A study on further characteristics of contingency tables with the R Package - RAP.

U. Sangeetha, M. Subbiah, M.R. Srinivasan

Abstract — The ubiquitous Chi square test statistic for association between two or more categorical variables provides ample scope to investigate its characteristic in terms of methodological or application point of view. Many studies have pointed out its appropriateness in case of sparse tables and also there are few attempts to understand the category wise associations through partitioning Chi square distribution using G^2 statistic. This work attempts to study the exhaustive possibilities of forming sub tables from the given contingency table to study the category wise association through Chi square test statistic; particularly the tables which exhibit reversal association pattern (RAP) when compared to original conclusion. This computer intensive effort necessitates developing an R package called RAP for complete enumeration of sub tables. Further, the simulation study has observed that this behavior of RAP persistently exists among 2 x 2 tables and this software can be used to understand one more characteristic of Chi square statistic and a supporting tool to fix sub tables for partitioning schema for an academic exercise or for typical application studies.

Index Terms — Chi-Square tests, Partitioning, Association, Categorical variables, R packages

1 INTRODUCTION

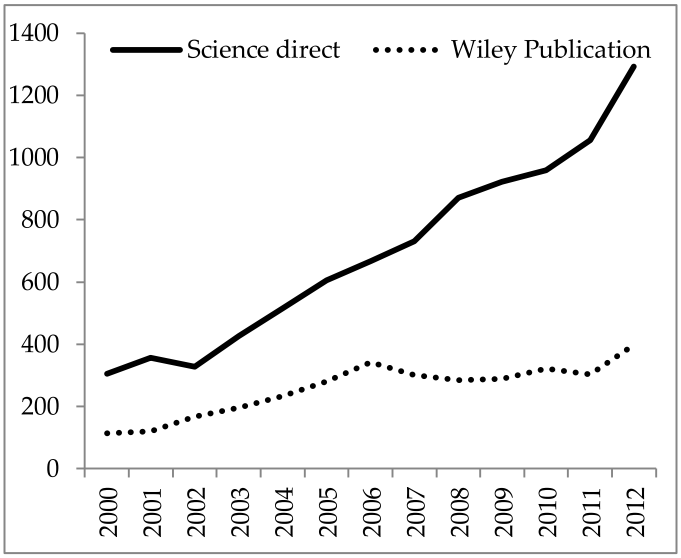
Categorical data analysis had found applications in many fields such as medicine and social science (Agresti, 1992, Tang, et al, 2012). Such data consist of frequency counts of observations occurring in the response categories. For two categorical variables with I and J levels respectively, a contingency or cross-classification table is generally used; each cell of the table counts the number of cases for the simultaneous occurrence of row and column variables. Most of the statistical analyses related to categorical data presented in a contingency tables deal with testing independence of the categorical variables. In this attempt many studies have focused the issues of sparse contingency tables especially the presence of zero or small counts (Koehler and Larntz, 1980, Brown and Fuchs 1983, Haberman 1988, Gorman et al 1990, Maiste and Weir, 2004, Burman 2004, Campbell 2007, Hashino 2012). Ratio of sample size of the table to the number of cells is invariably considered as a tool to understand sparseness beyond the presence of zeros and small counts.

Recently Rapallo (2012) has provided the methods for outlier patterns in contingency tables, using distance between the cell counts. Also such distance plays significant role in estimation of multinomial probabilities as noted in May and Johnson (2000) with respect to a method due to Sison and Glaz (1995). Apart from this structural metric of a contingency table, statistical studies have focused on the use of Chi-square test as a measure of association (Mirkin, 2001). Under multinomial sampling in two-way contingency tables, Pearson Chi square statistic has found an extensive usage to test the null hypothesis of statistical independence.

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i=1,2,…, I and j=1,2,…, J where $\left\{\pi_{ij}\right\}$ is the joint probability distribution of both categorical variables and marginal distributions are the row and column totals denoted by $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$.

The graph below shows the number of articles containing the phrase "Chi square Statistic for independence" between 2000 and 2012. The search is limited to Science Direct and Wiley Publications. It is also observed that numbers are doubled when the search phrase includes Chi square, Chi square test for association.

