IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728. Volume 5, Issue 6 (Mar. - Apr. 2013), PP 20-23
www.iosrjournals.org

Post Component Analysis of Categorical data

U. Sangeetha[1], M. Subbiah[2], M.R. Srinivasan[3]
[1](*Department of Management Studies, SSN College of Engineering, Chennai, India)*
[2] *(Department of Mathematics, L. N. Government College, Ponneri, India)*
[3] *(Department of Statistics, University of Madras, Chennai, India)*

Abstract: *Categorical data are often stratified into two dimensional I × J tables in order to test the independence of attributes. Agresti (1999) has discussed testing methods with partitioning property of chi square to extract components that describe certain aspects of the overall association in a table. In this work an attempt has been made to study the pattern in which sub-tables exhibit a sign of reverse association when compared to a significant association of attributes with regard to the original I × J table. Simulation studies and subsequent results of vote counting method indicate that 2 × 2 tables have the reversal component association when compared to higher order sub-tables; interestingly, in more than 90% cases. The computationally extensive and exhaustive procedure provide a better tool to understand the association between the categories that focus on the strongest differences among all comparisons, and could be practically and other aspects in experimental studies such as clinical trials.*

Keywords: *Contingency tables, Chi-Square test, Component Analysis.*

I. INTRODUCTION

Categorical data analysis has found a rapid development since 1960’s which might be due to the methodological sophistication of the social and biomedical sciences. It is not uncommon in social sciences for measuring attitudes and opinions on various issues and demographic characteristics such as gender race and social class. Categorical scales have been used extensively in biomedical sciences to measure many factors, such as severity of injury, degree of recovery from surgery, and stage of a disease etc. This progress has been found as an outcome of stimulus ties between social scientists and / or biomedical scientists and Statisticians. However, one of the most influential contributions by Karl Pearson and Yule on association between categorical variables has also been used extensively in many research works and by the practitioners in this large data driven era. The data for this analysis are displayed in a table with I rows and J columns and many software perform this analysis to provide a test of significance using a chi-square analysis testing the null hypothesis that there is no association between the two variables.

If the chi-square test results in a p-value of 0.05or smaller, the results are deemed significant, the null hypothesis is rejected. In such cases when there are only two groups being compared (a 2 × 2 table) it is easy to observe the association between them; however when there are three or more groups, there has been a need to identify the significant/non-significant association between pairs of groups. An extensive literature such as Berkson, (1938), Norton, (1945), Simpson, (1951), Cochran, (1954), Mantel and Haenszel, (1959), Goodman, (1964), Goodman, (1969), Blyth, 1972, Agresti, (1990), Agresti (1999), Sergio and Paulette (2011) provides a method based on portioning of chi-square it has laid a practical limitation to the users, in that partitioning the given I × J table satisfying the basic assumptions of the procedure.

This paper has identified a need based analysis between the practice and the theory and has provided a procedure based on the post component analysis which is similar to the multiple comparisons in ANOVA for continuous response variables. A primitive approach and subsequent analysis issues have been discussed with limited size of the underlying classification table; this paper includes a simulation study to understand the decisions on the Reversal Association Pattern (RAP) in a I × J table. Also, this attempt has laid a foundation to visualize this post component analysis under various model settings and with large contingency tables.

II. METHOD

A bivariate relationship has been defined by the joint distribution of the two associated variables. If X and Y denote two categorical response variables, X having I levels and Y having J levels then the IJ possible combinations of classification could be displayed in a contingency table, where the cells contains frequency counts of outcomes. Such a contingency table is referred to as an I × J table.

In a two-way contingency tables, the null hypothesis of statistical independence has been tested using Pearson chi-square statistic (Agresti 1990)

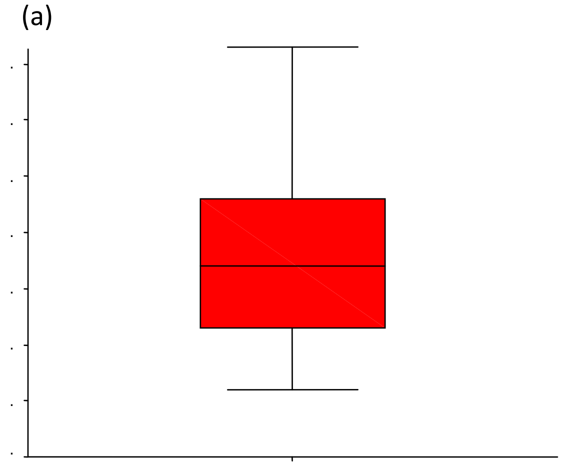
$\chi^2=\Sigma\Sigma\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$ where n_{ij} and m_{ij} are observed and expected frequencies of (i, j) cell (i=1,2,…,I; j=1,2,..,J). Interestingly m_{ij} used in χ^2 depend on row and column marginal totals, but not on the order in which the rows and columns are listed. Chi square does not change under permutations of rows or columns and a treats both classifications as nominal scales.

