CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

Volume 11 Number 1
April 2020
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (`www.soche.cl`). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS is an international scientific forum strongly committed to gender equality, open access of publications and data, and the new era of information. The ChJS covers a broad range of topics in statistics, data science, data mining, artificial intelligence, and big data, including research, survey and teaching articles, reviews, and material for statistical discussion. In particular, the ChJS considers timely articles organized into the following sections: Theory and methods, computation, simulation, applications and case studies, education and teaching, development, evaluation, review, and validation of statistical software and algorithms, review articles, letters to the editors.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

Contents

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 25–39

Categorical Data
Research Paper

Detecting outliers in $I \times J$ tables through the level of susceptibility

Thodur Parthasarathy Sripriya[1,*], Mamandur Rangaswamy Srinivasan[1], and Meenakshisundaram Subbiah[2]

[1]Department of Statistics, University of Madras, Chennai, India,
[2]CEO, Acroama Gnan Vikas Pvt. Ltd., Chennai, India

(Received: 15 November 2019 · Accepted in final form: 30 April 2020)

Abstract

Detecting outliers in two-way contingency tables is an important and interesting statistical problem. There is no clear objective procedure available in literature to handle outliers in categorical data unlike other data types. Therefore, this study envisages a two-step procedure, to first indicate and then to identify outliers in two-dimensional contingency tables. The approach deals with enhancing the summary measure to indicate the presence of possible outlying cells followed by residual approaches supplemented by boxplot in identifying the outliers. The fundamental definition of outlying cell as "markedly deviant" cell is clearly exploited in this two-step procedure. A simulation study has been carried out to examine the consistency of the proposed methods and later applied to a large collection of real datasets from various applications of social sciences.

Keywords: Boxplot · Contingency tables · Outlying cells · Residuals · Summary measures.

Mathematics Subject Classification: Primary 62H17 · Secondary 97K80.

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 (Barnett and Lewis (1978)). On the other hand, it is possible that an outlier is simply a manifestation of the inherent variability of the data.

Unlike in metric case, there exists no clarity in the definition of outliers for categorical data, as the cells are purely frequencies or counts of a contingency table. Hence, the problem is to first identify a pivotal element and then markedly deviated cells are detected as outliers. In continuous data, mean or quartiles are considered as pivot and the metric

*Corresponding author. Email: sri.chocho@gmail.com

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)

http://www.soche.cl/chjs

such as $Q_3 \pm 1.5Q_1$, $\mu \pm K\sigma$, etc., are used to identify the outliers (Park et al., 2019; Kim, 2015). However, it is challenging to establish exact criteria for deciding on an observation to be unusual, denoted as an outlier, in contingency tables. Hence, an attempt has been made to provide a set of statistical rules enabling the experimenter to look closely for causes of an outlier to really exist, and then to decide on its plausible acceptability.

The existence of one or two outliers in a sample can badly distort the summary indications and analyses of data. In the detection of outliers in contingency tables, residual based approach has been widely used (Haberman, 1973; Brown, 1974; Simonoff, 1988; Fuchs and Kenett, 1980; Bradu and Hawkins, 1982; Yick and Lee, 1998).

The use of residual approach may cause masking and swamping and a method resistant to it has been studied by Kotze and Hawkins (1984) and Lee and Yick (1999). But, residuals play an important role in detecting outliers in two-way contingency tables, and an extensive review is presented in Kateri (2014). 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. Velez and Marmolejo-Ramos (2017) proposed an extension of a graphical diagnostic test for contingency tables using polygraph. 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 (2018a) and Sripriya and Srinivasan (2018b) have suggested a new approach in the detection of outliers in categorical tables of order $I \times J$, based on Poisson log-linear model. Kuhnt et al. (2014) detected outliers through subsets of cell counts called minimal patterns for the independence model.

The principal interest in the analysis of $I \times J$ contingency tables is to test the independence between the two categories. The Pearson chi-square and the log-likelihood ratio statistics (Agresti (2002)) are the long standing techniques in testing independence under multinomial set up. Literature is abundant to show that the residual test statistic converges approximately to the chi-square distribution (Song, 2007; McCullagh and Nelder, 1989). Following Agresti (2002) and Sangeetha et al. (2014) proposed the reversal pattern of association (RAP) to understand deeply the association between attributes in high dimensional tables. Sripriya and Srinivasan (2018a), Sripriya and Srinivasan (2018b) adapted the RAP to detect the outliers based on chi-square statistic through an iterative algorithm. Indeed, there are many procedures like residuals based approach, pattern based approach, and test based approach which are more heuristic in nature as pointed by Simonoff (2003) leading us to the present study based on the characteristics of the contingency table.

In this paper, an attempt has been made to explain the fundamental meaning of "markedly deviant" by answering; which cell, from where and, by how much. To realize the definition, there is a need for a measure which captures the deviation from the pivotal element. Thus, a measure based on the generic characteristics of the table has been considered as a pivotal element for detection of outliers.

The purpose of the present study is to detect possible outliers for a two-way contingency table in a more generic way by a two-step procedure, firstly through an indicator followed by an exact identifier. The first step involves the enhancement of summary measures for categorical data, and a methodical way to indicate susceptibility to outliers by explaining the characterization of contingency tables through three different methods. In step two, potential outliers are detected by using theoretical approach of residuals supported by the boxplot in explaining the deviation of residuals. Lastly, a simulation study has been carried out by contaminating the cell values to determine the stability of the results for detecting outliers through the proposed method.

The paper is organized as follows. In Section 2, we define our two-step procedure and discuss the classification of level of susceptibility. The results of simulation study in Section 3 reveals that the two-step approach performs well in detecting outliers. Section 4 presents few applications to real data in detecting the outliers in two-way contingency tables. Finally, some concluding remarks are given in Section 5.

