Statistics and Applications {ISSN 2454-7395 (online)}
Volume 16 No.2, (New Series), pp 103-113

Detection of Outlying Cells in Two-Way Contingency Tables

Sripriya T.P. and M.R. Srinivasan
University of Madras, Chennai, India.

Received: February 11, 2018; Revised: October 24, 2018; Accepted: November 10, 2018

Abstract

Detecting outliers is one of the important statistical problems of immense practical interest. The problem of identification of outliers and testing for the same is extremely important in any data set as its presence could affect the inference. Thus, the present study proposed an iterative procedure to detect outliers in a two-dimensional contingency table especially for large tables (I, J>3). The approach deals with identifying the reversal of association in sub-tables using chi-squared statistics from the original table followed by residual approach supplemented by boxplot in detecting the outlying cells. Numerical examples show that the iterative procedure could be a feasible approach in detecting outliers.

Key words: Contingency table, Chi-square, Sub-tables, Reversal of association, Residuals, Boxplot, Outlier(s).

1. Introduction

The phenomenal growth of availability of data, in recent years, has drawn the attention of researchers in the identification of unusual observations (outliers) in data for its own significance and its impact on the data analysis. Outliers may be errors, or else accurate but unexpected observations which could shed new light on the phenomenon under study. In every statistical data analysis, somehow surprising observations or outlier(s) can occur which deviate strongly from the remaining observations or the assumed model. In metric data, measures such as cook's distance, leverage, DFFITS, DFBETA, etc., are used to identify the outliers (Barnett and Lewis 1994). There are two basic reasons to search for outliers: 1) the interest in the outliers for their own sake, and 2) the outliers could influence the results from the rest of the data. For example, where outliers themselves are of primary importance is in the problem of air safety (Kafadar and Morris, 2002) and the unusual gestational period of Mrs. Hadlum, as discussed in and Barnett and Lewis (1994). However, it is challenging to establish exact criteria for deciding on an observation to be unusual, denoted as an outlier, in contingency table. Unlike in metric case, there exists no clarity in the definition of outliers for categorical data as the cells are purely frequency or counts of a contingency table.

The existence of one or two outlying observations in a sample can badly distort the analyses of data. In an analysis of a two-way contingency table, the primary interest will be a hypothesis of independence between two categorical variables or a hypothesis of

Corresponding Author: M. R. Srinivasan
E-mail: mrsrin8@gmail.com

homogeneity, depending on the sampling scheme. In spite of the excellent advanced statistical techniques for modeling categorical data or using correspondence analysis, the most widely used statistic is still the classical Pearson's chi-squared statistics, since it is an omnibus one. The Pearson's chi-square and the log likelihood ratio statistics (Agresti 2002) are long standing techniques in goodness-of-fit testing under multinomial set up. Sangeetha et al. (2013) proposed the reversal pattern of association (RAP) to understand deeply the association between attributes following Agresti (1992) in higher dimensional tables.

Conscientious researchers rarely examined the application of this omnibus test statistics because the rejection of null hypothesis raises a question as to which of the cells are responsible for the rejection in a table. In this context, detecting a few outlying cells as the cause for deviation from the overall association pattern and thus may be responsible for the reversal of association. Such outliers may contain information on unknown dependencies in the data and its presence could potentially bias the inference. Polarization of cell frequencies and the sparseness in I x J table are the major issues in the detection of outliers. However, this study presents an algorithm for high-dimensional contingency table to detect such outlying observations using the concept of RAP.

The structure and nature of cell counts in a contingency table play an important role in the data analysis with the cell counts ranging from zero to very high values. Basu and Sarkar (1994) considered a more general class of goodness-of-fit test statistics to study the behavior of the disparities in controlling the "outliers" and "inliers". Kuhnt (2004) identified outliers based on the tails of the Poisson distribution. Rapallo (2012) introduces a notion of patterns of outliers in connection with goodness-of-fit tests by applying techniques from algebraic statistics. Kuhnt et al. (2014) detected outliers through subsets of cell counts called minimal patterns for the independence model.

Residual based techniques has been widely used to detect outliers in contingency table (Haberman 1973; Brown 1974; Gentleman and Wilk 1975a,b; Simonoff 1988; Fuchs and Kenett 1980; Bradu and Hawkins 1982; Kotze and Hawkins 1984; Yick and Lee 1998; and Lee and Yick 1999). Graphical display of contingency table can be made with plots such as association plot, sieve plot, and mosaic plot (Friendly 2000) which are based on independence of the row and column variables. In mosaic display, Friendly (1994) discussed the patterns of deviations in terms of residuals from various models for categorical data. Thus, residuals play an important role in detecting outliers in two-way contingency tables, and an extensive review is presented in Kateri (2014). Even though, the residual technique has been widely used by the researchers, but no cutoff is provided in choosing the maximum residuals and is more heuristic in nature (Simonoff 2003).

