On testing dependency for data in multidimensional contingency tables

Dominika Polko[1]

Abstract

Multidimensional data analysis has a very important place in statistical research. The paper considers the problem of testing dependency for data in multidimensional contingency tables, where variables are measured with ordinal scale. Permutation tests were used to verify hypothesis of significance of dependency between variables. Significance of correlation coefficients was studied to determine cumulative influence of several variables on dependent variable in contingency table or determine dependency between variables where influence of other variables is eliminated. Included empirical example contains data from Social Diagnosis.

Keywords: *contingency tables, permutation tests, dependencies analysis, multidimensional data, Monte Carlo study*

JEL Classification: C49

1. Introduction

In economics also in other researches it is often important to test dependency. The purpose of such researches is to assess relationship among set of variables. Choosing an appropriate method depends on the type of variables. When the data are quantitative commonly used measure is Pearson's correlation coefficient. Spearman's rank correlation coefficient or Kendall's rank correlation coefficient is applied when population is defined by ordinal variables and measures based on chi-square coefficient when variables are nominal.

Discovering relations between characteristics of populations is one of the most important areas in statistical data analysis. When data are qualitative (nominal or ordinal) frequent method of analysis is tabular. Often data composed of two or more economical, social, medical, psychological and biological variables can be presented in the form of contingency tables. The paper considers the problem of testing dependency for data in multidimensional contingency tables, where variables are measured with ordinal scale. Permutation tests were used to verify hypothesis of significance of dependency between variables. Proposed method can be applied irrespective of distribution of tested variables even when the sample size is small. Significance of correlation coefficients was studied to determine cumulative influence of several variables on dependent variable in contingency

[1] Corresponding author: University of Economics in Katowice, Department of Statistics, 1-go Maja 50, 40-287 Katowice, Poland, e-mail: dpolko@gmail.com.

table or determine dependency between variables where influence of other variables is eliminated. All calculations were performed in R program.

2. Testing dependency in contingency table with ordinal variables

Variables having categories have two types of scales. Categorical variables without a natural ordering are called nominal. For nominal variables the order is irrelevant. However, many categorical variables have ordered categories. Such variables are called ordinal. Ordinal variables have ordered categories, but distances between categories are unknown. The methods of analysis of nominal and ordinal variables also apply to interval variables having a small number of distinct values for which the values are grouped into ordered categories (Agresti, 2002). Categorical data are usually described in contingency tables. There are many publications concerning the analysis of categorical data for which the row and column variables are ordinal measurements. If such data are analysed by using a method for the nominal data, then ordering information is ignored (Gautam, 2002). Statisticians increasingly have recognized that many profits can achieve from using methods with ordering among categories in contingency tables (Basso et al., 2009).

In statistics, ordinal association is the association or relationship between two ordinal variables *X* and *Y* presented for example in contingency table. There are several measures such as Kendall's *tau*, Goodman and Kruskal's *gamma* or Sommers *d* which can test for significance in association two ordinal variables. These measures are based on classifying each pair of subjects as concordant or discordant. A pair is concordant if the subject ranked higher on *X* also ranks higher on *Y*. The pair is discordant if the subject ranking higher on *X* ranks lower on *Y*. The pair is tied if the subjects have the same classification on *X* and/or *Y* (Agresti, 2002). The following measures for testing significance in association are most often given in literature (Górniak and Wachnicki, 2000):

- Goodman and Kruskal's *gamma*:

$$\gamma = \frac{N_c - N_d}{N_c + N_d}, \tag{1}$$

- Kendall's *tau-a*:

$$\tau_a = \frac{N_c - N_d}{\frac{N(N-1)}{2}}, \tag{2}$$

- Kendall's *tau-b*:

$$\tau_b = \frac{N_c - N_d}{\sqrt{\frac{1}{2}N(N-1) - L_x}\sqrt{\frac{1}{2}N(N-1) - L_y}}, \tag{3}$$

- Kendall's and Stuart's *tau-c*:

$$\tau_c = \frac{N_c - N_d}{\frac{1}{2}N^2\left(\frac{m-1}{m}\right)} = \frac{2(N_c - N_d)}{N^2\left(\frac{m-1}{m}\right)} \tag{4}$$

where N – number of pairs, N_c – number of concordant pairs $N_c = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[n_{ij}\cdot\sum_{k=i+1}^{r}\sum_{m=j+1}^{c}n_{km}\right]$, N_d – number of discordant pairs $N_d = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[n_{ij}\cdot\sum_{k=i+1}^{r}\sum_{m=1}^{j-1}n_{km}\right]$, L_x – correction factor for the number of tied pairs for the variable X $L_x = \sum_{i=1}^{r}\frac{1}{2}n_{i\bullet}(n_{i\bullet}-1)$, L_y – correction factor for the number of tied pairs for the variable Y $L_Y = \sum_{j=1}^{c}\frac{1}{2}n_{\bullet j}(n_{\bullet j}-1)$, m – min(number of rows; number of columns).

For the details on other tests statistics for ordinal data, see Agresti (2002).

3. The measurement of partial and multiple association for ordinal data

Data collection settings often involve collecting information on multiple attributes of each object in the study (Oja and Randles, 2004). If data is characterized by more than two variables, a partial or multiple correlation coefficient measures the relationship between variables. In the case when variables are quantitative commonly used measures are partial and multiple Pearson's correlation coefficients. This correlation coefficients measure the strength of dependency. If multiple and partial correlation are studied together, a very useful analysis of the relationship between the different variables is possible. When tested data is ordinal test statistics that use the ordinality by considering ordinal variables as quantitative rather than qualitative (nominal scale) are usually more appropriate and provide greater power (Agresti, 2002). Nonparametric test of total independence based on Kendall's *tau* for multidimensional data was considered by Simon (1977). In order to determine partial and multiple dependencies Kendall's *tau* coefficient can be used as well.

Let consider the partial association which refers to the study of whether an association or correlation of variable X with variable Y is really due to the associations of each with a third

variable *Z*. Measurement of partial association usually involves studying relationship between two variables after eliminating the influence of one or more independent variables from both of them (see for example, Kendall, 1955). The partial association, for example between *X* and *Y*, when the variation of *Z* is eliminated, will be studied, in situations where the statistical analysis is to be based on the rank orders of the observed values of variables *X*, of *Y*, and of *Z*. To measure the association for data (three ordinal variables) partial correlation Kendall's *tau* coefficient based on (3) can be use:

$$\tau_{XY.Z} = \frac{\tau_{XY} - \tau_{XZ}\tau_{YZ}}{\sqrt{\left(1-\tau_{XZ}^2\right)\left(1-\tau_{YZ}^2\right)}}. \tag{5}$$

Kendall's partial *tau* calculates the strength of dependency between a pair of variables. The partial *tau* is a rank statistic and varies between -1 and +1.

On the other hand the problem of testing dependency to investigate cumulative influence of several variables on dependent variable in contingency table was considered too. The multiple association, for example between variable *Z* and conformation of the rest of variables *X* and *Y*, will be studied. To measure the association for data (three ordinal variables) multiple correlation Kendall's *tau* coefficient based on (3) can be use:

$$\tau_{Z.XY} = \sqrt{\frac{\tau_{ZX}^2 + \tau_{ZY}^2 - 2\tau_{ZX}\tau_{ZY}\tau_{XY}}{1-\tau_{XY}^2}}. \tag{6}$$

The multiple correlation coefficient takes value between 0 and +1. The partial and multiple *tau* are based on the simple pairwise Kendall's *tau-b* rank correlation coefficients ($\tau_{XY}, \tau_{XZ}, \tau_{YZ}$). Partial and multiple Kendall's *tau* are the nonparametric analogues of the usual Pearson's correlation coefficients. Presented formulas enable testing significance of correlation coefficients in three–dimensional contingency tables.

4. Permutation test of partial and multiple association in multidimensional contingency table

When consider data stored in multidimensional contingency table, due to the lack of information on the theoretical distribution of considered statistics, in order to compare the results a permutation test was used. Often for non–parametric methods tests are based on the permutation distribution. They are used where the assumption about conformity with normal distribution is not fulfilled. Permutation tests could be used in practice because of flexibility of the test statistic and minimal assumptions (Butar and Park, 2008). Permutation tests can be applied to all kind of data and to values of normal or non-normal density. Permutation

methods can be applied whenever parametric statistical methods fail (Good, 1994). The major drawback is that they can be computationally intensive and at times practically impossible to use without using a computer. However, nowadays, with the availability of fast computers, permutation tests are quite easy to carry out, and they are quite useful in many problems (Tran et al., 2014).

It is consider data characterized by more than two variables in a multidimensional contingency table. In a three–dimensional table, one of the variables is designated as the row variable, a second variable is designated as the column variable, and the third variable is designated as the layer variable. Zar (1999) employs the term *tier variable* to designate the third variable. Such contingency table can be represented by cube (see Polko, 2014). In a three–dimensional table there is a total of $r \times c \times l$ cells, where r represents the number of row categories, c the number of column categories, and l the number of layer categories (see Sheskin, 2003). Data presented in three–dimensional contingency table can be also written in three columns (see Table 1).

X	***Y***	***Z***
x_1	y_1	z_1
…	…	…
x_r	y_c	z_1
x_1	y_1	z_2
…	…	…
…	…	…
x_r	y_c	z_l

Table 1. The form of the data from contingency table.

In three-dimensional contingency table the issue of studying dependency to eliminate influence of distorting variable can be enrolled with hypotheses:

H_0: There is no dependency between X and Y variables where influence of Z variable has been eliminated.

H_1: There is dependency between X and Y variables where influence of Z variable has been eliminated.

In the case of permutation test of partial correlation, the value of test statistics T_0 on the basis of (5) for the sample data has been calculated, then N permutations of variables X and Y were performed and values of statistics T_i (i =1, 2, …, N) were determined. The decision

concerning a verified hypothesis is made on the basis of *ASL* (*achieving significance level*) value (see Efron and Tibshirani, 1994):

$$ASL = P\left(|T| \geq \left|T_0\right|\right). \tag{7}$$

ASL value is unknown and its evaluation is determined by empirical distribution of *T* statistic:

$$ASL \approx \frac{card\{\, i : |T_i| \geq |T_0|\,)}{N}. \tag{8}$$

ASL is another name for *p-value*. An example is for a two-tailed, alternative hypothesis, where H_0 denotes the null hypothesis, T_0 is the observed value of the test statistic (5) based on (3), and T_i is the random variable corresponding to the test statistic.

A random permutation of data can be obtained by recreating the original data (Fig. 1).

X	*Y*	*Z*
x_1	y_1	z_1
…	…	…
x_r	y_c	z_1
x_1	y_1	z_2
…	…	…
…	…	…
x_r	y_c	z_l

X	*Y*	*Z*
x_3	y_2	z_1
…	…	…
x_r	y_1	z_1
x_6	y_1	z_2
…	…	…
…	…	…
x_1	y_c	z_l

X	*Y*	*Z*
x_1	y_1	z_1
…	…	…
x_5	y_3	z_1
x_r	y_3	z_2
…	…	…
…	…	…
x_2	y_1	z_l

Fig. 1. Example of permutation *X* and *Y* variables in contingency table.

In case of issue concerning study of cumulative influence of two variables on third variable hypotheses can be enrolled as following:

H_0: There is no cumulative influence of *X* and *Y* on *Z* variable.

H_1: There is correlation between variables *X* and *Y* and variable *Z*.

In permutation test of multiple correlation, the value of test statistics T_0 on the basis of (6) has been calculated, then *N* permutations of variable *Z* were performed and values of statistics T_i (i =1, 2, …, *N*) were determined. The decision is made on the basis of *ASL* value (see Efron, Tibshirani, 1994):

$$ASL = P\left(T \geq T_0\right), \tag{9}$$

for which estimation is obtained on the basis of:

$$ASL \approx \frac{card\{\, i : T_i \geq T_0\,)}{N}. \tag{10}$$

An example is for an upper-tailed, alternative hypothesis, where H_0 denotes the null hypothesis, T_0 is the observed value of the test statistic (6) based on (3), and T_i is the random variable corresponding to this test statistic.

In both situations (permutation tests of partial and multiple association) when *ASL* is lower than the assumed level of significance α, then hypothesis H_0 is rejected in favor of hypothesis H_1.

5. Empirical example

Idea of application the proposed method presents following example. Data concerns evaluation of income situation of households in 2007-2013. Three variables that were analysed were ordinal, multicategorical variables. Total debt of household (variable *X*) was defined on five levels (1–5): debt lower than monthly income of household (1), debt greater than monthly income but less than three months income (2), greater than three months income but less than half-year income (3), greater than half–year but lower than annual income (4) and debt greater than annual income of household (5). Evaluation of income level of household (variable *Y*) were stored on nine levels (1–9): we can afford everything and we can save for the future (1), we can afford everything but we do not save for future (2), we live frugally and that is why we can afford everything (3), we live very frugally to save for considerable needs (4), we can afford cheapest food, clothes, accommodation, repayment (5), we can afford cheapest food, clothes, accommodation, we cannot afford repayment (6), we can afford cheapest food, clothes, but we cannot afford accommodation (7), we can afford cheapest food, but we cannot afford clothes (8), we cannot afford even cheapest food (9). Four time periods (1-4), which represents years: 2007, 2009, 2011, 2013 were analyzed (variable *Z*). Table 2 presents collected data from Social Diagnosis in multidimensional contingency table.

All calculations were conducted in R program. This program is well suited for multidimensional permutation testing. Empirical distributions of test statistics *T* (on the basis of (5) three distributions in first row and on the basis of (6) three distributions in second row) were presented on Figure 2. Significance level $\alpha = 0.05$ in all performed tests was assumed. $N = 10000$ permutations of respective variables were used. Permutation tests for all possible cases of studying dependency between two variables when influence of third variable is eliminated were performed. As test statistic (5) based on (3) was used. *ASL* values for hypothesis on independence of variables: *X* and *Y* where influence of *Z* variable has been eliminated, also *X* and *Z* where influence of *Y* variable has been eliminated equal 0. For both

cases dependency between two variables when influence of third variable is eliminated has been confirmed.

Time	Total debt of household	Evaluation of incomes level of household								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1	27	47	207	74	98	15	20	25	5
	2	23	67	226	116	167	27	17	29	7
	3	10	46	121	74	112	15	14	16	5
	4	17	27	99	35	73	12	10	11	2
	5	26	41	103	33	64	14	12	7	7
2	1	73	118	436	177	224	15	29	43	12
	2	66	111	435	198	245	28	31	40	13
	3	55	108	288	129	199	22	15	32	15
	4	48	71	248	84	143	21	19	26	4
	5	92	129	263	90	157	27	16	23	10
3	1	77	89	376	185	180	23	29	28	19
	2	55	91	353	215	222	23	36	48	13
	3	50	57	262	179	169	19	16	28	11
	4	37	80	198	116	137	24	23	18	10
	5	100	129	391	191	210	30	35	27	21
4	1	49	73	292	152	185	12	25	32	11
	2	45	63	288	211	190	25	14	32	13
	3	49	58	213	143	164	25	8	24	14
	4	32	52	160	84	128	33	11	20	10
	5	119	151	425	194	247	46	32	18	22

Table 2. Evaluation of income situation of households in years 2007-2013.

