Austrian Journal of Statistics
August 2020, Volume 49, 59–67.
http://www.ajs.or.at/
doi:10.17713/ajs.v49i5.938

Detection of Outlying Cells in Contingency Tables Using Model Based Diagnostics

Thodur P. Sripriya
University of Madras

Michele Gallo
University of Naples - L'Orientale

Mamandur R. Srinivasan
University of Madras

Abstract

Detecting outliers in contingency table is an interesting statistical problem and it poses additional difficulties due to the polarization of cell counts. The fundamental definition of 'markedly deviant' cell as an outlier is clearly exploited in this study by introducing a pivot element to capture the deviations. The present study considers a two-step confirmatory procedure to detect outliers in $I \times J$ contingency table. The procedure deals with (i) identifying the reliable set of candidate outliers using the deviation from the pivot element and then (ii) detect those set of outlying cells by examining different type of residuals of the suitable fitted model. The robustness of the procedure is investigated through a simulation study along with applications to real datasets.

Keywords: Poisson log-linear model, negative binomial model, diagnostics, residuals, boxplot, outlier(s).

1. Introduction

In recent years, a great deal of attention has been paid to the accommodation and identification of unusual observations (outliers) in the data. Outliers may be real errors, or else accurate but unexpected observations which could shed new light on the phenomenon under study (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. Hence, an attempt has been made to explain the fundamental meaning of 'markedly deviant' as a pivotal element by answering; which cell, from where and, by how much, based on the generic characteristics of the table.

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 study on independence. Graphical display such as biplots, mosaic plots, etc., can also be useful in studying the association between the I rows and J columns and could be useful in identifying the outlying cells in contingency table (Friendly (2000); Beh and Lombardo (2014)). 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 test the goodness of fit to specify the notion of outlier with the use of algebraic statistics. Kuhnt, Rapallo, and Rehage (2014) detected outliers through subsets of cell counts called minimal patterns for the independence model. Mignone and Rapallo (2018) identified the outlying cells based on a set of proportions in a contingency table. Sripriya and Srinivasan (2018) has proposed a new method to detect outliers in two-dimensional tables. However, this study presents an alternative approach to detect outliers based on the assumption of model independence.

Residual based techniques has been widely used to detect outliers in contingency tables (Haberman (1973); Fuchs and Kenett (1980); Bradu and Hawkins (1982); Yick and Lee (1998); Simonoff (1988); Lee and Yick (1999)). Even though, the residual technique has been used, no cutoff criteria is provided in choosing the maximum residuals and is more heuristic in nature.

Further, polarization of cell counts is one of the major problem when it comes to outlier detection. Polarization is basically an uneven distribution of counts in the $I \times J$ Table. Polarization in contingency tables involve 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 (Sripriya, Srinivasan, and Gallo (2019)). 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, Subbiah, Srinivasan, and Nandram (2014)). Following, Subbiah and Srinivasan (2008) on the sensitivity analysis of 2×2 tables, location of polarized counts in the table pose additional challenge in the detection of outliers.

In this paper, we propose a two-step confirmatory procedure to detect potential outliers in two-way contingency table. Firstly, the method identifies the reliable set of candidate outliers in $I \times J$ table through the deviation from the pivot element. Secondly, the model based diagnostics is used to obtain the results followed by boxplots to confirm the outlying cells.

2. Proposed method

Consider N sample observations that are cross-classified in an $I \times J$ $(=N)$ contingency table, and Y_k , $k = 1, \ldots, N$. are assumed to be the realizations of random variables. Once a contingency table is constructed, the first interest will be the hypothesis of either homogeneity or independence depending on the sampling scheme (Agresti (2002)). When the null hypothesis is rejected, the cell residuals are investigated to identify the cells which deviate greatly from others. The cell is considered to be an outlier when the observed frequency deviates markedly from the corresponding expected frequency under the null model.

Let n_{ij} be the observed cell frequencies of $I \times J$ table, $N = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$ be the total frequency, and let $T = N/k$; where $k = IJ$, be the pivot element through which the markedly deviant cells are obtained as the candidate set of outliers, denoted by the subset as S. For an $I \times J$ table, calculate the deviations $D_{ij} = |T - n_{ij}|$ and examine the deviations D_{ij} for each row and if any D_{ij} is markedly deviant from the neighbouring cells then that particular cell is said to be discordant and is included in the subset S. The steps involved in the confirmatory procedure are as follows:

Step 1: Given an $I \times J$ table, locate the set of candidate outliers S, using $D_{ij} = |T - n_{ij}|$.

Step 2: Fit a Poisson Log-Linear model for the data with S as the nature of the data is count. If the model fits well go to step 3, else step 4.

