Vol.8.Issue.3.2020 (July -Sept)

http://www.bomsr.com
Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

Detection of Outliers in Sparse $I \times J$ Contingency Tables

T. P. Sripriya*, M. R. Srinivasan
Department of Statistics, University of Madras, Chennai, India.
*Corresponding Author Email:sri.chocho@gmail.com
DOI: 10.33329/bomsr.8.3.1

ABSTRACT

Sparseness in contingency table often occurs in practice and detecting outliers in such sparse contingency table is an interesting statistical problem and poses additional difficulties due to the polarization of cell counts. The purpose of this article is to propose a new classification of sparseness and formalize a general identification procedure through modeling technique in sparse I x J tables. The procedure deals with fitting of suitable models for categorical data and examines different types of residuals supplemented by boxplot in identifying the exact outlying cells. We also test our proposed method through a simulation study followed by a practical example to examine the consistency of the results.

Keywords: Sparseness, Log-Linear models, Zero-inflated models, Hurdle models, Residuals, Outlier(s).

1. Introduction

Rapid advancements in the statistical data analysis have facilitated the use of developing various techniques to detect outliers. The term "outlier" is generally defined as the observations which deviate strongly from the remaining observations (Barnett and Lewis 1994). 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. Outliers are only vaguely described as such cell frequencies which deviate markedly from the expected value or cause a significant lack of fit.

Many classical statistical methods are extremely sensitive even to slight deviations from usual distributional assumptions. Until now research on outliers in $I \times J$ contingency tables has been restricted mainly to the independence model. Further, polarization of cell counts is one of the major problems when it comes to outlier detection. Polarization is basically an uneven distribution of counts in $I \times J$ tables. Polarization in contingency tables involves presence of counts/frequencies

of disparate nature, such as presence of zero counts, low counts, high counts, and extreme values, etc.

Suppose a table consists of more number of zero counts and very few high counts forming unusual clusters which could affect the inference of $I \times J$ table, in addition to the detection of outliers. Thus, 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 frequencies (Sangeetha et al 2014).

For an $I \times J$ contingency table, the measures of interest are (i) total frequency (N), (ii) order of the table $(I \times J)$, (iii) high cell frequencies, (iv) low frequencies and (v) cells with zero frequencies which in turn cause a problem of polarization and this leads to a major issue in detecting outliers. Sripriya and Srinivasan (2018a) have suggested a new approach in the detection of outliers in categorical tables of order $I \times J$, based on log - linear model.

Mignone and Rapallo (2018) introduced an algorithm to detect outliers in $2 \times K$ tables using the concept called outlying proportions and minimal patterns. Kuhnt (2004) described a procedure to identify outliers based on the tails of the Poisson distribution and declared a cell as outlier if the actual count falls in the tails of the distribution. Rapallo (2012) studied the pattern of outliers by fitting log-linear model and tests the goodness of fit to specify the notion of outlier with the use of algebraic statistics. Sripriya and Srinivasan (2018b) proposed an iterative algorithm to detect outliers in contingency table based on chi-square association measure. Kuhnt et al (2014) detected outliers through subsets of cell counts called minimal patterns for the independence model.

Agresti and Yang (1987), Reiser and Vanden-Berg (1994), and Jöreskog and Moustaki (2001) discuss a range of ideas that have been suggested in the literature to improve the measurement of model fit for sparse contingency tables. These ideas include (i) adding constants to cells in order to nullify the effect in estimating the parameters, (ii) collapsing cells, (iii) considering only cells with observed or expected frequencies that exceed a certain value, and (iv) deriving the small sample distribution of a fit statistic by means of the bootstrap. Consequences of sparseness for the evaluation of the model fit have been widely investigated and discussed in the literature (Larntz 1978 and Koehler and Larntz 1980). Modeling the count data with excess zero has been done by adopting various models such as zero-inflated Poisson model (Lambert 1992), Hurdle model (Germu et al 1996), two-part model (Heilbron 1994), and zero-modified distributions (Dietz and B¨ohning 2000). ZIP models are more widely used as all important statistical inferences can be carried out more easily and conveniently than the others. Applications of ZIP models can be found in many areas, such as, agriculture (Ridout et al 1998), epidemiology (B¨ohning et al 1999), biostatistics (Van den Broek 1995) and industry (Lambert 1992). However, this study classified the sparse nature of the table/proportion of zero cell frequencies into three groups which in turn helps to provide a general outliers identification procedure based on the modeling aspects.

Residual based techniques have been widely used to detect outliers in contingency table (Haberman 1973; Bradu and Hawkins 1982; Lee and Yick 1999; Yick and Lee 1998; Simonoff 1988). Even though, the residual technique has been widely used by the researchers, there is no specific cutoff criterion as in metric case for choosing the maximum limit of residuals and the method adopted in literature is more heuristic in nature. To overcome this limitation, we have introduced boxplot for residuals to detect the outlying cells in $I \times J$ tables.

