g+h+i**
Totals	**a+d+g**	**b+e+h**	**C+f+i**	**a+b+c+d+e+f+g+h+i = total number of samples (N)**

Table 4: General notation for a 2×2 contingency table (Expected values for the data)

	Column 1	Column 2
Row 1	(a+b)*(a+c)/N	(a+b)*(b+d)/N
Row 2	(c+d)*(a+c)/N	(c+d)*(b+d)/N

*Expected values= corresponding row total * corresponding column total)/total number of samples (N)*

2.5.2 Reject of fail to reject the null hypothesis

Chi square criterion

If the chi-square calculated value is greater than the chi-square critical value, then we reject the null hypothesis. If the chi-square calculated value is less than the chi-square critical value, then, we "fail to reject" the null hypothesis.

P Value criterion

If P-value is smaller than a pre-specified level (called significance level, 5% for example), then the null hypothesis is rejected. That is to say:- If the p-value is less than or equal to the significance level, we reject the null hypothesis. If the p-value is larger than the significance level, we fail to reject the null hypothesis because we do not have enough evidence to conclude that the data do not follow the distribution with specified proportions (That is Fail to reject H_0 scenario).

Suppose we want to find out that, whether there is an association between smoking and lung disease. In this case we use a 2×2 contingency table**.** Assume the following hypothetical data (Table 5).

We state the hypothesis as follows:
Null hypothesis (H_0): There is no association between smoking and lung disease
Alternative hypothesis (H_1): There is an association between smoking and lung disease.

Table 5: Hypothetical data containing observed values

	Smokers	Nonsmokers	**Totals**
Suffering from lung disease	39	18	**57**
Not Suffering from lung disease	34	29	**63**
Total	**73**	**47**	**120**

Then: How do we generate expected values from the observed frequencies? See Table 6 below.

Table 6: Expected values from observed values

	Smokers	Nonsmokers
Suffering from lung disease	57*73/120 =34.68	57*47/120 =22.33
Not Suffering from lung disease	63*73/120 =38.33	63*47/120 =24.68

*Expected values = corresponding row total * corresponding column total)/total number of samples (N)*

The Chi-square value for our example is 2.14, df = 1 as shown in Table 7 below. If we want to test our hypothesis at 5% level of significance then our predetermined alpha level of significance is 0.05. Looking into the Chi-square distribution table (Table 8) with 1 degree of freedom (calculated using equation 3) and reading along the row we find out that the chi-square critical value is 3.841.

Using Chi square criterion:
It can be observed that, the chi-square calculated value (2.14) is less than the chi-square critical value (3.841), then, in this case we "fail to reject" the null hypothesis.

Using P Value criterion: The chi-square calculated value (2.14) lies between 2.706 and 3.841. The corresponding probability is between the 0.10 and 0.05 probability levels. That means that the *P* value is above 0.05. Since the p-value is larger than the significance level (predetermined alpha level of significance was 0.05), we fail to reject the null hypothesis.

Conclusion: We conclude that smoking and lung disease are independent, or simply there is no relationship between smoking and lung disease. **Note that:** The two criteria must arrive to the same conclusion. But also take note that of the conclusion that, we used hypothetical data.

Table 7: Summarizing the data for calculating the Chi-square value

Observed count (O_i)	Expected count (E_i)	(O_i-E_i)	(O_i-E_i)^2	(O_i-E_i)^2/E_i
36	34.68	1.32	1.74	0.05
18	22.33	(4.33)	18.75	0.84
34	38.33	(4.33)	18.75	0.49
29	24.68	4.32	18.66	0.76
χ^2				2.14

We calculate the degree of freedom (df) using formula in equation 3 above as (2-1) * (2-1) = 1

Table 8: Extract from the Chi-square distribution table

df	0.995	0.9	0.1	0.05	0.025	0.01	0.005
1	0.000	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	1.610	9.236	11.070	12.833	15.086	16.750

3.0 Empirical Application of Chi-square Test

3. 1 Materials and Methods

3.2 Data for the study

We now use the empirical data from the study which was conducted between September 2017 and March, 2018 in two municipalities namely Dodoma and Morogoro, Tanzania. The sampling frame was division, wards, and finally a household with or without livestock. The study involved one urban division from Dodoma Municipality where eight (8) wards were selected; and Morogoro Urban Division which also constitute the Morogoro Urban District where seven (7) wards were selected based on livestock population densities; making a total of 15 wards. There were 345,884 households in the study area where 2,681of them were keeping livestock. A cross-sectional survey involving 298 households was conducted. The determination of this sample was based on the formula by Cochran (1977) as follows:

$$n = \frac{Z^2 (1-p)\, p}{(ME)^2}$$

Where,

n, is a sample size,

Z, is critical value (1.96 for 95% confidence interval);

p, is proportion of the livestock keeping households in the population; (2,681/345,884)

ME, is marginal error (1%).

Out of the 298 respondents, 158 were drawn from Dodoma Municipal Council and 140 were

from Morogoro Municipal Council

Data collection methods included interviews to household heads using semi-structured questionnaire, discussion with key informants and observation. Both closed and open-ended questions were included in the household questionnaires. The information sought included respondent's characteristics (age, gender, education, marital status and type of occupation), number of livestock, types of livestock (cattle, pigs, goat, sheep and poultry), grazing systems), bylaws, awareness of bylaws, number of extension staff, environmental pollution (odour, animal waste heaps, dust, noise plants' destruction), waste disposal, and occurrences of conflict.

