American Journal of Applied Sciences

Original Research Paper

The Analysis of High Dimensional Contingency Table with Interactions

**Keorapetse Sediakgotla, Lesego Gabaitiri,
Keamogetse Setlhare, Njoku Ola Ama and Dahud Kehinde Shangodoyin**

Department of Statistics, University of Botswana, Botswana

Article history
Received: 12-12-2016
Revised: 23-03-2017
Accepted: 14-07-2017

Corresponding Auhtor:
Keorapetse Sediakgotla
Department of Statistics,
University of Botswana,
Botswana
Email: sediakgo@mopipi.ub.bw

Abstract: This paper is focused on the analysis of categorical data in a $2 \times c \times K$ contingency table. The theoretical frame work of a $2 \times c$ design is extended to $2 \times c \times K$ with provision for testing interactions among subsets of either lower or upper columns of the designated table. The developed chi square tests for the total interactions as well as for the partitions are shown to be significant and the degrees of freedom additive.

Keywords: Contingency Table, Chi-Square Test Statistics, Interactions and Proportions

Introduction

The development of methods for analyzing categorical data begun about four decades ago after Bartlett (1935) made challenging comments on the lack of contributions to contingency tables of high dimensions. Most data in social sciences and more importantly survey data come handy in categorized forms and researchers often make use of chi-square tests to perform necessary statistical tests of associations. Adamu (1969) had commented that the common approach by many researchers in social sciences is to calculate an overall Chi-square for each contingency table and on the basis of the critical value for this test statistic decide whether to accept or reject the hypothesis of association between variables that form the basis of classification of the table. The decision on the computed statistic might be simple for a $2 \times c$ contingency table but the challenges faced with researchers on the interpretation of multi-way contingency tables are enormous, especially when considering interactions among some variables. A number of authors have presented some procedures in dealing with the analysis of interaction in multi-dimensional contingency tables Ostle (1954); Lewis (1962). Lancaster (1960) considered a canonical form, or rather a class of canonical forms, for three dimensional probability distributions subject to a rather mild restriction of fixed margins and developed suitable tests of independence and lead to a consideration of the partition of χ^2 in the analysis of complex contingency table. Ama (1992) developed one degree of freedom chi-square test for interaction in an r×c, two-way contingency table in a manner similar to the Turkey's one degree of freedom F-test of interaction in a two factorial experiment. The test was extended to an $r \times c \times K$ three-way contingency table by the same author and further developed a 1 d.f. chi-square test for the 3 way interaction after re-parameterization of the interaction Ama (1994).

Bishop *et al*. (2007) have shown that we can collapse over one or more classifications only if those classifications are independent of at least one of the remaining classifications. Seligman (n.d) in their paper caution that collapsing tables without due justification can lead to incorrect results. For instance in the paper, if the sex classification is viewed as unimportant and collapsed (summed) the male and female categories, the resulting 2×2 table would lead to the rejection of the null hypothesis of independence of classifications in the collapsed table. Goodman (1964) proposed a definition of the order interactions in an m-dimensional ($d_1 \times d_2 \times \ldots \times d_m$) contingency table ($r = 0, 1, 2, \ldots, m$-1) and presented methods for testing the hypothesis that any specified subset of these interactions is equal to zero. In addition, the author presented simple methods for obtaining simultaneous confidence intervals for these interactions or for any specified subset of them. However, to the best of our knowledge none of these authors gave a breakdown of the overall Chi-square test for interaction in a multidimensional contingency table (especially when this test is significant) to accommodate various scenarios of the interactions such as the cases where in an $r \times c \times K$ contingency one margin is fixed or the total frequency is fixed.

According to Lewis (1962), although several important papers appeared on this subject, the treatment of these contingency tables is widely neglected in

standard text books. The paper presented by Adamu (1969) accommodated this aforementioned challenge but fell short of extension to higher dimensional contingency tables usually encountered in problems in Education, Medicine Science and in the Social Science, particularly where the response variable is dichotomous while the other variables are multi-level.