2. Two-Step Procedure

Let X and Y denote two categorical response variables, X with I categories $R_1, \ldots, R_I$ and Y with J categories $C_1, \ldots, C_J$ leading to IJ possible combinations. When the cells contain frequency (n_{ij}) of outcomes from a sample, the table is called a contingency table, or cross-classification table.

Sparseness in contingency tables often occurs in practice and detecting outliers in the sparse data is a challenging one. The remedial actions for sparseness in categorical data such as collapsibility of cells with small frequencies, or dropping the tables altogether lead to loss of information (Baglivo et al. (1988)). However, this study considers the detection of outlier in $I \times J$ contingency tables without considering the sparseness index but in terms of polarization and its underlying issues.

Further, polarization of cell counts is one of the major problem when it comes to outlier detection. Polarization is basically a highly uneven distribution of counts in $I \times J$ tables. Polarization in contingency tables involves presence of counts/frequencies of disparate in nature, such as 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 $I \times J$ tables, in addition to 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)). The relevance of sparseness on summary measure and the sensitivity of analysis in 2×2 tables have been discussed by Subbiah and Srinivasan (2008).

The prevailing researches on the characteristics of $I \times J$ tables are: Order of k, numerical issues (aberration/zero width intervals ZWI), polarization of cell counts, low cell count, sparseness and computational complexity. However, the present study is concerned with the detection of unusual observations or outliers in contingency table. The two step process considered in this study as follows:

Step 1: Indicator – Identify whether the table contains outlier cells through the level of susceptibility
Step 2: Identifier – Detect the exact outlying cells using boxplot of residuals

---

The detailed two step procedure is as follows:
Step I: Contingency tables are often summarized by its size $I \times J$ $(= k)$ and total frequency $N = \sum_i \sum_j n_{ij}$ (Agresti and Yang (1987)). However, there can be other characteristics of contingency table which can be captured and included in the summary measures, such as

Z_C: Number of zero counts in a $I \times J$ table
P_Z: Proportion of zero counts in a table $= Z_C/k$
L_C: Number of low counts in a table
P_L: Proportion of low counts in a table $= L_C/k$
H_C: Number of high counts in a table
P_H: Proportion of high counts in a table $= H_C/k$
R: Range of the cell counts
T: $T = N/k$
Q: $Q = \text{Range}/k$

The three defined measures T, Q and P (P_Z, P_L, P_H) can be considered as an enhancement of the summary measures apart from k and N and could constitute an important component of contingency tables and in particular to indicate the presence of outliers in a table. In an ideal table, all the observations are expected to be closer to the pivot element and thereby expected values are closer with smaller residuals. Suppose all the k cells are quite closer to T, then one may not suspect outlier(s) to be present, except in the heuristic residual approach. Hence, T can be perceived as an Pivot element, for example, a table with $k = 36$ cells, $N = 366$, and $T = 10.16667$ yields all the cells counts to be pretty closer to T and the expected values are closer to each other. Following Agresti and Yang (1987), the present study considers the classification of P, T and Q for the detection of outliers with Low (L), Moderate (M), and High (H) categories as follows

$$P_Z = \begin{cases} \text{Low}, & 0 \le P_Z \le 0.10; \\ \text{Moderate}, & 0.10 < P_Z \le 0.20; \\ \text{High}, & P_Z > 0.20; \end{cases}$$

$$P_L(n_{ij} < 6) = \begin{cases} \text{Low}, & 0 \le P_L \le 0.20; \\ \text{Moderate}, & 0.20 < P_L \le 0.40; \\ \text{High}, & P_L > 0.40; \end{cases}$$

$$P_H(n_{ij} > T) = P_L(n_{ij} < T) = \begin{cases} \text{Low}, & 0 \le P_H, P_L \le 0.45; \\ \text{Moderate}, & 0.45 < P_H, P_L \le 0.55; \\ \text{High}, & P_H, P_L > 0.55. \end{cases}$$

Similarly, T and Q have been classified as

$$T = \begin{cases} \text{Low}, & 0 \le T \le 20; \\ \text{Moderate}, & 20 < T \le 250; \\ \text{High}, & T > 250; \end{cases}$$

$$Q = \begin{cases} \text{Low}, & 0 \le Q \le 10; \\ \text{Moderate}, & 10 < Q \le 100; \\ \text{High}, & Q > 100. \end{cases}$$

Table 1. Categorization of susceptibility

Susceptibility	(T, P_Z, P_L)	(Q, P_Z, P_L)	(P_Z, P_L, P_H)
High	8	12	12
Moderate	10	9	12
Low	9	6	3

Our study proposed three methods (i) (T, P_Z, P_L) (ii) (Q, P_Z, P_L) and (iii) (P_Z, P_L, P_H) based on the above classification to identify the susceptibility to outliers in $I \times J$ tables. Thus there will be a total of 27 combinations for each method under consideration. Suppose a table with (T, P_Z, P_L) is (L, L, L), then, there will be a less chance of outliers being present and hence denote the $I \times J$ table as of low susceptibility to outliers. Correspondingly, a table with (T, P_Z, P_L) is (H, L, L), then there may be few markedly deviant cells to exist in the table and denoted as highly susceptible to outliers. Similarly, the combination of M and L is taken to be moderately susceptible to outliers. Thus the 27 combinations of L, M, and H are categorized for susceptibility under the three proposed methods and presented in Table 1. The categorization of susceptibility is based on the direction provided by Agresti and Yang (1987), but could be suitably modified based on T, Q, and P

and accordingly susceptibility to outliers will also vary. In the same way, method 2 has been categorized under the three levels; H, M, and L, whereas in the third method, LLL is taken to be highly susceptible to outliers based on the above mentioned classification. Thus, 27 combinations are categorized into three methods as presented in the following table. Consider a 5×5 table constructed by Simonoff (1988) for the detection of outliers. Based on the approach outlined earlier, with $k = 25$, $N = 558$, $T = 22.32$, $Q = 1$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.8$, and $P_H = 0.2$ reveals the table is highly susceptible to outliers. Thus, the study basically affirms the approach to be capable of indicating the presence of outliers. After due classification of $I \times J$ table, the next step is to identify the outlying frequencies in the table.

