Real-Time Studying of Statistic Distributions of Non-Parametric Goodness-of-Fit Tests when Testing Complex Hypotheses

Boris Yu. Lemeshko, Stanislav B. Lemeshko, Andrey P. Rogozhnikov
Department of Applied Mathematics, Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: Lemeshko@fpm.ami.nstu.ru

Abstract

In present work, a “real-time” ability to simulate and research the distributions of tests statistics in the course of testing the complex goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown (before the testing procedure starts).

Keywords: goodness-of-fit test, composite hypotheses testing, Kolmogorov test, Cramer-Mises-Smirnov test, Anderson-Darling test, methods of statistical simulation.

Introduction

In composite hypotheses testing in the form $H_0 : F(x) \in \{F(x, \theta),\ \theta \in \Theta\}$, when the estimate $\hat{\theta}$ of scalar or vector distribution parameter θ is calculated by the same sample, the nonparametric goodness-of-fit Kolmogorov, ω^2 Cramer-Mises-Smirnov, and Ω^2 Anderson-Darling tests lose their distribution-free property.

The value

$$D_n = \sup_{|x|<\infty} |F_n(x) - F(x, \theta)|,$$

where $F_n(x)$ is the empirical distribution function, n is the sample size, is used in Kolmogorov test as a distance between the empirical and theoretical laws. When testing hypotheses, this statistic is usually used with Bolshev's correction (Bolshev, [3]) in the form (Bolshev and Smirnov, [4])

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}} \tag{1}$$

where $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max\limits_{1\le i\le n} \left\{\frac{i}{n} - F(x_i, \theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n} \left\{F(x_i, \theta) - \frac{i-1}{n}\right\}$, n is the sample size, $x_1, x_2, \ldots, x_n$ are sample values in an increasing order. The distribution of statistic (1) in testing simple hypotheses obeys the Kolmogorov distribution law $K(S) = \sum\limits_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}$.

In ω^2 Cramer-Mises-Smirnov test, one uses a statistic in the form

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left\{F(x_i, \theta) - \frac{2i-1}{2n}\right\}^2, \tag{2}$$

and in test of Ω^2 Anderson-Darling type (Anderson and Darling, [1, 2]), the statistic in the form

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_i,\theta))\right\}. \qquad (3)$$

In testing a simple hypothesis, statistic (2) obeys the distribution (see Bolshev and Smirnov, [4]) with the CDF

$$a1(S) = \frac{1}{\sqrt{2s}}\sum_{j=0}^{\infty}\frac{\Gamma(j+1/2)\sqrt{4j+1}}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2}{16S}\right\}\times\left\{I_{-\frac{1}{4}}\left[\frac{(4j+1)^2}{16S}\right]-I_{\frac{1}{4}}\left[\frac{(4j+1)^2}{16S}\right]\right\},$$

where $I_{-\frac{1}{4}}(\cdot)$, $I_{\frac{1}{4}}(\cdot)$ are modified Bessel functions, $I_\nu(z)=\sum_{k=0}^{\infty}\frac{(z/2)^{\nu+2k}}{\Gamma(k+1)\Gamma(k+\nu+1)}$, $|z|<\infty$, $|\arg z|<\pi$, and statistic (3) obeys the distribution (Bolshev and Smirnov, [4]) with the CDF

$$a2(S) = \frac{\sqrt{2\pi}}{S}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+1/2)(4j+1)}{\Gamma(1/2)\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8S}\right\}\times\int_0^\infty\exp\left\{\frac{S}{8(y^2+1)}-\frac{(4j+1)^2\pi^2y^2}{8S}\right\}dy.$$

1 Statistic distributions of the tests in testing composite hypotheses

In composite hypotheses testing, the conditional distribution law of the statistic $G(S|H_0)$ is affected by a number of factors: the form of the observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types and number of parameters to be estimated; sometimes, it is a specific value of the parameter (e.g., in case of gamma-distribution, inverse Gaussian law, generalized Weibull distribution, beta-distribution families); the method of parameter estimation.

