Biuletyn WAT
Vol. LXIV, Nr 2, 2015

Testing the Homogeneity of Sequences of Multivariate Gaussian Random Variables

ZBIGNIEW WESOŁOWSKI

Military University of Technology, Faculty of Cybernetics,
Institute of Computer and Information Systems,
2 Kaliskiego Str., 00-908 Warsaw, Poland, zwesolowski@wat.edu.pl

Abstract. The paper discusses the method of examining the homogeneity of sequences of random variables including testing the normality of multivariate random variables as well as testing the equality of many expected values and testing the equality of many covariance matrices. Some results of application of the method to study the homogeneity of sequences of multivariate Gaussian random variables have been presented.
Keywords: homogeneity of sequences of multivariate random variables, multivariate normality tests, multivariate location equivalence tests, covariance matrix equivalence tests
DOI: 10.5604/12345865.1156932

1. Introduction

Many methods of mathematical statistics require making an assumption that the observations are realizations of homogeneous sequences of Gaussian random variables. This assumption is of particular importance in the analysis of correlation and regression, the discriminant analysis, the factor analysis, and the analysis of variance [1, 4, 10, 23, 30, 32].

The concept of homogeneity of sequences of multivariate random variables has many interpretations in mathematical statistics. The most important are the concept of homogeneity due to the expected value and the concept of homogeneity on account of the covariance matrix. The work has been devoted to discussing the most common used methods for testing the multivariate normal random variables (section 4), of methods for testing the equality of several expected values (section 5) and of methods for testing the equality of several covariance matrices (section 6). Testing has been carried out on the basis of results of statistical survey (section 2). The survey is conducted based on the plan formulated this way in order to ensure the required accuracy of estimating expected values from observations (point 3).

Let us enter markings: $\mathbb{N}$ — the set of natural numbers, $\mathbb{R}$ — the set of real numbers, $\mathbb{R}_0 = [0,\infty) \subset \mathbb{R}$, $\|\mathbf{x}\|$ — the norm of the vector $\mathbf{x}$, $\mathrm{F}(m, n)$ — the F-distribution with the pair (m, n), $(m, n \in \mathbb{N})$, degree of freedom, $\chi^2(k)$ — the chi-square distribution with k $(k \in \mathbb{N})$ degree of freedom, $N(0, 1)$ — the standard normal distribution, $N(\mu, \Gamma)$ — the multivariate normal distribution with the expected value $\mu \in \mathbb{R}^{p\times 1}$ and the covariance matrix $\Gamma \in \mathbb{R}^{p\times p}$.

2. Statistical Survey

Let $X = \{X_1, \ldots, X_k\}$ be a studied general population, where X_i is a statistical unit, for $i = 1,\ldots, k$, $k \in \mathbb{N}$. Let us suppose the population X is examined on account of a certain measurable statistical characteristic $\mathbf{x} \in \mathbb{R}^p$.

Let us consider the statistical survey $E = \{e_1, \ldots, e_k\}$ compound of many experiments e_i, $i \in K$, where $K = \{1, \ldots, k\}$ is a set of numbers of experiments. The experiment e_i relies on the multiple registrations of values of the statistical characteristic $\mathbf{x}$ observed in the statistical unit X_i.

Let,

$$P_E = \begin{Bmatrix} e_1 & e_2 & \cdots & e_k \\ n_1 & n_2 & \cdots & n_k \end{Bmatrix} \tag{1}$$

be a discrete plan of the survey E, where n_i is the number of repetitions of the experiment e_i, for $i \in K$.

The matrix $\mathbf{X}_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}] \in \mathbb{R}^{p\times n_i}$ is called the result of the experiment e_i (or *sample*), where $\mathbf{x}_{ij} = [x_{ij_1}, \ldots, x_{ij_p}]^T \in \mathbb{R}^{p\times 1}$ is the value of the examined statistical characteristic $\mathbf{x}$ registered within the j-th execution of the experiment e_i (or *observation*). Let $M_i = \{1, \ldots, n_i\}$ be a set of numbers of components of sample $\mathbf{X}_i$. The matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k] \in \mathbb{R}^{p\times n}$, $n = \sum_{i=1}^{k} n_i$, is called the result of the survey E (or *sample*).