Outlier(s) in sparse contingency tables is a serious problem in statistical practice. There are many robust methods to analyze the sparseness in contingency tables but detecting outliers in the presence of zero and/or low cell frequencies is a challenging one. Hence, in this paper, we develop a rigorous and computationally efficient technique to detect potential outliers in sparse $I \times J$ table. It deals with fitting six different models namely Poisson log-linear model, Negative Binomial model, Zero-Inflated Poisson model, Zero-Inflated Negative Binomial model, Hurdle Poisson model and Hurdle Negative Binomial model and the usual diagnostic procedure called residuals supplemented by boxplot helps to detect the outlying cell in $I \times J$ tables. Further, this study provides a general outlier detection rule for $I \times J$ contingency tables in the presence of sparseness using simulation technique.

2. METHOD

Consider n sample observations that are cross-classified in a $I \times J$ contingency table, n_{i+} $(i = 1,2,...,I)$ be the i^{th} row total, n_{+j} $j = 1,2,...,J)$ be the j^{th} column total and $N = \sum_I \sum_J n_{ij}$ being the total frequency of the table, assumed to be the realizations of random variables Y_j, $j = 1,2,...,N$ (Agresti 2002; Kateri 2014). Under the independence model, the expected cell frequency is given by $e_{ij} = \frac{n_{i+} * n_{+j}}{N}$. In this context, the following classifications are considered in this study which helps in providing the more suitable outlier detection method for $I \times J$ tables under the condition of sparseness.

The major objective of statistical data analysis is to extract and explicate the informational content of a body of data. Techniques addressed to this objective involve summarization perhaps in terms of a statistic which is undergirded by some tightly specified model or in terms of a simple plot. It is all the more essential to have informal, informative summarization and exposure procedures. Following Agresti and Yang (1987), here, the number of cell frequencies k in $I \times J$ table is classified into three categories as follows:

$$k = \begin{cases} Low & if \quad 9 \le k \le 20 \\ Moderate & if \quad 20 < k \le 50 \\ High & if \quad k > 50 \end{cases}$$

Similar to the classification of k, the proportion of zero cells, $P_Z = \frac{zc}{k}$; where *zc* is the number of zero cells in $I \times J$ table, classified into three categories as :

$$P_Z = \begin{cases} Low & if \quad 0 \le P_Z \le 0.10 \\ Moderate & if \quad 0.10 < P_Z \le 0.20 \\ High & if \quad P_Z > 0.20 \end{cases}$$

The classification of k and P_Z provides a basic idea about the nature of $I \times J$ table and could be useful in choosing the suitable model based on the residuals. However, there is no general rule to classify the order of the table and the sparseness in literature. The ranges of those can vary and substantially the conclusions based on those classifications may vary. This study followed the idea of classification of Agresti and Yang (1987) and based on the above classification, a total of nine combinations of k and P_Z is obtained and each residual under each model is tested using the

method adopted in this study. As pointed out earlier, residual based techniques have been widely used in the detection of outliers. Before, deciding on the nature and type of residuals it is essential to decide on the most plausible models for fitting the categorical data.

2.1 Modeling of Categorical data

In recent decades, there are plenty of techniques available for modeling categorical data starting from logistic regression to log-linear model, but the problem of finding more suitable model still persists. In practice, many empirical data sets exhibit more number of zero observations. To deal with excess zero observations, Zero-Inflated models and Hurdle models has emerged. In this study, six different models are considered to examine the most plausible model for outlier detection under the condition of sparseness in $I \times J$ tables. Log-Linear models in contingency table are the most widely used method in analyzing categorical data. However, there are other possible alternatives along with log-linear model to model a contingency table which can be useful in modeling sparse tables. They are:

Model 1: Poisson Log-Linear Model with the density function, $f(y;\mu) = \frac{\exp(-\mu)\cdot\mu^{y}}{y!}$; $y = 0,1,2,\ldots$ with mean μ and $g(\mu) = \log(\mu)$ is the canonical link function results in log-linear relationship between mean and linear predictor.

Model 2: Negative Binomial model for $y_i \mid x_i$ with density function $f(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\cdot y!}\cdot\frac{\mu^{y}\cdot\theta^{\theta}}{(\mu+\theta)^{y+\theta}}$; $y = 1,2,\ldots,n$ with mean μ and the shape parameter θ.

Model 3: Zero-Inflated Poisson model (ZIP) for the data which are count in nature is a mixture of a point mass at zero $I_{\{0\}}(y)$ and a count distribution $f_{count}(y;x,\beta)$ and the density can be expressed as $f_{zero\inf}(y;x,z,\beta,\gamma) = f_{zero}(0;z,\gamma)\cdot I_{\{0\}}(y) + (1-f_{zero}(0;z,\gamma))\cdot f_{count}(y;x,\beta)$ where x and z are the covariate matrices, y is the dependent variable, β and γ are the unknown parameters, $f_{zero}(0;z,\gamma)$ is the observed zero inflated probability, that is it is the probability that produces only zeroes (structural zeroes), $(1-f_{zero}(0;z,\gamma))$ is the probability of the Poisson distribution and hence it is the chance of further zeroes, $I_{\{0\}}(y)$ is the indicator function, and $f_{count}(y;x,\beta)$ is the density of Poisson distribution..