In this paper, we propose an iterative procedure to detect potential outliers in two-way contingency table. It deals with the reversal of association in sub-tables from I x J table by an exact identifier through boxplot of residuals in measuring such deviations.

2. Iterative Procedure for Outlier detection

Consider N paired sample observations that are cross-classified in a IxJ contingency table and let n_{ij} be the observed cell frequency at the ith row and jth column of the table, $n_{i+} = \sum_{j=1}^{J} n_{ij}$ be the ith row marginal, $n_{+j} = \sum_{i=1}^{I} n_{ij}$ be the jth column marginal. Once a contingency table is constructed, the first interest will be the hypothesis of either homogeneity or independence depending on the sampling scheme. When the null hypothesis is rejected, the so called cell residuals are investigated to identify the cells which deviate greatly from others.

The test statistic for chi-square is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}$; where $e_{ij} = \frac{n_{i+}.n_{+j}}{N}$ be the expected frequency in the cell at the ith row and jth column and the Pearson (Chi-square) residual is denoted by $r_{ij} = \frac{n_{ij}-e_{ij}}{\sqrt{e_{ij}}}$.

Chi-square test of association is widely applied to understand the association between the variables. Mirkin (2001) listed different ways to look at chi-square tests and observed that it is an appealing measure for studying the association. However, the present study is concerned with the detection of unusual observations or outliers in contingency table via reversal pattern of chi-square association measure. Suppose that an initial set of sub-tables contrast the original decision of independence, and then the problem is to identify a “most outlying” cell or set of cells in those sub-tables. The sub-tables can be obtained iteratively with I x (J-1),…,I x 2 and (I-1) x J,…,2 x J, whence successively obtain the sub-tables which got reversed from original decision and stop the iteration process once the reversal happened. The stepwise algorithm for the same is given below.

Step 1: Examine the overall association using Chi-square statistics for its significance.

Step 2: Examine the sub-tables iteratively both row-wise [(I-1)xJ, (I-2)xJ,…,2xI] and column-wise [Ix(J-1), Ix(J-2),…,Ix2] and stop the iteration where the reversal of association in the sub table is observed .

Step 3: Obtain the Pearsonian residuals for those sub-tables with reversal of association and identify the exact outlying cells using box plot of residuals.

The above algorithm is highly computer intensive as the maximum number of sub tables to be examined for reversal pattern in I x J table is $(2^I\text{-I-1})\ (2^J\text{-J-1})\text{-1}$. However, advent of advanced computers makes the implementation of the algorithm faster and thereby identify the reversal pattern in sub tables. The flowchart of the proposed iterative procedure for detecting potential outliers is presented in appendix. The above algorithm has been implemented in R software for a set of over 10 distinct real data sets to examine its relevance and applicability.

3. Data Analysis

Literature is abundant with I X J contingency tables obtained from many experimental studies. The tables are of varied nature starting from 2 x 2 to 19 x 6 with highly variant cell frequencies. The present study has identified over 10 tables representing the cross-section of the real data sets to examine the validity of the proposed algorithm to identify the outliers, if any. Yick and Lee (1998) considered a study on student enrolment of seven community schools conducted in eight different periods of the year from Northern Territory, Australia as given in Table 1. The graphical description of the data in terms of association plot is presented in Figure 3.

Table 1: Student's Enrolment

School/Period	1	2	3	4	5	6	7	8
A	93	96	99	99	147	144	87	87
B	138	141	141	201	189	153	135	114
C	42	45	42	48	54	48	45	45
D	63	63	72	66	78	78	82	63
E	60	60	54	51	51	45	39	36
F	174	165	156	156	153	150	156	159
G	78	69	84	78	54	66	78	78

The study on association for the above table based on chi-square statistics leads to rejection of the null hypothesis of independence between schools and different time periods of student enrolment. Following the rejection, the iterative scheme discussed earlier was performed to examine the pattern of reversal in sub-tables. While performing column-wise iteration, the association gets reversed for two sub-tables of dimension 7x6 with higher residuals and each table identifies one outlying cell through boxplot of residuals namely (2,4) and (1,6). Under row-wise iteration, the reversal happens in a sub-table of dimension 5x8, but there is no outlying cell because all the expected values are closer to its corresponding observed values. When iterating further, 8 sub-tables of dimension 4x8 yields reversal of association, out of which 2 tables gives higher residuals and four cells have been identified as outliers [(2,4),(1,6),(4,7),(7,5)]. From this example, we observe that (2,4) and (1,6) are potential outlier cells.

