On Some New *k*-Samples Tests for Testing the Homogeneity of Distribution Laws

Boris Yu. Lemeshko, Irina V. Veretelnikova
Novosibirsk State Technical University, Novosibirsk, Russia

***Abstract* – Properties of *k*-samples tests for testing the homogeneity of distribution laws have been investigated. Tests have been proposed in which the maximum of the two-samples statistics of the Smirnov, the Lehmann-Rosenblatt and the Anderson-Darling tests as applied by pairs to the k-samples was utilized. Models of limiting statistics distributions for the proposed tests as well as for the Anderson-Darling *k*-sample test have been made. Comparative analysis of the power of the tests has been done.**

***Key words* – *k*-samples tests, homogeneity test, power of test, statistics modeling.**

I. Introduction

THE NECESSITY of solving the task of checking the hypotheses of two (or more) samples of random values belonging to the same universe population (the homogeneity test) may arise in different areas. For example, this task may arise naturally when checking the measurement means and trying to be certain that the random measurement errors distribution law has not undergone any serious changes within some time period.

The task of testing the homogeneity of k-samples can be stated as follows. Assume x_{ij} as the j^{th} observation in the set of order statistics of i^{th} sample $j=\overline{1,n_i}$, $i=\overline{1,k}$. Let us assume that the i^{th} sample correlates with the continuous distribution function of $F_i(x)$. It is required to check the hypothesis of $H_0: F_1(x)=F_2(x)=\ldots=F_k(x)$ type without defining the common distribution law.

As a rule, the two-sample Smirnov [1] and Lehmann-Rosenblatt [1, 2, 3] tests are used here. The national scientific letters fail to mention the use of the two-sample Anderson-Darling [4] (Anderson-Darling–Pettitt) or, moreover, the use of the k-sample Anderson-Darling version (AD) [5] or the Zhang tests (with statistics of Z_{K}, Z_{A}, Z_C) [6, 7, 8].

The present work presents and deals with some new k-samples tests which are based on the use of two sampling analogues as applied to the aggregate of k samples.

II. Problem Definition

Various approaches to formulating k-samples sampling analogues of the Smirnov, Lehmann–Rosenblatt and Anderson-Darling tests have been dealt with in [9]. The k-sample version of the Kolmogorov–Smirnov test, based on this approach, was formulated in [10] and is dealt with in the successive revisions of the book [11]. The same approach is the basis for the k-samples Anderson-Darling test viewed in [5]. In these tests, as well as the Zhang homogeneity test, a combined sample is created, and the statistics measure the deviation of empiric distributions of individual samples from the empiric deviation built on the aggregate of the samples under analysis.

It must be noted that the models of limiting statistics distribution models for the k-samples Anderson-Darling test [5] were built in [12–14]. The power of the k-samples Anderson-Darling test [5] and the Zhang tests with statistics of Z_{K}, Z_{A}, Z_C [6, 7, 8] was studied in [14, 15].

A different way of building a k-samples test is possible. It is possible to apply a two-samples test with the statistics of S (totaling $(k-1)k/2$ versions) to analyze the k samples, and the decision will be made on the cumulative result. Statistics of the type below can be taken as an example statistics for that k-samples test (when using the right-hand two-samples test)

$$S_{\max} = \max_{\substack{1\le i\le k \\ i<j\le k}} \{S_{i,j}\}, \tag{1}$$

where $S_{i,j}$ is the values of the two-samples test statistics as resulting from the analysis of i^{th} and j^{th} samples.

The H_0 hypothesis under verification will deviate with **greater** values of $S_{\max}$. Another advantage of this test is that as a result we will define a pair of samples the difference between which will be the most significant from the standpoint of the utilized two-samples test.

Statistics of the two-samples tests by Smirnov (preferably the modified versions thereof [16]), Lehmann–Rosenblatt and Anderson-Darling can be used as $S_{i,j}$. In this case the distributions of the corresponding statistics of $S_{\max}$ will converge to some limiting ones, and the models thereof can be found based of the results of statistic modeling.