The probabilistic space $(\Omega, \Xi, \mathrm{P})$ is called a probabilistic model of the survey E, where: $\Omega = \{\omega_j = \omega_{\mathbf{x}_j} : \mathbf{x}_j \in \mathbb{R}^{p\times 1}, j \in \mathbb{N}\}$ is a set of elementary events, ω_j is an elementary event meaning that the observed quantity assumes the value $\mathbf{x}_j \in \mathbb{R}^{p\times 1}$; Ξ is a set of random events being a σ-field of all subsets of the set Ω; $\mathrm{P} : \Xi \to \mathbb{R}_0$ is a probabilistic measure.

Let $\xi_i : \Omega \to \mathbb{R}^p$ be a random variable defined on the probabilistic space (Ω, Ξ, P). Lest us assume that the components $\mathbf{x}_{ij}$ of sample $\mathbf{X}_i$ are realizations of this random variable, i.e. $\xi_{i_{\omega_j}} = \mathbf{x}_{ij}$, for: $j \in M_i$, $i \in K$. The random variable $\xi_i = \left[\xi_{i_1}, \ldots, \xi_{i_p}\right]^T$ is called a stochastic model of the experiment e_i. The sequence of random variables $(\xi_i)_{i \in K}$ is called a stochastic model of the survey E.

Let $\mu_i = \mathbf{E}(\xi_i)$ and $\Gamma_i = \mathbf{E}\left[(\xi_i - \mu_i)(\xi_i - \mu_i)^T\right]$ are, respectively, the expected value and the covariance matrix of the random variable ξ_i.

Definition 1. The sequence of random variables $(\xi_i)_{i \in K}$ is *homogenous due to the expected value*, if $\|\mu_i\| < \infty$, for $i \in K$, and if $\mu_i = \mu_j$, for $i, j \in K$.

□

The sequence of random variables $(\xi_i)_{i \in K}$ is *heterogeneous due to the expected value*, if at least one random variable from this sequence differs than other with expected value or her expected value has an infinite norm.

Definition 2. The sequence of random variables $(\xi_i)_{i \in K}$ is *homogenous due to the covariance matrix* (or *homoscedastic*), if $\|\Gamma_i\| < \infty$, for $i \in K$, and if $\Gamma_i = \Gamma_j$, for $i, j \in K$.

□

The sequence of random variables $(\xi_i)_{i \in K}$ is *heterogeneous due to the covariance matrix* (or *heteroscedastic*), if at least one random variable from this sequence differs than other with covariance matrix or her covariance matrix has an infinite norm.

3. Estimating the Number of Repetitions of Experiments

Suppose that the aim of the experiment e_i is to designate an approximately 100(1-α)% confidence interval for the expected value $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}$ from sample $\mathbf{X}_i$, where α is a fixed level of significance.

Let

$$\hat{\mu}_i^{(0)} = \frac{1}{n_0}\sum_{j=1}^{n_0} \mathbf{x}_{ij}^{(0)} \text{ and } \hat{\Gamma}_i^{(0)} = \frac{1}{n_0 - 1}\sum_{j=1}^{n_0} \left(\mathbf{x}_{ij}^{(0)} - \hat{\mu}_i^{(0)}\right)\left(\mathbf{x}_{ij}^{(0)} - \hat{\mu}_i^{(0)}\right)^T =$$

$$= \begin{bmatrix} \hat{\gamma}_{i_{11}}^{(0)} & \hat{\gamma}_{i_{12}}^{(0)} & \cdots & \hat{\gamma}_{i_{1p}}^{(0)} \\ \hat{\gamma}_{i_{21}}^{(0)} & \hat{\gamma}_{i_{22}}^{(0)} & \cdots & \hat{\gamma}_{i_{2p}}^{(0)} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_{i_{p1}}^{(0)} & \hat{\gamma}_{i_{p2}}^{(0)} & \cdots & \hat{\gamma}_{i_{pp}}^{(0)} \end{bmatrix}$$

are, respectively, the expected value and the covariance matrix from the pilot sample $\mathbf{X}_i^{(0)} = \left[\mathbf{x}_{i1}^{(0)}, \ldots, \mathbf{x}_{in_0}^{(0)}\right] \in \mathbb{R}^{p \times n_0}$, where: $\hat{\gamma}_{i_{ab}}^{(0)} = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left(x_{ij_a}^{(0)} - \hat{\mu}_{i_a}^{(0)}\right)\left(x_{ij_b}^{(0)} - \hat{\mu}_{i_b}^{(0)}\right)$ is a covariance coefficient of random variables $\xi_{i_a} \in \xi_i$ and $\xi_{i_b} \in \xi_i$ from this sample, for $a, b = 1, \ldots, p$, $n_0 = 15$ is the cardinality of the pilot sample calculated by a method used to estimate a proportion with specified precision [7].

The number of repetitions $n_i \in \mathbb{N}$ of the experiment e_i can be determined from the formula,

$$n_i = \max\left\{n_{i_a}; a = 1, \ldots, p\right\}, \tag{2}$$

where $n_{i_a} \in \mathbb{N}$ is the number of repetitions of the experiment e_i needed to provide an approximately 100(1-α)% confidence interval for the expected value $\hat{\mu}_{i_a}^{(0)} = \frac{1}{n_0} \sum_{j=1}^{n_0} x_{ij_a}^{(0)}$ from sample $\mathbf{X}_i^{(0)}$. This number satisfies the inequality [7, 23, 25, 30, 32] of the form,

$$n_{i_a} \geq \lambda_{1-\alpha/2}^2 \hat{\gamma}_{i_{a,a}}^{(0)},$$

where: $\lambda_{1-\alpha/2}$ is a quantile of the rank (1-α/2) of the distribution $N(0, 1)$.

Because many tests applied for examining the homogeneity of sequences of multivariate random variables are characterized by low power when the cardinalities of the samples $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are different, and taking into account the fact that statistical analyzes are often conducted on the basis of data generated by the simulation programs [9, 16], the plan P_E, Eq. (1), should be constructed so that the number of repetitions of experiments $e_1, \ldots, e_k$ was the same. In this case, the plan of the survey takes the form,

$$P_E = \begin{Bmatrix} e_1 & e_2 & \cdots & e_k \\ m & m & \cdots & m \end{Bmatrix},$$

where $m \in \mathbb{N}$ is determined from the formula,

$$m = \max\left\{n_1, \ldots, n_k\right\}, \tag{3}$$

where n_i, Eq. (2), $i \in K$.

4. Testing the Normality of Distributions of Multivariate Random Variables

Let $\mathrm{F}_i(\mathbf{x}) = \mathrm{P}\left(x_i \le \xi_{i_1}, \ldots, x_p \le \xi_{i_p}\right) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_p} \mathrm{f}_i\left(y_1, \ldots, y_p\right) \mathrm{d}\, y_1, \ldots, \mathrm{d}\, y_p$ be a distribution function of the random variable ξ_i, where:

$$\mathrm{f}_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \sqrt{\det(\Gamma_i)}} \mathrm{e}^{\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Gamma_i^{-1} (\mathbf{x}-\mu_i)\right]}$$

is a density function of this random variable, $\mathbf{x} = \left[x_1, \ldots, x_p\right]^T$.

Let $\mathrm{F}_0(\mathbf{x}) = \mathrm{P}\left(x_1 \le \xi_{0_1}, \ldots, x_p \le \xi_{0_p}\right)$ be a theoretical distribution function, i.e. the distribution function of a random variable $\xi_0 : \Omega \to \mathbb{R}^p$ with the distribution $N(\mu_0, \Gamma_0)$, where: $\xi_0 = \left[\xi_{0_1}, \ldots, \xi_{0_p}\right]^T$.

The problem of testing the normality of the random variable ξ_i consists in the verification of hypotheses on the form,

$$\mathrm{H}_0 : \mathrm{F}_i(\mathbf{x}) = \mathrm{F}_0(\mathbf{x}), \tag{4}$$

$$\mathrm{H}_1 : \mathrm{F}_i(\mathbf{x}) \ne \mathrm{F}_0(\mathbf{x}). \tag{5}$$

To solve the problem, Eqs. (4)-(5) proposes a number of different tests. These include: Mardia's tests, Henze-Zirkle's test, Royston's H test, and the Baringhaus-Henze-Epps-Pulley (BHEP) test.

