ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 3(302), 2014

Dominika Polko[*], *Grzegorz Kończak*[**]

ON THE METHOD OF COMPARING POPULATIONS' STRUCTURES BASED ON THE DATA IN THE CONTINGENCY TABLES

ABSTRACT. Comparison of populations is one of the most important problems in statistics. The most common comparisons apply to two populations, but comparisons of *k* populations, where $k > 2$ are also carried out. Parametric methods allow to compare the means, variances or proportions. The non-parametric methods allow to compare the distributions of two or more populations. The problem of comparison structures based on data in contingency tables is analyzed in the paper. The permutation tests were applied in the multivariate nominal data structure comparison.

Keywords: comparing structures, contingency tables, permutation tests

I. INTRODUCTION

In economic research it is often important to compare the structures of populations. These comparisons may involve various social groups, regions or time periods. As part of a statistical analysis used in economic research these comparisons include both parametric and nonparametric tests. Parametric tests require the assumptions about the distribution of the characteristic in the population to be fulfilled. Among these there are, tests that enable to compare parameters based on samples taken from two or more populations. In the case of assumptions of verification the hypotheses are not fulfilled, non-parametric tests could be used. To ensure better power of tests, permutation tests can be used. This paper considers comparison of multidimensional structures based on the data presented in the contingency tables. Basso et all (2009) have analyzed the structures of populations that have ordered categories by means of permutation tests. Similar analyses are applied for the multivariate nominal data in the paper.

[*] M.A., Department of Statistics, Katowice University of Economics, dpolko@gmail.com.

[**] Ph.D., Associate Professor, Department of Statistics, Katowice University of Economics, grzegorz.konczak@ue.katowice.pl.

II. COMPARING POPULATIONS

It is often necessary to decide if compared populations are different in means, variances or proportions. To determine if any significant differences exist in the two populations based on random sampling methods statistical inference is used. Most commonly used statistical tests that allow to compare the population parameters include (see Wywiał J., 2004):

- *t*-test to compare the expected values in the two populations,
- ANOVA test for comparison of the expected values in k ($k > 2$) populations,
- equality tests of two or more variances,
- equality tests of proportions.

These tests, except for the last one, require the samples to be taken from a population with a normal distribution. For large samples the limit distributions of statistics can be used. In the case of small samples, if the following assumption is not fulfilled, appropriate non-parametric tests such as *U* Mann-Whitney test or Kruskal-Wallis test should be used.

Besides comparisons of parameters, it is often necessary to refer to the comparison of distributions in the two populations. In this case, Kolmogorov-Smirnov test is used most often. In this test statistic refers to the comparison of the empirical distribution functions.

Sometimes there is a need to compare the structures in two or more populations. In this case a similarity index is used for structure comparison. The next section presents a proposal to compare multidimensional structures based on data presented in two contingency tables.

III. COMPARING STRUCTURES

The result of comparison of the structures of two or more populations or one population for different time periods is to calculate the appropriate measure of compliance characterized by the degree of similarity (ω) of such structures. Assuming that the structure of two different populations described due to the characteristic of X are compared, such that k variants have been distinguished, the analyzed structures can be described using vectors

$$S_1 = \begin{bmatrix} w_{\backprime} \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \qquad S_2 = \begin{bmatrix} w_1^{'} \\ w_2^{'} \\ \vdots \\ w_k^{'} \end{bmatrix} \tag{1}$$

where $w_i = \frac{n_i}{\sum_{i=1}^{k} n_i}$ and $w_i^{'} = \frac{n_i}{\sum_{i=1}^{k} n_i}$, which fulfill conditions $0 \le w_i \le 1$, $0 \le w_i^{'} \le 1$ and $\sum_{i=1}^{k} w_i = 1$, $\sum_{i=1}^{k} w_i^{'} = 1$.