978-1-5386-7054-5/18/$31.00 ©IEEE

III. STATISTICAL TESTS

A. k -samples version of the Smirnov test

The two-samples homogeneity test by Smirnov was offered in [17]. It is assumed that the distribution functions of $F_1(x)$ and $F_2(x)$ are continuous. The Smirnov test statistics measures the distance between the empirical distribution functions as formulated upon the samples below:

$$D_{n_2,n_1} = \sup_x \left| F_{2n_1}(x) - F_{1n_2}(x) \right| .$$

In practice, it is recommended to calculate the value of D_{n_1,n_2} statistics in accordance with the following ratios [1]:

$$D^{+}_{n_2,n_1} = \max_{1 \le r \le n_2} \left[\frac{r}{n_2} - F_{1n_1}(x_{2r}) \right] = \max_{1 \le s \le n_1} \left[F_{2n_2}(x_{2s}) - \frac{s-1}{n_1} \right],$$

$$D^{-}_{n_2,n_1} = \max_{1 \le r \le n_2} \left[F_{1n_1}(x_{2r}) - \frac{r-1}{n_2} \right] = \max_{1 \le s \le n_1} \left[\frac{s}{n_1} - F_{2n_2}(x_{1s}) \right],$$

$$D_{n_2,n_1} = \max\left(D^{+}_{n_2,n_1}, D^{-}_{n_2,n_1} \right).$$

When the H_0 hypothesis is true then with the limitless increase of the amount of samples the statistics of

$$S_{\text{C}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_2,n_1} \qquad (2)$$

shall conform with the Kolmogorov distribution $K(S)$ [1].

In case of the k -samples version of the Smirnov test the Smirnov test statistics modification will be taken as $S_{i,j}$ in (1) [16]

$$S_{\text{mod}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(D_{n_2,n_1} + \frac{n_1 + n_2}{4.6 n_1 n_2} \right), \qquad (3)$$

its distribution always being closer to the limiting distribution $K(S)$ by Kolmogorov. In this case the $S_{\max}$ statistics will be presented as $S^{Sm}_{\max}$.

With equal volumes of the samples under comparison the $S^{Sm}_{\max}$ statistics distributions present considerable discreteness (similar to the two-samples version) (see Fig. 1) and are different from asymptotic (limiting) distributions (see Fig. 2). When possible, it is preferable to take co-primes as n_i then the $G(S \mid H_0)$ distributions of the $S^{Sm}_{\max}$ statistics will be almost similar to asymptotic ones.

Models of asymptotic (limiting) distributions of the $S^{Sm}_{\max}$ statistics with $k = \overline{3,11}$ formulated on empirical statistics distributions as obtained with the Monte-Carlo technique and the number of imitation experiments $N = 10^6$ are shown in Tab. I.

Good models are those of the beta-distributions III laws class with the density

$$f(x) = \frac{\theta_2^{\theta_0}}{\theta_3 \mathrm{B}(\theta_0, \theta_1)} \frac{\left(\frac{x-\theta_4}{\theta_3} \right)^{\theta_0 - 1} \left(1 - \frac{x-\theta_4}{\theta_3} \right)^{\theta_1 - 1}}{\left[1 + (\theta_2 - 1) \frac{x-\theta_4}{\theta_3} \right]^{\theta_0 + \theta_1}} \qquad (4)$$

with exact values of this law's parameters being $\mathrm{B}_{\text{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ as found in statistics samples of the $N = 10^6$ volume and found in the course of modeling.