Similarly, Bradu and Hawkins (1982) considered a study on prevalence rate of men aged 55-64 with hearing levels 16 decibels or more above the audiometric zero for the better ear at 500, 1000, 2000, 3000, 4000, 6000 cycles per second (cps) against the occupational status and the same is represented in 7 x 7 table given in Table 2 and the graphical description of the data in terms of association plot is presented in Figure 4.

Table 2: Prevalence rate of men

Frequency	Professional, Managerial	Farm	Clerical, Sales	Craftsmen	Operatives	Service	Labourers
Average 500 cps	2.1	6.8	8.4	1.4	14.6	7.9	4.8
1000 cps	1.7	8.1	8.4	1.4	12	3.7	4.5
2000 cps	14.4	14.8	27	30.9	36.5	36.4	31.4
3000 cps	57.4	62.4	37.4	63.3	65.5	65.6	59.8
4000 cps	66.2	81.7	53.3	80.7	79.7	80.8	82.4
6000 cps	75.2	94	74.3	87.9	93.3	87.8	80.5
Normal Speech	4.1	10.2	10.7	5.5	18.1	11.4	6.1

In our iterative procedure, the association leads to the acceptance of alternative hypothesis that is, there is an association between the level of hearing and occupational status. When performing the column-wise iteration, four tables of dimension 7x5 yields reversal association and a cell (3,2) have been detected through boxplot of residuals. In the same way, while performing row-wise reduction, the reversal occurs in a 5x7 sub-table and the cells [(1,5),(7,5) and (2,5)] have been identified as outlier using residual approach. From this results, we conclude that the cells (3,2),(1,5),(7,5) and (2,5) are the potential outliers.

However, literature has shown the existence of other methods for detection of outliers. Hence, a comparative study has been carried out to evaluate the proposed iterative method of identification of outliers in contingency tables.

4. Comparative Study

In literature, outliers have been identified either using residuals or a suitable cutoff criteria. This study has identified six more dataset for the comparison purpose. The following five methods are considered in this comparative study:

1. Haberman (1973) considered residuals greater than 2 as outliers (Method1).
2. Kuhnt et al (2014) proposed α-outlier region to detect the outliers and if the cell counts above/below the region are considered as outliers (Method 2).
3. Yick and Lee (1998) proposed perturbation diagnostics to detect outliers (Method 3).
4. Simonoff (1988) considered the cells deviating from the model of independence as outliers (Method 4).
5. Proposed iterative algorithm based on RAP to detect the outliers (Method 5).

Table 3: Comparative results for various datasets

Data set	Source	I x J	Method 1	Method 2	Method 3	Method 4	Method 5
I	Yick and Lee (1998)	7x8	(2,4),(1,5),(2,5),(6,5),(7,5),(1,6),(4,7),(2,8)	(2,4),(1,5),(2,5),(7,5),(1,6)	(1,5),(1,6),(2,4),(2,5)	(1,5),(1,6),(2,4),(7,5)	(2,4),(7,5),(1,6),(4,7)
II	Bradu and Hawkins (1982)	7x7	(3,2),(1,5),(2,5),(7,5)	(1,4),(1,3),(2,3),(3,2),5,3)	(3,2),(1,5),(2,5),(7,5)	(3,2),(1,5),(2,5),(7,5)	(3,2),(1,5),(7,5),(2,5)
III	Kotze and Hawkin (1984)	14x14	(5,1),(6,1),(10,1),(5,4),(12,7),(10,10),(11,13)	(12,1),(5,4),(10,1),(12,7),(10,10)	(10,1),(5,4),(12,7),(10,10),(8,8)	(5,4),(10,1),(12,7),(10,10	(12,1),(5,4)
IV	Kateri (2014)	5x7	(5,1),(1,4),(4,4),(1,6),(4,6)	(1,4),(4,4),(4,6)	(5,1),(1,4),(4,4),(1,6),(4,6)	(5,1),(1,4),(4,4),(5,4),(1,6),(4,6)	(5,1),(4,2),(1,6),(4,6)
V	Agresti (2002)	6x4	(1,1),(2,1),(5,1),(6,1),(1,4),(2,4),(5,4),(6,4)	(6,1),((1,1),(5,4),(6,4),(2,4)	(1,1),(6,1),(2,4),(6,4),(5,4)	(1,1),(6,1),(5,4),(2,4),(6,4)	(6,1),(6,3),(2,4)
VI	Kateri (2014)	5x7	(1,1),(2,2),(1,4),(1,6)	(1,1),(2,2),(3,5),(5,2)	(1,1),(2,2),(3,5),(5,2)	(1,1),(2,2),(1,4),(1,6)	(3,5),(1,6),(1,1),(1,4),(5,2)
VII	Data from SN Hospital (2014)	4x4	(2,1),(4,1),(2,2),(4,2),(2,4),(3,4)	(4,1),(2,4),(2,1)	(2,1),(4,1),(4,2),(2,4)	(2,1),(4,1),(4,2),(2,4)	(1,1),(1,3),(2,4),(4,1)
VIII	Mostellar and Parunak (1985)	19x6	(2,3),(11,5),(18,1),(18,2),(9,1),(14,1),(19,1),(19,2),(13,3),(14,3),(9,5),(13,5),(19,5),(3,6),(12,6)	(4,1),(19,1),(4,2),(7,2),(18,2),(11,5),(18,1),(2,3)	(2,3),(11,5),(18,1),(18,2),(9,1)	(2,3),(11,5),(18,1),(18,2),(9,1),(13,3),(13,5)	(9,1),(18,1),(3,6),(12,6),(7,6),(13,3),(13,5),(19,1),(14,1)