Step II: Residual techniques have been carried out by many researchers in order to identify the outlying cells in a table by considering "large" residual. But many of them failed to justify "how large" the residual should be considered for an observation as an outlier. The usual residual based methods of outlier detection methods are devoid of contingency table characteristics. In the heuristic approach, outliers are identified irrespective of the polarization of cell frequencies and order of the tables. To overcome this, the box plot of the following three types of residuals has been considered to identify the outlying cell:
(i) [Pearson residual]

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}, \quad e_{ij} = (n_{i+} \times n_{+j})/N.$$

(ii) [Adjusted residual; Haberman (1973)]

$$\tilde{r}_{ij} = \frac{r_{ij}}{AF}, \quad \text{AF} = (1 - n_{i+}/N)(1 - n_{+j}/N).$$

(iii) [Deleted residual; Simonoff (1988)]

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}, \quad e_{ij} = (n_{i+} - n_{ij})(n_{+j} - n_{ij})/(N - n_{i+} - n_{+j} + n_{ij}).$$

Thus, the two step process provides a systematic approach of identifying outliers under conditions of polarity for varying order of k. The following section deals with examining the robustness of susceptibility criteria as envisaged through a simulation study.

3. Simulation Study

Simulating a two way contingency table situation can be achieved using varying combinations of its total frequency, levels in each of the categories, cell probabilities, and the test statistic used to analyze the independence. Thus, the present study considers two scenarios of generating $I \times J$ tables where the cell entries are from (i) bi-variate normal distribution with the assumption of independence, and (ii) multinomial distribution as in Agresti (2002) since it models the probability of counts in each categories for n independent trails.

Bivariate normal distribution The simulation study starts with generating the entries of $I \times J$ table from bi-variate normal distribution with different correlation structures. In this scenario, the study considered correlation ρ and the size of the table $k(= I \times J)$ as the potential parameters. Here, we consider four different values of k, (9, 16, 25, and 100) with five different correlation structures (0.5, 0.6, 0.7, 0.8, and 0.9) to evaluate the performance

of proposed susceptibility methods by contaminating each cell at a time with a constant α $(=0.5, 1, 1.5, 2)$ and repeated 500 times. The results of this simulation are summarized in Table 2. The following are the observations based on the simulation presented in Table 2:

(i) The pattern of susceptibility level remains unchanged for $k = 9, 16$ and with changes in $k = 25$ and 100 irrespective of ρ and α in methods 1-2.
(ii) When k increases, the susceptibility level increases only when the correlation is 0.5 when $\alpha = 0.5$. However, it shows few fluctuations due to outliers in other correlation structures with different α considered.
(iii) Susceptibility level fluctuates largely for all k in all methods irrespective of ρ and α.
(iv) As α increases, the susceptibility level shows similar pattern for all order of k with $\rho = 0.5, 0.6, 0.8$, and 0.9. However, fluctuations are visible between the contamination α for all k with $\rho = 0.7$.
(v) The variability in the susceptibility is largely observed from method 3 as it gives poor results for all k irrespective of correlation structure and α.

Table 2. Susceptibility to outliers (in %) for scenario 1

Order		Method 1 (T, P_Z, P_L)				Method 2 (Q, P_Z, P_L)				Method 3 (P_Z, P_L, P_H)			
$I \times J$	ρ	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	0.5	76	75	73	72	78	77	75	74	45	44	42	39
	0.6	74	74	73	73	76	75	74	73	73	72	71	70
	0.7	75	52	60	54	79	53	59	52	68	53	52	51
	0.8	66	58	47	41	63	49	45	44	64	57	52	41
	0.9	78	65	51	49	74	62	53	47	72	63	51	50
4×4	0.5	77	74	72	70	75	74	72	71	47	45	40	38
	0.6	76	73	70	69	74	72	71	68	71	69	68	65
	0.7	77	55	62	53	75	51	59	52	69	51	49	49
	0.8	60	52	40	40	62	54	44	43	60	58	51	39
	0.9	76	60	55	52	72	61	52	50	70	64	55	54
5×5	0.5	78	74	71	70	76	71	69	68	49	46	41	37
	0.6	59	56	51	47	62	57	53	48	60	55	51	49
	0.7	72	54	56	52	74	51	58	49	65	50	45	43
	0.8	56	42	38	34	53	45	41	37	50	49	42	36
	0.9	66	50	45	42	62	51	42	40	60	54	49	44
10×10	0.5	81	79	76	71	79	74	71	69	51	49	47	46
	0.6	67	64	60	53	66	61	57	50	63	52	51	50
	0.7	75	57	63	55	72	53	65	51	69	54	49	42
	0.8	61	58	47	44	59	47	45	40	60	57	51	44
	0.9	76	64	54	51	69	57	54	48	67	55	47	46

Following susceptibility, Table 3 presents the results of the simulation involving the identification of outliers based on three residual methods under different levels of contamination. The following are the observations based on simulation presented in Table 3:

(i) The identification of exact outlying cell for all k shows similar trend irrespective of α and ρ in all the three residuals considered in this simulation scenario.
(ii) As α increases, the identification level also increases for all k irrespective of the correlation structure ρ.

(iii) Stability of level of identifying the outlier cell increases as ρ increases for $k = 9, 16, 25$. However, for $k = 100$, yields poorer results for all the three residual approaches.