The Mardia tests [20-23, 27] are based on multivariate extensions of skewness and kurtosis measures. The Henze-Zirkler test is based on a functional distance measure between two distribution functions $\mathrm{F}_i(\mathbf{x})$ and $\mathrm{F}_0(\mathbf{x})$. In the works [12, 14], there are demonstrated that if the observations have multivariate normal distribution, then the test statistic has approximately the log-normal distribution. This enables constructing a rejection region of the test. The BHEP test [8, 13, 27] is based on the weighted distance between the empirical characteristic function and the parametric estimate of the characteristic function.

The most commonly used are Mardia's tests. The Mardia skewness test uses the statistics of the form,

$$\mathrm{A}_i = \frac{1}{6n} \sum_{j=1}^{n_i} \sum_{l=1}^{n_i} \left[\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)^T \hat{\Gamma}_i^{-1} \left(\mathbf{x}_{il} - \hat{\mu}_i\right)\right]^3, \tag{6}$$

where: $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$ and $\hat{\Gamma}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)^T$ are, respectively, an evaluation of the expected value μ_i and an evaluation of the covariance matrix Γ_i from sample $\mathbf{X}_i$ of the random variable ξ_i.

The statistics A_i, Eq. (6), has, at the truth of the hypothesis H_0, Eq. (4), and for $n_i \to \infty$, the distribution χ_q^2, where $q = \frac{1}{6}p(p+1)(p+2)$. The rejection region of this test takes the form $C = \left\{\mathrm{A}_i : \mathrm{A}_i < \chi^2_{\alpha/2}(q) \vee \mathrm{A}_i > \chi^2_{1-\alpha/2}(q)\right\}$, where $\chi_u^2(q)$ is a quantile of the rank u of the distribution $\chi^2(q)$.

The Mardia kurtosis test uses the statistics of the form,

$$B_i = \sqrt{\frac{n_i}{8p(p+2)}}\left\{\frac{1}{n_i}\sum_{j=1}^{n_i}\left[\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)^T \hat{\Gamma}_i^{-1}\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)\right]^2 - p(p+2)\right\}. \tag{7}$$

The statistics B_i, Eq. (7), has, at the truth of the hypothesis H_0, Eq. (4), and for $n_i \to \infty$, the distribution $N(0, 1)$. The rejection region of this test takes the form $C = \{\mathrm{B}_i : |\mathrm{B}_i| > \lambda_{1-\alpha/2}\}$.

5. Testing the Homogeneity of Sequences of Multivariate Random Variables due to the Expected Value

The problem of testing the homogeneity of sequences of multivariate random variables $(\xi_i)_{i\in K}$ due to the expected value consists in the verification of hypotheses on the form,

$$H_0 : \mu_1 = \cdots = \mu_k, \tag{8}$$

$$H_1 : \mu_i \neq \mu_j \text{ for at least one pair } (i, j); i \neq j; i, j = 1,\ldots,k. \tag{9}$$

To solve the problem, Eqs. (8)-(9), proposes a number of different tests. In the paper [5] it is proposed a nonparametric multisample test based on the classical Kruskal-Wallis statistics [17, 18]. In the works [25, 26], there are proposed modifications of the Kruskal-Wallis statistics by using the componentwise ranking [11]. In the paper [6], the proposed test uses averaging the Euclidean norms of covariance matrices. The principal disadvantage of these tests is that their power depends on the form of the covariance matrix and it is very low when the random variables $\xi_1,\ldots,\xi_k$ are correlated.

The most commonly used is the T^2 Lawley-Hotelling test [15, 19, 33] using the statistics of the form,

$$T^2 = \sum_{i=1}^{k} n_i \left(\hat{\mu}_i - \hat{\mu}\right)^T \hat{A}^{-1} \left(\hat{\mu}_i - \hat{\mu}\right), \tag{10}$$

where:

$$\hat{\mu}_i = \frac{1}{n_i} \sum\nolimits_{j=1}^{n_i} \mathbf{x}_{ij} \text{ for } i = 1, \ldots, k;$$

$$\hat{\mu} = \frac{1}{n} \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{n_i} \mathbf{x}_{ij};$$

$$\hat{A} = \frac{1}{n-k} \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \hat{\mu}_i\right)\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)^T.$$

The statistics T^2, Eq. (10), has, at the truth of the hypothesis H_0, Eq. (8), and for: $n_1 \to \infty, \ldots, n_k \to \infty$, the distribution $\chi^2\left[p(k-1)\right]$. The rejection region of this test takes the form $C = \left\{T^2 : T^2 > \chi^2_{1-\alpha}\left[p(k-1)\right]\right\}$, where $\chi^2_{1-\alpha}\left[p(k-1)\right]$ is the quantile of the rank $(1-\alpha)$ of the distribution $\chi^2\left[p(k-1)\right]$.