Model 4: Another way to model the excess zero is to assume the data from the Negative Binomial population. The Zero-inflated Negative Binomial (ZINB) density function can be expressed as $f_{zero\inf}(y;x,z,\beta,\gamma) = f_{zero}(0;z,\gamma)\cdot I_{\{0\}}(y) + (1-f_{zero}(0;z,\gamma))\cdot f_{count}(y;x,\beta)$ where x and z are the covariate matrices, y is the dependent variable, β and γ are the unknown parameters, $f_{zero}(0;z,\gamma)$ is the observed zero inflated probability, that is it is the probability that produces only zeroes (structural zeroes), $(1-f_{zero}(0;z,\gamma))$ is the probability of the Negative Binomial distribution and hence it is the chance of further zeroes, $I_{\{0\}}(y)$ is the indicator function, and $f_{count}(y;x,\beta)$ is the density of Negative Binomial distribution.

Model 5: Hurdle Poisson models are another class of models to handle excess zero counts in the data. It is a two component model with truncated and hurdle components. Truncated component

handles the positive counts in the data and the hurdle component handles the zero counts in the data. The density function is given by

$$f_{hurdle}(y;x,z,\beta,\gamma)=\begin{cases} f_{zero}(0;z,\gamma) & if\ y=0 \\ (1-f_{zero}(0,z,\gamma))\cdot f_{count}(y;x,\beta)/(1-f_{count}(0;x,\beta)) & if\ y>0 \end{cases}$$

where x and z are the covariate matrices, y is the dependent variable, β and γ are the unknown parameters, $f_{zero}(0;z,\gamma)$ is the probability that y_i produces only zeroes, $(1-f_{zero}(0;z,\gamma))$ is the probability that y_i is greater than 0, $f_{count}(y;x,\beta)$ is the density of Poisson distribution and $f_{count}(0;x,\beta))$ is the cumulative distribution of Poisson distribution.

Model 6: Similar to Hurdle Poisson model, Hurdle Negative Binomial model is used to handle excess zeroes in the table and the density function can be expressed as

$$f_{hurdle}(y;x,z,\beta,\gamma)=\begin{cases} f_{zero}(0;z,\gamma) & if\ y=0 \\ (1-f_{zero}(0,z,\gamma))\cdot f_{count}(y;x,\beta)/(1-f_{count}(0;x,\beta)) & if\ y>0 \end{cases}$$

where x and z are the covariate matrices, y is the dependent variable, β and γ are the unknown parameters, $f_{zero}(0;z,\gamma)$ is the probability that y_i produces only zeroes, $(1-f_{zero}(0;z,\gamma))$ is the probability that y_i is greater than 0, $f_{count}(y;x,\beta)$ is the density of Negative Binomial distribution and $f_{count}(0;x,\beta))$ is the cumulative distribution of Negative Binomial distribution.

2.2 Residuals

Residual techniques have been carried out by many researchers in order to identify the outlying cells in a table by considering residuals greater than ±3 and are heuristic in nature. In this heuristic approach, outliers are identified irrespective of the polarization of cell frequencies and order of the tables. To overcome this, in this study, the box plot of different types of residuals has been considered to identify the outlying cell. The different diagnostic measures considered are, Response residual (R), Deviance residual (D), and Pearson residual (P).

With the above set of classifications, the procedure adopted in this study as follows:

Step 1: Given a sparse $I\times J$ table, fit the six different models considered in this study by assuming the nature of the data as nominal.

Step 2: Examine the residuals associated with the model.

Step 3: Detect the outlying cells with the help of boxplot of residuals.

Thus this procedure provides a systematic approach of identifying outliers under conditions of polarity for varying order of the table. The following section deals with examining the robustness of proposed procedure as envisaged through a simulation study.

3. SIMULATION STUDY

The study of over 100 real time datasets available in the literature has shown that polarization is largely observed in tables of order more than 2×2. However, the simulation study considered tables of different order namely, 3×3, 4×4, 5×4, 5×5, 7×8 and 10×10, with varying N from 50 to 4550 and contain at least one zero cell count for the detection of outliers. The

cell frequencies of the tables are assumed to follow $Multinomial(N,(p_1, p_2,...,p_k))$ where the probabilities are considered as

(i) $p_i \sim U(0,1); i = 1,2,...,k$

(ii) the number of cells in a table is divided into three parts and p_i are chosen from the considered minimum (0, 0.2, 0.6) and maximum (0.2, 0.6, 1) probabilities, i.e. for first part of the table, the minimum probability will be 0 and maximum will be 0.2, and the cell counts are generated within this probability limits. Similarly the other two parts are considered with mentioned minimum and maximum probabilities

(iii) the number of cells in a table is divided into five halves and p_i are chosen from the given minimum (0 to 0.6) and maximum (0.2 to 1) probabilities.