This paper, specifically, extends the derivations of Adamu (1969) to higher dimensional tables and presents an alternative statistic and test to the overall Chi-square test statistic for interaction in the case of $2 \times c \times K$ contingency table. The paper will be arranged in four sections. Section two immediately following this introduction will contain the derivation of the generalized chi-square statistic for the $2 \times c \times K$ contingency table and the partial test statistics. Section three will be the application of the methods to real life problem, while section four will contain the results and discussions**.**

Materials and Methods

Derivation of the 2 ×c ×K Chi-Square Test Statistic

In a general three-dimensional $r \times c \times K$ contingency table, r is the number of rows; c is the number of columns and K is the number of layers,($r > 2$; $c > 2$ and $K > 2$), let f_{ijk} denote the observed frequency in the cell of the i^{th} row, j^{th} column and k^{th} layer. Similarly, let P_{ijk} denote the probability of an observation belonging to the $(ijk)^{th}$ cell. The marginal totals over the row, column, layer, column × layer, row × layer and row × column are given respectively as:

$$m_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{K} f_{ijk}; \quad m_{.j.} = \sum_{i=1}^{r}\sum_{k=1}^{K} f_{ijk};$$
$$m_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} f_{ijk}; m_{.jk} = \sum_{i=1}^{r} f_{ijk}; m_{i.k} = \sum_{j=1}^{c} f_{ijk};$$
$$m_{ij.} = \sum_{k=1}^{K} f_{ijk} \text{ and grand total}, m_{...} = \sum_{ijk} f_{ijk}$$

The 'dot' notation indicates a summation over the subscripts. A similar notation can be written for the cell probabilities, P_{ijk}, with $P_{...} = 1$. When the row classification is dichotomous, the general $r \times c$ contingency table becomes $2 \times c \times K$ contingency table shown in Table 1.

One interest in the analysis of contingency table is usually to test whether there is mutual independence or association between the ways of classification, that is, to test the null hypothesis H_0: $P_{ijk} = P_{i..} P_{j.} P_{.k}$ against the alternative hypothesis H_1: $P_{ijk} \neq P_{i..} P_{j} P_{.k}$. The test rejects the null hypothesis in favour of the alternative hypothesis if the overall chi-square test statistic:

$$X^2 = \sum_{ijk} (Observed - Expected)^2 / Expected$$
$$\forall i = 1,2\ j = 1,...c \text{ and } k = 1,.....,l \quad (1)$$

Is greater than $X^2_{(r-1)(c-1)(l-1)}(a)$, at a- level of significance. Adamu (1969) had presented a general formula for the overall Chi-square in (1) when testing for proportion of success in a $2 \times c$ contingency table without consideration for the k-th upper columns. In this study we consider a $2 \times c$ contingency table as assumed in Adamu (1969) and then provide an alternative overall Chi-square test statistic:

$$Let\ e_{ij} = \hat{p}m_{.j}, \hat{p} = \frac{m_{1..}}{m_{..}} \text{ and } p_{ij} = \frac{f_{ij}}{m_{.j}} \quad (2)$$

where, e_{ij} is the expected frequency in the $(ij)^{th}$ cell; p_{ij} is the probability of an observation belonging to the $(ij)^{th}$ cell; $\hat{p}$ is the estimated total probability

The generalized form of Equation 1 for a $2 \times c$ contingency table is:

$$T = \sum_{i}\sum_{i}\left[\frac{f_{ij}^2 - 2\hat{p}f_{ij}m_{.j} + \hat{p}_2 m_{.j}^2}{\hat{p}m_{.j}}\right]$$

Using the quantities in Equation 2 we derive the overall Chi-square test statistic for the $2 \times c$ contingency table (a special case of Table 1 when the upper columns are neglected) is:

$$\begin{aligned} T &= \sum_{i}\sum_{i}\left[\frac{f_{ij}^2 p_{ij} m_{.j} - 2f_{ij}\hat{p}m_{.j} + \hat{p}^2 m_{.j}^2}{\hat{p}m_{.j}}\right] \\ &= \sum_{i}\sum_{i}\left[\left(\frac{f_{ij}^2 p_{ij} m_{.j}}{\hat{p}m_{.j}}\right) - \left(\frac{2f_{ij}\hat{p}m_{.j}}{\hat{p}m_{.j}}\right) + \hat{p}m_{.j}\right] \\ &= \sum_{i}\sum_{j}\frac{f_{ij}p_{ij}}{\hat{p}} - 2\sum\sum f_{ij} + \sum_{i}\sum_{j}\hat{p}m_{.j} \end{aligned} \quad (3)$$