Table 3. Identification of outliers (in %)

		Pearson				Adjusted				Deleted			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	0.5	33.3	34	34.3	34.5	35	35.5	36	36	34	35.5	36.3	36.5
	0.6	36	36.5	36.7	37	36.5	37	37.7	38	37	37.5	38	38.3
	0.7	36.2	37	37.5	37.7	36	37	37.7	38.2	38.2	38.7	39.2	39.7
	0.8	37	37.7	38.2	38.5	36.7	37.5	38	39	37.7	38.5	39.7	40
	0.9	38.2	38.5	39	39	38	38.7	39	39.7	38	38.7	39.5	40
4×4	0.5	43	44.5	45	45.5	45	45	46	46	46	46.7	47	47.5
	0.6	45	46	46.5	47.5	47	47.7	48	48.5	47	48.5	48.7	49
	0.7	46.5	47.2	48.5	48.7	46	47.7	48.2	48.7	48.5	49	49.5	49.5
	0.8	46.7	46.7	47.2	48	47.7	48	48.5	49	48.7	49	49.2	49.7
	0.9	48	48.7	49.2	49.5	48	48	49.5	49.7	48	48.5	49	50
5×5	0.5	44.4	45	45.4	46.5	45	45.5	46	46	45	45.5	46.4	46.5
	0.6	46	46.5	46.7	47	46.5	47	47.7	48	47	47.5	48	48.4
	0.7	46.2	47	47.5	47.7	46	47	47.7	48.2	48.2	48.7	49.2	49.7
	0.8	47	47.7	48.2	48.5	46.7	47.5	48	49	47.7	48.5	49	49.5
	0.9	48.2	48.5	49	49	48	48.7	49	49.7	48	48.7	49.5	49.7
10×10	0.5	25	25.5	26.2	26.5	25	25.5	26	26	24	25.5	26.2	26.5
	0.6	26	26.5	26.2	22	26.5	22	22.2	28	22	22.5	28	28.2
	0.7	26.2	22	22.5	22.2	26	22	22.2	28.2	28.2	28.2	29.2	29.2
	0.8	27	27.2	28.2	28.5	26.2	26.5	28	29	22.2	28.5	29	29.2
	0.9	28.2	28.5	29	29	28	28.2	29	29.2	28	28.2	29.5	30.5

The associations between the two categorical variables are identified generally using the chi-square distribution. Here, the p-value of the chi-square distribution is used to identify the independence of the two categorical outcomes and found that there is no change in the independence assumption even after contaminating the cell entries. Moreover, the data generation process in simulation in no way alters the independence assumption. The percentage of identification of outliers in this scenario yield poor results since the data generated from bi-variate normal distribution with the parameter lambda where lambda is the parameter used to change the continuous bi-variate normal random variables to count variables. Thus, a more appropriate data generation rule using multinomial distribution is considered and is explained below.

Multinomial distribution The simulation study considers two potential parameters k; the size and N; the total frequency of the table and $X_1, X_2, \ldots, X_k \sim$ Multinomial$(N, (p_1, \ldots, p_k))$ where the probability $p_i \sim U(0,1); i = 1, \ldots, k$. The probability range between 0 and 1 is automatically maintained in multinom function in R. The study of over 100 real time datasets from various fields of social sciences has shown that polarization is largely observed in tables of order more than 4 and larger tables $(I, J > 10)$ occurs occasionally and are not discussed in the simulation study. Hence our simulation study is restricted to $k = 9, 16, 20$ and 56 with $N = 50, 350, 950, 2150$, and 4550 providing a varied cross section of the contingency table to examine the susceptibility to outliers. The process starts by contaminating the cell frequencies with alpha (α) for each cell at a time and then covering the entire table k times. Four different level of contamination α

$(=0.5, 1, 1.5, 2)$ are considered and repeated 500 times. The results of simulation based on the above procedure are summarized in Table 4.

The following are the observations based on the simulation presented in Table 4:

(i) Susceptibility level remains unchanged for $k = 9, 16$ and minor fluctuations in $k = 20$ and 56 irrespective of N and α in method 1.
(ii) When k increases, irrespective of α, there exists small changes due to outliers in method 2 for moderate N of size 350 and 950.
(iii) Susceptibility level fluctuate largely for all k except for a lower order of k $(=9)$, in method 3 irrespective of N and α.
(iv) As α increases, the level of susceptibility remains constant for all order of k and for small and large values of N under method 1. However, fluctuations are visible for moderate values of N and higher order of k.
(v) Susceptibility level remains constant as α increases for all k and for large values of N under method 2. However, fluctuations are visible for low and moderate values of N irrespective of k.
(vi) In method 3, as α increases, the susceptibility level remains constant for a small order of k and moderate to large N and the instability in susceptibility are observed from rest of k and N.

Table 4. Susceptibility to outliers (in %)

Order		Method 1 (T, P_Z, P_L)				Method 2 (Q, P_Z, P_L)				Method 3 (P_Z, P_L, P_H)			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	50	98	92.6	92.8	87.2	63.2	62.2	62.2	61.6	71.2	71.2	71.2	71
	350	100	100	100	100	100	100	100	100	100	100	100	100
	950	100	98.6	98.6	98	100	100	100	100	100	100	100	100
	2150	100	100	100	100	100	100	100	100	100	100	100	100
	4550	100	100	100	100	100	100	100	100	100	100	100	100
4×4	50	100	100	100	100	69.8	69.8	69.8	69.8	50.2	50.2	48.6	49
	350	99.4	95.4	95.4	95.4	99.4	95.4	95.4	95.4	90.2	89.6	89.6	80
	950	100	100	100	100	100	100	100	100	100	100	100	100
	2150	100	100	100	100	100	100	100	100	99.4	99.4	99.4	99
	4550	100	100	100	100	100	100	100	100	100	100	100	100
5×4	50	100	100	100	100	86.4	86.4	86.4	86.4	65.2	55.4	55.4	55
	350	85.6	77	77	70.6	86.2	79.4	77.6	71.2	87	64.6	59.4	54
	950	96.8	94.8	94.8	94.6	96.8	94.8	94.8	94.8	69.8	69.8	64.8	64
	2150	100	100	100	100	100	100	100	100	95	90.6	88.2	85
	4550	100	100	100	100	100	100	100	100	100	100	100	100
7×8	50	100	100	100	100	100	100	100	100	100	100	99.2	99
	350	91	80	80	80	99.4	99.4	99.4	99.4	91.4	99	93.2	93
	950	97.6	97.6	89.2	97.6	97.6	97.6	97.6	97.6	55.8	55.8	55.8	52
	2150	100	100	100	100	100	100	100	100	94	94	90.8	91
	4550	100	100	100	100	100	100	100	100	87.4	87.4	87.4	87

As outlined in Section 2, following susceptibility, next step involves identification of outliers based on three residual methods under different levels of contamination. The results of the simulation are presented in Table 5.