- *U. Sangeetha – Asst. Professor, Department of Management Studies, SSN College of Engineering, Kalavakkam, Chennai. Email: usangee19@gmail.com*
- *M. Subbiah – Asst. Professor, Department of Mathematics, L.N.Govt. College, Ponneri. Email: sisufive@gmail.com*
- *M. R. Srinivasan – Professor & Head, Department of Statistics, University of Madras, Chennai. Email:mrsrin8@gmail.com*



In spite of its theoretical popularity (Mirkin 2001) and computational ease, Chi square test faces a warning about its usage for small samples or sparse large contingency table (Berkson 1938, Cochran 1954, Campbell, 2007, Agresti, 1992). Also such situations are indicated in many statistical software with inbuilt warning messages.

Hence the objective of this work is twofold; to work out a metric for classifying a contingency table based on the polarized cell counts and to develop an R package to understand the category-to-category association supplementing statistical inference for contingency tables. This work could help the practitioners to classify the sparseness of the given contingency table and compare with the association results of its all possible tables. Since the exhaustive enumeration involves $(2^{I} - I - 1)(2^{J} - J - 1) - 1$ attempts (I: no. of rows; J: no. of columns), a convenient procedure in the R package shortlists the sub table which reverse the association compared to the original table.

This article has brought out one such feature of Chi square test for independence based on the way category to category association behaves when compared to over all association. Along to the pair wise comparisons in ANOVA models, this work attempts to observe the relationship between possible association that could be exhibited between the levels of categorical bivariates. A systematic R package has been developed to implement the study that involves an exhaustive enumeration and calculations.

2 Motivation

Agresti (1992) and few many studies have indicated the partitioning of Chi squared statistic. This is mainly to understand the component wise association aspects. A partition could help to show an association to indicate the differences between certain categories. Two studies can be considered for illustrating the notion of partitioning Chi square statistic; Example 1 deals with most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia (Agresti, 1992) and following table presents the actual data.

	Biogenic	**Environmental**	**Combination**
Eclectic	90	12	78
Medical	13	1	6
Psycho-analytic	19	13	50

Example 1

Example 2 investigates whether there is evidence to indicate a difference in the distribution of preference across the four state universities; this can be accessed from www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_10.pdf and the details are provided in the table

State	**Bargaining agent**		
University	**101**	**102**	**103**
1	42	29	12
2	31	23	6
3	26	28	2
4	8	17	37

Example 2

However, the partitioning procedure need not be unique combinations of sub tables yet it requires a careful way of construction of sub tables. Hence an attempt has been made to obtain all possible sub tables exhaustively and a scope to pick sub tables for a suitable partitioning schemes.

3 Reversal Association Pattern (RAP)

An exhaustive enumeration of all possible sub tables will bring more information about category to category associations together with overall association. This approach needs a large number of sub tables that are reckoned using following details. Let I, J be the number of rows and number of columns. The problem is to find the number of sub tables with $2 \le i \le I$, $2 \le j \le J$ with the assumption that original table is also considered as a sub table of itself.

The number of ways 2 rows can be selected is $\binom{I}{2}$. For each of these choices there are $\binom{J}{2},\binom{J}{3},...,\binom{J}{J}$ ways of selecting 2 columns, 3 columns etc. Hence number of sub tables with exactly 2 rows

$$= \binom{I}{2}\binom{J}{2} + \binom{I}{2}\binom{J}{3} + \ldots + \binom{I}{2}\binom{J}{J}$$

$$= \binom{I}{2}\sum_{j=2}^{J}\binom{J}{j} = \binom{I}{2}\left(2^{J} - J - 1\right)$$

Similarly the number of sub tables with exactly 3 rows $= \binom{I}{3}\left(2^{J} - J - 1\right)$.

Hence the number of sub tables (including original)

$$= \sum_{i=2}^{I}\binom{I}{i}\left(2^{J} - J - 1\right) = \left(2^{I} - I - 1\right)\left(2^{J} - J - 1\right)$$

Therefore, number of sub tables required for finding reversal pattern $= (2^{I} - I - 1)(2^{J} - J - 1) - 1$.

Naturally, the number of sub tables will increase in the order of I + J; however, advances in computations make such task achievable and hence a tool in the open source environment R (R core development team), has been developed. This R package RAP will help the user to identify the sub tables which reflect an association that reverse the overall association between the given two variables. Reversal aspects are based on the usual level of significance (5%) followed in tests for independence.