The nature and behavior of different types of residuals with contaminating cells has been observed in the process of diagnostics for outlier detection. Here, contamination is restricted to single cell at a time and the number of cells to be contaminated are selected using $\min\{I, J\}$ where I and J be the number of rows and columns respectively. The concept of α -outlier region (Davies and Gather 1993) is considered for the contamination and repeated 500 times. First we assume that no outliers is present in the table and we contaminate a cell with a small positive quantity, say α, based on the outlier region and detected the contaminated cell as outliers with the help of residuals under different models as mentioned in section 2.1. It is to be noted that the residuals under the six models behaves similar irrespective of the chosen probabilities p_i and are tabulated in Table 1-3.

Table 1: **Percentage of correct identification of outliers with P_Z as low**

k	N	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
		R	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*
3 x 3	50	20	21	18	18	20	17	19	22	20	18	20	18	17	19	18	19	21	16
	350	19	20	18	20	22	19	16	18	15	17	21	19	16	22	20	19	23	19
	950	15	17	14	14	17	13	16	18	14	14	17	16	13	18	15	14	18	13
	2150	10	14	11	09	12	10	11	13	10	12	15	10	12	14	10	12	15	11
	4550	08	10	07	10	14	11	09	13	10	10	15	12	11	14	13	11	14	13
4 x 4	50	20	21	18	18	20	17	19	22	20	18	20	18	17	19	18	19	21	16
	350	19	20	18	20	22	19	16	18	15	17	21	19	16	22	20	19	23	19
	950	15	17	14	14	17	13	16	18	14	14	17	16	13	18	15	14	18	13
	2150	10	14	11	09	12	10	11	13	10	12	15	10	12	14	10	12	15	11
	4550	08	10	07	10	14	11	09	13	10	10	15	12	11	14	13	11	14	13
5 x 4	50	20	21	18	18	20	17	19	22	20	18	20	18	17	19	18	19	21	16
	350	19	20	18	20	22	19	16	18	15	17	21	19	16	22	20	19	23	19
	950	15	17	14	14	17	13	16	18	14	14	17	16	13	18	15	14	18	13
	2150	10	14	11	09	12	10	11	13	10	12	15	10	12	14	10	12	15	11
	4550	08	10	07	10	14	11	09	13	10	10	15	12	11	14	13	11	14	13
5 x 5	50	25	23	21	23	20	23	26	28	24	28	26	30	37	39	36	38	40	36
	350	30	28	27	33	30	34	36	38	34	38	36	40	38	41	46	48	39	46
	950	39	36	34	43	40	46	47	49	44	48	43	40	48	51	56	52	49	48
	2150	51	49	46	53	50	55	59	56	53	58	52	51	61	59	52	64	53	59
	4550	48	51	49	56	53	61	65	68	57	63	59	64	68	62	69	62	55	58

k	N	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
		R	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*
7 x 8	50	42	37	41	48	43	47	43	51	47	43	42	40	48	53	51	42	52	49
	350	53	49	50	54	49	53	55	58	53	57	51	49	53	58	52	52	56	49
	950	58	56	52	56	50	57	61	67	63	64	62	61	67	69	64	63	69	62
	2150	57	55	49	64	59	67	63	69	61	68	64	63	69	71	63	67	72	64
	4550	53	49	47	56	52	64	57	62	58	61	56	53	62	69	57	62	71	59
10 x 10	50	43	39	42	50	45	46	47	52	49	44	41	39	46	50	49	44	50	48
	350	53	49	50	54	49	53	55	58	53	57	51	49	53	58	52	52	56	49
	950	58	56	52	56	50	57	61	67	63	64	62	61	67	69	64	63	69	62
	2150	57	55	49	64	59	67	63	69	61	68	64	63	69	71	63	67	72	64
	4550	53	49	47	56	52	64	57	62	58	61	56	53	62	69	57	62	71	59

Table 2: **Percentage of correct identification of outliers with P_Z as moderate**