To understand the notion that the components represent certain aspects of the associations, the present attempt has relaxed the assumptions and include all possible a x b sub-tables. ($2 \leq a \leq I$ and $2 \leq b \leq J$). Chi square statistic have been based on this exhaustive and extensive list of sub-tables [$(2^I - I - 1)\ (2^J - J - 1)$-1] to portray the association pattern compared to the original table. In particular interest lies on the I × J tables which have significant over all association. By fixing size of the tables, the cell counts are simulated using resampling technique by considering arbitrary yet meaningful limits for lower and upper limits and suitable probability limits.

A long sequence of tables of required size has been generated to study the nature of associations in sub-tables compared to the original I × J table. Decisions are made based on vote counting procedures; by calculating the proportions of sub-tables exhibiting RAP to the number of possible sub-tables. While in such calculations it has been emphasized for classifying the sub-tables as basic 2 × 2 categories and rest of them as higher order through the original I × J table has not been considered as a sub table (as opposed to the notion of subsets)

III. Results

The RAP routine calculates the necessary values for the post component analysis and outputs a standard table of results. This includes the number of sub-tables with RAP, corresponding row and column number together with the conclusion of chi square test. The required proportions are calculated based on these numbers and conclusions are drawn there in. It further provides the distribution of cell counts in the given I × J table so as to understand the spread of the data that would be helpful in conjugating the new hypothesis pertain to RAP. However, due to paucity of space, this current simulation study has been limited to size 3 ×4 and 4 x 4. Figure 1 provides the box plot for the distribution of proportions of 2 × 2 sub-tables which have shown RAP; it could be observed that the pattern has been consistently high varies over the range of 60 % to 80 % in all the three cases considered and no extreme observation has been recorded. This pattern has a consistent behavior over the designs of the study.



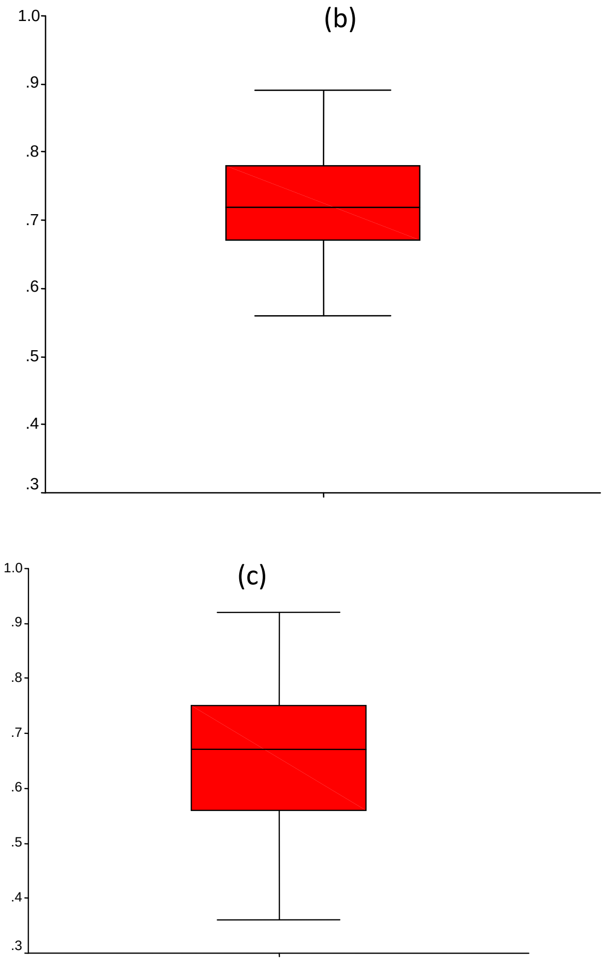


Figure 1: Box plot for the proportions of 2 x 2 tables which have RAP based on the simulation of three types of contingency tables (a) 3 × 4 (1 – 80) (b) 3 × 4 (101 – 160) (c) 4 × 4 (0 – 40)