The following are the observations based on simulation presented in Table 5:

(i) Identification of exact outlying cell remains same for all k irrespective of α and N and a few fluctuations are observed in moderate to high N in Pearson

and Adjusted residual approach whereas in the case of Deleted residual, the simulation yields inconclusive results.

(ii) As α increases, the identification level decreases for all k and it remains constant when N varies from moderate to high in Pearson and Adjusted residual approach whereas in Deleted residual approach, the identification level decreases as α increases for all k except for $k = 16$ irrespective of N.

(iii) Stability of level of identifying the outlier cell oscillates as N increases irrespective of k and α for all the three residual approaches.

Table 5. Identification of outliers (in %)

		Pearson				Adjusted				Deleted			
$I \times J$	N	α											
		0.5	1	1.5	2	0.5	1	1.5	2	0.5	1	1.5	2
3×3	50	95.8	93.8	91.4	86	96.4	93.8	92	89.2	83	72	92	54.4
	350	92.6	89.6	87.5	85	94.4	89.4	86.3	84.2	99.4	96.4	94.2	91.6
	950	99.2	99	99	99	99.3	99	99	99	91.2	92.3	90	90
	2150	100	100	100	100	100	100	100	100	97	96	95	95
	4550	95	94.9	100	95	97	98	100	98	93	90	100	91
4×4	50	94.8	92.8	91	88	95.8	92.8	92	89	90	87	70	69
	350	93	93	89	89	94	92	90	88	89	88	89	80
	950	98.8	99	99	99	99	99	99	99	92	92.3	90	91
	2150	99	99	99	99	99	99	99	99	98	97	95	96
	4550	96	94	93	91	97	95	92	90	100	100	96	99
5×4	50	96	94	90	87	95	93	89	89	92	87	83	86
	350	93	93	90	87	94	92	91	88	89	86	85	80
	950	100	100	100	100	100	100	100	100	89	86	82	78
	2150	92	90	91	92	93	92	93	92	83	79	77	75
	4550	93	92	94	90	94	93	95	91	90	91	89	89
7×8	50	94	93	91	89	95	94	92	92	89	82	77	76
	350	93	93	89	90	94	92	90	91	89	88	85	83
	950	95	94	90	89	96	95	91	90	88	85	83	78
	2150	100	100	100	100	100	100	100	100	89	87	83	72
	4550	92	100	97	96	93	95	98	97	84	86	88	78

In summary, even though the level of susceptibility fluctuate in few cases in all the methods, the identification level of exact outlying cells in all the residual approaches show that our two-step procedure could be a best alternative in the detection of outliers in $I \times J$ tables. The results based on the simulation study have paved the way to examine the application of two-step process of detection of outliers in contingency table to real time datasets.

4. Data Analysis

In this section, we illustrate our two-step procedure to six datasets from literature by assuming the nature of the data as nominal. Kotze and Hawkins (1984) considered a dataset with $k = 196$, $N = 775$ and identified 15 most outlying cells by adding 0.5 to zero cells using elimination method. The mosaic display of the data is presented in Figure 1.

The present approach, with $T = 3.95$, $Q = 0.27$, $P_Z = 0.26$, $P_L(n_{ij} < 6) = 0.52$, $P_L(n_{ij} < T) = 0.39$, and$P_H = 0.35$, shows low susceptibility in method 1 and 2 and high susceptibility in method 3. Also, boxplot for residuals as presented in Figure 2 identified

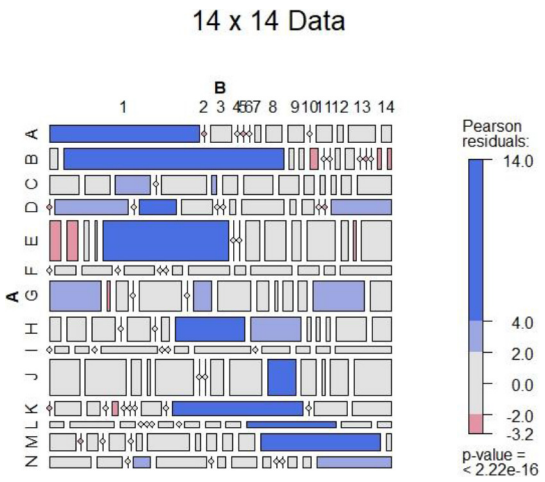

Figure 1. Mosaic Plot for 14×14 data

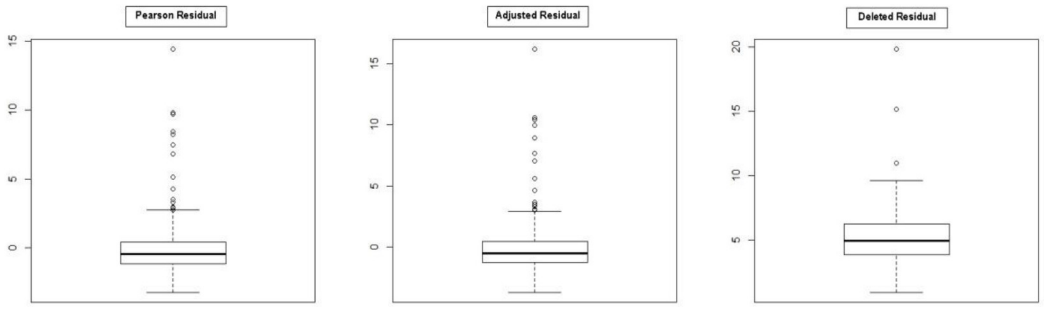

Figure 2. Boxplots for Kotze and Hawkins Data

the same 14 cells as possible outliers in the case of Pearson and Adjusted residuals and only 3 cells in the case of Deleted residuals.