k	N	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
		R	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*
3 x 3	50	24	19	18	27	23	21	24	28	26	25	23	22	23	20	19	27	29	24
	350	31	27	26	34	32	30	34	38	29	32	30	27	31	26	23	34	27	21
	950	37	31	29	41	35	31	43	37	32	46	43	37	47	41	37	49	45	34
	2150	21	23	19	24	27	23	27	25	21	29	24	26	28	26	21	24	23	19
	4550	24	20	18	27	25	20	27	26	21	24	27	23	19	24	20	27	23	18
4 x 4	50	25	17	17	26	20	19	20	27	24	23	20	19	19	19	17	24	26	20
	350	32	25	24	31	35	26	32	35	23	29	29	24	27	24	22	31	24	33
	950	34	29	27	39	36	32	39	32	30	42	41	32	43	39	34	44	41	17
	2150	23	21	16	26	23	22	24	23	23	27	23	23	26	23	17	23	24	15
	4550	24	17	14	24	20	16	23	26	24	21	26	21	16	21	16	26	22	17
5 x 4	50	23	15	14	22	24	19	24	23	21	21	18	21	23	17	19	21	24	17
	350	30	23	21	29	19	21	19	21	24	28	27	26	25	23	24	29	23	30
	950	30	24	24	34	33	26	30	32	31	39	38	34	40	37	37	43	42	16
	2150	21	19	19	24	26	31	35	30	26	24	20	26	28	21	21	21	21	17
	4550	20	15	16	23	21	20	20	26	27	19	25	20	21	18	18	27	19	19
5 x 5	50	35	26	31	32	27	29	37	32	39	41	33	37	38	29	34	36	25	32
	350	41	35	29	43	35	41	45	37	42	47	43	48	51	50	53	49	36	51
	950	39	32	24	37	31	42	43	35	37	43	46	42	49	47	54	51	49	54
	2150	44	38	26	29	26	29	38	34	36	40	47	49	39	38	43	53	51	56
	4550	41	36	32	34	30	33	43	37	35	37	36	38	42	39	44	47	44	46
7 x 8	50	47	44	43	48	37	49	57	52	51	51	53	51	58	53	54	56	55	49
	350	51	49	41	53	45	51	55	57	49	53	54	53	61	54	52	59	56	48
	950	57	55	53	57	51	53	53	55	47	54	56	52	59	51	50	61	59	51
	2150	61	59	51	69	66	59	58	54	46	50	52	48	53	49	47	63	54	52
	4550	63	62	59	64	60	53	53	57	49	57	55	51	49	47	43	57	49	43
10 x 10	50	25	23	21	23	20	23	26	28	24	28	26	30	37	39	36	38	40	36
	350	45	39	48	51	49	47	55	49	51	53	49	51	57	53	51	58	56	51
	950	52	48	42	53	51	46	61	51	53	61	53	53	66	61	59	64	55	47
	2150	61	55	50	62	56	55	67	58	61	68	57	52	68	59	62	72	65	51
	4550	64	57	49	67	58	53	69	61	64	71	62	61	69	63	64	73	62	61

Table 3: **Percentage of correct identification of outliers with P_Z as high**

k	N	Model 1			Model 2			Model 3			Model 4			Model 5			Model 6		
		R	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*	*R*	*D*	*P*
3 x 3	50	22	21	18	19	23	21	23	20	19	21	20	18	30	21	19	29	24	20
	350	18	16	14	14	12	11	17	16	12	15	16	13	15	12	11	12	11	10
	950	16	12	11	15	13	12	15	12	09	12	11	10	12	10	12	11	09	08
	2150	13	09	07	09	07	06	12	09	06	11	08	07	10	07	06	06	05	04
	4550	10	06	05	06	04	04	11	05	04	07	06	06	11	06	06	05	06	05
4 x 4	50	21	20	17	21	19	19	21	19	17	19	18	16	21	19	17	21	19	16
	350	16	15	13	15	10	11	16	15	11	15	14	11	14	11	10	10	09	05
	950	15	12	12	13	12	11	14	11	07	11	10	09	10	09	05	09	06	04
	2150	12	11	09	11	09	07	11	08	05	09	07	05	05	05	03	05	03	02
	4550	11	07	06	07	06	05	09	04	03	03	05	04	03	04	03	04	03	02
5 x 4	50	21	18	15	20	18	18	19	18	16	17	16	15	17	18	17	21	19	16
	350	15	14	11	14	11	12	15	14	12	14	12	10	13	10	10	10	09	05
	950	14	12	10	12	10	11	13	10	08	10	10	08	09	08	05	09	06	04
	2150	10	11	08	09	08	07	10	07	06	08	08	07	06	05	03	05	03	02
	4550	09	08	07	08	07	06	08	05	04	04	06	05	04	04	03	04	03	02
5 x 5	50	23	21	19	20	19	20	23	21	20	23	21	21	19	20	23	25	23	21
	350	21	19	16	17	15	16	21	19	17	18	19	18	17	16	18	21	19	18
	950	11	09	08	14	11	12	12	14	13	12	13	12	13	12	14	12	14	10
	2150	11	08	08	12	10	09	10	12	11	11	09	10	12	11	13	10	12	09
	4550	10	11	08	09	08	07	10	07	06	08	08	07	06	05	03	05	03	02
7 x 8	50	21	19	20	24	23	21	23	19	21	27	23	21	24	21	19	25	27	26
	350	35	32	29	41	42	38	43	39	37	46	41	38	42	38	36	45	38	36
	950	26	24	20	27	21	19	24	28	24	28	26	22	27	24	26	31	28	22
	2150	20	19	17	21	18	13	19	15	13	17	16	14	18	15	11	19	17	13
	4550	21	20	18	23	20	17	19	15	14	23	22	19	23	20	17	21	20	16
10 x 10	50	26	28	27	31	30	25	37	32	29	34	31	28	37	34	32	36	31	25
	350	34	32	30	37	32	31	39	37	36	41	37	32	39	35	36	34	31	29
	950	31	30	25	32	29	27	23	34	31	28	37	35	31	38	32	30	28	26
	2150	34	36	34	32	37	35	32	31	30	29	34	31	28	32	30	27	26	21
	4550	21	18	17	24	23	18	26	23	21	28	23	17	27	22	18	24	23	20