The paper Kac [13] was a pioneer in investigating statistic distributions of the nonparametric goodness-of-fit tests with composite hypotheses. Then, various approaches to the solution to this problem where used (Darling [6, 7], Durbin [8, 9, 10], Gihman [12], Martynov [27], Pearson and Hartley [30], Stephens [31, 32], Chandra [5], Tyurin [33], Tyurin [34], Dzhaparidze and Nikulin [11], Nikulin [28, 29]).

In our research (Lemeshko and Postovalov [14, 15, 16], Lemeshko and Maklakov [17], Lemeshko [18, 24, 25], Lemeshko and Lemeshko [19, 20, 21], Lemeshko S. [26]), statistic distributions of the nonparametric goodness-of-fit tests are investigated by the methods of statistical simulation, and approximate models of the laws are found for constructed empirical distributions. The most complete list of the constructed models of statistic distributions and tables of percentage points for nonparametric goodness-of-fit tests is provided in Lemeshko [18, 24, 25]. These models and tables are usable when testing complex hypotheses if maximum likelihood estimators were applied.

For a number of distributions often used in applications for description of random variates, distributions of statistics of nonparametric goodness-of-fit tests only have a limited set of dependences: the form of the observed law $F(x,\theta)$ that corresponds to the true hypothesis H_0; types and number of parameters to be estimated; the method of parameter estimation. In these cases, there are no impediments for studying test statistic distributions by means of statistical simulation and further construction of approximate models for them when testing complex hypothesis (Lemeshko [18, 24, 25]).

Complications arise in case the statistic distributions $G(S\,|H_0)$ of nonparametric goodness-of-fit tests depend on a certain value of parameter/parameters of the distribution $F(x,\theta)$ when testing complex hypotheses (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families).

The existing dependence on parameters values should not be neglected. For example, in composite hypotheses testing subject to gamma-distribution with the density function $f(x,\theta) = \frac{x^{\theta_0-1}}{\theta_1^{\theta_0}\Gamma(\theta_0)}\exp\left(-\frac{x}{\theta_1}\right)$, limiting statistics distributions of the nonparametric goodness-of-fit tests depend on value of the form parameter θ_0. Figure 1 illustrates the dependence of the Kolmogorov statistic distribution upon the value θ_0 in testing a composite hypothesis only in the case of calculating maximum likelihood estimates (MLE) for the scale parameter of gamma-distribution.

The most serious impediment to a complete solution of the problem of testing complex hypotheses with the use of non-parametric goodness-of-fit tests is that the distributions of the test statistics depend on specific values of shape parameters of the observed laws. In papers (Lemeshko [18, 19, 20, 21, 24, 25]) models of distributions of statistics were obtained for a limited set of combinations of (integer) values of shape parameters (for gamma distribution, two-sided exponential law, inverse Gaussian law, generalized Weibull distribution, and beta-distribution families). It is unrealistic to build the models for an infinite set of combinations of the parameters values.

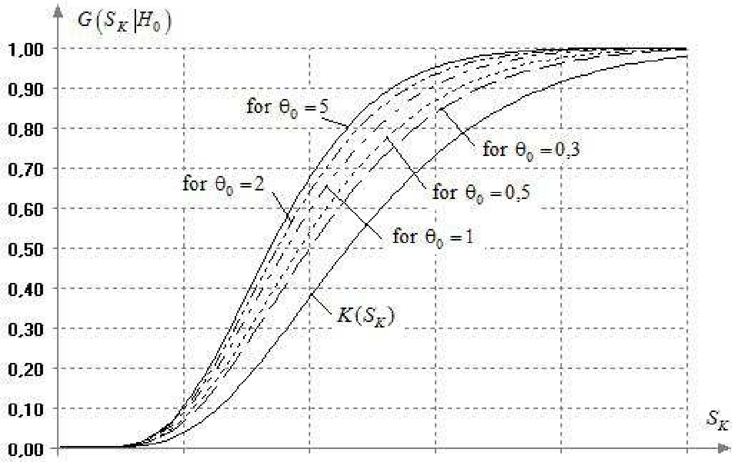

Figure 1: The Kolmogorov statistic (1) distributions for testing composite hypotheses with calculating MLE of scale parameter

In present work, a “real-time” ability to simulate and research the distributions of tests statistics in the course of testing the complex goodness-of-fit hypothesis (for distributions with estimated parameters) is implemented by the use of parallel computing. It makes it possible to make correct statistical inferences even in those situations when the distribution of the test statistic is unknown (before the testing procedure starts).