In dataset I, method 1 identified eight cells, method 2 detected five cells, method 3 and 4 each detected four cells as outliers. In our iterative procedure, four cells (2,4), (1,6), (7,5) and (4,7) are considered as outliers and the same cells are identified in other methods with few other outlying cells. In dataset II, methods 1, 3, 4 and 5 identified the same cells as outliers whereas method 2 identified four other cells as outliers except the cell (3,2). Similarly, all other dataset considered in this study detected few outlying cells in each method considered and the results are presented in Table 3. From the results, it is evident that method 1 contains all the other methods and the results revealed that the proposed iterative procedure could be a best alternative method to detect outliers in two-way contingency table.

5. Conclusions

Detection of outliers in a two-dimensional contingency table has been of great interest over the past few decades but the notion of outlier in contingency tables is not well defined. Outliers are only vaguely described as such cell frequencies which deviate markedly from the others or a cause for a significant lack of fit. The prevailing research on the characteristics of I × J table are: Order of the table, numerical issues (aberration/zero width intervals), polarization of cell counts, low cell count, sparseness and computational complexity leading to major issues in the detection of outliers. The residual approach of identifying outlier is a more appealing procedure but limitation in identifying the exact number of outlying cells persists. To overcome this limitation, detection of outliers through the boxplot approach for the chi-square residual is considered in this study. This paper deals with finding outliers in large contingency tables through an iterative procedure on association pattern. The main idea is to identify the outliers through sub-tables which got reverted in association pattern from I x J table by using chi-square association measure. This computer intensive effort developed using R program has provided scope for the complete enumeration of the results. The use of boxplot provides a good indicator of the actual number of outliers in a table. It is to be noted that zero cells in the table is in no way complicate or invalidate the use of this iterative procedure. The proposed method has been applied to more than 10 real time data sets of nominal nature from the literature to examine the feasibility of the approach. The experimental results showed that the proposed iterative procedure could be a viable approach to identify outliers in large contingency table. Future work includes the identification of alternative methods to detect outliers in higher dimensional tables and also examine the robustness of the approach against odds.

Acknowledgement

The authors are grateful to an anonymous referee for the constructive and insightful comments which led to improvements over the earlier manuscript.

References

Agresti, A. (2002). *Categorical Data Analysis (Second Edition).* Wiley, New York.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science,* **7**, 131-177.

Barnett, V. D. and Lewis, T. (1994). *Outliers in Statistical Data (Third Edition).* Wiley, New York.

Basu, A. and Sarkar, S. (1994). On disparity based goodness-of-fit tests for multinomial models. *Statistics and Probability Letters,* **19**, 307-312.

Bradu, D. and Hawkins, D.M. (1982). Location of multiple outliers in two-way tables using tetrads. *Technometrics,* **24**, 103-108.

Brown, B.M. (1974). Identification of the sources of significance in two-way contingency tables. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **23**, 405.

Fienberg, S.E. (2007). *The Analysis of Cross-Classified Categorical Data (Second Edition).* Springer.