Yick and Lee (1998) considered the archaeological data and artificial data by Simonoff (1988) in identifying outliers. For the artificial 5×5 data, three cells (2, 1), (1, 2) and (1, 3) are identified as outliers and the cell (1, 1) being swamped in the perturbation approach. In our method, with $k = 25$, $N = 558$, $T = 22.32$, $Q = 1$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.8$, and $P_H = 0.2$, this dataset is found to be moderately susceptible to outliers and the residual boxplot identifies exactly the same cells as outliers as in perturbation approach.

For the archeological data, the perturbation approach identified three cells (2, 3), (11, 5) and (18, 1) as outliers out of which two cells have extreme residuals and these two extreme cells are identified correctly in our two step procedure with $k = 114$, $N = 3297$, $T = 28.92$, $Q = 3.42$, $P_Z = 0.07$, $P_L(n_{ij} < 6) = 0.21$, $P_L(n_{ij} < T) = 0.65$, $P_H = 0.72$ and the method show that the data is moderately susceptible to outliers. The mosaic display and boxplot of residuals for these two data is presented in Figures 3, 4 and 5.

Yick and Lee (1998) considered the 7×8 student enrolment data from seven community schools from Northern Territory, Australia and identified the cells (1, 5), (1, 6), (2, 4) and (2, 5) as potential outliers using perturbation diagnostics. The mosaic display of the data is presented in Figure 6.

In our proposed method, the datasets is highly susceptible to outliers with $k = 56$, $N = 5248$, $T = 93.71$, $Q = 2.9$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.625$, $P_H = 1$ and identified the cells (2, 4) and (1, 6) as potential outliers using boxplot of all the residuals and boxplot are presented in Figure 7.

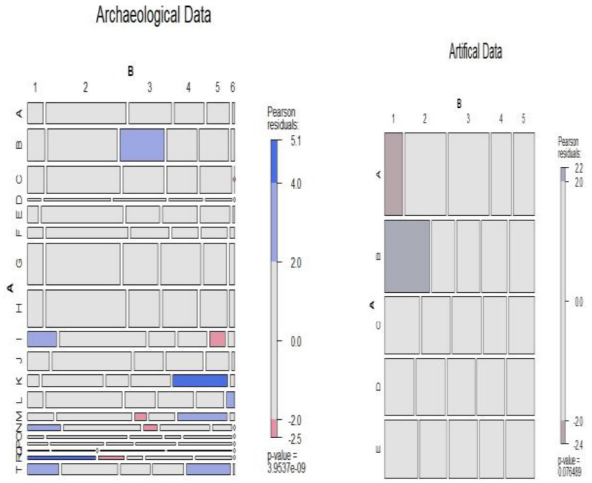

Figure 3. Mosaic Plot for Archaeological and Artificial data

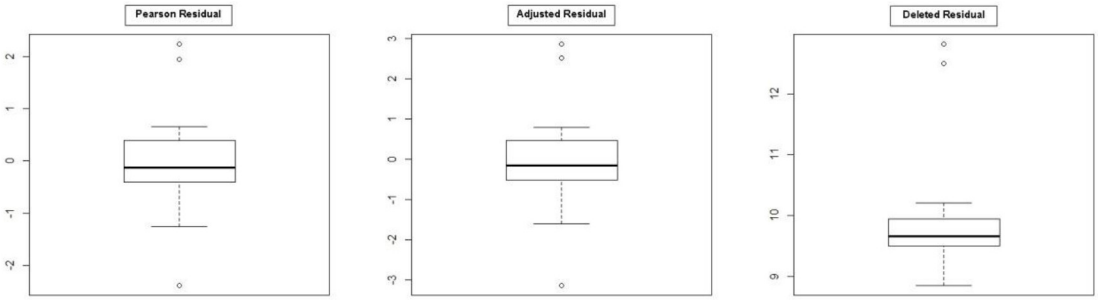

Figure 4. Boxplots for Artificial Data

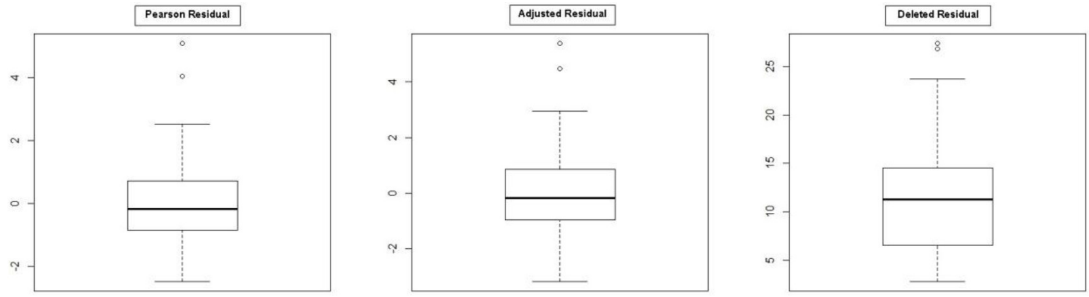

Figure 5. Boxplots for archeological data

Kuhnt et al. (2014) considered 3×3 table of social mobility in Britain and 4×4 table of artifacts in Nevada and detected outliers using three different algorithms. For the social mobility data, all the three algorithms doesn't give satisfactory results and detected, (i) all the cell counts, (ii) only diagonal cells and (iii) cells (1, 1), (3, 1), (1, 3) and (3, 3) as outliers, whereas in our method the table shows highly susceptible to outliers with $k = 9$, $N = 3494$, $T = 366.33$, $Q = 67$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0$, $P_L(n_{ij} < T) = 0.44$, $P_H = 0.56$, and detected the cells (1, 1), (3, 1) and (2, 2) as outliers with the help of boxplot of residuals and the mosaic display is presented in Figure 8.