3.3 Research Hypotheses

We state the hypothesis as follows:

Null hypothesis (H_0): There is no association between keeping livestock and environmental pollution

Alternative hypothesis (H_1): There is an association between keeping livestock and environmental pollution

4.0 Empirical Results and Discussion

The chi-square tests were conducted to ascertain whether the two categorical variables under the study (keeping livestock and environmental pollution) are independent or not. Table 9 shows the test results on independence between livestock keeping (cattle, pig, goat, sheep, poultry) and environmental pollution (odour, noise, heaps of wastes, dust, plant destruction and conflict).

The chi-square test of association between keeping cattle and environmental pollution rejected the null hypothesis of independence at 5% level of significance on pollution variables except one (dust), implying that keeping cattle could result into noise, heaps of waste, odour, and plant destruction. However, the null hypothesis of independence between cattle and conflict were also rejected at 5% level of significance indicating that keeping cattle could not results into conflict among community members in the study area. The fact that the Chi-square test failed to reject null hypothesis of independence at 5% level of significance between keeping cattle and environmental pollution resulting to dust implies that there is little or no evidence to suggest that keeping cattle could cause dust among the community in the study area.

The chi-square test of association between keeping pig and environmental pollution rejected the null hypothesis of independence at 5% level of significance on all cases variables except one (dust), implying that keeping pig in urban areas could result into environmental pollution namely odour, noise, plant destruction and heaps of waste. Further analysis indicated that keeping pig in urban areas could also result into conflict among the community in the study areas at 5% level of significance.

Similarly, the chi-square test of independence between keeping goats and environmental

pollution rejected null hypothesis of independence at 5% level of significance on all cases variables except one (dust), implying that keeping goats in urban areas also could result into environmental pollution namely odour, noise, plant destruction and heaps of waste. Further analysis indicated that keeping goat in urban areas could also result into conflict among the community in the study areas at 5% level of significance.

Following a Chi-square test of independence conducted to ascertain whether keeping sheep could results into environmental pollution, the test results rejected the null hypothesis of independence at 5% level of significance on two cases (plant destruction and heaps of waste); implying that keeping sheep in urban areas could result into environmental pollution namely plant destruction and heaps of waste. The test statistic failed to reject null hypothesis of independence between keeping sheep in urban areas and environmental pollution namely, odor, noise and dust respectively, also test statistic failed to reject null hypothesis of independence between keeping sheep in urban areas and social conflict.

In this category of there is little or no evidence to suggest that keeping sheep could cause odor, noise, dust and social conflict among the community in the study area on the basis of the data provided With regards to poultry keeping the chi-square test of independence rejected null hypothesis of independence at 5% level of significance on all cases, except two cases (noise and dust) implying that keeping poultry in urban areas could also result into environmental pollution namely, odour, plant destruction and heaps of waste. In this category of livestock the analysis indicated that keeping poultry does not result into noise and dust respectively; but could result into conflict among the community in the study areas at 5% level of significance. While these results cannot be taken on absolute terms as voiced out by the respondents, they are nevertheless an important reflection on how people feel bad to see the problems that are caused by urban livestock keeping in their areas on daily basis.

Based on the foregoing discussion on effects of urban livestock keeping in the two Municipal cities of Dodoma and Morogoro, it can be argued that as much as livestock keeping has continued to be integral part of urban life, its management has continued to fall short of proper urban development dynamics. There is poor animal waste disposal resulting into absurd heaps, livestock cause noise, destructs infrastructure and gardens, cause dusty conditions, nasty smell and, diseases to urban dwellers. Generally, all types of livestock cause certain types of challenges with varying degrees of magnitude. The gravity of each type of environmental problem will certainly differ with the type of livestock involved. The main conclusion is that environmental effects of urban livestock keeping are demonstrated by all types of livestock at varying degrees. Livestock keeping of any type in urban areas has negative environmental and health consequences that can only be mitigated through effective enforcement of relevant municipal bylaws.

5.0 Conclusions and Recommendations

The article has demonstrated theoretically and empirically the hypothesis testing for the association between categorical variables using Chi-square Test. It can be concluded that the

Chi-square test, only tells us the probability of independence of a distribution of data or in simple terms it does only test that whether two categorical variables are associated with each other or not. It does not tell us that how closely they are associated. We recommend that, once we got to know that there is a relation between these two variables, we need to explore other methods to calculate the amount of association between them.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing Interests: Authors declared that they have no conflict of interests

Acknowledgement
The authors are grateful to households in Dodoma and Morogoro Municipals who were involved during the survey.

References

Plackett, R.L. (1983). Karl Pearson and the Chi-squared test, *International Statistical Review*;*51:59-72.*

Magnello M.E (2006). Karl Pearson and the origin of modern statistics: An electrician becomes a statistician, *Rutherford J, Vol.* 1, 2005-2006.

Pearson K. (1992) 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. In: Kotz S., Johnson N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY

Table 9: Association between keeping livestock and environmental pollution

	Environmental Pollution											
Keeping	**Odor**		**Noise**		**Heaps**		**Dust**		**Plant Destruction**		**Conflict**	
Livestock	**chi2**	**P-value**	**chi2**	**P-value**	**chi2**	**P-value**	**chi2**	**P-value**	**chi2**	**P-value**	**chi2**	**P-value**
Cattle	64.03	0.000	108.91	0.000	59.42	0.000	0.95	0.330	53.67	0.000	45.80	0.000
Pig	209.45	0.000	185.09	0.000	159.53	0.000	0.7382	0.390	172.09	0.000	163.49	0.000
Goat	275.13	0.000	191.75	0.000	185.49	0.000	2.1552	0.142	98.31	0.000	206.79	0.000
Sheep	0.0535	0.817	0.2982	0.585	186.44	0.000	1.8262	0.177	63.19	0.000	0.5716	0.450
Poultry	158.55	0.000	1.1645	0.281	242.80	0.000	0.7634	0.382	117.88	0.000	84.92	0.000