Friendly, M. (2000). *Visualizing Categorical Data.* Cary, NC: SAS Institute.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association,* **89**, 190-200.

Fuchs, C. and Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association- Theory and Methods Section*, **75**, 395-398.

Gentleman, J.F. and Wilk, M.B. (1975a). Detecting outliers in a two-way tables: I statistical behavior of residuals. *Technometrics*, **17**,1-14.

Gentleman, J.F. and Wilk, M.B. (1975b). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics*, **31**, 387-410.

Haberman, S.J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, **29**, 205-220.

Kafadar, K. and Morris, M. (2002). Data-based detection of potential terrorist attacks. White Paper.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R.* Birkhauser Basel.

Kotze, T.J.vW. and Hawkins, D.M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Applied Statistics,* **33**, 215-223.

Kuhnt, S. (2004). Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. *Scandinavian Journal of Statistics,* **31**, 431–442.

Kuhnt, S., Rapallo, F. and Rehage, A. (2014). Outlier detection in contingency tables based on minimal patterns. *Statistical Computing*, **24**, 481-491.

Lee, A.H. and Yick, J.S. (1999). A perturbation approach to outlier detection in two-way contingency tables. *Australian & New Zealand Journal of Statistics,* **41(3),** 305–314.

Mignone, F. and Rapallo, F. (2018). Detection of outlying proportions. *Journal of Applied Statistics*, **45(8)**, 1382-1395.

Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table: Probabilistic and exploratory approaches. *In D.C. Hoaglin, F. Mosteller, and J.W. Tukey, eds., Exploring Data Tables, Trends and Shapes,* John Wiley and Sons, New York, 189-224.

R Core Team. (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* URL http://www.R-project.org/.

Rapallo, F. (2012). Outliers and patterns of outliers in contingency tables with algebraic statistics. *Scandinavian Journal of Statistics*, **39(4),** 784–797.

Sangeetha, U., Subbiah, M., Srinivasan, M.R. and Nandram, B. (2014). Sensitivity analysis of Bayes factor for categorical data with emphasis on sparse multinomial data. *Journal of Data Science*, **12**, 339-357.

Simonoff, J.S. (1988). Detecting outlying cells in two-way contingency tables via backwards stepping. *Technometrics*, **30,** 339-345.

Simonoff, J.S. (2003). *Analyzing Categorical data.* Springer Texts in Statistics, Springer-Verlag New York.

Yick, J.S. and Lee, A.H. (1998). Unmasking outliers in two-way contingency tables. *Computational Statistics and Data Analysis,* **29**, 69-79.

Appendix

Flowchart of the proposed iterative procedure:

