ARGUMENTA
OECONOMICA
CRACOVIENSIA

No 1 (22) • 2020
ISSN 1642-168X
e-ISSN 2545-3866
AOC, 2020; 1 (22): 31–45
https://doi.org/10.15678/AOC.2020.2203

Grzegorz Kończak

APPLICATIONS OF PERMUTATION METHODS IN THE ANALYSIS OF ASSOCIATIONS

Abstract

Objective: The permutation model in hypothesis testing was introduced by R. A. Fisher in 1925. These methods permit us to test hypotheses with as minimal assumptions as possible. The tests require high computing power and therefore have found greater application in recent years. However, the concept of permutation methods is much wider than the issue of permutation testing. In 1923 J. Spława-Neyman introduced a permutation model for the analysis of field experiments. The purpose of the article is to present the possibilities of applying permutation methods in the analysis of dependencies. The article presents selected possibilities of data rearranging in dependency analysis.

Research Design & Methods: The study considered the analysis of multivariate data. The paper presents theoretical considerations and refers to the Monte Carlo simulation.

Findings: A proposed method to allow investigation of the significance of the relationship between two data sets is presented. The considerations are supplemented by comparing the size and power of the proposed test with tests known from canonical correlation analysis.

Implications/Recommendations: The proposal is most powerful for non-normally distributed variables and small samples.

Contribution: The proposed test can be used in the analysis of multidimensional economic and social phenomena.

Grzegorz Kończak, University of Economics in Katowice, Department of Statistics, Econometrics and Mathematics, ul. 1 Maja 50, 40-287 Katowice, Poland, e-mail: grzegorz.konczak@ue.katowice.pl, ORCID: https://orcid.org/0000-0002-4696-8215.

Keywords: permutation methods, data labels permutation, association analysis, canonical correlations.
JEL Classification: C12, C15, C18.

1. Introduction

The permutation model in hypothesis testing was introduced by R. A. Fisher in 1925. These methods permit us to test hypotheses with as minimal assumptions as possible. The tests require high computing power and therefore have found greater application in recent years. However, the concept of permutation methods is much wider than the issue of permutation testing. In 1923, J. Spława-Neyman introduced a permutation model for the analysis of field experiments (Spława-Neyman 1923, Berry, Johnston & Mielke 2014). This paper was published in Polish. Neyman's work was translated into English in 1990 and published in *Statistical Science* (Ledwina 2012), which allowed it to reach the international statistician community (Lehmann & Romano 2005, p. 210). The paper was recognized as a pioneering achievement in the field of statistical methodology for the analysis of causal relationships.

The purpose of this article is to present the possibilities of applying permutation methods in the analysis of dependencies. Berry, Johnston and Mielke Jr. (2018) present a permutation approach for generating resampling probability values for various measures of association. They show many examples of the practical use of permutation methods in association analysis. The article presents selected possibilities of applying permutation methods in the analysis of association between variables. The main focus is on various possibilities of using permutation methods and the possibilities of data labels rearranging. A proposed method to allow investigation of the significance of the relationship between two data sets is presented. The presented example illustrates the use of permutation methods for testing the significance of the relationship between two sets of variables. The proposed method is compared to the well-known canonical correlation analysis.

2. Permutation Methods

In mathematics, the term "permutation" denotes the act of rearranging objects or values in an ordered fashion. In statistics, the idea of permutation objects is used in several methods, especially in testing hypotheses (permutation tests).

Permutation tests permit us to choose the form of the test statistic. Through sample size reduction, permutation tests can reduce the costs of experiments and surveys. Permutation tests are the most powerful of statistical procedures. There are five steps in the process of permutation testing (Good 2005):

1. Identify the null and the alternative hypothesis.
2. Choose the form of the test statistic.
3. Calculate the test statistic for the sample data.
4. Determine the frequency distribution of the test statistic using data permutations.
5. Make a decision using this empirical distribution as a guide.

The basic idea in permutation testing is to generate a reference distribution by recalculating a test statistic for many permutations of the data. The term "permutation methods" should not, however, be limited to the problem of testing hypotheses. In recent years, various statistical methods have been proposed that refer to the idea of the permutation model introduced by J. Spława-Neyman (1923).