For the Artifacts in Nevada data, the author identified two cells as outliers but our methods gave inconclusive decision in susceptibility with $k = 16$, $N = 164$, $T = 10.25$, $Q = 3.77$, $P_Z = 0$, $P_L(n_{ij} < 6) = 0.43$, $P_L(n_{ij} < T) = 0.68$, $P_H = 0.32$, and no outliers are detected using boxplot of residuals. The boxplot for these datasets are presented in Figure 9 and 10.

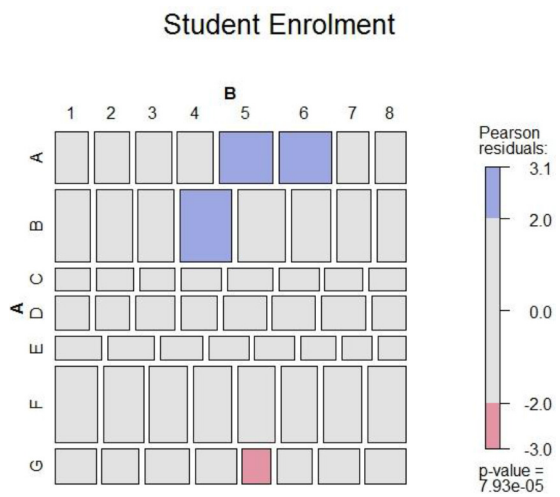

Figure 6. Mosaic Plot for Student Enrolment data

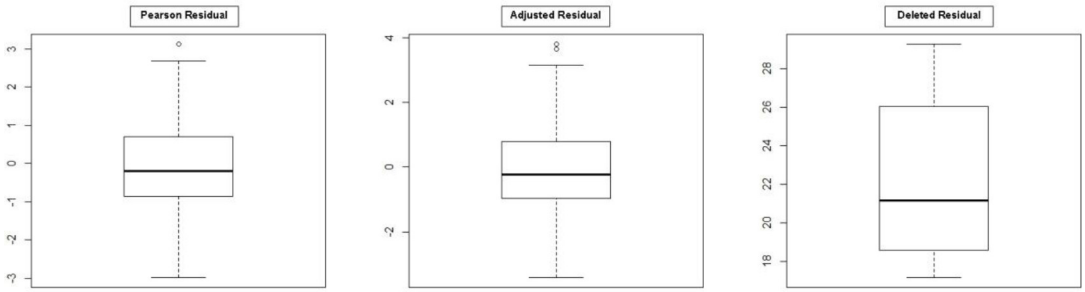

Figure 7. Boxplots for Student Enrolment Data

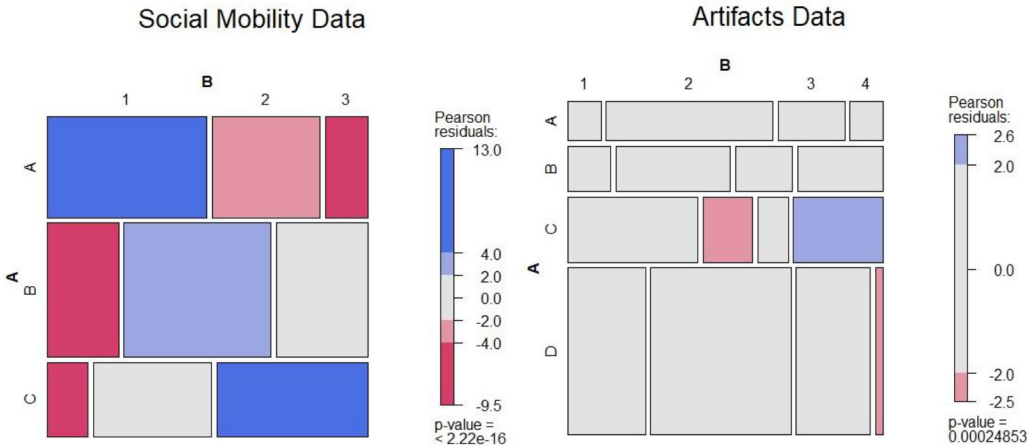

Figure 8. Mosaic Plot for Social Mobility and Artifacts data

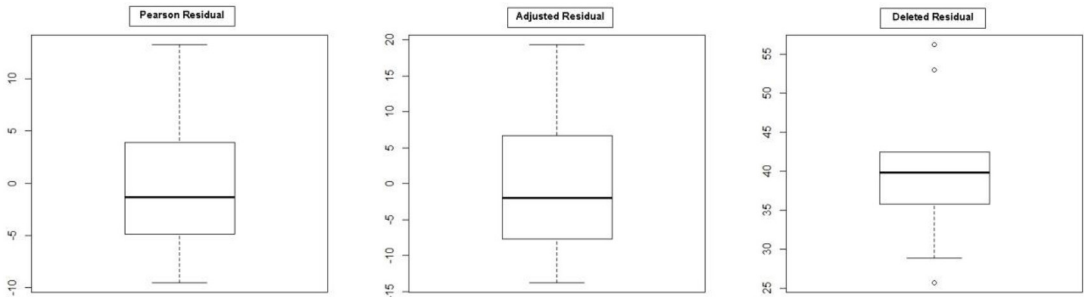

Figure 9. Boxplots for social mobility data

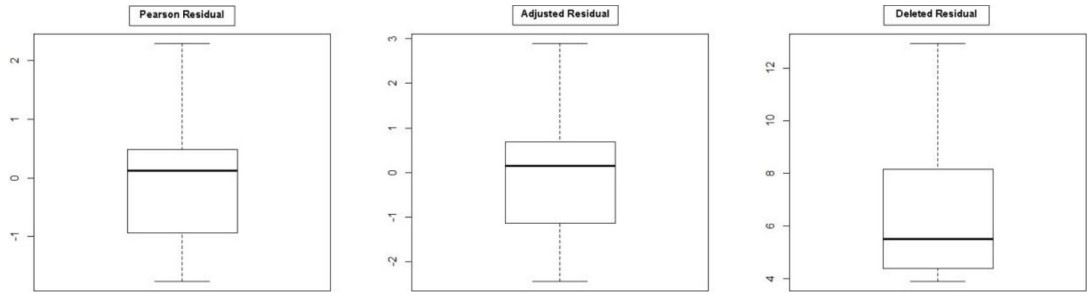

Figure 10. Boxplots for Artifacts in Nevada data

In addition, our study considered 50 other datasets of varied characteristics ranging from $k = 6$ to 196 cells based on the literature to identify the feasibility of our methods and the results are presented in Table 6. Most of the researchers nullified the zero cells in a table by adding constants, but our two-step method helps to identify the outlying cells even in the presence of zero cells in a table.

Table 6. Identification of outlying cells through Boxplot

	(T, P_Z, P_L)			(Q, P_Z, P_L)			(P_Z, P_L, P_H)		
	T	I	NI	T	I	NI	T	I	NI
Highly susceptible	25	25(100%)	–	27	25(92.5%)	2	41	38(92.6%)	3
Moderately susceptible	7	6(85.7%)	1	9	8(88.8%)	1	9	2(22.2%)	7
Low susceptible	18	9(50)	9	14	7(50)	7	–	–	–

T–Total; I–Identified; NI–Not-Identified

The above table clearly shows that method 1 performs better in highly susceptible category and method 2 performs better in moderately susceptible category, method 1 & 2 equally performs better in low susceptible category. The classification of datasets under method 2 also contains the datasets under method 1. On the whole, method 3 appears to be more stringent in identifying outliers since it classifies almost all datasets as highly susceptible to outliers.

5. Conclusions

The problem of identification of outliers in $I \times J$ contingency tables has been examined through the ambiguous notion of "markedly deviant" nature of cells from which the other cell values deviate greatly. However, in this paper a simple measure T has been introduced as a pivotal element to explain the deviation of other cells in the table. In this direction, a two-step procedure is devised to first examine the nature of the table through susceptibility followed by identification of outliers through box plot techniques. The stability of our proposed methods towards the identification of outliers is examined through a simulation study. The results have revealed that methods (T, P_Z, P_L) and (Q, P_Z, P_L) are found to be more consistent based on two simulation scenarios. Moreover, it is evident from the results that a triplet with the pivot element along with proportion of zero and low counts provide an idea 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 approach as a combination of summary measures and boxplot for residuals could be a feasible approach to identify outlier cells in contingency table. However, as pointed out in the earlier section, a judicious choice is necessary in some cases of ambiguity. Further, even if the boxplot or the residual approach fails in some cases, summary measure will indicate clearly whether the table contains high, moderate, or low outlying cells. The practicality of two pronged approach has been well corroborated by an extensive amount of data sets for its efficacy and its usefulness in identifying outlying cells.