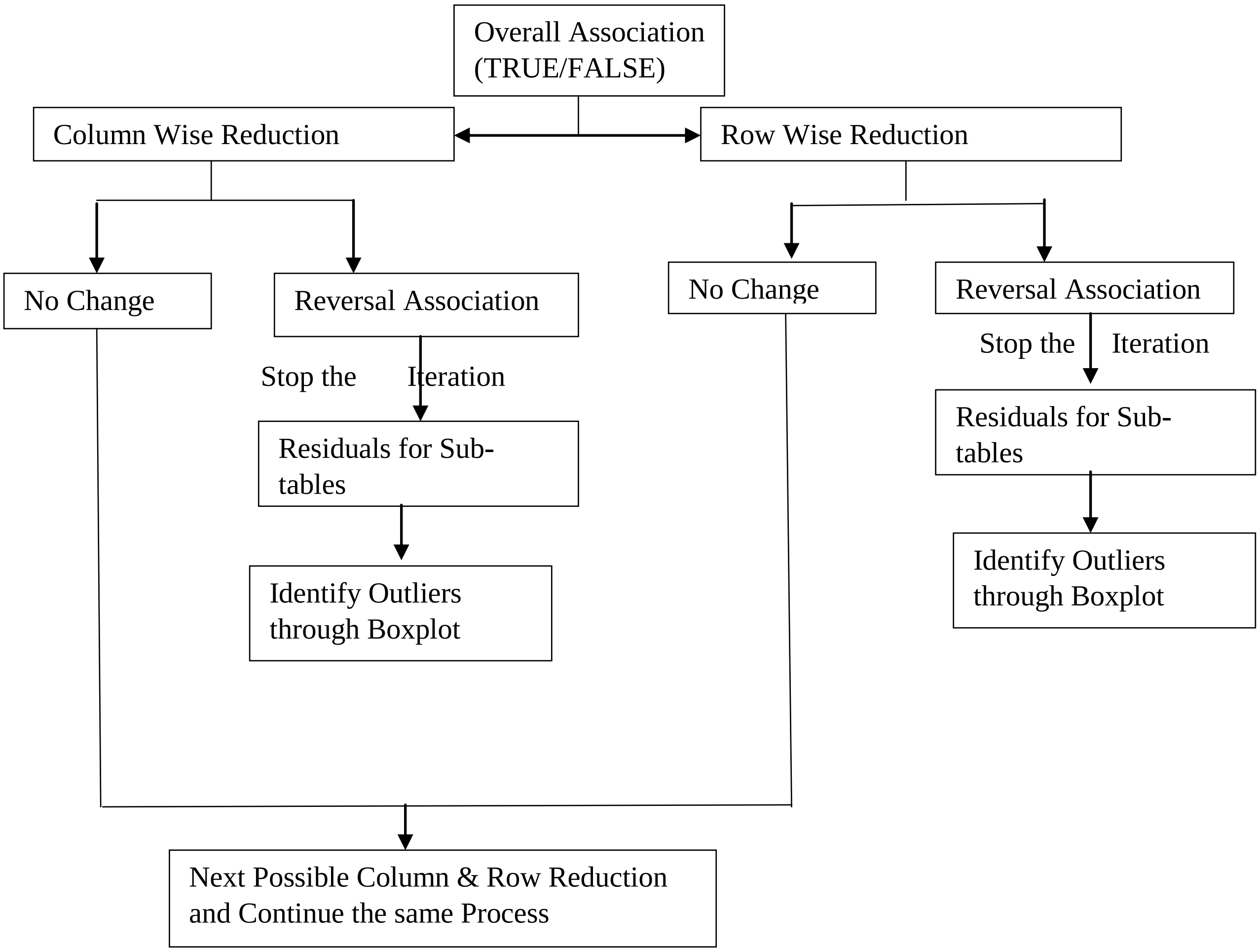


Figure 1: Column Iteration

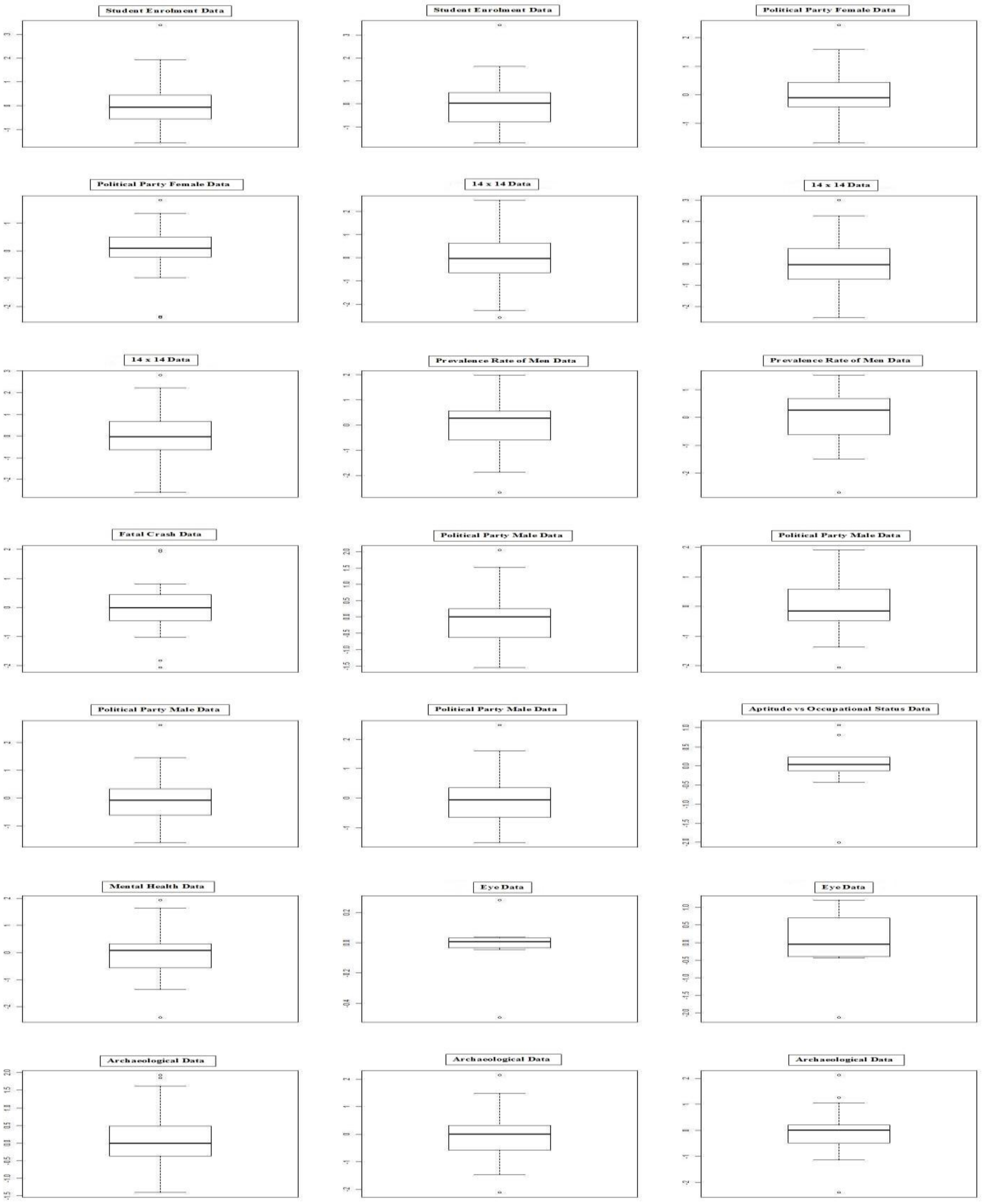


Figure 2: Row Iteration

