APPLIED METHODS OF STATISTICAL ANALYSIS.

STATISTICAL COMPUTATION AND SIMULATION

PROCEEDINGS

OF THE INTERNATIONAL WORKSHOP

18-20 September 2019

Novosibirsk

2019

UDC 519.22(063)
A 67

E d i t o r s:

Prof. Boris Lemeshko, Prof. Mikhail Nikulin,
Prof. Narayanaswamy Balakrishnan

A 67 **Applied Methods of Statistical Analysis. Statistical Computation and Simulation** - AMSA'2019, Novosibirsk, Russia, 18-20 September, 2019: Proceedings of the International Workshop. - Novosibirsk: NSTU publisher, 2019. - 574 pp.
ISSN 2313-870X

ISSN 2313-870X **UDC 519.22(063)**

Applied Methods of Statistical Analysis. Statistical Computation and Simulation

Scientific Program Committee:

A. Vostretsov	Novosibirsk State Technical University, Russia
B. Lemeshko	Novosibirsk State Technical University, Russia
N. Balakrishnan	McMaster University, Canada
A. Medvedev	Siberian Federal University, Russia
G. Koshkin	Tomsk State University, Russia
A. Antonov,	Institute of Nuclear Power Engineering, Russia
E. Chimitova,	Novosibirsk State Technical University, Russia
Yu. Dmitriev,	Tomsk State University, Russia
M. Krnjajić,	National University of Ireland, Ireland
H. Liero,	University of Potsdam, Germany
N. Limnios,	Université de Technologie de Compiègne, France
G. Mikhailov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
V. Melas,	St. Petersburg State University, Russia
V. Ogorodnikov,	Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia
B. Ryabko,	Siberian State University of Telecommunications and Information Sciences, Russia
V. Rykov,	Institute of Computational Technologies SB RAS, Russia
V. Solev,	St.Petersburg Department of Steklov Mathematical Institute RAS, Russia
F. Tarasenko,	Tomsk State University, Russia
V. Timofeev	Novosibirsk State Technical University, Russia

Local Organizing Committee:

Ekaterina Chimitova, Evgenia Osintseva, Mariia Semenova

Preface

The Fifth International Workshop "Applied Methods of Statistical Analysis. Statistical Computation and Simulation" – AMSA'2019 is organized by Novosibirsk State Technical University.

The first two Workshops AMSA'2011 and AMSA'2013, as well as AMSA'2019, took place in Novosibirsk. AMSA'2015 was held in the resort Belokurikha located at the foothills of Altai. AMSA'2017, organized together with Siberian State University of Science and Technologies called after academician M.F. Reshetnev, took place in Krasnoyarsk.

The First Workshop "Applied Methods of Statistical Analysis" AMSA'2011 was focused on Simulations and Statistical Inference, AMSA'2013 – on Applications in Survival Analysis, Reliability and Quality Control, AMSA'2015 – on Nonparametric Approach and AMSA'2017 – on Nonparametric Methods in Cybernetics and System Analysis.

The Workshop AMSA'2019 was mainly oriented to the discussion of problems of Statistical Computation and Simulation, which are crucial for the development of methods of applied mathematical statistics and their effective application in practice.

The Workshop proceedings would certainly be interesting and useful for specialists, who use statistical methods for data analysis in various applied problems arising from engineering, biology, medicine, quality control, social sciences, economics and business. The Proceedings of International Workshop "Applied Methods of Statistical Analysis" are indexed in Scopus starting with 2017 materials.

The organization of the Fifth International Workshop "Applied Methods of Statistical Analysis. Statistical Computation and Simulation" – AMSA'2019 was supported by the Russian Ministry of Education and Science (project 1.1009.2017/4.6).

Prof. Boris Lemeshko

Contents

Yu. Grigoriev
Actuarial risk theory: becoming in Russia, main problems, and development of concepts **11**

Yu. Dmitriev, O. Gubina, G. Koshkin
Estimation of the present values of net premiums and life annuities for the different actuarial models **30**

Z. Warsza, J. Puchalski
Method of the estimation of uncertainties in multiparameter measurements of correlated quantities **47**

Z. Warsza, J. Puchalski, A. Idzikowski
Application of the vector method for estimating characteristic function based on measurements uncertainty at two control points **60**

I. Malova, S. Malov
On estimation algorithms in nonparametric analysis of the current status right-censored data **74**

A. Abdushukurov
Survival function estimation from fixed design regression model in the presence of dependent random censoring **85**

N. Nurmukhamedova
Asymptotics of chi-square test based on the likelihood ratio statistics under random censoring from both sides **90**

L. Kakadjanova
Empirical processes of independence in presence of estimated parameter **96**

D. Zakhidov, D. Iskandarov
Empirical likelihood confidence intervals for truncated integrals **102**

A. Popov, V. Karmanov
Construction of basic durability model of drilling with using fuzzy regression models **105**

E. Chetvertakova, E. Chimitova, E. Osintseva, R. Snetkov
The Wiener degradation model in the analysis of the laser module ILPN-134 **114**