Adopting the quantities in Equation 2 and assuming the summand is a linear operator for which i and j could be used without restriction, then we have the right hand side of expression (3) as:

$$\begin{aligned} T &= \sum_{i}\sum_{j}\frac{f_{ij}^2 p_{ij}}{\hat{p}} - 2m_{..} + \sum_{i}\hat{p}m_{..} \\ &= \sum_{i}\sum_{j}\frac{f_{ij}^2 p_{ij}}{\hat{p}} - 2m_{..} + 2\hat{p}m_{..} \\ &= \frac{\sum\sum f_{ij}p_{ij} - 2\hat{p}m_{..} + 2\hat{p}2m_{..}}{\hat{p}} \\ &= \frac{\sum\sum f_{ij}p_{ij} - 2m_{..}\hat{p}\hat{q}}{\hat{p}} \end{aligned} \quad (4)$$

Table 1. Layout for the 2 × *c* × *K* three-way contingency table

	Columns																
	1				2				j				c				
	Layers				Layers				Layers				Layers				
Rows	1	2	……	K	1	2	…..	K	1	2	……	K	1	2	….	K	Total
1	f_{111}	f_{112}	…..	f_{11K}	f_{121}	f_{122}	……	f_{12K}	f_{1j1}	f_{1j2}	……	f_{1jK}	f_{1c1}	f_{1c2}	……	f_{1cK}	$m_{1..}$
2	f_{211}	f_{212}	…..	f_{21K}	f_{221}	f_{222}	……	f_{22K}	f_{2i1}	f_{2i2}	……	f_{2jK}	f_{2c1}	f_{2c2}	……	f_{2cK}	$m_{2..}$
Total	$m_{.11}$	$m_{.12}$	…..	$m_{.1K}$	$m_{.21}$	$m_{.22}$	…..	$m_{.2K}$	$m_{.j1}$	$m_{.j2}$	…..	$m_{.jK}$	$m_{.c1}$	$m_{.c2}$	….	$m_{.cK}$	

Equation 4 is an alternative expression of the Chi-square test statistic provided by Adamu (1969) and provides an overall criterion for testing proportions in a 2 × *c* table.

We shall consider the test statistic for interactions in a 2 × *c* × *K* contingency table. Let f_{ijk} denote the observed frequency in the i^{th} row, j^{th} lower column and k^{th} upper layer of the 2 × *c* × *K*, three-dimensional contingency table, where, *i* = 1, 2; *j* = 1, 2, …., c and *k* = 1, 2, …., *K*. We denote *x*, *y* and *z* as the 3-variable with values in natural order; f_{ijk} is the cell frequency for (*i*, *j*, *k*)th cell where *i* = 1,2 *j* = 1,2,...,*c* and *k* = 1,2,.....1.

The general test statistic for interaction in a 2 × *c* × *K* is:

$$T=\sum_i\sum_j\sum_k\left(\frac{f_{ijk}^2-2f_{ijk}e_{ijk}-e_{ijk}^2}{e_{ijk}}\right) \tag{5}$$

where:

$$e_{ijk}=\frac{m_{i.k}m_{.jk}}{m_{..k}},\ p_{ijk}=\frac{f_{ijk}}{m_{..k}},$$
$$\hat{p}_{(ik)}=\frac{m_{i.k}}{m_{..k}}\ and\ \hat{p}=\frac{\sum_k m_{i.k}}{\sum_k m_{..k}}=\frac{m_{i..}}{m_{...}}$$

(For more details of these notations see Lewis (1962). Thus we have:

$$\begin{aligned}T&=\sum_i\sum_j\sum_k\left[\left(\frac{f_{ijk}^2}{e_{ijk}}\right)-2f_{ijk}+e_{ijk}\right]\\&=\sum_i\sum_j\sum_k\left[\frac{f_{ijk}p_{ijk}m_{..k}}{m_{..k}\hat{p}_{(ik)}}-2f_{ijk}+\hat{p}_{(ik)}m_{..k}\right]\\&=\sum_i\sum_j\sum_k\frac{f_{ijk}p_{ijk}}{\hat{p}_{(ik)}}-2\sum_i\sum_j\sum_k f_{ijk}+\sum_i\sum_k\hat{p}_{(ik)}m_{..k}\end{aligned} \tag{6}$$