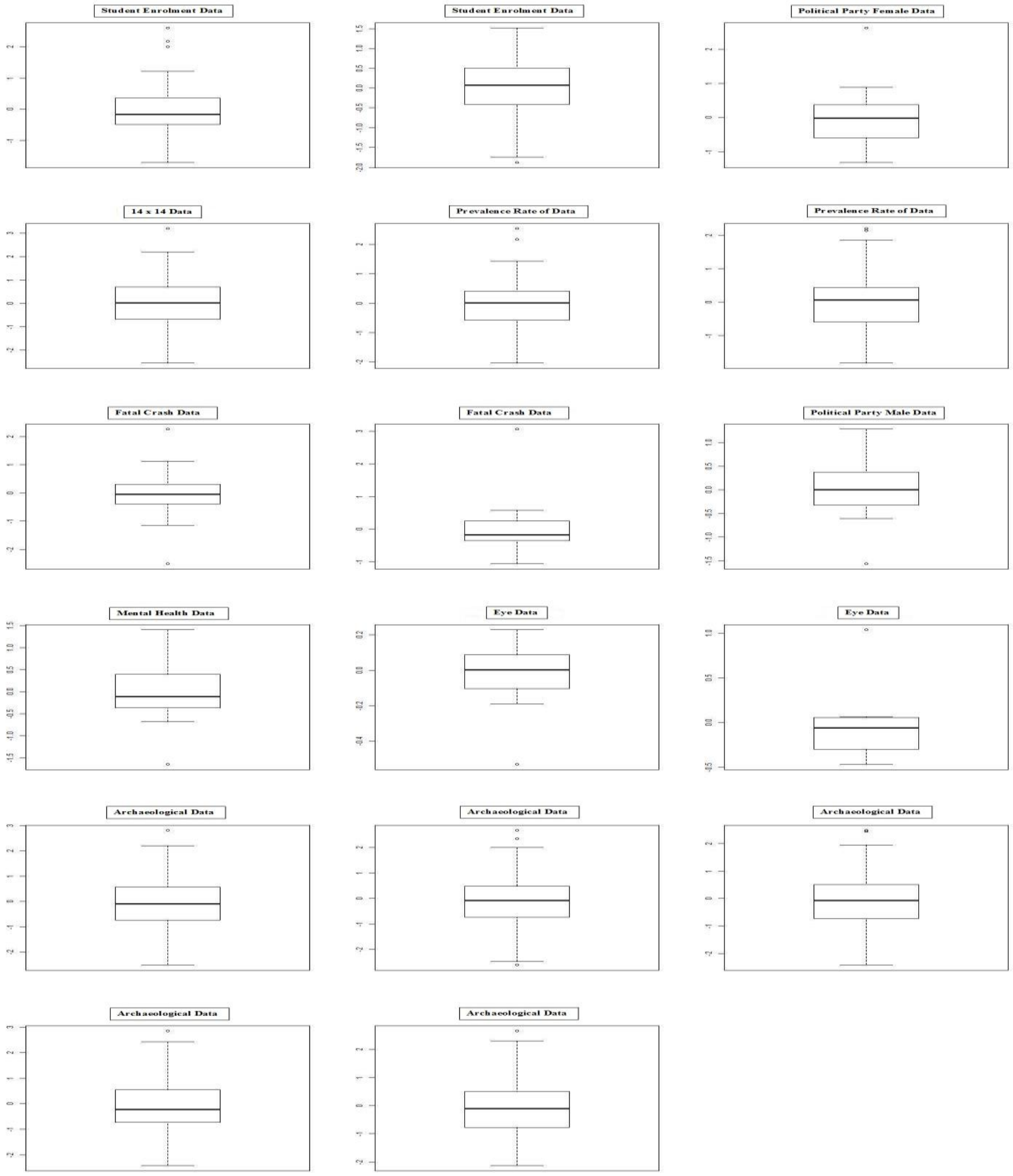


Figure 3: Graphical Representation of Student Enrolment Data

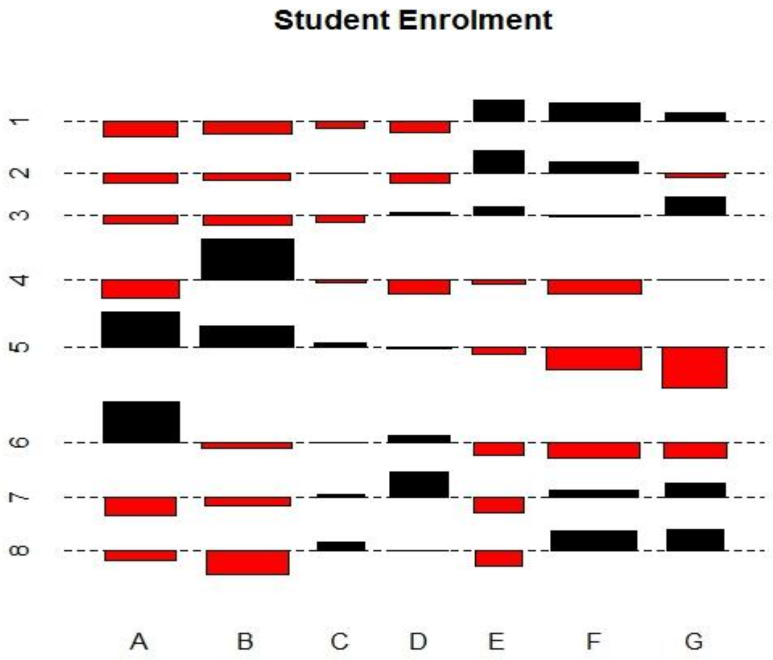