4 Results

Initially the procedure of RAP has been obtained for the two illustrative data sets and the results are displayed in Tables 1 and 2.

S.No	No. of rows	No. of cols	Selected rows	Selected cols	Pvalue	P value significant at 5%?
1	3	3	1,2,3	1,2,3	0.0002	TRUE
2	2	2	1,2	1,2	0.9503	FALSE
3	2	2	1,2	1,3	0.3221	FALSE
4	2	2	1,2	2,3	0.6138	FALSE
5	2	2	1,3	2,3	0.3271	FALSE
6	2	2	2,3	1,2	0.0545	FALSE
7	2	2	2,3	2,3	0.9207	FALSE
8	2	3	1,2	2,3	0.4437	FALSE
9	3	2	1,2,3	2,3	0.4789	FALSE

Table 1 Details of sub tables extracted from the original table of Example 1 that have the pattern of reversal association.

S.No	No. of rows	No. of cols	Selected rows	Selected cols	Pvalue	P value Significant at 5%
1	4	3	1,2,3,4	1,2,3	0	TRUE
2	2	2	1,2	1,2	0.9895	FALSE
3	2	2	1,2	1,3	0.6609	FALSE
4	2	2	1,2	2,3	0.5952	FALSE
5	2	2	1,3	1,2	0.2971	FALSE
6	2	2	1,3	1,3	0.1581	FALSE
7	2	2	2,3	1,2	0.4407	FALSE
8	2	2	2,3	1,3	0.4707	FALSE
9	2	2	2,3	2,3	0.233	FALSE
10	2	2	2,4	1,2	0.063	FALSE
11	2	2	3,4	1,2	0.2696	FALSE
12	2	3	1,2	1,2,3	0.7161	FALSE
13	2	3	1,3	1,2,3	0.0523	FALSE
14	2	3	2,3	1,2,3	0.2473	FALSE
15	3	2	1,2,3	1,2	0.4402	FALSE
16	3	2	1,2,3	1,3	0.2218	FALSE
17	3	2	1,2,3	2,3	0.0626	FALSE
18	3	2	1,2,4	1,2	0.0528	FALSE
19	3	2	1,3,4	1,2	0.0588	FALSE
20	3	2	2,3,4	1,2	0.1089	FALSE
21	3	3	1,2,3	1,2,3	0.1929	FALSE
22	4	2	1,2,3,4	1,2	0.0935	FALSE

Table 2 Details of sub tables extracted from the original table of Example 2 that have the pattern of reversal association.

Also, a simulation study has been carried out to investigate whether any systematic combinations of tables exhibit a pattern of reversal associations as these two examples and other similar cases indicate more number of 2 x 2 tables in the outcome of RAP. The choices of parametric representations for 1000 bootstrap sample of each case include the number of cells (k) and sample size (n) is studied and the median proportion is give in Table 3.

S.No	K	Size	n	Median of the proportion of 2x2 sub tables
1	9	3x3	160	1
2	9	3x3	2726	1
3	16	4x4	926	0.6
4	24	6x4	2714	1
5	25	5x5	3600	0.625

Table 3: Results from the bootstrap samples to understand the nature of sub tables that have exhibited the reversal association.

	RM1	SM1	RM2	SM2	RM3	SM3	RM4	SM4	RM5	SM5
Mean	0.80	6.05	8.36	56.69	3.17	13.27	4.23	69.16	23.99	144.07
S.E of Mean	0.01	0.08	0.12	0.69	0.01	0.34	0.06	0.35	0.42	2.39
2.5 Percentiles	0.63	3.89	3.38	38.07	2.70	5.20	2.42	60.51	12.48	76.30
97.5 Percentiles	0.88	8.31	9.19	77.48	3.30	23.40	5.00	79.00	28.44	221.71

Table 4: Summary of the bootstrap samples for the stability of two measures based on range (RM) and sum (SM) of counts in the contingency tables. 1 to 5 indicate the six combinations of datasets that are used for the bootstrap exercise.