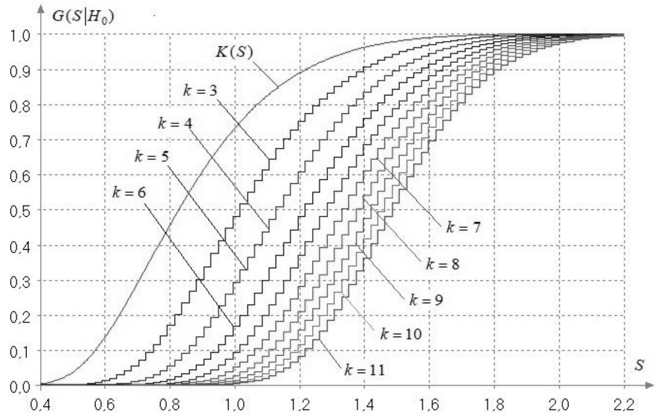

Fig. 1. Distribution of statistics $S^{Sm}_{\max}$, $n_i = 1000$

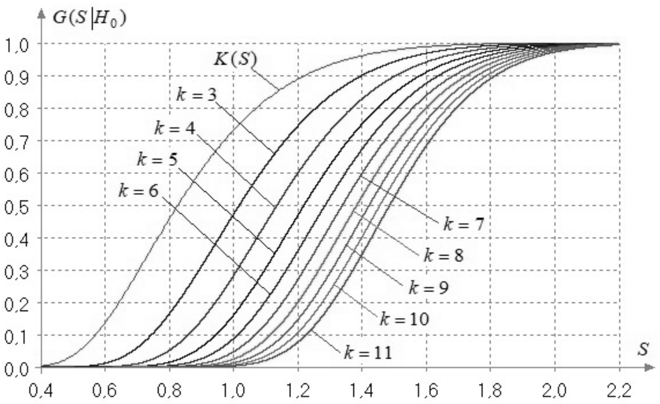

Fig. 2. Asymptotic distributions of statistics $S^{Sm}_{\max}$

TABLE I

MODELS OF LIMITING STATISTICS DISTRIBUTIONS $S^{Sm}_{\max}$

k	Model
2	$K(S)$
3	B_{III} (6.3274, 6.6162, 2.8238, 2.4073, 0.4100)
4	B_{III} (7.2729, 7.2061, 2.6170, 2.3775, 0.4740)
5	B_{III} (7.1318, 7.3365, 2.4813, 2.3353, 0.5630)
6	B_{III} (7.0755, 8.0449, 2.3163, 2.3818, 0.6320)
7	B_{III} (7.7347, 8.6845, 2.3492, 2.4479, 0.6675)
8	B_{III} (7.8162, 8.9073, 2.2688, 2.4161, 0.7120)
9	B_{III} (7.8436, 8.8805, 2.1696, 2.3309, 0.7500)
10	B_{III} (7.8756, 8.9051, 2.1977, 2.3280, 0.7900)
11	B_{III} (7.9122, 9.0411, 2.1173, 2.2860, 0.8200)

Models $\mathrm{B}_{\mathrm{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ shown in Tab. I and with the parameter values as described, allow finding p_{value} estimates with the corresponding k number of samples, based on the statistics values as calculated per ratio (1) when using statistics (3) or (2) as $S_{i,j}$.

B. k-samples version of the Lehmann–Rosenblatt test

Statistics of the two-samples Lehmann–Rosenblatt test proposed in [2] is used in the form of [1]

$$T = \frac{1}{(n_1+n_2)}\left[n_2\sum_{i=1}^{n_2}(r_i - i)^2 + n_1\sum_{j=1}^{n_1}(s_j - j)^2\right] - \frac{4n_1n_2 - 1}{6(n_1+n_2)}, \quad (5)$$

where r_i is the numerical order (rank) of x_{2i}; s_j is the numerical order (rank) of x_{1j} in the combined set of order values. It was shown in [3] that the statistics at the limit (5) is distributed as $a1(t)$ [1].

Statistics (5) is used as $S_{i,j}$ in the $S_{\max}^{LR}$ statistics of the type (1) in the case of k-samples version of the Lehmann–Rosenblatt test. Dependence of statistics distributions on the number of samples with H_0 being correct is illustrated in Fig. 3.

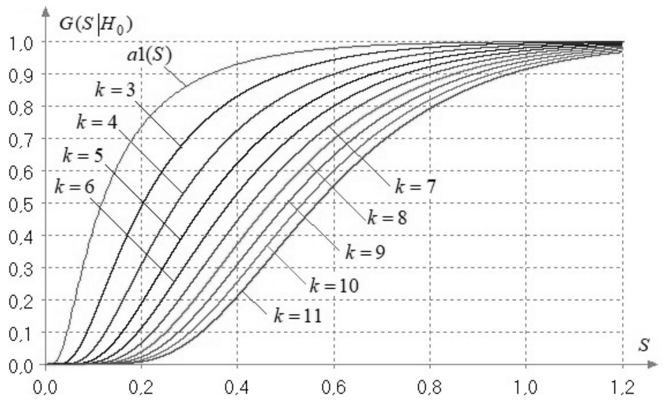

Fig. 3. Distribution of statistics $S_{\max}^{LR}$

In Tab. II the models of asymptotic (limiting) statistics distributions $S_{\max}^{LR}$ with the number of samples under comparison being $k = \overline{3,11}$ are shown. In this case the best models are the *Sb*-Johnson distributions with the density of

$$f(x) = \frac{\theta_1\theta_2 \exp\left\{-\frac{1}{2}\left[\theta_0 - \theta_1 \ln\frac{x-\theta_3}{\theta_2+\theta_3-x}\right]^2\right\}}{\sqrt{2\pi}(x-\theta_3)(\theta_2+\theta_3-x)}$$

with exact values of this law's parameters being shown in Tab. II as $Sb(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$. The presented models allow finding p_{value} estimates with the corresponding k number of samples by the values of statistics $S_{\max}^{LR}$.

TABLE II
MODELS OF LIMITING STATISTICS DISTRIBUTIONS $S_{\max}^{LR}$

k	Model
2	$a1(t)$
3	Sb(3.2854, 1.2036, 3.0000, 0.0215)
4	Sb(2.5801, 1.2167, 2.2367, 0.0356)
5	Sb(3.1719, 1.4134, 3.1500, 0.0320)
6	Sb(2.9979, 1.4768, 2.9850, 0.0380)
7	Sb(3.2030, 1.5526, 3.4050, 0.0450)
8	Sb(3.2671, 1.6302, 3.5522, 0.0470)
9	Sb(3.4548, 1.7127, 3.8800, 0.0490)
10	Sb(3.4887, 1.7729, 3.9680, 0.0510)
11	Sb(3.4627, 1.8168, 3.9680, 0.0544)

C. k-samples version of the Anderson-Darling test

The two-sample Anderson-Darling test has been dealt with in [4]. This test's statistics is defined by the expression

$$A^2 = \frac{1}{n_1n_2}\sum_{i=1}^{n_1+n_2-1}\frac{\left(M_i(n_1+n_2) - n_1 i\right)^2}{i(n_1+n_2-i)}, \quad (6)$$

where M_i is the number of elements in the first sample, smaller than or equal to the ith element of the combined set of order values. Distribution $a2(t)$ is the limiting distribution of statistics (6) when the hypothesis H_0 under investigation is correct [1].

In the case of the k-samples version of the Anderson-Darling test statistics (6) is used as $S_{i,j}$ in the $S_{\max}^{AD}$ statistics of type (1). Dependence of statistics distributions on the number of samples with H_0 being correct is illustrated in Fig. 4.

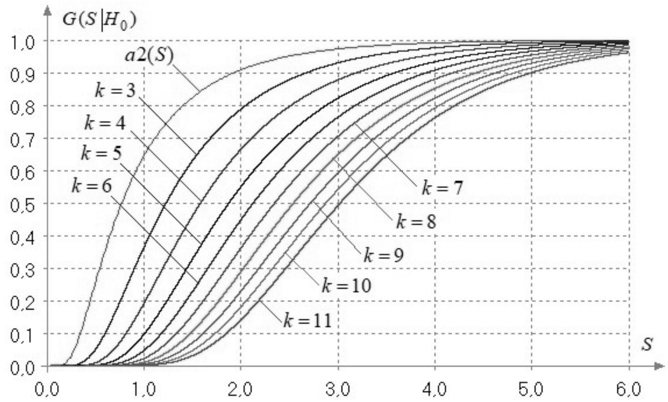

Fig. 4. Distribution of statistics $S_{\max}^{AD}$

Models of asymptotic (limiting) distributions of the $S_{\max}^{AD}$ statistics with $k = \overline{3,11}$ have been produced for distributions $G(S_{\max}^{AD} | H_0)$ and they are shown in Tab. III. In this case the best models are the beta-distributions type III (4) which are shown in Tab. III as $\mathrm{B}_{\mathrm{III}}(\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$ with exact parameters and can be used for evaluating p_{value} with the k number of samples.