2 Testing complex hypotheses in "real-time"

In present work, an approach is proposed and implemented that is based upon authors' evolving software and the use of simulation (Lemeshko [23]). Computational processes in the simulation of statistics of various tests can be parallelized rather easily by the use of available resources of nearby computer network. This makes it possible to dramatically reduce the time required for simulation (studying) an unknown distribution of the statistic $G(S|H_0)$. Statistical analysis is carried out by the following scheme (Fig. 2) in case of the use of nonparametric goodness-of-fit tests for testing complex hypotheses in regard to laws with characteristic dependence of statistic distribution on parameter values. Such an approach was used in Lemeshko [22]. Here the studying of $G(S|H_0)$ is carried out in "real-time" of testing the hypothesis.

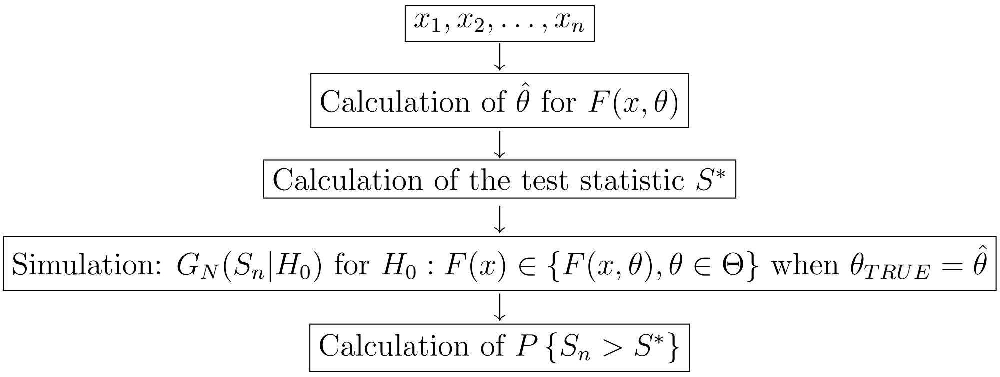
$x_1, x_2, \ldots, x_n$
Calculation of $\hat{\theta}$ for $F(x,\theta)$
Calculation of the test statistic S^*
Simulation: $G_N(S_n|H_0)$ for $H_0 : F(x) \in \{F(x,\theta), \theta \in \Theta\}$ when $\theta_{TRUE} = \hat{\theta}$
Calculation of $P\{S_n > S^*\}$

Figure 2: Testing the complex hypothesis $H_0 : F(x) \in \{F(x,\theta), \theta \in \Theta\}$

When testing complex the hypothesis $H_0 : F(x) \in \{F(x,\theta),\ \theta \in \Theta\}$ by an existing sample $x_1, x_2, \ldots, x_n$, the parameter vector estimate $\hat{\theta}$ for the law $F(x,\theta)$ is found in accordance with the selected method. Then, the value of statistic S^* of the goodness-of-fit test in use is calculated in accordance with the estimate $\hat{\theta}$ found. For making an inference on whether to reject or to accept the hypothesis H_0 under test, it's necessary to know the distribution $G(S\,|H_0)$ of the test statistic that corresponds to the parameter value $\hat{\theta}$.

After that, statistical simulation procedure is started that results in obtaining empirical distribution $G_N(S_n\,|H_0)$ of the test statistic for the corresponding sample volume n and the given number of simulations N and $F(x,\theta)$ with the parameters vector $\theta = \hat{\theta}$. One can find an estimate of an achieved significance level $P\{S_n > S^*\}$ or estimates of percentage points by the use of empirical distribution $G_N(S_n\,|H_0)$. The hypothesis is not rejected if $P\{S_n > S^*\} > \alpha$, where α is a given type I error probability.

The value of N defines the required accuracy of simulation of $G(S_n\,|H_0)$: the greater N the better. However, time spent for simulation increases along with growth of N, therefore, one can determine N during parallelization of simulation process basing upon available computer resources (number of processors and cores) that could be used for the problem under solution.