Corain, Arboretti and Bonnini (2016) present a novel permutation-based nonparametric approach for ranking several multivariate populations. Using data collected from both experimental and observational studies, it covers some of the most useful designs widely applied in research and industry investigations, such as the multivariate analysis of variance and multivariate randomized complete block designs.

Berry, Johnston and Mielke Jr. (2018) use rearranging data to generate probability values and measures of effect size for various measures of association. They define association for two interval-level variables, measures of association for two nominal-level variables or two ordinal-level variables, and measures of agreement for two nominal-level or two ordinal-level variables.

Berry, Mielke Jr. and Johnston (2016) provide a synthesis of many statistical tests and measures, which, at first sight, appear disjointed and unrelated. Numerous comparisons of permutation and classical statistical methods are presented, and the two classes are compared via probability values and, where appropriate, measures of effect size.

Mielke and Berry Jr. (2007) offer a broad treatment of statistical inference using permutation techniques. Its purpose is to make available to practitioners a variety of useful and powerful data analysis tools that rely on very few distributional assumptions. Although many of these procedures have appeared in journal articles, they are not readily available to practitioners.

3. Permutational Methods in Hypothesis Testing

3.1. General Remarks

The testing of statistical hypotheses is a major branch of study in classical statistical inference. Based on a relatively small sample, one can infer the characteristics of a much larger population. There are two basic kinds of statistical hypotheses: parametric and non-parametric. According to the parametric or non-parametric hypothesis, a parametric statistical test or a nonparametric test is used. Parametric tests require that the sample is taken from a specified distribution, usually the normal distribution. Permutation tests such as nonparametric tests do not require specific population distributions of the variables such as the normal distribution. These tests use data labels rearranging. A typical application of permutation tests is to compare the distributions of two or more populations based on two samples taken independently.

3.2. Methods of Data Permutations

The permutation model is based on the data labels permutation. O'Gorman (2012, p. 78) lists some permutation methods which can be used for testing the significance of the coefficient in the linear regression model. In the linear regression model there is one dependent variable Y and k independent variables $X_1, X_2, \ldots, X_k$. Some methods of rearranging data labels include:

– permute the dependent variable,
– permute the independent variable for the considered variable,
– permute the residuals from the reduced model,
– permute the residuals from the complete model.

To test the significance of a parameter in the linear regression model, various methods of rearranging data labels can be used. The results of testing the significance of the coefficient for different permutation methods are not equivalent. O'Gorman (2012) points out that it is not clear which method is superior.

Let us assume that $\boldsymbol{y} = [y_1, y_2, \ldots, y_k]^T$ and we want to shuffle this vector. To permute this vector we can use the square matrix of dimension $k \times k$

$\mathbf{P}_k = [a_{ij}]$ where a_{ij} are only zeros and ones for $i, j = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} a_{ij} = 1$ for $j = 1, 2, \ldots, k$ and $\sum_{j=1}^{k} a_{ij} = 1$ for $i = 1, 2, \ldots, k$.

The examples of a permutation matrix in the case $k = 4$ are:

$$\mathbf{P}_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ \mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \ \mathbf{P}_3 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

It is easy to notice that for the first matrix we have

$$\mathbf{P}_1 \boldsymbol{y} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} y_2 \\ y_3 \\ y_1 \\ y_4 \end{bmatrix}.$$

So vector $[y_2, y_3, y_1, y_4]^T$ is a rearrangement of vector $[y_1, y_2, y_3, y_4]^T$.
For the above-given permutation matrices $\mathbf{P}_1$, $\mathbf{P}_2$ and $\mathbf{P}_3$ we have

$$\mathbf{P}_1 \boldsymbol{y} = [y_2, y_3, y_1, y_4]^T,$$

$$\mathbf{P}_2 \boldsymbol{y} = [y_1, y_4, y_2, y_3]^T,$$

$$\mathbf{P}_3 \boldsymbol{y} = [y_2, y_3, y_4, y_1]^T.$$

3.3. Properties of Permutations

Let $\mathbf{A}$ be the vector given by $\mathbf{A} = [y_1, y_2, \ldots, y_k]^T$. The permutation of vector $\mathbf{A}$ is function α: $\mathbf{A} \to \mathbf{A}$ that is bijective (i.e. both one-to-one and onto).