Step 3: Examine different types of residuals associated with the model and detect the outliers through boxplot of residuals.

Step 4: Fit a Negative Binomial model and do step 3.

Residual techniques have been carried out by researchers in order to identify the outlying cells in a table by considering residuals greater than ∓ 3. In this heuristic approach, outliers are identified irrespective of the polarization of cell frequencies and order of the contingency tables. To overcome this, the box plot of different types of residuals has been considered to identify the outlying cell. The different diagnostic measures considered are, (i) Response residual, (ii) Deviance residual, (iii) Pearson residual, and (iv) Deleted residual.

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 study considered tables of order (3×3), (4×4) and (5×5) with N varying from 50 to 350 for the detection of outliers. The cell frequencies of the tables are assumed to follow Mult $(N, (p_1, p_2, \ldots, p_k))$ where $p_i \sim$ U$(0, 1)$; $i = 1, 2, \ldots, k \ldots$ The behaviour of different types of residuals with contaminating the 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. Different level of contamination α (10% to 100% of row total) are considered and repeated 500 times. We examined the consistency of correctly identified cells among four different residuals in this simulation study.

The six different scenarios described below are carried out through a simulation study and the results are presented in Table 1-6.

Generate 500 tables of size 3×3 and N ranges from 50 to 100. The results reveals that the response and deleted residuals performs well in detecting the outliers in Poisson model and the response residuals performs well in Negative Binomial model. The Pearson residuals yield a poor results in detecting outliers in this approach.

Generate 500 tables of size 3×3 and N ranges from 100 to 350. The residual analysis shows that response and deleted residual identified the outliers to a greater level in both the models and also the four residuals yields better performance in Poisson model than Negative Binomial model.

Generate 500 4×4 tables and N ranges from 50 to 100. The results reveals that all the four residuals performed poorly in detecting the outliers except response residuals in Negative Binomial model.

Generate 500 4×4 tables and N ranges from 100 to 350. The results reveals that all the four residuals performed poorly in detecting the outliers in both the models due to the behaviour of the neighbouring cells and probably even distribution of counts in the table generated.

Generate 500 5×5 tables and N ranges from 50 to 100. The results reveals that all the four residuals performed poorly in detecting the outliers except response residuals in Poisson Log-Linear model.

Finally, simulation is carried out by considering 500 tables of size 5×5 and N ranges from 50 to 350. The result showed that all the four residuals performed poorly in detecting the outliers due to the behaviour of the neighbouring cells and probably even distribution of counts in the table generated.

Table 1: Number of outliers detected in 3×3 with N lying between 50 and 100 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	08	06	07	05	05	00	00	02
20	11	10	10	198	09	02	11	53
30	13	12	12	210	10	03	13	66
40	18	17	16	248	14	04	16	98
50	24	20	19	256	28	17	27	114
60	200	24	21	267	232	29	30	125
70	253	27	205	311	242	31	112	148
80	256	29	245	340	249	35	144	181
90	256	33	256	364	258	37	178	242
100	290	36	268	402	267	39	187	249

Table 2: Number of outliers detected in 3×3 with N lying between 100 and 350 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	04	05	03	06	10	02	03	05
20	05	08	05	86	15	11	12	43
30	06	17	08	135	21	13	13	99
40	09	21	11	176	30	15	16	123
50	15	29	14	188	42	18	19	137
60	27	32	32	230	56	31	34	141
70	93	39	88	255	68	83	96	149
80	114	87	146	264	93	115	127	157
90	153	175	162	299	142	139	138	163
100	238	198	195	368	189	142	144	174

Table 3: Number of outliers detected in 4×4 with N lying between 50 and 100 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	08	05	06	08	09	01	02	06
20	14	09	12	19	16	05	11	12
30	25	15	17	56	31	10	18	44
40	29	18	23	76	58	11	21	50
50	32	32	29	88	74	33	30	76
60	37	62	56	96	88	55	55	92
70	46	85	104	106	91	70	101	135
80	67	108	118	152	153	102	114	139
90	86	127	122	154	181	124	129	146
100	93	136	126	162	279	128	131	152

Table 4: Number of outliers detected in 4×4 with N lying between 100 and 350 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	01	01	02	03	10	02	00	01
20	04	02	05	06	15	07	02	03
30	08	06	09	10	19	11	07	05
40	10	11	15	17	26	15	10	12
50	18	20	24	26	34	17	16	15
60	22	27	35	38	48	22	22	26
70	37	34	47	59	57	38	34	41
80	42	48	65	63	68	54	43	54
90	76	65	78	84	74	71	57	67
100	89	72	92	91	79	78	68	81