The probability that elements of $\hat{\theta}$ are integer is zero. Thus, one should cautiously use models and percentage points of test statistic distributions for values of parameters close to integer ones provided in (Lemeshko [18, 19, 20, 21, 24, 25]) as, with interpolation applied, results obtained can be far from the true distribution $G(S\,|H_0)$ with the given $\hat{\theta}$.

Let us consider an example where a complex hypothesis is tested in regard to the inverse Gaussian law with the density function $f(x) = \left(\frac{\theta_1}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\theta_1(x-\theta_0)^2}{2\theta_0^2 x}\right)$. In this case, distributions $G(S\,|H_0)$ of the nonparametric tests depend on specific values of θ_0 and θ_1.

The sample under analysis is presented in Table 1 ($\theta_0 = \theta_1 = 2.5$). Maximum likelihood estimates of the parameters: $\hat{\theta}_0$=2.4706, $\hat{\theta}_1$=2.5769. In Table 2, values of the tests statistics and achieved significance levels (P-values) obtained by test statistic distributions simulated (in "real time") under different values of N are given.

Table 1: 100 pseudorandom numbers from the inverse Gaussian distribution

0.278	0.633	0.928	1.078	1.334	1.937	2.297	2.630	3.554	5.674
0.312	0.686	0.933	1.080	1.497	1.965	2.362	2.919	3.593	5.989
0.358	0.716	0.936	1.089	1.612	1.991	2.364	2.995	3.948	6.284
0.361	0.776	0.938	1.113	1.671	2.012	2.417	3.002	3.996	6.863
0.362	0.777	0.956	1.119	1.680	2.026	2.467	3.120	4.053	7.580
0.374	0.789	0.996	1.159	1.687	2.027	2.566	3.149	4.141	7.644
0.403	0.796	1.038	1.165	1.731	2.069	2.577	3.166	4.363	7.874
0.590	0.805	1.053	1.166	1.735	2.146	2.599	3.224	4.597	9.236
0.597	0.822	1.060	1.192	1.763	2.210	2.621	3.278	5.022	11.704
0.599	0.849	1.066	1.245	1.898	2.213	2.628	3.528	5.201	20.069

It should be noted, that distributions of nonparametric goodness-of-fit test statistics (1)–(3) for $\hat{\theta}_0$ =2.4706, $\hat{\theta}_1$=2.5769 differ substantially from corresponding distributions under different combinations of integer values of θ_0 and θ_1.

Another example is generalized Weibull distribution with the density function

$$f(x;\theta_0,\theta_1) = \frac{\theta_0}{\theta_1} x^{\theta_0-1}\left(1+x^{\theta_0}\right)^{\frac{1}{\theta_1}-1} exp\left\{1-\left(1+x^{\theta_0}\right)^{\frac{1}{\theta_1}}\right\},$$

$\theta_0 = \theta_1 = 2.5$ (Table 3). Maximum likelihood estimates of the parameters: $\hat{\theta}_0 = 2.4718$, $\hat{\theta}_1 = 2.5187$. Values of the tests statistics and P-values obtained by simulation are given in Table 4.

Table 2: P-values of the tests for different volumes of simulations (inverse Gaussian distribution)

Test	S^*	$P\{S_n > S^*\}$				
		N=1000	N=5000	N=10000	N=100000	N=1000000
K	0.59361	0.656	0.668	0.668	0.670	0.671
ω^2	0.05380	0.562	0.576	0.574	0.578	0.578
Ω^2	0.35021	0.556	0.570	0.568	0.566	0.566

