ABACUS, Pages 193 – 199, Vol. 43, No. 2, 2016

Application of Pearson's X^2 Test of Independence with Small Expected Cell Frequencies

[1]Sanni, O. O. M.; [2]Abidoye, A. O.; [3]Ikoba, N. A.
[1,2,3]Department of Statistics, University of Ilorin, Ilorin, Nigeria

ABSTRACT

Pearson's X^2 is re-examined within the context of small expected cell frequency (e_{ij}< 5), for the Pearson's X^2statistic to satisfy the asymptotic approximation to Chi-square distribution. This paper proposes scalar multiplier 'θ', θ> 0, such that $\theta e_{ij}' \geq 5$, where e_{ij}' is the smallest expected cell count in the contingency table under consideration. The product of the sample size 'n' and 'θ' results in each cell count becoming θn_{ij}, which does not cause any change in the cell probabilities, p_{ij}.Thus the assumption of independence is thereby satisfied. This approach guarantees the safe application of Pearson's X^2 for test of independence under small expected cell counts with the degrees of freedom also multiplied by θ.

Keywords: *Goodness-of-fit; chi-square; scalar multiplier; small expected cell frequency; independence hypothesis.*

1. INTRODUCTION

In many field of studies, such as medicine, social sciences, education, and so on, we come across categorical data for which a number of observations are cross-classified by some categorical variables that satisfy the assumption of independence.Consequently the hypothesis testing is made most frequently by the use of goodness-of-fit tests like the Pearson's X^2, the likelihood ratio statistic, Y^2 and Freeman-Turkey statistic, T^2. Slight variations in the data collection scheme usually permits the use of some other assumptions regarding the underlying distribution without changing the estimates of the expected cell counts (Birch,1963).

The most commonly used test statistic in goodness-of-fit tests is Pearson's X^2 defined as

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \qquad (1.1)$$

where n_{ij} are the observed frequencies and the e_{ij} are the corresponding frequencies expected in cell (i,j). Under the hypothesis of independence of row and column classifications, then the expected cell frequency (e_{ij}) has an estimate given by

$$e_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad (1.2)$$

where $n_{i.} = \sum_j n_{ij}$, $n_{.j} = \sum_i n_{ij}$, $and\ n.. = \sum_{ij} n_{ij}$

It is well known that when all the e_{ij} are not small, the X^2 statistic is distributed approximately as chi-square with (r-1)(c-1) degrees of freedom . In cases where not all expected cell counts are large enough to meet the assumption, the traditional practice adopted in order to meet the requirements of minimum expected cell count was to collapse such cell(s) with the neighbouring cell. This method has serious disadvantages in that information on the collapsed cell (s) are lost and the structure of the table may be disrupted. There is a wide difference of opinion about how small the e_{ij} can be without invalidating the chi-square approximation. Many computer programs for goodness-of-fit tests such as Dixion's BMDP (1981), Dean et. al. Epi-Info (1990), Minitab (1991), etc, caution users when expected count of any cell is less than five. Fisher (1925) recommends that no expectation be less than five. Cramer (1946) emphasizes the need for expectation of at least ten, Kendall (1953) states that the approximation may "confidently be applied when all the theoretical cell frequencies are, such that, none is less than 20". Cochran (1954) states that in goodness-of-fit tests of unimodal distributions, such as the Normal or Poisson, expectations at one or both tails should be at least one. Lewontin and Felsenstein (1965) found that the test was satisfactory with some expected counts equal to 0.5, Yarnold (1970) suggested that the minimum cell expected frequency should be greater than 5s/rc, where 's' is the number of cells with expectations less than 5 and 'r x c' is the table dimension. Agresti (1990) maintained that it would be sufficient to caution on the use of goodness-of-fit test statistics when the contingency table's sparseness index (Is), which is the ratio of sample size to the dimension of the table, (i.e, Is = n/rc), is low. Sanni (1997) found that X^2 can still be used to approximate chi-square when the minimum expected cell frequency is as low as 0.1. Sanni and Jolayemi (1997) observed that the Pearson's X^2 achieved correctly any perceived significance level when the sparseness index (Is) and the minimum expected cell count are as small as 0.3 and 0.1, respectively. In this paper, we re-examined the Pearson's X^2 as an approximation to chi-square distribution using a two dimensional r×c (r≥2, c≥2) contingency table under the small expected cell counts from another perspective.

In section 2, we discuss briefly the Pearson's X^2 test statistic in an r × c contingency table. In section 3 a proposal is suggested that deals with some few illustrative numerical examples. This paper is concluded with conclusions and recommendation in section 4.