The closer are the components of two vectors, the more similar are these vectors (Żwirbla A., 2006). As the degree of dissimilarity of the two structures S_1 and S_2 increases, the index $\omega(S_1, \mathrm{S}_2)$ leads to zero. When compared structures are the same, the value of index $\omega(S_1, \mathrm{S}_2)$ is 1. In literature, the following characteristics of similarity measure structures are given (see Kukuła K., 1986)

$$\omega(S_1, S_2) = 1 \Leftrightarrow S_1 = S_2 \tag{2}$$

$$\omega(S_1, S_2) = \omega(S_2, S_1) \tag{3}$$

$$\omega(S_1, S_2) \in \langle 0,1 \rangle . \tag{4}$$

To compare the structure, the structure similarity index is frequently used. The classic approach to similarity structures can be represented as follows

$$\omega_1 = \sum_{i=1}^{k} \min(w_i, w_i^{'}) . \tag{5}$$

The similarity index that is constructed on the basis of indicators of the structure has values of the interval $\langle 0,1 \rangle$. Other measures that enable the assessment of the similarity of population structures are as follows:

$$\omega_2 = 1 - \frac{\sum_{i=1}^{k} \left| w_i - w_i^{'} \right|}{2} \tag{6}$$

$$\omega_3 = cos(w_i, w_i^{'}) = \frac{S_1^{\ T} S_2}{|S_1||S_2|} \tag{7}$$

$$\omega_4 = \frac{\sum_{i=1}^{k} min(w_i, w_i^{'})}{\sum_{i=1}^{k} max(w_i, w_i^{'})} \tag{8}$$

The data composed of two nominal variables usually is presented in a contingency table. Contingency tables are arrays of non-negative integers that arise from the crossclassification of a sample or a population of N objects based on a set of categorical variables of interest. The entries n_{ij} (i = 1,2,…, r, j = 1,2,…, c) are the counts for every two-way combination of rows and columns. Table 1 presents the model for such kind of data

Table 1. Contingency table

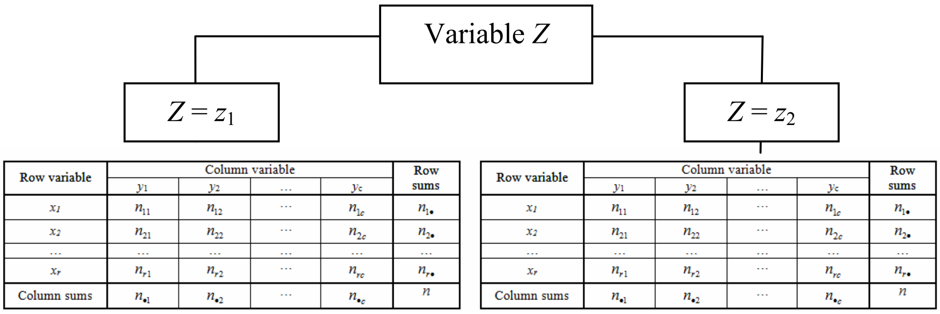
Row variable	Column variable				Row sums
	y_1	y_2	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\bullet}$
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\bullet}$
Column sums	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet c}$	n

where $n_{i\bullet} = \sum_{j=1}^{c} n_{ij}$, $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$ and $n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j}$.

Source: own work.

In the next part of this article the problem of comparing two populations will be considered, for which it has the results from the sample in the form of two-dimensional contingency tables.

Let us assume that data from the samples is given in two two-dimensional contingency tables. Let us compare the structure of the populations based on the data samples given in these contingency tables. To compare these structures based on the data in two contingency tables (classification variables X and Y) let us consider a third classifying, dichotomous variable Z, which takes the value of z_1 for the elements of the first table and the value of z_2 for the elements of the second table. The data could be written in the form presented in figure 1. In general, it is possible to compare s contingency tables. Then the variable that identifies contingency tables will take the values z_1, z_2, …, z_s. In the next part of this paper structures of populations based on the data in two contingency tables will be considered.



Row variable	Column variable				Row sums
	y_1	y_2	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\bullet}$
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\bullet}$
Column sums	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet c}$	n

Row variable	Column variable				Row sums
	y_1	y_2	…	y_c	
x_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\bullet}$
…	…	…	…	…	…
x_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\bullet}$
Column sums	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet c}$	n

Figure. 1. Representation of the three-way contingency table

Source: own work.

In the case of data presentation in a three-dimensional contingency table, to compare the structures based on the data in two contingency tables the index (5) can be used as a natural extension of the two-dimensional case

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \min_{s} (w_{ij}^{(s)}) \tag{9}$$

where

$$w_{ij}^{(s)} = \frac{n_{ij}^{(s)}}{n^{(s)}}$$

and s ($s = 1, 2$) is a contingency table number.

Taking into account that the data presented in the two contingency tables can be stored as one three-dimensional contingency table (see Fig. 1) another possible comparison of these structures provides a chi-square ratio calculated for a multi-dimensional contingency table. In the case of a three-dimensional table, where the variable Z can take two values, this can be represented as follows

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{2} \frac{\left(n_{ijk} - \hat{n}_{ijk}\right)^2}{\hat{n}_{ijk}} \tag{10}$$

where $\hat{n}_{ijk} = \dfrac{n_{ij\bullet} \cdot n_{\bullet\bullet k}}{n}$.

Data stored in two contingency tables can be written in three columns (see table 2), where the third column identifies the number of the contingency table.

Table 2. The form of the data from two contingency tables

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

Source: own work.

The problem of comparing two contingency tables comes down to investigate the effect of variable Z (contingency table index) on variables X and Y. Verified hypothesis H_0 says that the structures of the two contingency tables are identical to hypothesis H_1, which is the negation of H_0. Distributions of statistics (9) and (10) are unknown, hence to make the decision, the permutation test will be used (Efron B., Tibshirani R., 1993). The value of statistic $\hat{\theta}_0$ that measures similarity (or dissimilarity) of structures has been calculated, N permutations of variable Z were performed and values $\hat{\theta}_i$ ($i = 1, 2, \ldots, N$) were determined. The decision concerning a verified hypothesis is made on the basis of ASL (*achieving significance level*) value

$$ASL = P\left(\hat{\theta} \geq \hat{\theta}_0\right),$$

for which estimation is obtained on the basis of

$$ASL \approx \frac{card\{i : \hat{\theta}_i \geq \hat{\theta}_0)}{N}.$$

This notation applies, where the H_0 rejection area is right-sided. In the case of left-sided rejection area in above notation inequality should be changed.

When *ASL* is lower than the assumed level of significance α, then H_0 is rejected in favor of hypothesis H_1.

The permutation test procedure that is used for the verification of the hypothesis on the compliance of contingency tables structures is as follows

1. Assume the level of significance α.
2. Calculate the value of statistics for the sample data.
3. Perform the permutation of variable Z N-times, then calculate the statistics test value.
4. On the basis of empirical distribution of statistics, the *ASL* value is determined. If $ASL < \alpha$, then H_0 is rejected, otherwise H_0 hypothesis cannot be rejected.

IV. MONTE CARLO STUDY

The properties of the statistics described above have been analyzed in the Monte Carlo study. Three versions of compared structures were analyzed:

a) The same structures

b) Similar structures

c) Different structures

Theoretical structures of compared populations are presented in Figure 2.

a)

	$Z = z_1$			$Z = z_2$		
	$Y = y_1$	$Y = y_2$	$Y = y_3$	$Y = y_1$	$Y = y_2$	$Y = y_3$
$X = x_1$	0.1	0.1	0.05	0.1	0.1	0.05
$X = x_2$	0.1	0.3	0.1	0.1	0.3	0.1
$X = x_3$	0.05	0.1	0.1	0.05	0.1	0.1

b)

	$Z = z_1$			$Z = z_2$		
	$Y = y_1$	$Y = y_2$	$Y = y_3$	$Y = y_1$	$Y = y_2$	$Y = y_3$
$X = x_1$	0.1	0.1	0.05	0.1	0.1	0.1
$X = x_2$	0.1	0.3	0.1	0.05	0.3	0.05
$X = x_3$	0.05	0.1	0.1	0.1	0.1	0.01

c)

	***Z* = *z*$_1$**			***Z* = *z*$_2$**		
	***Y* = *y*$_1$**	***Y* = *y*$_2$**	***Y* = *y*$_3$**	***Y* = *y*$_1$**	***Y* = *y*$_2$**	***Y* = *y*$_3$**
***X* = *x*$_1$**	0.1	0.1	0.05	0.2	0.05	0.1
***X* = *x*$_2$**	0.1	0.3	0.1	0.1	0.05	0.1
***X* = *x*$_3$**	0.05	0.1	0.1	0.1	0.1	0.2

Figure 2. Theoretical structures of compared populations

Source: own work.

In simulation analyses, samples of a distribution were generated (Fig. 2) of count n = 100, 150, 200, 250. On the basis of obtained data, hypothesis H_0 on identity of compared distributions was verified using the above described permutation test. In the analyses the significance level α = 0,05 was assumed. Estimated probabilities of rejection H_0 are presented in table 3.

Table 3. Estimated probabilities of rejection H_0

	Chi-square statistic				*T*-statistic			
n	100	150	200	250	100	150	200	250
a)	0.06	0.05	0.05	0.02	0.05	0.06	0.04	0.01
b)	0.15	0.19	0.18	0.28	0.13	0.22	0.22	0.28
c)	1	1	1	1	0.98	1	1	1

Source: own work.

In case a) probability of rejection the hypothesis on identity of distributions is near α. In case c) hypothesis H_0 is rejected with probability near 1. In the case of both statistics, the obtained results are similar.

V. CONCLUDING REMARKS

The paper deals with the problem of comparing population structures. It presents the procedure that enables comparing multi-dimensional structures on the basis of the data stored in contingency tables. Due to the lack of information on the theoretical distribution of considered statistics, in order to compare the results - the permutation test was used. In the case of chi-square statistic verification of hypothesis on identity of the structures is equal to verify the hypothesis on independence of variables (variable that identifies the table and conformation of the rest of variables). The *T*-statistic, on the other hand, is the extension of the classic proportions onto two-dimensional structures.

ACKNOWLEDGEMENTS

The part of the second author was supported by Polish National Science Centre grant DEC-2011/03/B/HS4/05630.

REFERENCES

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

Efron B., Tibshirani R. (1993), *An Introduction to the Bootstrap*, Chapman & Hall. New York

Kukuła K. (1986), *Przegląd wybranych miar zgodności struktur*, Przegląd Statystyczny, R.XXXIII zeszyt 4.

Sheskin D.J. (2003), *Handbook of Parametric and Nonparametric Statistical Procedures,* Chapman & Hall, Boca Raton.

Wywiał J. (2004), *Wprowadzenie do wnioskowania statystycznego*, Wydawnictwo Akademii Ekonomicznej w Katowicach.

Żwirbla A. (2006), *Próba konstrukcja mierników struktury oraz zmian strukturalnych*, Wiadomości Statystyczne, nr 10.

Dominika Polko, *Grzegorz Kończak*

O PEWNEJ METODZIE PORÓWNYWANIA STRUKTUR POPULACJI NA PODSTAWIE DANYCH W TABLICACH WIELODZIELCZYCH

Do najważniejszych zagadnień rozważanych w statystyce należy porównywanie zbiorowości. Najczęściej porównania takie dotyczą dwóch populacji, ale niekiedy prowadzi się porównania k populacji, gdzie $k > 2$. Metody parametryczne pozwalają na porównywanie wartości przeciętnych, wariancji lub wskaźników struktury a metody nieparametryczne na porównywanie postaci rozkładów w dwóch lub większej liczbie populacji. W artykule podjęto zagadnienie porównywania struktur tablic wielodzielczych. Zaproponowano metodę pozwalającą na porównanie takich struktur z wykorzystaniem testu permutacyjnego.