TABLE III

MODELS OF LIMITING STATISTICS DISTRIBUTIONS $S_{\max}^{AD}$

k	Model
2	$a2(t)$
3	B_{III} (4.4325, 2.7425, 12.1134, 8.500, 0.1850)
4	B_{III} (5.2036, 3.2160, 10.7792, 10.000, 0.2320)
5	B_{III} (5.7527, 3.3017, 9.7365, 10.000, 0.3000)
6	B_{III} (5.5739, 3.4939, 7.7710, 10.000, 0.3750)
7	B_{III} (6.4892, 3.6656, 8.0529, 10.500, 0.3920)
8	B_{III} (6.3877, 3.8143, 7.3602, 10.800, 0.4800)
9	B_{III} (6.7910, 3.9858, 7.1280, 11.100, 0.5150)
10	B_{III} (6.7533, 4.2779, 6.6457, 11.700, 0.5800)
11	B_{III} (7.1745, 4.3469, 6.6161, 11.800, 0.6100)

IV. COMPARATIVE ANALYSIS OF THE POWERS OF TESTS

The power of k-samples tests was investigated by methods of statistic modeling of situations when the testes hypothesis H_0 was whether all samples belonged to the standard normal law, the competing hypothesis H_1 being if all samples but the last one belonged to the standard normal law and the last sample belonged to the normal law with the shift parameter $\theta_0 = 0.1$ and he scale parameter $\theta_1 = 1$, hypothesis H_2 being that the last sample belonged to the normal law with the shift parameter $\theta_0 = 0$ and the scale parameter $\theta_1 = 1.1$, the competing hypothesis H_3 being the last sample belonged to the logistic law with the density of

$$f(x) = \frac{1}{\theta_1\sqrt{3}} \exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\} \Bigg/ \left[1+\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2$$

and parameters being $\theta_0 = 0$ and $\theta_1 = 1$. Cases of samples of equal volume were investigated. Evaluation of power was done by results of modeling statistics distributions with the tested hypothesis $G(S|H_0)$ and competing hypotheses $G(S|H_1)$, $G(S|H_2)$ and $G(S|H_3)$ being true at equal volumes n_i of the tested samples.

As an example power estimates only with the significance level $\alpha = 0.1$ and $k = 3$ are shown in Tab. IV for comparable tests.

Results of research and the analysis of power allow making the following conclusions: As concerns the competing hypotheses and the changes in the shift parameters, the power criteria can be ranged as follows:

$$S_{\max}^{AD} \succ \mathrm{AD} \succ S_{\max}^{LR} \succ S_{\max}^{Sm} \succ Z_C \succ Z_{\mathrm{A}} \succ Z_{\mathrm{K}}.$$

As concerns the changes in scale parameters –

$$Z_C \succ Z_{\mathrm{A}} \succ Z_{\mathrm{K}} \succ \mathrm{AD} \succ S_{\max}^{AD} \succ S_{\max}^{Sm} \succ S_{\max}^{LR}.$$