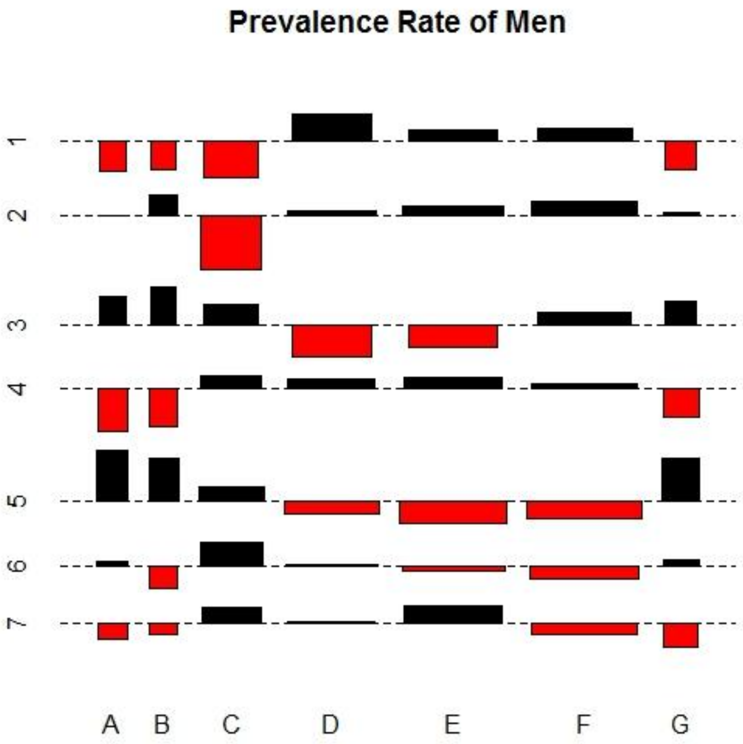


Figure 4: Graphical Representation of Prevalence Rate of Men Data


Prevalence Rate of Men
A
B
C
D
E
F
G