Acknowledgement

The authors wish to thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript.

References

Agresti, A. and Yang, M.C., 1987. An empirical investigation of some effects of sparseness in contingency tables. Computational Statistics and Data Analysis, 5, 9–21.

Agresti, A., 2002. Categorical Data Analysis. Wiley, New York.

Baglivo, J. Olivier, D. and Pagano, M., 1988. Methods for the analysis of contingency tables with data, large and small cell counts. Journal of the American Statistical Association, 83, 1006–1013.

Barnett, V.D. and Lewis, T., 1978. Outliers in Statistical Data. Wiley, New York.

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 C, 23, 405–413.

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

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, 75, 395–398.

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

Kateri, M., 2014. Contingency Table Analysis. Springer, New York.

Kim, S.S., 2015. Variable selection and outlier detection for automated K-means clustering. Communications for Statistical Applications and Methods, 22, 55–67.

Kotze, T.J.W. and Hawkins, D.M., 1984. The identification of outliers in two-way contingency tables using 2 x 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. Statistics and Computing, 24, 481–491.

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

McCullagh P. and Nelder, J., 1989. Generalized Linear Models. Chapman and Hall/CRC, Boca Raton, FL, US.

Park, J.S. Park, C.G. and Lee, K.E., 2019. Simultaneous outlier detection and variable selection via difference-based regression model and stochastic search variable selection. Communications for Statistical Applications and Methods, 26, 149-161.

Rapallo, F., 2012. Outliers and patterns of outliers in contingency tables with algebraic statistics. Scandinavian Journal of Statistics, 39, 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, New York.

Song, P.X.K., 2007. Correlated Data Analysis: Modeling, Analytics, and Applications. Springer, New York

Sripriya, T.P. and 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, 2018, 35-43.

Sripriya, T.P. and Srinivasan, M.R., 2018b. Detection of outlying cells in two-way contingency tables. Statistics and Applications, 16,103–113.

Subbiah, M. and Srinivasan, M.R., 2008. Classification of 2 x 2 sparse data with zero cells. Statistics and Probability Letters, 78, 3212–3215.

Velez, J.I. and Marmolejo-Ramos, F., 2017. Extension of a graphical diagnostic test for contingency tables. Chilean Journal of Statistics, 8, 53-65.

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

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF file of the manuscript in a free format to Editors of the ChJS (E-mail: `chilean.journal.of.statistics@gmail.com`). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at `http://chjs.mat.utfsm.cl/files/ChJS.zip`. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at `http://www.ams.org/mathscinet/msc/`. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright