Dominika Polko–Zając[1]

Permutation test for equality of distributions in two populations

Abstract

A comparison of multidimensional populations is very interesting and common statistical problem. The most often way is to verify the hypothesis about the equality of mean vectors in two populations. The classic test for verification of this hypothesis is the Hotelling's T^2 test. Another solution is the use of simulation and randomization methods to study the significance of differences between the studied populations. Permutation tests are to enable statistical inference in situations where it is not possible to use classical parametric tests. These tests are supposed to provide comparable power to parametric tests with simultaneous reduction of assumptions, e.g. regarding the sample size taken or the distribution of the tested variable in the population. The article presents a permutational, complex procedure for assessing the overall *ASL* (*achieved significance level*) value. The applied nonparametric statistical inference procedure uses combining function. A simulation study was carried out to determine the size and power of the test under normality. A Monte Carlo simulation let to compare empirical power of this test with Hotelling's T^2 test power. The advantage of the proposed method is that the method can be used even when samples are taken from any type of continuous distributions in population.

Keywords: permutation tests, comparing populations, power of test, Monte Carlo simulation, R software.
JEL Classification: C30, C150, C880.

1. Introduction

Population comparisons most often involve comparison of characteristics in these populations. If it is assumed that population distributions differ only in the central tendency there are various parametric and nonparametric tests to verify this hypothesis. Many authors undertake to study both the power and size of tests for the significance of differences between means or medians in two or more populations using for this

[1] Dominika Polko–Zając, University of Economics in Katowice, Faculty of Management, Department of Statistics, Econometrics and Mathematics, 1 Maja 50, 40-287 Katowice, Poland, e-mail: dominika.polko@ue.katowice.pl

purpose simulation methods based on bootstrap or permutation tests [Janssen and Pauls 2005, Chang and Pal 2008, Kończak 2016, Anderson et al. 2017].

In a situation where the statistical test for certain measurable variables is conducted in several multidimensional populations the hypothesis about the equality of mean vectors in these populations may need to be verified. A special case is the study of differences in the means of variables ${}^1X, {}^2X, \ldots, {}^PX$ in two populations. The problem is to test the hypothesis about the equality of the mean vectors of *P*–dimensional random variable in the first and the second population respectively in the form of

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2, \tag{1}$$

against alternative hypothesis

$$H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{2}$$

The classic test for verification of this hypothesis (1) is the Hotelling's T^2 test. The method using the T^2 test was proposed by Hotelling [1931, 1947] and Mahalanobis [1930, 1936] and is a generalization of the Student's *t* test for many variables. For the test to be used the assumption is made that the samples were taken from a population with multidimensional normal distributions [Rencher 2002].

In the Hotelling's T^2 test two populations are considered from which two samples are taken independently from the distribution $N_P(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and from the distribution $N_P(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Assuming that covariance matrices are unknown, but the same $(\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma})$ in order to verify the null hypothesis (1) on the equality of the mean vectors, the statistics can be used

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{3}$$

where:

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1i} - \bar{\mathbf{x}}_1)^T + \sum_{i=1}^{n_2} (\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2i} - \bar{\mathbf{x}}_2)^T \right).$$

If the H_0 hypothesis is true, the statistics (3) has a Hotelling's T^2 distribution with *P* and $n_1 + n_2 - 1$ degrees of freedom, where *P* is the number of variables (dimensions) examined and n_1, n_2 are the sizes of samples taken form populations. It is also possible to determine the critical values for this statistics using statistics of form [Krzyśko 2009]

$$F = \frac{n_1 + n_2 - P - 1}{(n_1 + n_2 - 2)P} T^2, \tag{4}$$

which has a Snedecor's F distribution of P and $n_1 + n_2 - P - 1$ degrees of freedom.

The Hotelling's T^2 test can only be used if the variables in each population have a multidimensional normal distribution. The article presents a method of testing the difference between two vectors of mean values that can be used also when the assumption regarding the occurrence of a multidimensional normal distribution in populations is not met. A simulation, randomization approach was proposed to investigate the significance of differences occurring between the studied populations. The method of solving the problem known from the literature and methods based on Monte Carlo simulations were compared. An approach was considered in accordance with the nonparametric statistical inference procedure using two-stage *ASL* (*achieved significance level*) determination. All simulations were carried out in the R statistical computing environment [R Core Team 2016].

2. Nonparametric combination procedures

It is assumed that there are two samples ${}^1X_1,...,{}^pX_1,...,{}^PX_1$ and ${}^1X_2,...,{}^pX_2,...,{}^PX_2$ independently taken from the population with distribution F_1 and F_2. These populations have continuous, *P*–dimensional distributions F_i for $i = 1, 2$ with unknown parameters. The zero hypothesis is verified claiming that two samples were taken from populations with identical distributions in the form of $H_0 : F_1(x) = F_2(x)$. Data taken from two populations can be noted [Marozzi 2008].

$$\underline{\underline{X}} = \begin{bmatrix} {}^1X_1 & {}^1X_2 \\ \vdots & \vdots \\ {}^pX_1 & {}^pX_2 \\ \vdots & \vdots \\ {}^PX_1 & {}^PX_2 \end{bmatrix} = \begin{bmatrix} {}^1X_{11} & \cdots & {}^1X_{1n_1} & {}^1X_{21} & \cdots & {}^1X_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ {}^pX_{11} & \cdots & {}^pX_{1n_1} & {}^pX_{21} & \cdots & {}^pX_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ {}^PX_{11} & \cdots & {}^PX_{1n_1} & {}^PX_{21} & \cdots & {}^PX_{2n_2} \end{bmatrix} =$$

$$= \begin{bmatrix} {}^1X_1 & \cdots & {}^1X_{n_1} & {}^1X_{n_1+1} & \cdots & {}^1X_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ {}^pX_1 & \cdots & {}^pX_{n_1} & {}^pX_{n_1+1} & \cdots & {}^pX_n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ {}^PX_1 & \cdots & {}^PX_{n_1} & {}^PX_{n_1+1} & \cdots & {}^PX_n \end{bmatrix} = \begin{bmatrix} {}^1\underline{X} \\ \vdots \\ {}^p\underline{X} \\ \vdots \\ {}^P\underline{X} \end{bmatrix}$$

where ${}^{p}X_{ij}$ indicate the *i*–th $\left(i=1,\ldots,n_j\right)$ sample for the *p*–th variable $\left(p=1,\ldots,P\right)$ in the *j*–th $\left(j=1,2\right)$ population and ${}^{p}\underline{X}$ is combined sample for the *p*–th variable.

The problem of testing equality of means in multidimensional populations can be considered in accordance with the Pesarin [2001] proposal. When the study deals with the problem of comparing the *P* means in two populations, partial *P* hypotheses are taken into account. The null hypothesis about the identity of the means vectors is in the form of

$$H_0 : \bigcap_{p=1}^{P} {}^{p}\mu_1 = {}^{p}\mu_2 , \tag{5}$$

against the alternative hypothesis

$$H_1 : \bigcup_{p=1`}^{P} {}^{p}\mu_1 \neq {}^{p}\mu_2 . \tag{6}$$

The study considered test statistics in the form of

$${}^{p}T = {}^{p}\overline{X}_1 - {}^{p}\overline{X}_2 . \tag{7}$$

The decision was made using the empirical distribution of the test statistic obtained on the basis of permutation of the data set. A nonparametric, complex procedure was used to assess the overall *ASL* values. In the first stage of separate testing of each of the *P* partial hypotheses considered, the *ASL* values are determined in accordance with the traditional permutation method used during verification of the hypothesis for one-dimensional data, i.e.:

1. The significance level α is determined.
2. The statistics values are calculated on the basis of the sample data (${}^{p}T_0$).
3. Perform the permutation of variables *N*–times then calculate the statistics test value (T_k).
4. Based on the empirical distribution of statistics the *ASL* value for each of the compared variables is estimated according to the formula

$$\hat{ASL}_{{}^{p}T}\left({}^{p}T_0\right) = \frac{0{,}5 + \sum_{k=1}^{N} I\left(\left|{}^{p}T_k\right| \geq \left|{}^{p}T_0\right|\right)}{N+1} . \tag{8}$$

The method of permutation of multidimensional variables is shown in Figure 1.

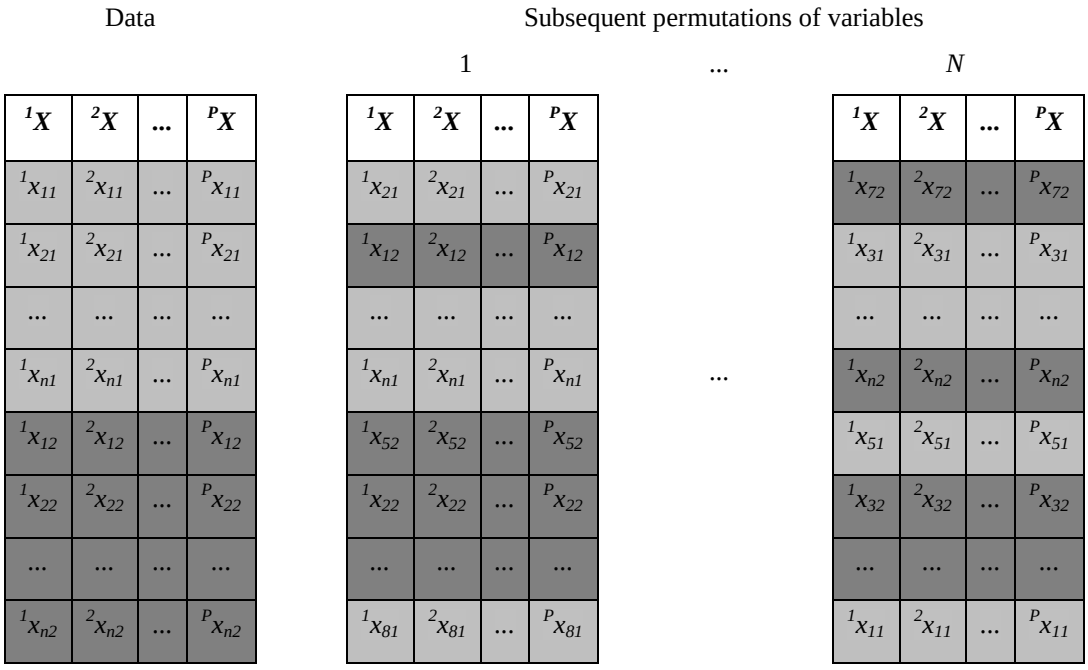

Fig. 1. Scheme of permutations of variables

Source: author's own work.

The second stage of the nonparametric statistical inference procedure involves the determination of the overall *ASL* value using combining functions [Pesarin 2001]

$$_{\varphi}T = \varphi\left(ASL_{_1T}\ ,...,ASL_{_P T}\ \right).$$

There are many forms of combining functions for determining the overall *ASL* value, however the authors the most often point to functions:

- the Fisher omnibus combining function [Fisher 1932]

$$C^{(F)} = -2 * \sum_{p=1}^{P} \log\left(\hat{ASL}\left({}_{P}T\right)\right),$$

- the Liptak combining function [Liptak 1958]

$$C^{(L)} = \sum_{p=1}^{P} \Phi^{-1}\left(1 - \hat{ASL}\left({}_{P}T\right)\right),$$

where Φ denotes the standard normal distribution function,

- the Tippet combining function [Tippet 1931]

$$C^{(T)} = \max\left\{1 - \hat{ASL}\ \left({}_{1}T\right),...,1 - \hat{ASL}\left({}_{P}T\right)\right\}.$$

The observed statistic value for the sample data using Fisher combining functions can be determined as

$$\underline{T}_0 = -2 * \sum_{p=1}^{P} \log\left(A\hat{S}L_{{}^pT}\left({}^pT_0\right)\right), \tag{9}$$

whereas the distribution of this statistic is determined on the basis of the same permutations as in the first step, for example for *k*–th permutation

$$\underline{T}_k = -2 * \sum_{p=1}^{P} \log\left(A\hat{S}L_{{}^pT}\left({}^pT_k\right)\right). \tag{10}$$

The total *ASL* value for the test under consideration is estimated using the formula

$$A\hat{S}L_{\underline{T}} = \frac{\sum_{k=1}^{N} I\left(\underline{T}_k \geq \underline{T}_0\right)}{N}. \tag{11}$$

If $ASL < \alpha$, the hypothesis H_0 is rejected, otherwise there is no basis for rejecting the H_0 hypothesis.

3. Monte Carlo simulation

Considering the nonparametric procedure based on the Fisher combining function, the size and power of the test were estimated by simulation study. A Monte Carlo analysis was carried out allowing comparison of two populations with three-dimensional normal distributions with parameters $\boldsymbol{\mu}_1 = [0,0,0]$, $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\boldsymbol{\mu}_2 = [x,x,x]$, $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, where $x \in (-1,1)$ with the increment 0.2. In the simulations samples of sizes (n_1,n_2) = (10,10), (20,20), (30,30), (50,50), (100,100) were generated. The results of the simulations carried out to determine the size and power of the tests are presented in Table 1 (small sample sizes) and Table 2 (large sample sizes). For comparative purposes, the tables also include the results obtained for the parametric Hotelling's T^2 test and its permutation equivalent. The procedure for conducting each

test included 1000 Monte Carlo simulations and 1000 permutations of variables and the assumed level of significance was $\alpha = 0.05$.

Table 1. Hotelling's T^2 test power and estimation of permutation tests power (small sample sizes)

Test statistic			
x	T^2	T^2 *(perm)*	$\underline{T}$
(10,10)			
-1.0	0.828	0.829	0.900
-0.8	0.632	0.628	0.701
-0.6	0.381	0.387	0.440
-0.4	0.187	0.193	0.199
-0.2	0.079	0.079	0.083
0	0.048	0.046	0.048
0.2	0.075	0.077	0.076
0.4	0.157	0.155	0.177
0.6	0.389	0.385	0.447
0.8	0.629	0.625	0.715
1.0	0.846	0.843	0.902
(20,20)			
-1.0	0.996	0.995	0.998
-0.8	0.958	0.957	0.972
-0.6	0.747	0.753	0.790
-0.4	0.390	0.394	0.408
-0.2	0.102	0.105	0.105
0	0.045	0.043	0.045
0.2	0.117	0.119	0.125
0.4	0.373	0.380	0.401
0.6	0.760	0.759	0.801
0.8	0.947	0.948	0.963
1.0	0.996	0.996	0.998
(30,30)			
-1.0	1.000	1.000	1.000
-0.8	0.998	0.997	0.998
-0.6	0.913	0.911	0.927
-0.4	0.548	0.556	0.586
-0.2	0.158	0.160	0.170
0	0.055	0.059	0.054
0.2	0.160	0.160	0.170
0.4	0.552	0.560	0.596
0.6	0.916	0.917	0.943
0.8	0.992	0.992	0.995
1.0	1.000	1.000	1.000

Source: computer simulations in the R program.

In the case of the analysis of multidimensional, equinumerous samples, the sizes of the presented tests are close to the assumed level of significance. The values of estimated probabilities of rejecting the hypothesis H_0, when it was true only slightly differed from $\alpha = 0.05$. The three considered tests reached comparable assessments of the

Table 2. Hotelling's T^2 test power and estimation of permutation tests power (large sample sizes)