Now assume the summands are operated on *i*,*j* and *k* without restrictions on ordering we have the test statistic as:

$$\begin{aligned}T&=\sum_i\sum_j\sum_k\frac{f_{ijk}p_{ijk}}{\hat{p}_{(ik)}}-2m_{..}+\hat{p}_{(ik)}\sum_i\left[\sum_k{}^m..k\right]\\&=\sum_i\sum_j\sum_k\frac{f_{ijk}p_{ijk}}{\hat{p}_{(ik)}}-2m_{...}+\hat{p}_{(ik)}2m_{...}\\&=\frac{\sum_i\sum_j\sum_k fi_{jk}p_{ijk}-2m_{...}\hat{p}_{(ik)}(1-\hat{p}_{(ik)})}{\hat{p}_{(ik)}}\\\therefore T_{Overall}&=\sum_i\sum_k\left[\frac{\sum_i\sum_j\sum_k f_{ijk}p_{ijk}-2m_{...}\hat{p}_{(ik)}\hat{q}_{(ik)}}{\hat{p}_{(ik)}}\right]\\&\approx\chi^2(c-1)(K-1)\end{aligned} \tag{7}$$

In this study we interpret χ^2 test in the context of significant difference between proportions. Many practical situations might demand to know whether the proportion of one group made up of the first *X* (lower columns) say X_1 to X_τ of the table is different from the remaining $X_{\tau+1}$ to X_c lower columns. Similarly we may wish to compare the difference between the first *Z* (upper columns) say Z_1 to Z_t and the remaining Z_{t+1} to Z_k. To circumvent the aforementioned scenarios, we breakdown the general expression in Equation 5 into different scenarios as:

- To test the variations among the first subdivision of say lower columns or upper columns we derive the test statistic from the overall Chi-square as:

$$\chi^2_{(I)}=\sum_i\sum_k\left[\frac{\sum_{i=1}^{2}\sum_{j=1}^{\tau}\sum_{k=1}^{K}f_{ijk}p_{ijk}-2m_{...}\hat{p}_{(ik)}^{(I)}\hat{q}_{(iK)}^{(I)}}{\hat{p}_{(ik)}}\right]\approx\chi^2_{(\tau-1)(K-1)} \tag{8}$$

- To test the variations among the last subdivision of say lower columns or upper columns we derive the test statistic from the overall Chi-square as:

$$\chi^2_{(II)}=\sum_i\sum_k\left[\frac{\sum_{1}^{2}\sum_{j=\tau+1}^{c}\sum_{k=1}^{K}f_{ijk}p_{ijk}-2m_{...}\hat{p}_{(ik)}^{(II)}\hat{q}_{(iK)}^{(II)}}{\hat{p}_{(ik)}}\right] \tag{9}$$
$$\approx\chi^2_{(c-\tau-2)(K-1)}$$

- To test for interaction, we derive the test statistic from the overall Chi-square as:

$$\chi^2_{(III)} = T_{\text{Overall}} - \chi^2_{(I)} - \chi^2_{(II)} \text{ follows } \chi^2_{(c-\tau-2)(K-1)-(c-1)(K-1)} \quad (10)$$

It is important to note that the partitioning of the various variables in the lower and upper columns is convenient and arithmetically logical if the variables are in their natural homogeneous order and it would be easy to make comparison between and within each group.

Data Analysis, Results and Discussion

The data below taken from Woodward (2013) represents the number of children who have caries and those who do not have caries (a control group) classified by age of child (in months) and age of mother (in years).

From the Table 2, let row i = 1,2 denote the dental caries and control respectively, column; j = 1,2,3 be the age of the child in months (<36, 36-47, ≥48). The age of the mother gives layers, for k = 1,2,3 (<25, 25-34, ≥35). The total chi-squared value provides a gross measure of the extent to which cell frequencies depart from expectation (Lewis, 1962) and was found to be, T = 3410.12 from our data. This statistic can be partitioned by using Equation 8-10. To test for interactions among the age of child for j = 1,2 (age groups of <36 and 36-47) with respect to the age of the mother and the dichotomous variable, caries and control (partition 1), we found a highly significant chi-squared value at a = 0.001, hence a significant interaction between the variables as given, $\chi_{(1)}$ = 1499.06. A similar argument applies for the interaction between the other level of the age of the child (≥ 48) and the other two sets of variables (partition II), which again shows highly significant interactions between the variables, $\chi^2_{(II)}$ = 603.53. The overall test of interactions between the three variables is obtained by subtraction (partition III, Equation 9). With $\chi^2_{III} = 1307.5$, the results show significant interaction between the age of the mother, age of the child and the dichotomous variable (caries and control).