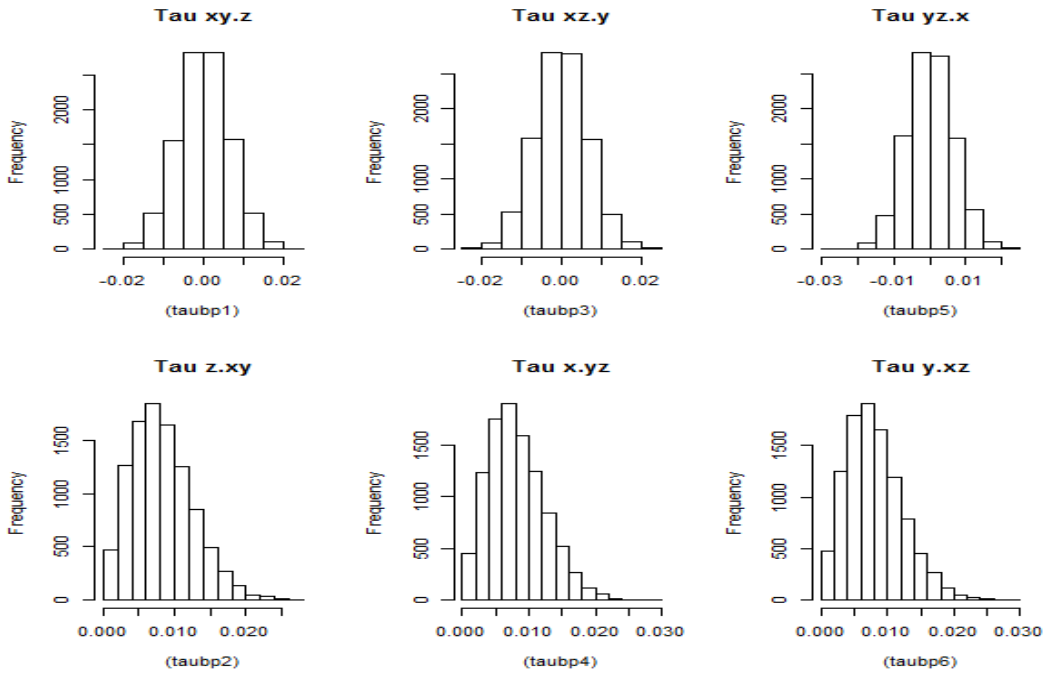

Fig. 2. Empirical distributions of test statistics.

In case of testing hypothesis on independence of variables Y and Z where influence of X variable has been eliminated value of ASL is 0.3116 therefore hypothesis H_0 cannot be rejected. Performing permutation tests for all cases of studying cumulative influence of two variables on third variable cause rejection of H_0 with assumed level of significance α. As test statistic (6) based on (3) was used. For all considered cases dependency between variables has been confirmed. ASL values calculated with empirical distributions of statistics are lower than assumed significance level. ASL values for hypothesis on independence of variables: there is no cumulative influence of X and Y on Z variable, also there is no cumulative influence of Y and Z on X variable equal 0 and there is no cumulative influence of X and Z on Y variable equals 0.0003. For all cases cumulative influence of two variables on third variable has been confirmed.

Conclusion

In dependency analysis studying significance of association between variables is often essential. When the rows and the columns are ordinal in contingency table, the chi-squared test of independence ignores the ordering information. In this case Kendall's *tau* rank correlation coefficient determines the strength of dependency. If data is characterized by more than two variables, a partial or multiple correlation coefficient can measures the relationship between variables. Article deals with a permutation approach to multidimensional problem of

hypothesis testing with regard to ordered data in a nonparametric framework. Multiple and partial correlation studied together give a very useful analysis of the relationship between the different variables.

References

Agresti, A. (2002). *Categorical data analysis*. 2nd ed. New York: Wiley.

Basso, D., Pesarin, F., Salmaso, L., & Solari, A. (2009). *Permutation Tests for Stochastic Ordering and ANOVA*, Heidelberg: Springer Science Business Media.

Butar, F. B., & Park, J. W. (2008). Permutation tests for comparing two populations. *Journal of Mathematical Science & Mathematics Education V3*, (2), 19-30.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Gautam, S. (2002). Analysis of mixed categorical data in 2× K contingency tables. *Statistics in medicine*, *21*(10), 1471-1484.

Good, P. (1994). *Permutation Tests: A practical guide for testing Hypotheses*, New York: Springer-Verlag.

Górniak, J., & Wachnicki, J. (2000). SPSS PL for Windows. *Pierwsze kroki w analizie danych, SPSS Polska.* Kraków.

Kendall, M. G. (1955). *Rank Correlation Methods*. 2nd ed. London: Griffin.

Oja, H., & Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Science*, 598-605.

Polko, D. (2014). On testing the similarity of multivariate populations structures, In: *Proceedings of 32nd International Conference Mathematical Methods in Economics*. Olomouc: Palacký University, 813-818.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. CRC press.

Simon, G. (1977). A nonparametric test of total independence based on Kendall's tau. *Biometrika*, *64*(2), 277-282.

Tran, L., Chu, B., Huang, C., & Huynh, K. P. (2014). Adaptive permutation tests for serial independence. *Statistica Neerlandica*, *68*(3), 183-208.

Zar, J. H. (1999). *Biostatistical analysis*. 4th ed. Upper Saddle River: Prentice Hall.