Identity permutation. Let $\mathbf{D} = [d_{ij}]$ be the identity matrix such that $d_{ii} = 1$, and $d_{ij} = 0$ for $i \neq j$ then matrix $\mathbf{D}$ leads to the identity permutation. If $\mathbf{D}$ is the matrix of permutation α, then for each vector $\mathbf{A}$: $\alpha\mathbf{A} = \mathbf{A}$.

Composition of permutations. Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be two matrices of permutations α and β, then $\mathbf{S}_2\,\mathbf{S}_1$ is the matrix of the composition of permutations $\gamma = \beta\alpha$.

Inverse permutation. Let $\mathbf{S}$ be the matrix of permutation α, then $\mathbf{S}^{\mathbf{T}}$ is the matrix of inverse permutation β, that $\beta\alpha\mathbf{A} = \alpha\beta\mathbf{A} = \mathbf{A}$ for each vector $\mathbf{A}$.

Let $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$ be the matrices of permutation α, β and γ. Permutation composition is associative, e.g. $\mathbf{S}_3\,\mathbf{S}_2\,\mathbf{S}_1$ is the matrix of permutation $\gamma(\beta\alpha) = (\gamma\beta)\alpha$.

4. Permutation Methods for Regression Models

Pearson's correlation coefficient ρ tells us about the strength of the linear relationship between two variables *X* and *Y*. To perform a test of the significance of the correlation coefficient a random sample of size *n* should be taken. The sample data is used to calculate *r*, the correlation coefficient for the sample. Permutation tests use data labels rearranging to obtain the empirical distribution of the coefficient. The data should be permuted *N* times (*N* is usually greater or equal to 1000). The empirical distribution of *r* is obtained from *N* permuted sets of data. The decision of hypothesis H_0 is based on the value of *r* and its empirical distribution. The original two-dimensional data and three examples of randomly permuted data are presented in Figure 1.

X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
*x*1	*y*1	*x*1	*y*2	*x*1	*y*6	*x*1	*y*5
*x*2	*y*2	*x*2	*y*1	*x*2	*y*5	*x*2	*y*1
*x*3	*y*3	*x*3	*y*9	*x*3	*y*8	*x*3	*y*9
*x*4	*y*4	*x*4	*y*6	*x*4	*y*7	*x*4	*y*7
*x*5	*y*5	*x*5	*y*10	*x*5	*y*9	*x*5	*y*6
*x*6	*y*6	*x*6	*y*4	*x*6	*y*4	*x*6	*y*8
*x*7	*y*7	*x*7	*y*7	*x*7	*y*1	*x*7	*y*3
*x*8	*y*8	*x*8	*y*8	*x*8	*y*10	*x*8	*y*2
*x*9	*y*9	*x*9	*y*5	*x*9	*y*2	*x*9	*y*4
*x*10	*y*10	*x*10	*y*3	*x*10	*y*3	*x*10	*y*10

Fig. 1. Example of a Two-dimensional Data Set (*X*, *Y*) and Three Random Data Sets (*X*, *Y*) with Randomly Rearranged Variable *Y*

Source: author's own elaboration.

There are more possibilities of rearranging data labels in the case of a multiple regression model than in the case of testing the significance of the correlation coefficient. Let us consider the multiple regression linear model:

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_k X_k + e. \tag{1}$$

There are many ways of data labels rearranging for one dependent variable Y and k independent variables X_1, X_2, ..., X_k. There are some examples of possible ways of permutation of one dependent variable and k ($k > 2$) independent variables:

– permute the dependent variable Y (it is the same result as permute all independent variables $X_1, X_2, \ldots, X_k$ in the same way),

– permute only one independent variable X_i ($i = 1, 2, \ldots, k$),

– permute two independent variables X_i and X_j ($i, j = 1, 2, \ldots, k, i \neq j$) in the same way,

– permute two independent variables X_i and X_j ($i, j = 1, 2, \ldots, k, i \neq j$) independently,

– permute two independent variables X_i and X_j ($i, j = 1, 2, \ldots, k, i \neq j$) in the same way and the variable X_s ($s = 1, 2, \ldots, k$) independently.

Labels of original data should be rearranged to test the significance of the parameters of the linear model. The empirical distribution of the parameter is obtained based on the results for the N permuted models. Let S, S_1, S_2,..., S_k are the permutation matrices. Some typical methods of permutation for the multiple regression model (1) for $k = 3$ are:

$$Y = a_1 SX_1 + a_2 SX_2 + \cdots + a_k SX_k + e$$

$$Y = a_1 S_1 X_1 + a_2 XS_2 X_2 + \cdots + a_k S_k X_k + e$$

$$Y = a_1 X_1 + a_2 SX_2 + \cdots + a_k SX_k + e.$$

5. The Significance of the Association of Two Sets of Variables

In addition to the methods of studying the relationship between two variables or a dependent variable and many dependent variables, there are statistical methods that measure the association between two sets of variables **X** and **Y**. One of these methods is the canonical correlations analysis. Let **X** = (X_1, ..., X_p) and **Y** = (Y_1, ..., Y_q) be two sets of variables. To determine the strength of the association the permutation method could be used. The use of permutation methods requires data labels to be rearranged. These two sets can be rearranged in many ways. The variables in one set or two sets can be permuted independently or dependently.

The variables in one set (or two sets) could be grouped and permuted within these groups independently or dependently. The original set of variables for $p = q = 3$ and two examples of permuted sets are presented in Figure 2. In the first case (on the left), the variables X_1, X_2 and X_3 are permuted dependently. In the second case (on the right), all variables are permuted independently.

*Y*1	*Y*2	*Y*3	*X*1	*X*2	*X*3
*y*11	*y*21	*y*31	*x*11	*x*21	*x*31
*y*12	*y*22	*y*32	*x*12	*x*22	*x*32
*y*13	*y*23	*y*33	*x*13	*x*23	*x*33
*y*14	*y*24	*y*34	*x*14	*x*24	*x*34
*y*15	*y*25	*y*35	*x*15	*x*25	*x*35
*y*16	*y*26	*y*36	*x*16	*x*26	*x*36
*y*17	*y*27	*y*37	*x*17	*x*27	*x*37
*y*18	*y*28	*y*38	*x*18	*x*28	*x*38
*y*19	*y*29	*y*39	*x*19	*x*29	*x*39

*Y*1	*Y*2	*Y*3	*X*1	*X*2	*X*3
*y*11	*y*21	*y*31	*x*12	*x*22	*x*32
*y*12	*y*22	*y*32	*x*14	*x*24	*x*34
*y*13	*y*23	*y*33	*x*16	*x*26	*x*36
*y*14	*y*24	*y*34	*x*15	*x*25	*x*35
*y*15	*y*25	*y*35	*x*11	*x*21	*x*31
*y*16	*y*26	*y*36	*x*18	*x*28	*x*38
*y*17	*y*27	*y*37	*x*19	*x*29	*x*39
*y*18	*y*28	*y*38	*x*13	*x*23	*x*33
*y*19	*y*29	*y*39	*x*17	*x*27	*x*37

*Y*1	*Y*2	*Y*3	*X*1	*X*2	*X*3
*y*11	*y*23	*y*35	*x*12	*x*21	*x*31
*y*12	*y*22	*y*34	*x*14	*x*27	*x*32
*y*13	*y*21	*y*39	*x*13	*x*28	*x*36
*y*14	*y*25	*y*37	*x*18	*x*24	*x*33
*y*15	*y*24	*y*31	*x*15	*x*26	*x*34
*y*16	*y*26	*y*33	*x*17	*x*22	*x*38
*y*17	*y*29	*y*36	*x*16	*x*29	*x*39
*y*18	*y*28	*y*32	*x*11	*x*25	*x*37
*y*19	*y*27	*y*38	*x*19	*x*23	*x*35

Fig. 2. The Original Set of Two Sets of Variables (Top) and Two Examples of Possible Permutations of These Sets

Source: author's own elaboration.

6. Canonical Correlations

Pearson's correlation coefficient measures the strength of the linear dependency for two variables X and Y. The multiple linear regression model could be used to describe the dependency between the dependent variable Y and a set of independent variables $X_1, X_2, \ldots, X_k$. Sometimes the association between the two sets of variables **X** and **Y** should be considered. Canonical correlation analysis could be used in this case. This method was proposed by H. Hotelling in 1935–36. Canonical correlation analysis is employed to study the relationships between two variable sets when each variable set consists of at least two variables. The main objectives of canonical correlations are as follows (Thompson 1984):

– determining the strength of the relationships that may exist between the two sets,

– deriving a set of weights for each set of dependent and independent variables so that the linear combinations of each set are maximally correlated,

– explaining the nature of any relationships existing between the sets of dependent and independent variables.

Canonical correlation analysis develops several independent canonical functions that maximize the correlation between the linear composites, also known as canonical variates, which are sets of dependent and independent variables.

The form of Pearson's correlation coefficient:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}. \tag{2}$$

The form of the canonical correlation coefficient:

$$\rho = \max_{a \in R^p, b \in R^q} \frac{Cov(\boldsymbol{a}'\mathbf{X}, \boldsymbol{b}'\mathbf{Y})}{\sqrt{Var(\boldsymbol{a}'\mathbf{X})Var(\boldsymbol{b}'\mathbf{Y})}}. \tag{3}$$

Canonical correlation analysis constructs such vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ based on the following criteria:

1. The first canonical variate $\boldsymbol{U}_1 = \boldsymbol{a}'_1\boldsymbol{X}, \boldsymbol{V}_1 = \boldsymbol{b}'_1\boldsymbol{Y}$ is constructed from the maximization of (3).

2. The second canonical variate pair $\boldsymbol{U}_2 = \boldsymbol{a}'_2\boldsymbol{X}, \boldsymbol{V}_2 = \boldsymbol{b}'_2\boldsymbol{Y}$ is constructed from the maximization of (3) with the restriction that $\boldsymbol{D}^2(\boldsymbol{U}_2) = \boldsymbol{D}^2(\boldsymbol{V}_2) = 1$ and pairs $(\boldsymbol{U}_1, \boldsymbol{V}_1)$ and $(\boldsymbol{U}_2, \boldsymbol{V}_2)$ are uncorrelated.

3. At the k step, the k-th canonical variate pair $\boldsymbol{U_k} = \boldsymbol{a}'_k \boldsymbol{X}, \boldsymbol{V_k} = \boldsymbol{b}'_k \boldsymbol{Y}$ is obtained from the maximization of (3) with the restriction that $\boldsymbol{D}^2(\boldsymbol{U_k}) = \boldsymbol{D}^2(\boldsymbol{V_k}) = 1$ and $(\boldsymbol{U_k}, \boldsymbol{V_k})$ are uncorrelated with the previous $(k - 1)$ canonical variate pairs.

4. Repeat step 3 until the number of canonical variates $s = \min(p, q)$.

The first canonical correlation coefficient is denoted by r_1, the second by r_2, and so on. The number of calculated CCA coefficients is equal to the minimum of the number of variables in two considered sets $s = \min(p, q)$. For testing the significance of the first canonical correlation coefficient, the following statistics can be used: Wilks' Lambda, the Hotelling-Lawley Trace, the Pillai-Bartlett Trace, and Roy's Largest Root.

7. Testing the Significance of the Association of Two Sets of Variables

One problem to be considered in the canonical correlation analysis is to test the hypothesis that none of the canonical correlation coefficients r_1, r_2, … , r_s is significant. To test such a hypothesis we usually use Wilks' lambda statistic, the Hotelling-Lawley trace, the Pillai trace or Roy's largest root. The form of these statistics is as follows:

Wilk's lambda statistic:

$$\Lambda_1 = \prod_{i=1}^{m} (1 - r_i^2). \tag{4}$$

This statistic is distributed as the Wilks Λ-distribution. Rejection of the null hypothesis is for small values of Λ_1.

Pillai trace statistic:

$$V^{(m)} = \sum_{i=1}^{m} r_i^2. \tag{5}$$

Rencher and Christensen (2012, pp. 391–95) has tables providing critical values for this statistic.

Hotelling-Lawley trace statistic:

$$U^{(m)} = \sum_{i=1}^{m} \frac{r_i^2}{1 - r_i^2}. \tag{6}$$

Rencher and Christensen (2012, pp. 391–95) has tables providing critical values for this statistic.

Roy's largest root statistic:

$$\theta = r_1^2. \tag{7}$$

Rencher and Christensen (2012, pp. 391–95) present tables providing critical values for Roy's largest root statistic. This statistic is based only on one largest canonical correlation coefficient and the other statistics use all the canonical correlation coefficients. For this reason, Roy's largest root statistic will not be included in the computer simulation studies.