Test statistic			
x	T^2	T^2 *(perm)*	$\underline{T}$
(50,50)			
-1.0	1.000	1.000	1.000
-0.8	1.000	1.000	1.000
-0.6	0.987	0.987	0.989
-0.4	0.835	0.836	0.850
-0.2	0.255	0.258	0.271
0	0.045	0.049	0.048
0.2	0.257	0.257	0.264
0.4	0.805	0.805	0.825
0.6	0.992	0.993	0.995
0.8	1.000	1.000	1.000
1.0	1.000	1.000	1.000
(100,100)			
-1.0	1.000	1.000	1.000
-0.8	1.000	1.000	1.000
-0.6	1.000	1.000	1.000
-0.4	0.997	0.997	0.997
-0.2	0.528	0.528	0.536
0	0.044	0.044	0.041
0.2	0.520	0.528	0.535
0.4	0.986	0.986	0.988
0.6	1.000	1.000	1.000
0.8	1.000	1.000	1.000
1.0	1.000	1.000	1.000

Source: computer simulations in the R program.

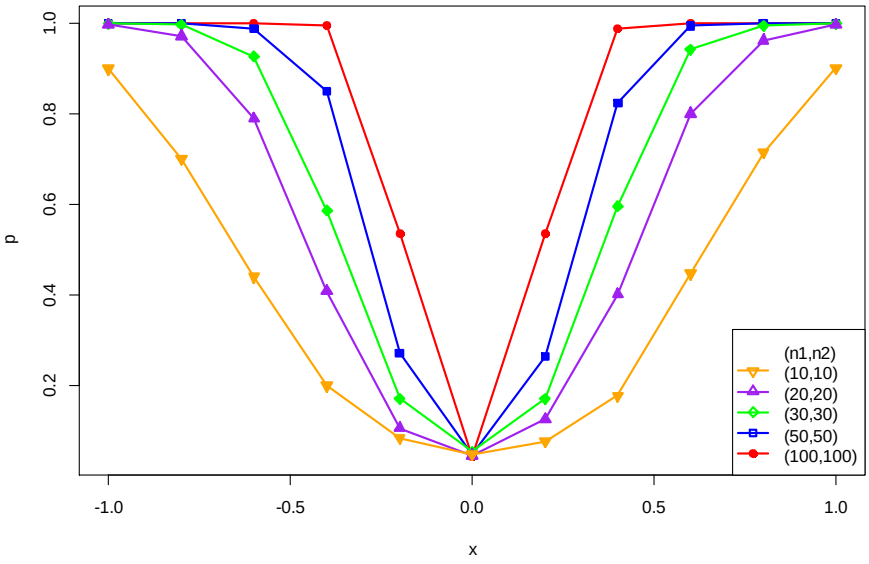

Fig.2. Graphs of the empirical power of the permutation test $\underline{T}$ for different sample sizes

Source: author's own work in the R program.

probabilities of rejecting the H_0 hypothesis when it was false. In the majority of analyzed cases, however, the most powerful test was the proposed permutation test based on a two-stage *ASL* determination method using the Fisher combining function. The probabilities of recognizing differences between means vectors increased as the differences between the considered three-dimensional models of the populations increased. Analyzing the graphs of the empirical power of the permutation test depending on the sample sizes taken from the populations (Figure 2), it can be seen that for 10 observations the differences in means at level 1 are detected with a probability of around 90% by the permutation test. For samples with 50 observations, this probability was obtained for the difference in means of around 0.5.

4. Conclusions

The aim of simulation research was to determine the ability of presented permutation test to maintain the nominal probability of committing the type I error and the ability to obtain a high probability of rejecting a false zero hypothesis in the conditions of changing distribution parameters in populations from which samples were taken. Simulation tests to determine the size and power of tests were carried out using permutation tests.

The tests that verify the hypothesis about the identity of distributions in the studied populations are presented. The results obtained in the simulation confirm the effectiveness of the method and the possibility of its application in order to infer differences between vectors of means in two populations with multidimensional normal distributions. All testing procedures (under normality) ensured control of type I error at the assumed level of significance. Higher power of presented tests was achieved thanks to the use of a nonparametric combination procedure that uses Fisher's combining functions to evaluate the overall *ASL* value. The advantage of the proposed method is that the method can be used even when samples are taken from any type of continuous distributions in population.