TABLE IV

THE POWER ESTIMATES OF TESTS

Test	n_i					
	20	50	100	300	500	10^3
Concerning the alternative H_1						
$S_{\max}^{AD}$	0.113	0.134	0.171	0.314	0.450	0.712
AD	0.113	0.134	0.171	0.313	0.449	0.711
$S_{\max}^{LR}$	0.114	0.134	0.168	0.306	0.437	0.694
$S_{\max}^{Sm}$	0.110	0.128	0.155	0.272	0.383	0.622
Z_C	0.113	0.131	0.160	0.273	0.380	0.612
Z_{A}	0.112	0.130	0.158	0.268	0.371	0.599
Z_{K}	0.110	0.125	0.144	0.231	0.321	0.525
Concerning the alternative H_2						
Z_C	0.107	0.125	0.160	0.319	0.475	0.771
Z_{A}	0.107	0.126	0.162	0.319	0.470	0.767
Z_{K}	0.107	0.123	0.147	0.263	0.376	0.621
AD	0.104	0.111	0.124	0.191	0.273	0.509
$S_{\max}^{AD}$	0.102	0.107	0.114	0.165	0.231	0.446
$S_{\max}^{Sm}$	0.103	0.104	0.114	0.136	0.164	0.253
$S_{\max}^{LR}$	0.103	0.104	0.108	0.127	0.152	0.241
Concerning the alternative H_3						
Z_{A}	0.103	0.108	0.116	0.181	0.279	0.580
Z_C	0.103	0.108	0.116	0.176	0.270	0.568
Z_{K}	0.104	0.110	0.117	0.170	0.233	0.423
AD	0.103	0.107	0.114	0.148	0.189	0.315
$S_{\max}^{Sm}$	0.102	0.105	0.111	0.148	0.183	0.288
$S_{\max}^{AD}$	0.102	0.104	0.110	0.134	0.166	0.272
$S_{\max}^{LR}$	0.103	0.104	0.107	0.124	0.145	0.218

At that, the Zhang tests with statistics Z_{A} and Z_C are practically equally powerful, and the Anderson-Darling test is markedly weaker than the Zhang tests.

As concerns the case when all samples but one belong to the same normal law and the last sample belongs to the logistic law, the tests were ranged in following order, power-wise:

$$Z_{\mathrm{A}} \succ Z_C \succ Z_{\mathrm{K}} \succ \mathrm{AD} \succ S_{\max}^{Sm} \succ S_{\max}^{AD} \succ S_{\max}^{LR}.$$

With the increase of the number of tested samples of the same volume the power of the test diminishes with respect to similar competing hypotheses, as a rule, which is perfectly natural. For example, it is somewhat more difficult to pinpoint a situation and give preference to a competing hypothesis when only one of the tested samples belongs to a different law.

We cannot but note that that the Zhang tests with statistics Z_{K}, Z_{A}, Z_C manifest obvious advantage in power with reference to some alternatives. The distribution of Zhang test statistics depends on a sample size that makes it difficult to use in applications.

That is why these tests can be applied only if software make it possible to simulate the distribution of Zhang statistics by the Monte Carlo method. After that, estimates p_{value} can be found using the resulting distributions of statistics $G(Z_K|H_0)$, $G(Z_A|H_0)$, $G(Z_C|H_0)$. For example, this is exactly how it is done in the ISW software system [18], using which the present studies were carried out.

V. Conclusion

The proposed k-samples tests and the formulated models of limiting statistics distributions $S_{\max}^{Sm}$, $S_{\max}^{LR}$, $S_{\max}^{AD}$ allow obtaining estimates p_{value} in the course of checking the homogeneity hypotheses, thus providing more information capacity and supportability of statistics conclusions. In that respect the advantage rests with the Zhang tests [8].

The conducted power analysis allows choosing more preferable criteria depending on several alternatives.

This study has been conducted with support of the Ministry of Education and Science of the Russian Federation within the framework of government order "The provision of scientific research" (No. 1.4574.2017/6.7) and the project part of the government order (No. 1.1009.2017/4.6).