2. THE SCALED PEARSON'S X^2 STATISTIC

Consider a two-dimensional r × c contingency table and let n_{ij} be the cell frequency (i,j), i=1,2,…,r and j=1,2,…,c. Without loss of generality with respect to any appropriate underlying sampling distribution, see for example, Birch (1963). We assume that the cell counts have the multinomial distribution. That is, the probability mass function of the set of counts is given by

$$P_r(n_{1.}, n_{2.}, n_{r.}; p_{1,}, p_{2.}, p_{r.}) = \binom{n}{n_{1.}, n_{2.}, \ldots\ldots\ldots, n_{r.}} \prod_{ij}^{rc} p_{ij}^{n} \quad (2.1)$$

where $n_{i..} = (n_{i1}, n_{i2}, n_{ic})$, $p_{i..} = (p_{i1,} p_{i2,}, p_{ic})$ and $\sum_{ij} n_{ij} = n$

The cell counts n_{ij} were simulated such that

$$E(n_{ij}) = np_{ij}. \quad (2.2)$$

where p_{ij} ,the cell probability satisfied the independence condition

$$p_{ij} = p_{i.} p_{.j} \quad (2.3)$$

where $p_{i.} = \sum_{j} p_{ij}$ and $p_{.j} = \sum_{i} p_{ij}$

In real life situation we often come across situations in which the sample sizes are not large enough to guarantee minimum expected cell count of at least five. For example, in medicine and biological experiments, where the nature of sample units do not permit the use of large sample sizes. Thus the asymptotic approximation of chi-square by Pearson's X^2 is threatened. In order to ensure that the minimum expected cell count satisfies the conventional minimum expected cell count of at least five, we choose an integer "a" such that $ae^*_{ij} = e'_{ij} \geq 5$,where now the new expected cell count is e^*_{ij}. By multiplying vector n by "a" such that the π_{ij} remain unchanged. Thus we have

$$N = \theta \underline{n} = \begin{bmatrix} \theta n_{11} & \theta n_{12} & \theta n_{13} & \ldots\ldots\ldots\ldots\ldots\ldots & \theta n_{1c} \\ \theta n_{21} & \theta n_{22} & \theta n_{23} & \ldots\ldots\ldots\ldots\ldots\ldots & \theta n_{2c} \\ . & . & . & \ldots\ldots\ldots\ldots\ldots\ldots & . \\ . & . & . & \ldots\ldots\ldots\ldots\ldots\ldots & . \\ \theta n_{r1} & \theta n_{r2} & \theta n_{r3} & \ldots\ldots\ldots\ldots\ldots\ldots & \theta n_{rc} \end{bmatrix} \quad (2.4)$$

When the independence hypothesis (H_0) is true, the expected cell count becomes

$$e_{ij}' = \theta n_{i.} \theta n_{.j} / \theta n = \theta e_{ij} \quad (2.5)$$

From (2.4) and (2.5) the Pearson's X^2 is given by

$$X^{2'} = \sum_{ij} (n_{ij}' - e_{ij}')^2 / e_{ij}' \quad (2.6)$$

where n_{ij}' and e_{ij}' are the observed and the corresponding expected cell counts under the scaled factor. Thus $X^{2'}$ has an approximate chi-square distribution with θ(r-1)(c-1) degrees of freedom. This approach ensures that π_{ij} 's are unaltered when the overall sample size is now scaled up to 'θn..' instead of n..

2. NUMERICAL EXAMPLES

Two examples are selected such that their minimum original expected cell counts are less than 5 in order to demonstrate the application of Pearson's X^2 under these conditions. The first example is taken from Sanni (1997) and the second is the example in Gupta (2013) page .The decision rule under this test statistic is such that H_0 is rejected if $X^2_{cal} > \chi_{\theta(r-1)(c-1)}\alpha$, otherwise accept H_0 .

Given the null hypothesis H_0 : $p_{i..} = p_{.j} = 0.25, 0.25, 0.25, 0.25$ one of the sample configuration from a simulation of a 4×4 contingency table with the associated marginal proportions under the null hypothesis of the independence of the rows and columns classification variables, is given in table 3.1 see Sanni (1997).

Table 3.1: A 4 x 4 simulated contingency table under equal marginal probabilities

Rows	Columns				Total
	1	2	3	4	
1	4	2	3	5	14
2	3	5	4	2	14
3	3	3	5	3	14
4	4	4	2	4	14
Total	14	14	14	14	56

From the above table the minimum expected cell count under independence hypothesis is

e_{ij} = 14 x 14/56 = 3.5 < 5

The scaling multiplier 'θ' required such that $e_{ij}' \geq 5$ is 'θ' = 5/3.5 ≈ 2 (rounded up to the nearest integer), this ensures that the minimum expected cell count is not less than 5. As a result of factor 'θ' the new transformed table with the expected cell frequencies in the parenthesis are as presented in the table below.

Table 3.2: Modified Table of Observed and Expected Frequencies in the Parentheses

Rows	Columns 1	2	3	4	Total
1	8 (7.0)	4(7.0)	4(7.0)	10(7.0)	28
2	6(7.0)	10(7.0)	8(7.0)	4(7.0)	28
3	6(7.0)	6(7.0)	10(7.0)	6(7.0)	28
4	8(7.0)	8(7.0)	4(7.0)	8(7.0)	28
Total	28	28	28	28	112

The Pearson's X^{2}' is

$$X^{2'} = \sum_{i=1}^{4} \sum_{j=1}^{4} (n_{ij}' - e_{ij}')^2 / e_{ij}' = 9.143$$

with $\theta(r-1)(c-1) = 2(4-1)(4-1) = 18$ degrees of freedom.