Table 5: Number of outliers detected in 5×5 with N lying between 50 and 100 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	11	01	00	00	03	00	00	01
20	27	03	01	01	05	01	00	03
30	34	10	03	05	11	03	01	08
40	39	18	17	08	16	05	04	13
50	48	21	26	18	27	08	07	23
60	52	28	29	32	39	11	17	34
70	69	31	39	48	48	18	22	47
80	65	39	44	52	52	21	34	53
90	91	41	53	64	67	24	40	61
100	129	42	61	69	72	37	54	68

Table 6: Number of outliers detected in 5×5 with N lying between 100 and 350 (out of 500)

α	Poisson Model				Negative Binomial			
(in %)	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
10	03	02	01	01	04	01	01	02
20	08	04	01	04	07	05	02	04
30	16	10	03	13	15	11	05	07
40	29	15	05	17	28	13	09	13
50	37	21	09	25	35	26	14	16
60	49	27	11	33	42	32	21	19
70	53	35	18	40	57	36	29	38
80	69	42	24	49	68	48	37	51
90	73	58	38	51	72	55	41	63
100	82	64	49	53	84	61	51	79

Table 7: Percentage of correct classification of non-outlying cells

Scenarios		Poisson Model				Negative Binomial			
$I \times J$	N	Response	Pearson	Deviance	Deleted	Response	Pearson	Deviance	Deleted
3×3	50 - 100	72	75	79	82	76	79	81	84
	100 - 350	83	87	89	91	81	85	91	94
4×4	50 - 100	77	79	83	87	78	82	88	90
	100 - 350	86	89	93	94	83	87	94	95
5×5	50 - 100	79	81	86	89	76	79	84	89
	100 - 350	85	88	92	95	82	87	91	94

Table 8: 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

Source: Yick and Lee (1998)

The study has also considered the percentage of correct classification of non-outlying cells or inliers in different scenarios considered in the simulation study irrespective of the contamination level and the same are presented in Table 7. The results revealed that the percentage of correct classification of inliers improves when N increases irrespective of the order of the table. Also, the inliers classified in deviance and deleted residuals are comparatively higher in both Poisson and Negative Binomial model as compared to response and Pearson residuals.

The simulation study has shown that polarization of cell counts is a major issue in the detection of outliers in $I \times J$ contingency tables. Indeed, the use of residuals as a suitable diagnostic measure under the suitable model with boxplot turns out to be a good choice in detecting the outlying cells. The present simulation study is restricted to smaller tables and could be extendedn to modelling higher dimensional tables for detecting outliers. Further to simulation, the study explored certain well known data to establish the results of simulation.

4. Data analysis

4.1. Student's enrolment data

The study consist of Student's enrolment data of Northern Territory, Australia conducted in seven community schools in eight different periods of the year and presented in the following table (Yick and Lee (1998)). The primary interest lies in detecting the outlying cells from the data, if any, before carrying out further analysis.

Following the method outlined in Section (2), deviations from the pivot element identified the candidate set $S = (1,5),(1,6),(2,4),(2,5)$ as outliers. In the confirmatory procedure, Negative Binomial model fits the data well and the four types of residuals detected (1, 5), (1, 6), (2, 4) and (7, 5) as potential outliers and the boxplot of residuals are presented in Fig 1. The non-aberrant cell (7, 5) is identified as outlier in the residual diagnostic approach and the cell (2, 5) is not detected due to masking in the identification stage. Upon application of perturbation diagnostics (Yick and Lee (1998)) produces the cells (1, 5), (1, 6), (2, 4) and (2, 5) as potential outliers. Thus, the identification of outliers from the pivot element appears to be a resistant to masking effect.

Table 9: Artificial data

18	41	41	20	21
39	20	20	22	22
24	20	20	16	18
20	20	19	19	19
23	19	20	17	20

Source: Simonoff (1988)

4.2. Artificial data

As a second illustration, we consider 5×5 contingency table from Simonoff (1988) an artificial data presented in the following table, with induced outliers.

In our method, the deviations from the pivot element identified the candidate set $S = (1,2),(1,3),(2,1)$ as outliers. In the confirmatory procedure, Poisson model fits the data well and the four types of residuals detected the cells (1, 1), (1, 2), (1, 3) and (2, 1) as outlying cells and the boxplots are presented in Fig 2. Here the non-aberrant cell (1, 1) is identified as outliers using the residual approach. The same cells (1, 2), (1, 3), and (2, 1) are found to be outliers via deleted residuals and (1, 1) being swamped in adjusted residuals by Simonoff (1988). Thus, the detection with deviation from pivot element is resistant to swamping effect also.