The computed values for the derived Equation 7-10 are:

$$\chi^2_{overall} = \chi^2_{(r-1)(c-1)(l-1)} = \chi^2_{(2-1)(3-1)(3-1)} = \chi^2_4 = 3410.12$$

$$\chi^2_{(I)} = \chi^2_{(r-1)(\tau-1)(l-1)} = \chi^2_{(2-1)(2-1)(3-1)} = \chi^2_2 = 1499.06$$

$$\chi_2 \chi^2_{(r-1)((c-\tau)-1)(K-1)} = \chi^2_{(2-1)((3-2)-1)(3-1)} = \chi^2_0 = 603.53$$

$$\chi^2_{(III)} = \chi_{[(r-1)(c-1)(K-1)]-\{[(r-1)(\tau-1)(K-1)]+[(r-1)((c-\tau))-1)(K-1]\}}$$

$$= \chi_2 = 1307.53$$

Table 2. Number of children with and without caries by age of child and age of mother

	Age of mother (years)	Age of a child (months) <36	36-47	≥ 48	Total
Caries	<25	1	1	8	10
	25-34	4	18	46	68
	≥35	1	3	10	14
	Total	6	22	64	92
Control	<25	1	5	9	15
	25-34	16	67	113	196
	≥35	2	14	35	51
	Total	19	86	157	262

Conclusion

The paper, therefore, has given a breakdown of the overall Chi- square which can then be used to test for interactions on the multidimensional contingency tables. The chi square values for the partitions/ interactions as well as their degree of freedom should be additive.

Acknowledgement

The authors acknowledge the contributions of the referees to this paper.

Funding Information

We are grateful to the University of Botswana for providing facilities that enabled the completion of this paper.

Author's Contributions

Keorapetse Sediakgotla: Data Analysis, interpretations and conclusion writing, as well as the general revision and coordination of the manuscript.

Lesego Gabaitiri: Data Analysis and interpretations

Keamogetse Setlhare: Literature review. Read and edited the final manuscript.

Njoku Ola Ama: Formulation of the problem, development of the article including review of literature and analysis.

Dahud Kehinde Shangodoyin: Principal investigator and derived the $2 \times c \times K$ Chi-Square test statistics.

Ethics

This paper, to best of our knowledge has not been published elsewhere.

References

Adamu, S.O., 1969. Statistical analysis of some determinants of entrepreneurial success: A Nigerian case study-A theoretical consideration and extension. Niger. J. Econom. Soc. Stud., 11: 29-42.

Ama, N.O., 1992. Test for interaction in two way contingency tables. Stat. Anno 9: 589-606.

Ama, N.O., 1994. One parameter model for three-factor interactions in contingency tables. Biom. J., 36: 855-864.

Bartlett, M.S., 1935. Contingency tables interactions. J. Royal Stat. Society, Suppl., 2: 248-252.

Bishop, Y.M., S.E. Fienberg and P.W. Holland, 2007. Discrete Multivariate Analysis: Theory and Practice. 1st Edn., Springer Science and Business Media, New York, ISBN-10: 0387728058, pp: 559.

Goodman, L.A., 1964. Interactions in multidimensional contingency tables. Ann. Math. Stat., 35: 632-646

Lancaster, H.O., 1960. On tests of independence in several dimensions. J. Aust. Math. Soc., 1: 241-254.

Lewis, B.N., 1962. On the analysis of interaction in multi-dimensional contingency table. J. Royal Stat. Society Series A, 125: 88-117.

Ostle, 1954. Statistics in Research: Basic Concepts and Techniques for Research Work. 1st Edn., Iowa State College Press, Ames I.A.

Seligman, E.J. (n.d). Applications of multidimensional contingency tables to the analysis of termination counts in disability income claim data.

Woodward, M., 2013. Epidemiology: Study Design and Data Analysis. 1st Edn., CRC Press, ISBN-10: 1584880090, pp: 330.