As an illustration the following exercise is extracted from Gupta (2013) (Exercise 18.1, number 27). The table below gives the distribution of Mathematics and Economic scores in a sample of 25 students.

Table 3.3: Distribution of observed Mathematics scores against Economics scores.

Economics Scores	Mathematics Scores <70	≥ 70	Total
<70	5	2	7
≥ 70	7	11	18
	12	13	25

From the above information the following 2 x 2 contingency table can be summarized:

Table 3.4: Distribution of observed Mathematics scores against Economics scores with Expected values in parentheses

Economics Scores	Mathematics Scores		Total
	<70	≥ 70	
<70	5 (3.36)	2 (3.64)	7
≥ 70	7 (8.64)	11 (9.36)	18
	12	13	25

It is observed that the minimum expected cell count e_{11}=3.36, which is less than the minimum asymptotic value of 5, thus violating assumption for X^2 to be approximated by χ^2. In order to meet this condition our suggested scaled multiplier is 'θ' = 5/3.36=1.5 ≈ 2 (rounded up to next integer). This guarantees the minimum expected cell count $e_{ij}' \geq 5$ with the cell probabilities e_{ij}'s remain unchanged. The new resulting table is as reported in table 3.5.

Table 3.5: Scaled Table of Distribution of Mathematics scores against Economic scores with the Expected Frequency in parentheses

Economics Scores	Mathematics Scores		Total
	<70	≥ 70	
<70	10 (6.72)	4 (7.28)	14
≥ 70	14 (17.28)	22 (18.72)	36
Total	24	26	50

Under the independence hypothesis, $X^{2'}$= 4.276

This leads to the decision that we fail to reject the null hypothesis of independence at the 5% level of significance. Hence Mathematics and Economics may be regarded as unrelated on the basis of the sample, as was also concluded in exercise 18.1, number 27 in Gupta (2013).

4. CONCLUSION

By scaling the expected minimum cell frequency to meet the asymptotic conventional minimum expected cell count of at least 5, this ensures the approximation of Pearson's X^2 by chi-square distribution with the degrees of freedom. Thus scaled by same factor .Thus the obtained value of X^2 produces the same expected conclusion. The approach presented in this paper has been shown to have considerably ameliorated the problem of expected cell count not meeting the minimum asymptotic condition. We recommend that to obtain good approximation of Pearson's X^2 by chi–square distribution, we are suggesting a scaling factor $\theta, such\ that\ \theta e_{ij} \geq 5\ (e_{ij}')$ where 'θ' is an integer (rounded up) and e_{ij}' is the minimum expected cell count.

REFERENCES

Agresti, A. (1990): Categorical Data Analysis.Wiley, New York.

Birch, M. W. (1963): Maximum likelihood in three-way contingency tables. Journal of the Royal Statistical Society Ser. B25, 220-233.

Cochran, W.G. (1954): Some Methods of Strengthening the Common Tests. Biometrics, 10, 417 – 451.

Cramer, H. (1946): Mathematical Methods of Statistics. Princeton, N.J., University Press.

Dean, A.G.; Dean, J.A.; Burlon, A.N. and Dicker, R.C. (1990): Epi-Info. Version 5: A Word Processing Database and Statistics Program for Epidemiology or Microcomputer (USD). Incorporated Stone.Mountain Georgia.

Dixion, W.J.(1981): BMDP Staistical Software, University of California Press Berkeley; Los Angeles.

Fisher, R.A. (1925): The Significance of Deviations for Expectations in Poisson Series. Biometrics, 6, 17 – 24.

Gupta, S. C.(2013): Fundamentals of Statistics, Seventh Edition, Himalaya Publishing House PVT LTD, Mumbai, India.

Kendall, M.G. (1953): The Advanced Theory of Statistics, Vol.1, 5th Edition, Griffin,London.

Lewontin, R.C. andFelsenstein, J. (1965): The Robustness of Homogeneity Tests in 2 x N tables, Biometrics 21, 19 – 33.

Minitab Reference Manual PC. Version 8, 1981.

Sanni, O.O.M. (1997): A Study of Some Goodness-of–fit Statistics for Contingency Tables with Sparse data. Unpublished Ph.D Thesis, University of Ilorin, Ilorin, Nigeria.

Sanni, O. O. M.and Jolayemi, E.T. (1997): On the Use of Some Categorical Test Statistics on Sparse Contingency Table. Journal of Pure and Applied Sciences, Faculty of Science, University of Ilorin, Nigeria.

Yarnold, J. K. (1970): The minimum expectation of χ^2 goodness-of-fit tests and the accuracy of approximations for the null distribution. Journal of American Statistical Association, 65, 864-866.