References

[1] Bolshev L.N. Tables of mathematical statistics / L.N.Bolshev, N.V.Smirnov. – M. : Nauka, 1983. – 416 pp.
[2] Lehmann E. L. Consistency and unbiasedness of certain nonparametric tests / E. L. Lehmann // Ann. Math. Statist. 1951. Vol. 22, No. 1. – pp. 165–179.
[3] Rosenblatt M. Limit theorems associated with variants of the von Mises statistic / M. Rosenblatt // Ann. Math. Statist. 1952. Vol. 23. – pp. 617–623.
[4] Pettitt A.N. A two-sample Anderson-Darling rank statistic // Biometrika. 1976. Vol. 63. No.1. pp. 161-168.
[5] Scholz F.W., Stephens M.A. k-sample Anderson–Darling Tests // Journal of the American Statistical Association. 1987. Vol. 82. No. 399. – pp. 918-924.
[6] Zhang J. Powerful goodness-of-fit and multi-sample tests / J. Zhang // PhD Thesis. York University, Toronto. 2001. – 113 pp. URL: http://www.collectionscanada.gc.ca/obj/s4/f2/dsk3/ftp05/NQ66371.pdf (date of the request 28.01.2013).
[7] Zhang J. Powerful Two-Sample Tests Based on the Likelihood Ratio / J. Zhang // Technometrics. 2006. V. 48. No. 1. – pp.95-103.
[8] Zhang J., Wu Y. k-Sample tests based on the likelihood ratio // Computational Statistics & Data Analysis. – 2007. – V. 51. – No. 9. – pp. 4682-4691.
[9] Kiefer J. K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-v. Mises Tests // Annals of Mathematical Statistics, 1959. Vol. 30. No. 2. – pp. 420-447.
[10] Conover W. J. Several k-sample Kolmogorov-Smirnov tests // The Annals of Mathematical Statistics. 1965. Vol. 36, No. 3. – pp.1019-1026.
[11] Conover W. J. Practical Nonparametric Statistics / W.J. Conover. – 3d ed. – Wiley, 1999. – 584 pp.
[12] Lemeshko B.Yu. Homogeneity hypotheses tests. Manual. M: INFRA–M, 2017. – 207 pp.
[13] Lemeshko B.Yu. On the use of distribution laws homogeneity tests / B.Yu.Lemeshko, S.B.Lemeshko, I.V.Veretelnikova // Bulletin of the Tomsk State University. Management, computer science and information science. 2017. No.41. – pp. 24-31.
[14] Lemeshko B.Yu. On the Application of Homogeneity Tests / B.Y. Lemeshko, I.V. Veretelnikova, S.B. Lemeshko, A.Y. Novikova // Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Nonparametric methods in cybernetics and system analysis". 18-22 September 2017, Krasnoyarsk. pp. 181-195.
[15] Lemeshko B.Y. Application of Homogeneity Tests: Problems and Solution / B.Y. Lemeshko, I.V. Veretelnikova, S.B. Lemeshko, A.Y. Novikova // In: Rykov V., Singpurwalla N., Zubkov A. (eds) Analytical and Computational Methods in Probability Theory. ACMPT 2017. Lecture Notes in Computer Science. : monograph. - Cham : Springer, 2017. - 10684. - pp. 461-475.
[16] Lemeshko B.Yu. On convergence of statistics distributions and the powers of the Smirnov and Lehmann–Rosenblatt homogeneity tests / B.Yu.Lemeshko, S.B.Lemeshko // Measurement techniques. 2005. No. 12. – pp. 9–14.
[17] Smirnov N.V. Evaluation of divergence between empirical distribution curves in two independent samples / N.V.Smirnov // Bulletin of MGU, Series A. 1939. Vol. 2, No. 2. – pp. 3–14.

Boris Yurievich Lemeshko
Chief research scientist, chair for applied and theoretical information science, Novosibirsk State Technical University, Doctor of Engineering Science, professor
(20 Karla Marksa Prospect, Novosibirsk, 630073),
phone: (383) 346-06-00,
e-mail: Lemeshko@ami.nstu.ru,
http://www.ami.nstu.ru/~headrd/

Irina Viktorovna Veretelnikova
Post-graduate student, chair for applied and theoretical information science, Novosibirsk State Technical University
(20 Karla Marksa Prospect, Novosibirsk, 630073),
phone (383) 346-06-00,
e-mail: ira-veterok@mail.ru