Further, the proportion of sub-tables that exhibit the characteristic of reversal association has been summarized in Table 1. This includes the simulation parameters together with some summary measures to understand the notion of RAP among the three contingency tables considered for the study. Numerical values for the mean as well as minimum and maximum for the proportions pertain to 2 × 2 sub-tables show that the dominance in terms of reversal associations. Also, it could be observed that the mean proportions of other higher order sub–tables that have RAP, have been systematically less than that of corresponding 2 × 2 sub-tables. However, maximum proportion does not differ as much as minimum proportions do as small as 0%, which indicates that in some cases could have higher order sub-tables without RAP though there could be no sign of such pattern observed in 2 × 2 sub-tables

Table 1: Summary measures for the proportions of sub-tables which exhibit the Reversal Association Pattern (RAP) based on the simulation study

Table Size	Resample Limits		Summary Measure	Size of sub-tables with RAP		Proportion of sub-tables with RAP
	Lower	Upper		2 x 2	Other	
3 x 4	101	160	Mean	0.7222	0.5091	0.5983
			S E	0.0095	0.0142	0.0118
			Minimum	0.3889	0.1200	0.2558
			Maximum	0.8889	0.8800	0.8372
3 x 4	1	80	Mean	0.4592	0.1894	0.3023
			S E	0.0095	0.0097	0.0094

			Minimum	0.1667	0.0000	0.0930
			Maximum	0.8889	0.6800	0.7674
4 x 4	0	40	Mean	0.6506	0.3589	0.4464
			S E	0.0095	0.0128	0.0115
			Minimum	0.3056	0.0476	0.1333
			Maximum	0.9167	0.7500	0.8000

IV. CONCLUSION

In general a multi comparison procedure that has been applied in many statistical methods such as ANOVA could share a common aim. If a significant main effect or interaction is found, then there could be a significant difference amongst the levels of attributes somewhere. Yet still this significant effect has to be isolated exactly where the differences lie. However, in the case of categorical data, an attempt has been to understand the pattern in which such analysis could be performed; understanding the extensive possibilities of multiple comparisons between different levels of categories, this study has attempted an initial investigation to describe the pattern of reversal association compared to the original I × J table.

The study used a non - parametric simulation procedure to generate the required contingency tables that has varied cell frequencies including zeros in few cells. It could be observed that in more than 90% cases 2 × 2 tables have the reversal component association when compared to higher order sub-tables and interestingly, higher order tables that have this pattern have strong relationships with their 2 × 2 counterparts. Though, this approach might lack some theoretical exactness but in terms of consistency and ease and breadth of application this would enable to understand and draw general conclusions regarding the general purpose of any multi comparison procedures. However, similar attempts could be made with higher order contingency tables from possible parametric simulations over a range of values that would yield low or high cell frequencies, so that the method of Post Component could be applied in many practical experimental studies such as clinical trials to design a drug efficacy study and the associations between different levels of categories could be explicitly analyzed to have cost effective yet pragmatic decisions.

References

[1] J.Berkson,. Some difficulties of interpretation encountered in the application of the chi-square test, *Journal of the American Statistical Association 33,* 1938, 526-536.

[2] H.W,Norton, Calculation of chi-square for complex contingency tables, *J. Amer. Statist. Assoc. 40,* 1945, 251-258.

[3] E.H.Simpson, The interpretation of interaction in contingency tables, *J. Roy. Statist. Soc. B 13,*1951, 238-241.

[4] W.G. Cochran, Some methods for strengthening the common chi-square tests, *Biometrics 24,* 1954, 315-327.

[5] N. Mantel, and W.Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, *J. Nat. Cancer Inst. 22,* 1959, 719-748.

[6] L.A. Goodman, Simple methods for analyzing three-factor interaction in contingency tables, *J Amer. Statist. Assoc. 59,* 1964, 319-352.

[7] L.A. Goodman, On partitioning χ^2 and detecting partial association in three-way contingency tables, *J. Roy. Statist. Soc. B 31,* 1969, 486-498.

[8] C.R. Blyth, On Simpson's paradox and the sure thing principle. *J Amer. Statist. Assoc. 67,* 1972, 364-366.

[9] A.Agresti, *Categorical Data Analysis,* (New York: Wiley & Sons 1990) pp 51-54

[10] A.Agresti, Exact inference for categorical data: recent advances and continuing controversies, *Statistical Methods and Applications, 20,* 1999, 2709-2722.

[11] A.Sergio Estay , I. PauletteI Naulin, Data analysis in forest sciences: why do we continue using null hypothesis significance tests? *BOSQUE 32(1),*2011, 3-9.