Bibliography

Anderson M.J., Walsh D.C.I., Clarke K.R. Gorley R.N., Guerra-Castro E. [2017], *Permutational Multivariate Analysis of Variance (PERMANOVA)*, Wiley StatsRef: "Statistics Reference Online", 1-15.

Chang C.-H., Pal N. [2008], *A Revisit to the Behrens–Fisher Problem: Comparison of Five Test Methods*, "Communications in Statistics - Simulation and Computation", 37:6, 1064-1085.

Fisher R. A. [1932], *Statistical Methods for Research Workers*, 4 edn, Oliver & Boyd, Edinburgh.

Hotelling H. [1931], *The generalization of Student's ratio*, "Annals of Mathematical Statistics" 2 (3): 360–378.

Hotelling H. [1947], *Multivariate Quality Control*. In C. Eisenhart, M. W. Hastay, and W. A. Wallis, eds. Techniques of Statistical Analysis, McGraw-Hill, New York.

Janssen A., Pauls T. [2005], *A monte carlo comparison of studentized bootstrap and permutation tests for heteroscedastic two-sample problems*, "Computational Statistics", 20(3), 369-383.

Kończak G. [2016], *Testy permutacyjne. Teoria i zastosowania*. Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice.

Krzyśko M. [2009], *Podstawy wielowymiarowego wnioskowania statystycznego*, Wydawnictwo naukowe UMA, Poznań.

Liptak I. [1958], *On the combination of independent tests*, Magyar Tudomanyos Akademia Matematikai Kutato Intezenek Kozlomenyei 3, 127–141.

Mahalanobis P. C. [1930], *On tests and measures of group divergence*, "Journal of the Asiatic Society of Bengal" 26: 541–588.

Mahalanobis P. C. [1936], *On the generalized distance in statistics*, National Institute of Science of India 12: 49–55.

Marozzi M. [2008], *The Lepage location-scale test revisited*, "Far East Journal of Theoretical Statistics" 24: 137–155.

Pesarin F. [2001], *Multivariate Permutation Test with Applications in Biostatistics*, Wiley, Chichester.

R Core Team [2016], *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, https://www.R-project.org/.

Rencher A.C. [2002], *Methods of Multivariate Analysis*, John Wiley & Sons, New York.

Tippett L. H. C. [1931], *The Methods of Statistics*, Williams and Norgate, London.

Abstract

Permutacyjny test identyczności rozkładów w dwóch populacjach

Porównanie populacji wielowymiarowych jest bardzo interesującym i często rozważanym problemem statystycznym. Najczęściej weryfikowana jest hipoteza o równości wektorów wartości średnich w dwóch populacjach. Klasycznym testem do weryfikacji tej hipotezy jest test T^2 Hotellinga. Innym rozwiązaniem jest wykorzystanie metod symulacyjnych, randomizacyjnych do badania istotności różnic między badanymi populacjami. Testy permutacyjne mają umożliwić wnioskowanie statystyczne w sytuacjach, w których nie jest możliwe zastosowanie klasycznych testów parametrycznych. Testy te mają

zapewnić porównywalną moc do testów parametrycznych przy jednoczesnym ograniczeniu założeń, np. w odniesieniu do przyjętej wielkości próby lub rozkładu badanej zmiennej w populacji. W artykule przedstawiona zostanie permutacyjna, złożona procedura do oceny łącznej wartość *ASL* (*achieved significance level*). Zastosowana nieparametryczna procedura wnioskowania statystycznego wykorzystuje funkcje łączące (*combining function*). Przeprowadzono badanie symulacyjne w celu określenia rozmiaru i mocy testu w warunkach normalności. Symulacja Monte Carlo pozwoliła porównać empiryczną moc tego testu z mocą testu T^2 Hotellinga. Zaletą proponowanej metody jest to, że metoda może być stosowana, gdy próby są pobierane z dowolnego, ciągłego rozkładu w populacji.

Słowa kluczowe: testy permutacyjne, porównanie populacji, moc testu, symulacja Monte Carlo, program R.