For instance, if a table falls under low category in both k and P_Z, then from the simulation study it is clear that the deviance residuals identified the outlying cells to a greater extent and is the more suitable diagnostics method across the models. When examining the preference of the suitable model under the three residuals, all the six different models performs equally in response and deviance residuals and Hurdle Poisson and the Hurdle Negative Binomial model is preferred in the case of Pearson residuals. Similarly, when the table with low k and moderate number of zero cells (P_Z), response residuals identified the outlying cells more precisely and is preferred among the three residuals. While examining the suitability of the models, Negative Binomial model and ZINB performs better in the response residuals, ZIP and Hurdle NB is more suitable in the case of deviance

residuals and all the six models performs equally in Pearson residuals. The results in finding the suitable detection technique under the classification of k and P_Z are tabulated in Table 4.

Table 4: **General rule for outlier detection based on the simulation study**

S.No.	k	P_Z	Residuals	Model		
				Response	*Deviance*	*Pearson*
1	L	L	Deviance	All	All	5 and 6
2	L	M	Response	2 and 4 are good 5 and 6 are partially good	3 and 6	All
3	L	H	Response and Deviance	5 and 6	3, 4, 5, and 6	3, 4 and 5
4	M	L	All	All	3 and 5	2 and 4
5	M	M	Response and Pearson	All	3, 4, 5 and 6	All
6	M	H	Response	All	4 and 6	2
7	H	L	Response performs better Deviance and Pearson are fairly good	All	3, 5 and 6	1, 3 and 5
8	H	M	Response and Deviance	All	3, 4, 5 and 6	All
9	H	H	All	All	3, 4, 5 and 6	All

The polarization of the cell counts becomes a major issue in the detection of outliers in $I \times J$ contingency tables. Indeed, the use of residuals under the suitable model with the use of boxplot turns out to be a good choice in detecting the outlying cells. Further to simulation, the study explored certain well known data to establish the results of simulation.

4. PRACTICAL EXAMPLE

Consider a 14×14 contingency table representing father and son occupation (see Table 5) with $k = 196$, $N = 775$ and the proportion of zero cell frequency $P_Z = 0.26$ tells us that this table contains large number of cells and also more zero cells.

Table 5: **Contingency table involving father and son's occupation**

Father/Son's Occupation	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**
A	28	0	4	0	0	0	1	3	3	0	3	1	5	2
B	2	51	1	1	2	0	0	1	2	0	0	0	1	1
C	6	5	7	0	9	1	3	6	4	2	1	1	2	7
D	0	12	0	6	5	0	0	1	7	1	2	0	0	10
E	5	5	2	1	54	0	0	6	9	4	12	3	1	13
F	0	2	3	0	3	0	0	1	4	1	4	2	1	5
G	17	1	4	0	14	0	6	11	4	1	3	3	17	7
H	3	5	6	0	6	0	2	18	13	1	1	1	8	5
I	0	1	1	0	4	0	0	1	4	0	2	1	1	4
J	12	16	4	1	15	0	0	5	13	11	6	1	7	15
K	0	4	2	0	1	0	0	0	3	0	20	0	5	6
L	1	3	1	0	0	0	1	0	1	1	1	6	2	1
M	5	0	2	0	3	0	1	8	1	2	2	3	23	1
N	5	3	0	2	6	0	1	3	1	0	0	1	1	9

The categories are: A: army; B: arts; C: teacher, clerk, civil servant; D: crafts; E: divinity; F: agriculture; G: landownership; H: law; I: literature; J: commerce; K: medicine; L: navy; M: politics and court; and N: scholarship and science

This table was also studied by Kotze and Hawkins (1984) in identifying outliers by adding half to the empty cells and detected 15 cells as outliers using half normal plot. In our method, the table falls under high category in k and P_Z. The response residual under the six models detected the cells (1,1), (7,1), (1,2), (2,2), (5,2), (7,2), (13,2), (1,5), (5,5), (8,8), (8,9), (10,10), (11,11), (12,12), (4,13), (6,13), (13,13), and (14,14) as outliers, the deviance residuals identified (1,1), (5,5), (2,2), (11,11), and (13,13) as most outlying cells and Pearson residuals detected (1,1), (2,2), (4,4), (5,5), (8,8), (10,10), (11,11), (12,12), and (13,13) as outliers under the six models considered. When examining the residuals, all the three residuals yield better identification of outliers. Also, ZIP, ZINB, Hurdle Poisson and Hurdle NB models are preferred in the case of deviance residuals and all the models behaves similar in identification of outliers in the case of Response and Pearson residuals and the boxplot of the residuals under the six different models is presented in Figures 1-3.