8. Monte Carlo Study

The proposed method of testing the significance of dependency between two sets of variables is based on the permutation method. In the proposed method, Wilk's lambda statistic (4) is used and the dependent rearrangement of one of the variable set labels is used (see Figure 2). The proposal was compared to the three well-known methods of testing the significance of correlations in the canonical analysis in the Monte Carlo study. The permutation had $N = 1000$ random rearrangements of data labels. The significance level $\alpha = 0.05$ was assumed in computer simulations.

In the first part of a computer simulation, two sets of variables **X** and **Y** were considered. All variables were normally distributed. The sample size was $n = 100$. The above assumptions are typical for the canonical correlation analysis.

The first set of variables is given by $\mathbf{X} = (X_1, X_2, X_3)$, where variables X_1, X_2, X_3 are independent and normally distributed with mean 10 and variance 1. The second set of variables is given by $\mathbf{Y} = (1 - \delta)\mathbf{Z} + \delta\mathbf{X}$, where $\mathbf{Z} = (Z_1, Z_2, Z_3)$, and variables Z_1, Z_2, Z_3 are independent and normally distributed with mean 10 and variance 1. The parameter δ ($\delta = 0.00, 0.05, \ldots, 0.25$) describes the strength of the association of two sets of variables. For $\delta = 0$ the sets **X** and **Y** are not associated.

The estimated probabilities of rejection H_0 (there is no association between **X** and **Y**) are presented in Table 1. The probabilities were estimated based on 1000 random samples. In the case of normally distributed variables with the sample of size $n = 100$, the size of all considered tests is close to the significance level α. The power of all tests is quite similar. The proposal has no advantage in this case.

Table 1. Estimated Probability of Rejection H_0 – Normal Distribution, $n = 100$

δ	Method			
	Permutation	Wilk's Lambda	Hotelling--Lawley trace	Pillai trace
0.00	0.051	0.053	0.054	0.049
0.05	0.081	0.080	0.082	0.075
0.10	0.199	0.202	0.204	0.193
0.15	0.497	0.495	0.499	0.487
0.20	0.866	0.861	0.861	0.862
0.25	0.993	0.991	0.990	0.989

Source: author's own elaboration.

The advantages of permutation tests are for the application of small samples and non-normally distributed variables. The next three simulations are based on normal, beta and gamma-distributed variables for small samples where $n = 20$.

The association of two sets **X** and **Y** of the three-dimensional variables were analysed in this part of the study. The considered sets are $\mathbf{X} = (X_1, X_2, X_3)$ and $\mathbf{Y} = (1 - \delta)\mathbf{Z} + \delta\mathbf{X}$ where $\mathbf{Z} = (Z_1, Z_2, Z_3)$, variables Z_1, Z_2, Z_3 are independently distributed and the parameter δ ($\delta = 0.00, 0.05, \ldots, 0.40$) describes the strength of association of two sets.

The parameters of random variables X_1, X_2, X_3 and Z_1, Z_2, Z_3 are as follows:

– normal distribution with expected value 10 and variance 1,
– beta distribution with shape parameters $s_1 = 2$ and $s_2 = 2$,
– gamma distribution with shape parameter $s = 2$.

The probabilities of rejection of H_0 are presented in Tables 2, 3, and 4. The probabilities were estimated based on 1000 random samples. In the permutation method for each sample, the data labels were randomly rearranged $N = 1000$ times. In the first case, samples were taken from the normal distribution but the sample size was $n = 20$. The results are presented in Table 2. In the case of normally distributed variables and small samples (Table 2), the test based on the permutation method has the greatest power. The three other tests have similar power. The results of the computer study for beta distributed variables are presented in Table 3.

Table 2. Estimated Probability of Rejection H_0 – Normal Distribution, $n = 20$

δ	Method			
	Permutation	Wilk's Lambda	Hotelling--Lawley trace	Pillai trace
0.00	0.046	0.041	0.052	0.033
0.05	0.069	0.064	0.078	0.047
0.10	0.070	0.074	0.087	0.046
0.15	0.088	0.090	0.110	0.056
0.20	0.186	0.173	0.187	0.144
0.25	0.285	0.263	0.278	0.226
0.30	0.496	0.468	0.479	0.438
0.35	0.726	0.688	0.682	0.664
0.40	0.907	0.873	0.858	0.859

Source: author's own elaboration.