Table 3: 100 pseudorandom numbers from the generalized Weibull distribution

0.199	0.647	0.932	1.059	1.253	1.648	1.855	2.033	2.482	3.356
0.248	0.703	0.937	1.060	1.367	1.664	1.891	2.180	2.500	3.474
0.311	0.734	0.939	1.067	1.444	1.680	1.892	2.218	2.658	3.583
0.316	0.793	0.941	1.086	1.482	1.692	1.920	2.221	2.679	3.791
0.317	0.794	0.956	1.091	1.488	1.700	1.948	2.279	2.703	4.040
0.333	0.806	0.991	1.122	1.493	1.701	2.000	2.293	2.741	4.062
0.373	0.812	1.025	1.127	1.521	1.725	2.006	2.301	2.835	4.139
0.600	0.821	1.038	1.128	1.523	1.770	2.017	2.328	2.932	4.587
0.608	0.837	1.043	1.147	1.541	1.807	2.029	2.354	3.104	5.351
0.611	0.862	1.049	1.188	1.624	1.808	2.032	2.470	3.174	7.676

Table 4: P-values of the tests for different volumes of simulations (generalized Weibull distribution)

Test	S^*	$P\{S_n > S^*\}$				
		N=1000	N=5000	N=10000	N=100000	N=1000000
K	0.60473	0.670	0.672	0.670	0.673	0.675
ω^2	0.05519	0.596	0.599	0.594	0.597	0.597
Ω^2	0.35462	0.577	0.580	0.580	0.580	0.580

Conclusions

In this work, software is implemented that makes it possible to test complex hypotheses with the use of nonparametric goodness-of-fit test in cases when statistic distributions depend on specific values of the observed distributions.

This research was partially supported by the Russian Foundation for Basic Research (Project № 09-01-00056-a), the Analytical Departmental Targeted Program "Development of the Potential of Institutes of Higher Education" (Project № 2.1.2/11855), and the Federal Targeted Program of the Ministry of Education and Science of the Russian Federation "Academic and Teaching Staff of an Innovative Russia".

References

[1] Anderson T.W., Darling D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23, 193-212.

[2] Anderson T.W., Darling D.A. (1954). A test of goodness of fit. *J. Amer. Statist. Assoc.*, 29, 765-769.

[3] Bolshev L.N. (1987). On the question on testing some composite statistical hypotheses. *Theory of Probability and Mathematical Statistics. Selected Works.* Nauka, Moscow, 5-63.

[4] Bolshev L.N., Smirnov N.V. (1983). *Tables of Mathematical Statistics.* Nauka, Moscow. 1983. (in Russian)

[5] Chandra M., Singpurwalla N.D., Stephens M.A. (1981). Statistics for Test of Fit for the Extreme-Value and Weibull Distributions. *J. Am. Statist. Assoc.* Vol. 76, No. 375, pp. 729-731.

[6] Darling D.A. (1955). The Cramer-Smirnov test in the parametric case. *Ann. Math. Statist.*, 26, 1-20.

[7] Darling D.A. (1957). The Cramer-Smirnov test in the parametric case. *Ann. Math. Statist.*, 28, 823-838.

[8] Durbin J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.*, 1, 279-290.

[9] Durbin J. (1975). Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests of spacings. *Biometrika*, 62, 5-22.

[10] Durbin J. (1976). Kolmogorov-Smirnov Test when Parameters are Estimated. *Lect. Notes Math.* 566, 33-44.

[11] Dzhaparidze K.O., Nikulin M.S. (1982). Probability distribution of the Kolmogorov and omega-square statistics for continuous distributions with shift and scale parameters. *J. Soviet Math.*, 20, 2147-2163.

[12] Gihman I.I. (1953). Some remarks on the consistency criterion of A.N. Kolmogorov. *Dokl. Akad. Nauk SSSR*, 91(4), 715-718.

[13] Kac M., Kiefer J., and Wolfowitz J. (1955). On Tests of Normality and Other Tests of Goodness of Fit Based on Distance Methods. *Ann. Math. Stat.*, 26, 189-211.

[14] Lemeshko B.Yu., Postovalov S.N. (1998). Statistical distributions of nonparametric goodness-of-fit tests as estimated by the sample parameters of experimentally observed laws. *Industrial laboratory (Ind. lab.)*, 64, 3, 197-208. (Consultants Bureau, New York)

[15] Lemeshko B.Yu., Postovalov S.N. (2001). Application of the nonparametric goodness-of-fit Tests in testing composite hypotheses. *Optoelectronics, Instrumentation and Data Processing*, 37, 2, 76-88.