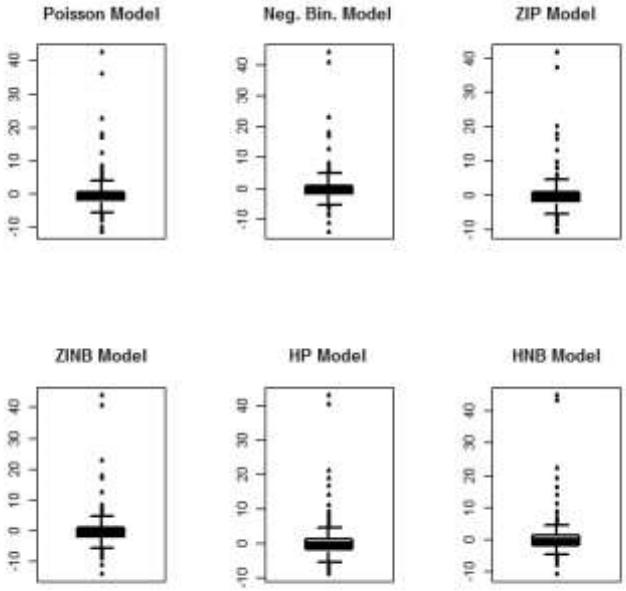

Figure 1. Boxplots of Response residuals

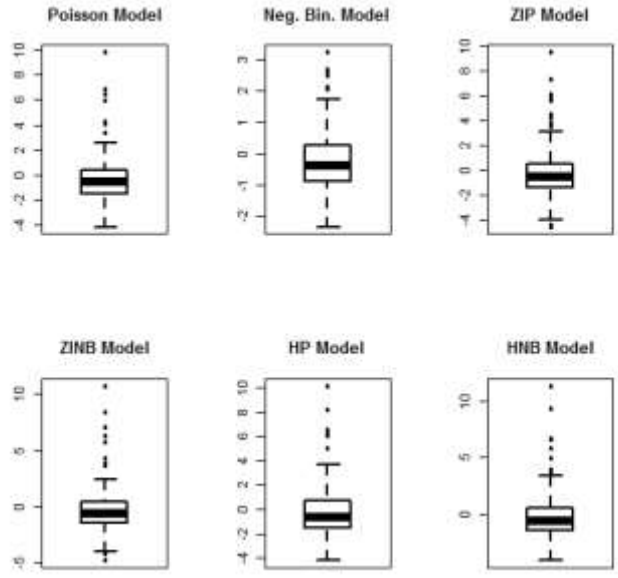

Figure 2. Boxplots of Deviance residuals

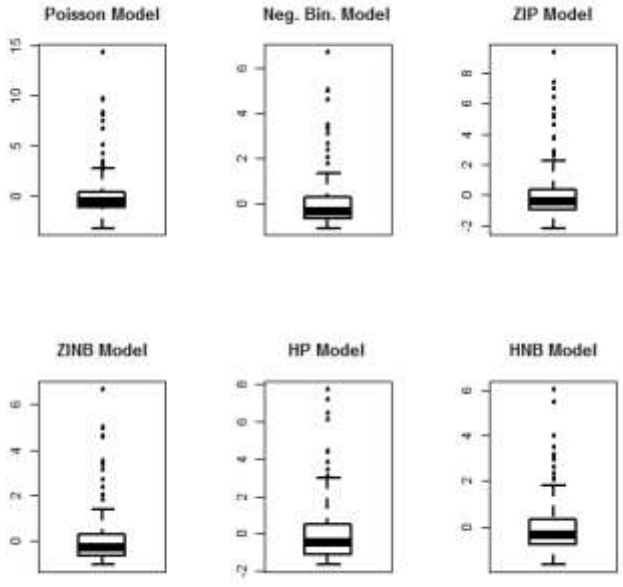

Figure 3. Boxplots of Pearson residuals

5. DISCUSSION AND CONCLUSION

Diagnostics in sparse $I \times J$ contingency table has drawn a great deal of attention to the statisticians for many years. There is no general agreement among the statisticians about the detection of outliers due to the polarization of cell frequencies in contingency tables. Such polarized cells in sparse $I \times J$ contingency tables has been examined through modeling aspect. Sparseness in $I \times J$ table and the total number of cell frequencies in the table are classified into three categories in order to arrive at a general identification rule across the models and residuals. The models considered in this study are Poisson, Negative Binomial, ZIP, ZINB, Hurdle Poisson, and Hurdle NB models along with three kinds of residuals namely Response, Deviance and Pearson residuals. The procedure deals with fitting six different models and the usual diagnostic measures such as residuals supplemented by boxplot are used to identify the exact outlying cells. The stability of our proposed methods towards the identification of outliers is examined through a simulation study. Moreover, it is evident that the results provide an idea on impact of polarization in sparse tables, and is found to be useful in detecting outliers. Based on the results, we conclude that the proposed technique provides a better way to choose the model and residuals under the proposed condition of sparseness and this could be a viable approach in detecting outlier cells in $I \times J$ contingency tables.