6. Testing the Homogeneity of Sequences of Multivariate Random Variables due to the Covariance Matrix

The problem of testing the homogeneity of sequences of multivariate random variables $\left(\xi_i\right)_{i\in K}$ due to the covariance matrix consists in the verification of hypotheses on the form,

$$H_0 : \Gamma_1 = \cdots = \Gamma_k, \tag{11}$$

$$H_1 : \Gamma_i \neq \Gamma_j, \text{ for at least one pair } (i, j); i \neq j; i, j = 1, \ldots, k \tag{12}$$

To solve the problem, Eqs. (11)-(12), it is commonly applied the M Box test [1-3, 24, 30] using the statistics of the form,

$$M = (n-k)\log\left[\det(\mathbf{A})\right] - \sum_{i=1}^{k} (n_i - 1)\log\left[\det(\mathbf{A}_i)\right], \tag{13}$$

where:

$$\mathbf{A}_i = \frac{1}{n_i - 1} \sum\nolimits_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \hat{\mu}_i\right)\left(\mathbf{x}_{ij} - \hat{\mu}_i\right)^T, \hat{\mu}_i = \frac{1}{n_i} \sum\nolimits_{j=}^{n_i} \mathbf{x}_{ij}, \mathbf{A} = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)\mathbf{A}_i,$$

$i = 1, \ldots, k$, $n = \sum\nolimits_{i=1}^{k} n_i$, for: $n_i > 1$, $\det(\mathbf{A}_i) > 0$, $i \in K$, $\det(\mathbf{A}) > 0$, $k > 1$, $n > k$.

The rejection region for the statistics M, Eq. (13), takes, at the truth of the hypothesis H_0, Eq. (11), and for: $n_1 \to \infty$, …, $n_k \to \infty$, the form $C = \{\gamma \mathrm{M} : \gamma \mathrm{M} \geq \mathrm{F}_{1-\alpha}(f_1, f_2)\}$, where $\mathrm{F}_{1-\alpha}(f_1, f_2)$ is the quantile of the rank (1-α) of the distribution $\mathrm{F}(f_1, f_2)$, where:

$$\gamma = \frac{a - f_1/f_2}{f_1},$$

$$a = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum_{i=}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k} \right),$$

$$f_1 = \frac{1}{2} p(p+1)(k-1),$$

$$f_2 = \frac{f_1 + 2}{\left| b - (1-a)^2 \right|},$$

$$b = \frac{(p-1)(p+2)}{6(k-1)} \left(\sum_{i=}^{k} \frac{1}{(n_i - 1)^2} - \frac{1}{(n-k)^2} \right).$$

Using the M Box test, one should take into account the fact that it features a low power when the numbers of samples $\mathbf{X}_1, \ldots, \mathbf{X}_k$ are different and when the observed distributions of these samples are different from the multivariate normal distribution [32].

Example 1. Let us consider the issue of testing the homogeneity of the sequence of multivariate random variables $(\xi_i)_{i \in K}$ due to the expected value and on account of the covariance matrix, where: $\xi_i : \Omega \to \mathbb{R}^p$, $p = 3$, $i \in K$, $K = \{1, \ldots, k\}$, $k = 3$.

Testing the hypotheses, Eqs. (4)-(5), Eqs. (8)-(9), and Eqs. (12)-(13) will be conducted on the basis of three samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(3)}$, where: $\mathbf{X}^{(v)} = \left[\mathbf{X}_1^{(v)}, \ldots, \mathbf{X}_k^{(v)}\right] \in \mathbb{R}^{p \times n^{(v)}}$, $\mathbf{X}_i^{(v)} = \left[\mathbf{x}_{i1}^{(v)}, \ldots, \mathbf{x}_{im^{(v)}}^{(v)}\right] \in \mathbb{R}^{p \times m^{(v)}}$, $n^{(v)} = m^{(v)}k$, $\mathbf{x}_{ij}^{(v)} \in \mathbf{X}_i^{(v)}$ is a realization of the random variable $\xi_i^{(v)} : \Omega \to \mathbb{R}^p$, $\xi_{i_{\omega_j}}^{(v)} = \mathbf{x}_{ij}^{(v)}$, for: $j = 1, \ldots, m^{(v)}$, $m^{(v)}$, Eq. (3), $v \in V$, $V = \{1, 2, 3\}$ is a set of numbers of samples, $m^{(1)} = 12$, $m^{(2)} = 18$, $m^{(3)} = 26$. The components of sample $\mathbf{X}_i^{(1)} \in \mathbf{X}^{(1)}$ are realizations of the random variable $\xi_i^{(1)} : \Omega \to \mathbb{R}^p$ of the distribution $N\left(\mu_i^{(1)}, \Lambda_i^{(1)}\right)$, $\mu_i^{(1)} = [0,0,0]^T \in \mathbb{R}^{p \times 1}$, $\Lambda_i^{(1)} = \mathbf{I}_p \sigma^2$, $\sigma^2 = 1$, for $i \in K$. The components of sample $\mathbf{X}_i^{(2)} \in \mathbf{X}^{(2)}$ are realizations of the random variable $\xi_i^{(2)} : \Omega \to \mathbb{R}^p$ of the distribution $N\left(\mu_i^{(2)}, \Lambda_i^{(2)}\right)$, $\mu_i^{(2)} = [i,i,i]^T \in \mathbb{R}^{p \times 1}$, $\Lambda_i^{(2)} = \mathbf{I}_p \sigma^2$, $\sigma^2 = 1$, for $i \in K$. The components of the sample

$\mathbf{X}_i^{(3)} \in \mathbf{X}^{(3)}$ are realizations of the random variable $\xi_i^{(3)} : \Omega \to \mathbb{R}^p$ of the distribution $N\left(\mu_i^{(3)}, \Lambda_i^{(3)}\right)$, $\mu_i^{(3)} = [0,0,0]^T \in \mathbb{R}^{p\times 1}$, $\Lambda_i^{(3)} = \mathbf{I}_p \sigma_i^2$, $\sigma_i^2 = i$, for $i \in K$.

Table 1 shows the results of the Mardia kurtosis test, where: $\hat{\mathrm{B}}\left(\mathbf{X}_i^{(v)}\right)$ is an evaluation of the statistics B_i, Eq. (7), from sample $\mathbf{X}_i^{(v)}$, $\hat{\alpha}^*\left[\hat{\mathrm{B}}_i\left(\mathbf{X}_i^{(v)}\right)\right]$ is an evaluation of the p-value α^* for the fixed value of the statistics B_i, for $i \in K$, $v \in V$. This table appears that at the level of significance $\alpha = 0.05$ there is no reason to reject the hypothesis H_0, Eq. (4), for all samples.

TABLE 1

The results of the Mardia kurtosis test

$v \in V$	$i \in K$	$\hat{\mathrm{B}}_i(\mathbf{X}_i^{(v)})$	$\hat{\alpha}^*\left[\hat{\mathrm{B}}_i(\mathbf{X}_i^{(v)})\right]$
1	1	−0.915425	0.359968
	2	−0.210487	0.833287
	3	−0.330638	0.740918
2	1	1.26529	0.205766
	2	0.174891	0.861165
	3	0.357914	0.720408
3	1	−0.719469	0.471852
	2	0.160924	0.872153
	3	−0.925162	0.354882

Table 2 shows the results of the T^2 Lawley-Hotelling test, where $\hat{\mathrm{T}}^2\left(\mathbf{X}^{(v)}\right)$ is an evaluation of the statistics T^2, Eq. (10), from the sample $\mathbf{X}^{(v)}$, for $v \in V$. This table appears that at the level of significance $\alpha = 0.05$ there is no reason to reject the hypothesis H_0, Eq. (8), for the sample $\mathbf{X}^{(1)}$, however one should reject this hypothesis for samples $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$.

TABLE 2

The results of the T^2 Lawley-Hotelling test

$v \in V$	$\hat{\mathrm{T}}^2(\mathbf{X}^{(v)})$	$\hat{\alpha}^*\left[\hat{\mathrm{T}}^2(\mathbf{X}^{(v)})\right]$
1	5.08816	0.467443
2	567.804	$2.04808 \cdot 10^{-119}$
3	14.1219	0.0283039

Table 3 shows the results of the M Box test, where $\hat{\mathrm{T}}^2\left(\mathbf{X}^{(v)}\right)$ is an evaluation of the statistics M, Eq. (13), from the sample $\mathbf{X}^{(v)}$, for $v \in V$. This table appears that at the level of significance $\alpha = 0.05$ there is no reason to reject the hypothesis H_0, Eq. (11), for sample $\mathbf{X}^{(1)}$, however one should reject this hypothesis for samples $\mathbf{X}^{(2)}$ and $\mathbf{X}^{(3)}$.

TABLE 3

The results of the M Box test

$v \in V$	$\hat{\mathrm{M}}(\mathbf{X}^{(v)})$	$\hat{\alpha}^*\left[\hat{\mathrm{M}}(\mathbf{X}^{(v)})\right]$
1	6.36104	0.393843
2	14.3148	0.0284822
3	81.3157	$3.383 \cdot 10^{-15}$

7. Summary and Future Work

The article discusses the method of testing the homogeneity of sequences of random variables that uses multivariate normality tests, tests for examining the homogeneity of many expected values and tests for examining the homogeneity of many covariance matrices. Hypothesis testing is performed on the basis of observations gathered in the execution of statistical surveys conducted by the plan P_E, Eq. (1). The work proposes a method for estimating the number of repetitions of experiments providing an approximately 100(1-α)% confidence interval for the expected value of the sequence of multivariate random variables.

Directions of further work should address the issue of examining the power of new multivariate normality tests [32, 33] and of examining the homogeneity of many expected values [33]. In particular, emphasis should be placed on developing tests for examining the homogeneity of many covariance matrices, because commonly used the M Box test is extremely inefficient when the samples, on the basis of which covariance matrices are evaluated, have different sizes and the observed distributions of these samples are different from the multivariate normal distribution.

Received June 13 2014. Revised April 22 2015.

BIBLIOGRAPHY

[1] ANDERSON T.W., *An Introduction to Multivariate Statistical Analysis*, Wiley-Interscience, Hoboken, 2003.

[2] BOX G.E.P., *A General Distribution Theory for a Class of Likelihood Criteria*, Biometrika, 36, 1949, pp. 317-346.

[3] BOX G.E.P., *Problems in the Analysis of Growth and Wear Curves*, Biometrics, 6, 1950, pp. 362-389.

[4] Box G.E.P., Jenkins G.M., Reinsel G.C., *Time Series Analysis: Forecasting and Control*, Wiley, Hoboken, 2011.

[5] Chenouri S., Small C.G., *A Nonparametric Multivariate Multisample Test Based on Data Depth*, Electronic Journal of Statistics, vol. 6, 2012, pp. 760-782.

[6] Choi K., Marden J., *An Approach to Multivariate Rank Tests in Multivariate Analysis of Variance*, Journal of the American Statistical Association, vol. 92, no. 440, Dec., 1997, pp. 1581-1590.

[7] Desu M.M., *Sample Size Methodology*, Academic Press, New York, 1990.

[8] Epps T.W., Pulley L.B., *A Test for Normality Based on the Empirical Characteristic Function*, Biometrika 70 (3), 1983, pp. 723-726.

[9] Fishman G.S., *Discrete-Event Simulation: Modelling, Programming, and Analysis*, Springer, New York, 2001.

[10] Gentle J.E. (ed.), Härdle W.K. (ed.), MORi Y. (ed.), *Handbook of Computational Statistics: Concepts and Methods*, Springer, Berlin, 2004.

[11] Gonzalez T.F. (ed.), *Handbook of Approximation Algorithms and Metaheuristics*, Chapman and Hall/CRC, Boca Raton, 2007.

[12] Henze N., Zirkler B., *A Class of Invariant and Consistent Tests for Multivariate Normality*, Communications in Statistics — Theory and Methods 19, Issue 10, 1990, pp. 3595-3617.

[13] Henze N., Wagner T., *A New Approach to the BHEP Tests for Multivariate Normality*, Journal of Multivariate Analysis, vol. 62, issue 1, July 1997, pp. 1-23.

[14] Henze N., *Invariant Tests for Multivariate Normality: a Critical Review*, Statistical Papers, 43, 2002, pp. 467-506.