[16] Lemeshko B.Yu., Postovalov S.N. (2002). The nonparametric goodness-of-fit tests about fit with Johnson distributions in testing composite hypotheses. *News of the SB AS HS*, 1(5), 65-74. (in Russian)

[17] Lemeshko B.Yu., Maklakov A.A. (2004). Nonparametric Test in Testing Composite Hypotheses on Goodness of Fit Exponential Family Distributions. *Optoelectronics, Instrumentation and Data Processing*, 40, 3, 3-18.

[18] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. (2010). Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses. *Comm. Stat. - Theory and Methods*, 39, 3, 460-471.

[19] Lemeshko B.Yu., Lemeshko S.B. (2007). Statistic distributions of the nonparametric goodness-of-fit tests in testing hypotheses relative to beta-distributions. *News of the SB AS HS*, 2, 9, 6-16. (in Russian)

[20] Lemeshko B.Yu., Lemeshko S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *Measurement Techniques*, 52, 6, 555-565.

[21] Lemeshko B.Yu., Lemeshko S.B. (2009). Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques*, 52, 8, 799-812.

[22] Lemeshko B.Yu., Lemeshko S.B., Akushkina K.A., Nikulin M.S., Saaidia N. (2011). Inverse Gaussian Model and Its Applications in Reliability and Survival Analysis. *Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and*

Quality Control (Edited by V. Rykov, N. Balakrishnan and M. Nikulin), 433-453. Birkhäuser, Boston.

[23] Lemeshko B.Yu., Lemeshko S.B., Chimitova E.V., Postovalov S.N., Rogozhnikov A.P. (2011). Software System for Simulation and Research of Probabilistic Regularities and Statistical Data Analysis in Reliability and Quality Control. *Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control* (Edited by V. Rykov, N. Balakrishnan and M. Nikulin), 417-432. Birkhäuser, Boston.

[24] Lemeshko B.Yu., Lemeshko S.B., Nikulin M.S., Saaidia N. (2010). Modeling statistic distributions for nonparametric goodness-of-fit criteria for testing complex hypotheses with respect to the inverse Gaussian law. *Automation and Remote Control*, 71, 7, 1358-1373.

[25] Lemeshko B.Yu., Lemeshko S.B., Akushkina K.A. (2010). Models of statistical distributions of nonparametric goodness-of-fit tests in testing composite hypotheses of the generalized Weibull distribution. *Proceed. 3^{rd} Int. Conf. on Accel. Life Testing (ALT'2010). Clermont-Ferrand, France*, 125-132.

[26] Lemeshko S.B. (2007). Expansion of applied opportunities of some classical methods of mathematical statistics. *The dissertation on competition of a scientific degree of Cand. Tech. Sci.* Novosibirsk State Technical University. Novosibirsk. (in Russian)

[27] Martynov G.V. (1978). *Omega-Square Tests.* Nauka, Moscow. (in Russian)

[28] Nikulin M.S. (1992). Gihman and goodness-of-fit tests for grouped data. *Mathematical Reports of the academy of Science of the Royal Society of Canada*, 14, 4, 151-156.

[29] Nikulin M.S. (1992). A variant of the generalized omega-square statistic. *J. Soviet Math.*, 61, 4, 1896-1900.

[30] Pearson E.S., Hartley H.O. (1972). *Biometrica Tables for Statistics. Volume 2.* University Press, Cambridge.

[31] Stephens M.A. (1970). Use of Kolmogorov-Smirnov, Cramer - von Mises and Related Statistics - Without Extensive Table. *J. R. Stat. Soc.*, 32, 115-122.

[32] Stephens M.A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons. *J. Am. Statist. Assoc.*, 69, 730-737.

[33] Tyurin Yu.N. (1984). On the Limiting Kolmogorov-Smirnov Statistic Distribution for Composite Hypothesis. *News of the AS USSR. Ser. Math.*, 48, 6, 1314-1343. (in Russian)

[34] Tyurin Yu.N., Savvushkina N.E. (1984). Goodness-of-Fit Test for Weibull-Gnedenko Distribution. *News of the AS USSR. Ser. Techn. Cybernetics*, 3, 109-112. (in Russian)