Table 3. Estimated Probability of Rejection H_0 – Beta Distribution, $n = 20$

δ	Method			
	Permutation	Wilk's Lambda	Hotelling--Lawley trace	Pillai trace
0.00	0.056	0.058	0.074	0.042
0.05	0.057	0.055	0.071	0.041
0.10	0.069	0.062	0.073	0.045
0.15	0.101	0.095	0.111	0.066
0.20	0.155	0.149	0.172	0.114
0.25	0.252	0.242	0.253	0.202
0.30	0.495	0.449	0.457	0.411
0.35	0.701	0.653	0.649	0.639
0.40	0.904	0.887	0.870	0.883

Source: author's own elaboration.

The size of the Hotelling-Lawley test is greater than the significance level α in the case of beta distributed variables. The most powerful test is the test based on the permutation method. The results of the computer study for gamma-distributed variables are presented in Table 4.

Table 4. The Estimated Probability of Rejection H_0 – Gamma Distribution, $n = 20$

δ	Method			
	Permutation	Wilk's Lambda	Hotelling--Lawley trace	Pillai trace
0.00	0.055	0.061	0.068	0.039
0.05	0.059	0.074	0.083	0.043
0.10	0.075	0.080	0.091	0.057
0.15	0.122	0.127	0.141	0.089
0.20	0.170	0.173	0.193	0.136
0.25	0.313	0.291	0.316	0.258
0.30	0.498	0.474	0.480	0.436
0.35	0.725	0.701	0.695	0.681
0.40	0.903	0.887	0.876	0.879

Source: author's own elaboration.

The size of the Hotelling-Lawley test is greater than the significance level α in the case of gamma-distributed variables. The most powerful test is the test based on the permutation method.

9. Conclusions

Permutation tests are one type of permutation method. They are very useful for non-normally distributed data and small samples. However, permutation methods can be used not just for testing hypotheses. They can be used for constructing rankings of multivariate data or measuring the association between variables and sets of variables. In these methods, the form of data labels rearrangement is very important. The data labels can be permuted in many ways. Some of these methods of rearranging labels which can be used in the association analysis are described in the paper.

As an example of the use of the permutation method in association analysis, the method of testing the significance of the association between two sets of variables is presented in the paper. This method uses Wilk's lambda statistic and is based on the dependent rearranging of data labels in variables in one of the considered sets. The properties of the proposal were compared to the well-known tests in canonical correlation analysis: Wilk's lambda, the Lawley-Hotelling trace, and Pillai trace. The Monte Carlo study shows that for normally distributed variables and big samples the power of all tests is quite similar. The proposal is most powerful for non-normally distributed variables and small samples.

Bibliography

Berry, K. J., Johnston, J. E. and Mielke Jr., P. W. (2014) *A Chronicle of Permutation Statistical Methods*. New York: Springer International Publishing.

Berry, K. J., Johnston, J. E. and Mielke Jr., P. W. (2018) *The Measurement of Association. A Permutation Statistical Approach.* Cham: Springer Nature Switzerland.

Berry, K. J., Mielke Jr., P. W. and Johnston, J. E. (2016) *Permutation Statistical Methods. An Integrated Approach.* Cham: Springer Nature Switzerland.

Corain, L., Arboretti, R. and Bonnini, S. (2016) *Ranking of Multivariate Populations. A Paermutation Approach with Applications.* Boca Raton: CRC Press.

Good, P. (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer Science Business Media.

Ledwina, T. (2012) "Neyman Jerzy (1894–1981)" in: *Statystycy polscy.* Warszawa: Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne.

Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses*. New York: Springer-Verlag.

Mielke, P. W. and Berry K. J. Jr. (2007) *Permutation Methods. A Distance Function Approach*. New York: Springer Science+Business Media.

O'Gorman, T. W. (2012), *Adaptive Tests of Significance Using Permutations of Residuals with R and SAS.* John Wiley and Sons.

Rencher, A. C. and Christensen, W. F. (2012) *Methods of Multivariate Analysis*. Hoboken: Wiley.

Spława-Neyman, J. (1923) "Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych". *Rocznik Nauk Rolniczych* 10: 1–51.

Thompson, B. (1984) *Canonical Correlation Analysis: Uses and Interpretation*. London: Sage Publications.