[15] Hotelling H., *A Generalized T Test and Measure of Multivariate Dispersion*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1951, University of California Press, Berkeley, pp. 23-41.

[16] Kleijnen J.P.C., *Design and Analysis of Simulation Experiments*, Springer, 2010.

[17] Kruskal W.H., *A Nonparametric Test for the Several Sample Problem*, The Annals of Mathematical Statistics, vol. 23, no. 4, Dec., 1952, pp. 525-540.

[18] Kruskal W.H., Wallis W.A., *Use of Ranks in One-Criterion Variance Analysis*, Journal of the American Statistical Association, vol. 47, no. 260, Dec., 1952, pp. 583-621.

[19] Lawley D.N., *A Generalization of Fisher's z-Test*, Biometrika 30, 1938, pp. 180-187.

[20] Mardia K.V., *Measures of Multivariate Skewness and Kurtosis with Applications*, Biometrika 57, 1970, pp. 519-530.

[21] Mardia K.V., *Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies*, Sankhyā: The Indian Journal of Statistics, Series B (1960-2002), vol. 36, no. 2, May, 1974, pp. 115-128.

[22] Mardia K.V., Zemroch P.J., *Algorithm AS 84: Measures of Multivariate Skewness and Kurtosis*, Journal of the Royal Statistical Society, Series C (Applied Statistics), vol. 24, no. 2, 1975, pp. 262-265.

[23] Mardia K.V., Kent J.T., Bibby J.M., *Multivariate Analysis*, Academic Press, New York, 1979.

[24] Pederzoli G., Rathie P.N., *The Exact Distribution of Bartlett's Criterion for Testing the Equality of Covariance Matrices*, Metron. International Journal of Statistics, issue: 3/4, 1983, pp. 83-89.

[25] Puri M.L., Sen P.K., *Nonparametric Methods in Multivariate Analysis*, John Wiley and Sons, New York, 1971.

[26] Randles R.H., Peters D., *Multivariate Rank Tests for the Two-Sample Location Problem*, Communications in Statistics — Theory and Methods, vol. 19, issue 11, 1990, pp. 4225-4238.

[27] Romeu J.L., Ozturk A., *A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality*, Journal of Multivariate Analysis, vol. 46, issue 2, August 1993, pp. 309-334.

[28] Royston J.P., *Approximating the Shapiro-Wilk W-test for Non-Normality*, Statistics and Computing, 2, 1992, pp. 117-119.

[29] Royston J.P., *Remark AS R94*: *A Remark on Algorithm AS 181: the W Test for Normality*, Applied Statistics, 44, 1995, pp. 547-551.

[30] Seber G.A.F., *Multivariate Observations*, New York: John Wiley & Sons, 1984.

[31] Székelya G.J., Rizzob M.L., *A New Test for Multivariate Normality*, Journal of Multivariate Analysis, vol. 93, issue 1, March 2005, pp. 58-80.

[32] Tabachnick B.G., Fidell L.S., *Using Multivariate Statistics*, Pearson, Harlow, 2013.

[33] Um Y., Randles R.H., *Nonparametric Tests for the Multivariate Multi-Sample Location Problem*, Statistica Sinica, 8, 1998, pp. 801-812.

Z. WESOŁOWSKI

Testowanie jednorodności ciągów wielowymiarowych gaussowskich zmiennych losowych

Streszczenie. W pracy omówiono metodę badania jednorodności ciągów wielowymiarowych zmiennych losowych obejmującą badanie normalności rozkładów zmiennych losowych oraz badanie równości wielu wartości oczekiwanych i równości wielu macierzy kowariancji. Przedstawiono przykładowe wyniki zastosowania prezentowanej metody do badania jednorodności ciągów wielowymiarowych gaussowskich zmiennych losowych.

Słowa kluczowe: jednorodność wielowymiarowych zmiennych losowych, testy normalności rozkładów wielowymiarowych zmiennych losowych, testy równości wielu wartości oczekiwanych wielowymiarowych zmiennych losowych, testy równości wielu macierzy kowariancji

<u>Badania</u>
Publikacja została częściowo sfinansowana przez NCBiR w ramach projektu Nr BIO4/006/13143/2013 pt. „Elektroniczny system zarządzania cyklem życia dokumentów o różnych poziomach wrażliwości".