REFERENCES

[1] Agresti, A., & Yang, M. C. (1987) An empirical investigation of some effects of Sparseness in Contingency tables. Computational Statistics & Data Analysis, 5, 9-21.

[2] Agresti, A. (2002) Categorical Data Analysis. Wiley, New York.

[3] Barnett, V. D., & Lewis, T. (1978) Outliers in statistical data. Wiley: New York.

[4] Bäohning, D., Dietz, E., & Schlattmann, P. (1999) The zero-inflated Poisson model and decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society, Series A, 162, 195-209.

[5] Bradu, D., & Hawkins, D.M. (1982) Location of Multiple Outliers in Two-Way Tables Using Tetrads. Technometrics, 24, 103-108.

[6] Davis, L., & Gather, U. (1993) The identification of multiple outlier. Journal of the American Statistical Association, 88, 782 – 792.

[7] Dietz, E., & Bäohning, D. (2000) On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics and Data Analysis, 34, 441-459.

[8] Germu, S., & Trivedi, P. K. (1996) Excess zeros in count models for Recreational Trips. Journal of Business and Economic Statistics, 14, 469-477.

[9] Haberman, S. J. (1973) The Analysis of Residuals in Cross-Classified Tables. Biometrics, 29, 205-220.

[10] Heilbron, D. (1994) Zero-altered and other regression models for count data with added zeros. Biometrical Journal, 36, 531-547.

[11] Joreskog, K.G., & Moustaki, I. (2001) Factor analysis of ordinal variables: A comparison of three approaches. Multivariate Behavioural Research, 36, 347-387.

[12] Kateri, M., (2014) Contingency Table Analysis. Springer.

[13] Koehler, K., & Larntz, K. (1980) An empirical investigation of goodness–of–fit statistics for sparse multinomials. Journal of the American Statistical Association, 75, 336-344.

[14] Kotze, T.J.vW., & Hawkins, D.M. (1984) The Identification of Outliers in Two-Way Contingency Tables using 2×2 Subtables. Applied Statistics, 33, 215-223.

[15] Kuhnt, S. (2004) Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. Scand. J. Stat., 31, 431–442.

[16] Kuhnt, S., Rapallo, F., & Rehage, A. (2014) Outlier Detection in Contingency Tables based on Minimal Patterns. Stat. Comput., 24, 481-491.

[17] Lambert, D. (1992) Zero-inflated Poisson regression with application to defects in manufacturing. Technometrics, 41, 29-38.

[18] Larntz, K. (1978) Small sample comparisons of exact levels for chi-squared goodness-of-fit statistics. Journal of the American Statistical Association, 73, 253-263.

[19] Lee, A. H., & Yick J. S. (1999) A Perturbation Approach to Outlier Detection in Two-Way Contingency Tables. Australian & New Zealand J. Statist., 41, 305–314.

[20] Mignone, F., & Rapallo, F. (2018) Detection of outlying proportions. Journal of Applied Statistics, 45, 1382 – 1395.

[21] Rapallo, F. (2012) Outliers and patterns of outliers in contingency tables with algebraic statistics. Scand. J. Stat., 39, 784–797.

[22] Reiser, M., & Vandenberg, M. (1994) Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. British Journal of Mathematical and Statistical Psychology, 47, 85-107.

[23] Ridout, M. S., Demetrio, C. G. B., & Hinde, J. P. (1998) Model for count data with many zeros. International Biometric Conference. 179-190.

[24] Sangeetha, U., Subbiah, M., Srinivasan, M.R., & Nandram, B. (2014) Sensitivity Analysis of Bayes Factor for Categorical Data with Emphasis on Sparse Multinomial Data. Journal of Data Science, 12, 339-357.

[25] Simonoff, J. S. (1988) Detecting Outlying Cells in Two-Way Contingency Tables via Backwards Stepping. Technometrics, 30, 339-345.

[26] Sripriya, T. P., & Srinivasan, M. R. (2018a), Detection of outliers in categorical data using model based diagnostics. Special Proceedings of 20th Annual Conference of SSCA held at Pondicherry University, Puducherry during January 29-31, 2018, 35-43.

[27] Sripriya, T. P., & Srinivasan, M. R. (2018b) Detection of outlying cells in Two-Way Contingency Tables. Statistics and Applications, 16, 103-113.

[28] Subbiah, M., & Srinivasan, M. R. (2008) Classification of 2×2 sparse data with zero cells. Statistics & Probability Letters, 78, 3212-3215.

[29] Van Den Broek, J. (1995) A score test for zero inflation in a Poisson distribution. Biometrics, 51, 738-743.

[30] Yick, J.S., & Lee, A.H. (1988) Unmasking Outliers in Two-Way Contingency Tables. Computational Statistics and Data Analysis, 29, 69-79.