From Table 4 it could be observed that the sparseness measure based on sum of the cell counts is more sensitive to that is derived from range based measure; this could be mainly due to the fact that when the given data set is sampled through bootstrap method, cell counts would change that directly affect their sum; however the study is focused to replicate the given data sets samples are not based on random sampling techniques such as Monte Carlo samples. Hence the re-sampling counts tend to repeat the values and range is mostly a robust measure in such repeated circumstances. This is to substantiate the notion that when data likelihood is based on independent identical samples its characteristics are closely fixed by the respective samples with replacements except the sum of the counts which is quite sensitive to a slight change. There by the present work through the extensive simulation study has made an attempt to establish that range based measure can be considered as a better tool to portray the distinct pattern of data dispersion in a contingency table.

Conclusion

Usual contingency table is limited to chi square or probability estimation. Sparse nature which plays a role in statistical inference theory is also a relatively important area. But beyond that nature of cell counts and association at micro level has motivated to propose a measure and develop a computing tool to understand the micro association. However beyond the usual sparseness metric using sample sizes and cell counts, the distance between the cell counts do draw active attention to classify the tables. A new metric has been proposed for understanding the extent to which the cell counts are dispersed with respect to the size of the table; this is further exemplified by a set of motivating examples and bootstrap samples for its sensitivity.

Also, another objective of this paper includes a main feature of Chi square test for independence between bivariate categorical variables. This attempt can be considered as a primitive approach for understanding category to category association that is similar to the exhaustive enumeration involved in LSD approach for pair wise comparisons in ANOVA models. Also, similar reversal effect in summary measure has provided a procedure to classify the sparse 2 x 2 data (Subbiah and Srinivasan, 2008). However, the practical implementation in general I x J table requires a large number of sub tables and their association. Hence the notion is supplemented. With RAP, R Package to understand the category wise association more easily studies in academic class room examples or typical association could make use of this tool to understand the important aspect of Chi squared test statistic for independence.

Reference

[1] Agresti, A. (1990), Categorical Data Analysis, John Wiley & Sons: New York.

[2] Campbell, I. (2007), “Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations“, Statistics in Medicine, 26, 3661-3675.

[3] Cochran, W.G. (1954), “Some methods for strengthening the common χ2 tests”, Biometrics, 4, 417– 451.

[4] Brown, M.B. and Fuchs, C. (1983), “On maximum likelihood estimation in sparse contingency tables”, Computational Statistics & Data Analysis, 1, 3-15

[5] Burman, P. (2004), “On some testing problems for sparse contingency tables”, Journal of Multivariate Analysis, 88, 1-18.

[6] Gorman, T.W., Woolson, R.F., Jones, M.P. and Lemke, J.H. (1990), “Statistical Analysis of K 2x2 Tables: A Comparative study of Estimators/Test Statistics for Association and Homogeneity”, Environmental Health Perspectives, 87, 103-107.

[7] Haberman, S.J. (1988), “A warning on the use of Chi-squared statistics with Frequency tables with small expected cell counts”, Journal of the American Statistical Association, 83,555-560

[8] Hoshino, N.(2012), “Random partitioning over a sparse contingency table”, Ann Institute of Statistical Mathematics, 64, 457-474

[9] Koehler, K.J. and Larntz, K. (1980), “An empirical investigation of Goodness-of –Fit statistics for sparse Multinomials”, Journal of the American Statistical Association, 75, 336-344.

[10] Maiste, P.J. and Weir, B.S. (2004), “Optimal testing strategies for large, sparse multinomial models”, Computational Statistics & Data Analysis, 46, 605-620

[11] May, W.L. and Johnson W.D. (2000), “Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells”, Journal of statistical software, 5(6).

[12] Mirkin, B. (2001), “Eleven ways to look at the Chi-squared Coefficient for Contingency Tables”, The American Statistician, 5, 111-120.

[13] R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.x.x. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[14] Rapallo, F. (2012), “Outliers and patterns of outliers in

contingency tables with algebraic statistics", Scandinavian Journal of Statistics, 39, 784-797.

[15] Sison, P.C. and Glaz J. (1995), "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions", Journal of the American Statistical Association 90: 366-369.

[16] Subbiah, M. and Srinivasan, M.R. (2008), "Classification of 2 x 2 sparse data with zero cells", Statistics & Probability Letters, 78, 3212 – 3215.

[17] Tang, W., He, H., and Tu, X.M. (2012), Applied Categorical and Count Data Analysis, Chapman & Hall / CRC Texts in Statistical Science.