5. Conclusions

Diagnostics in $I \times J$ contingency table has drawn a great deal of attention by the statisticians for many years but the notion of outliers is not well defined. 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 $I \times J$ contingency tables has been examined through the independence of attributes. In this direction, a two phase objective is devised with the identification of pivot element to examine their deviations and then a confirmatory approach to identify the outliers a model based diagnostics.

The procedure deals with finding the reliable set of candidate outliers through a distance measure $D_{ij} = |T - n_{ij}|$ and then applying the confirmatory procedure by fitting a suitable model and the usual diagnostic measures (residuals) followed by boxplot to identify the outlying cells. The stability of our proposed methods towards the identification of outliers is examined through a simulation study. The results have revealed that response and deleted residuals approach identifies the outliers to a greater extent than compared to other residual methods. Moreover, it is evident that the results provide an idea on impact of polarization in the table, and is found to be useful in detecting outliers.

Based on the numerical results, we conclude that the two-step confirmatory procedure as a combination of suitable diagnostic measure and an appropriate graphical approach through boxplots could be a viable approach in detecting outlier cells in $I \times J$ contingency tables. The proposed pivot element detection technique is resistant to masking and swamping effects. The results based on fitting of other generalised linear models with the presence of zero cell frequencies to detect outliers is under investigation.

Acknowledgements

The authors would like to thank the anonymous reviewers for their useful comments and suggestions. This work was financially supported by 2018 funds of the University of Naples - L'Orientale (I).

References

Agresti A (2002). *Categorical Data Analysis*. John Wiley & Sons.

Barnett V, Lewis T (1994). *Outliers in Statistical Data*. John Wiley & Sons.

Beh EJ, Lombardo R (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. John Wiley & Sons.

Bradu D, Hawkins DM (1982). "Location of Multiple Outliers in Two-way Tables, Using Tetrads." *Technometrics*, **24**(2), 103–108.

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

Fuchs C, Kenett R (1980). "A Test for Detecting Outlying Cells in the Multinomial Distribution and Two-way Contingency Tables." *Journal of the American Statistical Association*, **75**(370), 395–398.

Haberman SJ (1973). "The Analysis of Residuals in Cross-classified Tables." *Biometrics*, pp. 205–220.

Kuhnt S (2004). "Outlier Identification Procedures for Contingency Tables Using Maximum Likelihood and L1 Estimates." *Scandinavian Journal of Statistics*, **31**(3), 431–442.

Kuhnt S, Rapallo F, Rehage A (2014). "Outlier Detection in Contingency Tables Based on Minimal Patterns." *Statistics and Computing*, **24**(3), 481–491.

Lee AH, Yick JS (1999). "Theory & Methods: A Perturbation Approach to Outlier Detection in Two-Way Contingency Tables." *Australian & New Zealand Journal of Statistics*, **41**(3), 305–315.

Mignone F, Rapallo F (2018). "Detection of Outlying Proportions." *Journal of Applied Statistics*, **45**(8), 1382–1395.

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 MR, Nandram B (2014). "Sensitivity Analysis of Bayes Factor for Categorical Data with Emphasis on Sparse Multinomial Data." *Journal of Data Science*, **12**(2), 339–357.

Simonoff JS (1988). "Detecting Outlying Cells in Two-way Contingency Tables via Backwards-stepping." *Technometrics*, **30**(3), 339–345.

Sripriya TP, Srinivasan MR (2018). "Detection of Outlying Cells in Two-Way Contingency Tables." *Statistics and Applications*, **16**(2), 103–113.

Sripriya TP, Srinivasan MR, Gallo M (2019). "Robust Distance Measure to Detect Outliers for Categorical Data." *Soft Computing*, pp. 1–8.

Subbiah M, Srinivasan MR (2008). "Classification of 2× 2 Sparse Data Sets with Zero Cells." *Statistics & Probability Letters*, **78**(18), 3212–3215.

Yick JS, Lee AH (1998). "Unmasking Outliers in Two-way Contingency Tables." *Computational statistics & data analysis*, **29**(1), 69–79.

Affiliation:

Thodur P. Sripriya
Department of Statistics
University of Madras
Chennai, Tamilnadu, India
E-mail: sri.chocho@gmail.com

Michele Gallo
Department of Human and Social Sciences
University of Naples - L'Orientale
I-80134 Naples, Italy
E-mail: mgallo@unior.it
URL: http://docenti2.unior.it/index2.php?user_id=mgallo&content_id_start=1

Mamandur R. Srinivasan
Department of Statistics
University of Madras
Chennai, Tamilnadu, India
E-mail: mrsvasan8@hotmail.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 49
August 2020

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2019-02-11
Accepted: 2019-06-11