B. Lemeshko, S. Lemeshko, M. Semenova
Features of testing statistical hypotheses under big data analysis **122**

B. Lemeshko, I. Veretelnikova
On application of k-samples homogeneity tests **138**

A. Voytishek, T. Bulgakova
On conditional optimization of "kernel" estimators of densities **152**

O. Makhotkin
Investigation of the chi-squared test errors **160**

P. Peresunko, K. Pakhomova, E. Soroka, S. Videnin
Comparison of generalisation error's methods on case of linear regression **165**

P. Philonenko, S. Postovalov
On the distribution of the $MIN3$ two-sample test statistic **173**

P. Philonenko, S. Postovalov
The research of the two-sample test statistics convergence rate **181**

D. Politis, V. Vasiliev, S. Vorobeychikov
Optimal index estimation of log-gamma distribution **188**

Yu. Dmitriev, G. Koshkin
Estimation of present value of deffered life annuity using information about expectation of life **195**

V. Smagin, G. Koshkin, K. Kim
Robust extrapolation in discrete systems with random jump parameters and incomplete information **203**

T. Dogadova, V. Vasiliev
Adaptive prediction of Ornstein-Uhlenbeck process by observations with additive noise **212**

Yu. Burkatovskaya, V. Vasiliev
Parameter estimation with guaranteed accuracy for AR(1) by noised observations **219**

D. Lisitsin, A. Usol'tsev
Minimum gamma-divergence estimation for non-homogeneous data with application to ordered probit model **227**

E. Pchelintsev, S. Perelevskiy
Asymptotically efficient estimation of a drift coefficient in diffusion processes **235**

A. Medvedev
On controlled processes of multidimensional discrete-continuous systems **243**

A. Medvedev
On levels of a priori information in the of identification and control problems **251**

V. Branishti
Applying the method of moments to build the orthogonal series density estimator **257**

O. Cherepanov
Robust correlation coefficients based on weighted maximum likelihood method **263**

S. Andoni, V. Andoni, A. Shishkina, D. Yareschenko
About non-parametric algorithms identification of inertialess systems **271**

E. Mangalova, O. Chubarova, D. Melekh, A.Stroev
Acute pancreatitis severity classification: accuracy, robustness, visualization **278**

E. Mihov, M. Kornet
Non-parametric control algorithms for multidimensional H-processes **286**

A. Medvedev, D. Melekh, N. Sergeeva, O. Chubarova
Adaptive algorithm of classification on the missing data **292**

A. Tereshina, M. Denisov
Adaptive models for discrete-continuous process **299**

A. Raskina, E. Chzhan, V. Kukartsev, A. Karavanov, A. Lonina
Nonparametric dual control algorithm for discrete linear dynamic systems **306**

M. Akenteva, N. Kargapolova, V. Ogorodnikov
Numerical study of the bioclimatic index of severity of climatic regime based on a stochastic model of the joint meteorological time series **311**

A. Medvyatskaya, V. Ogorodnikov
Approximate numerical stochastic spectral model of a periodically correlated process **320**

O. Soboleva
Modeling of dispersion in a fractal porous medium **327**

T. Averina, K. Rybakov
Maximum cross section method in estimation of jump-diffusion random processes **335**

T. Averina, I. Kosachev, K. Chugai
A stochastic model of an unmanned aerial vehicle control system **342**

M. Shakra, Yu. Shmidt, I. Almosabbeh
Evaluating the impact of tourism on economic growth in Tunisia **349**

E. Gribanova
Algorithm for regression equation parameters estimation using inverse calculations **357**

L. Shiryaeva
On rotated versions of one parameter Grubbs's copula **365**

A. Timofeeva, A. Borisova
Logistic regression model of student retention based on analysis of the Bolasso regularization path **371**

V. Timofeev, A. Veselova, K. Teselkina
Analysis of the methods of the Kriging family and GWR for transport speeds prediction models development **379**

N. Oleinik, V. Shchekoldin
Study of the properties of geometric ABOD-approach modifications for outlier detection by statistical simulation **389**

Yu. Mezentsev, O. Razumnikova, I. Tarasova, O. Trubnikova
On the clustering task of Big Data in medicine and neurophysiology **396**

T. Sumskaya
Problems of Sub-Federal budget policy in Russian Federation (The case of municipalities of the Novosibirsk Oblast) **404**

A. Feldman, N. Molokova, D. Rusin, N. Nikolaeva
Data analysis in studying the geological section **413**

M. Karaseva
Computer-aided approach to synthesis the specialized frequency dictionaries **421**

K. Pakhomova, P. Peresunko, S. Videnin, E. Soroka
The income prediction module of the retail store's network **428**

V. Stasyshin
Research of educational business processes in the decision making support system of University **436**

N. Antropov, E. Agafonov
Adaptive kernel identification of nonlinear stochastic dynamical systems **445**

A. Popov, V. Volkova
An optimal design of the experiment in the active identification of locally

adaptive linear regression models **453**

A. Imomov, E. Tukhtaev, N. Nuraliyeva
On invariant properties of critical Galton-Watson branching processes with infinite variance **461**

M. Krnjajić, R. Maslovskis
On some practical approaches of data science applied in forecasting and personalization **468**

A. Vostretsov, V. Vasyukov
Effect of sampling jitter in devices for discrete signal processing **482**

N. Zakrevskaya, A. Kovalevskii
An omega-square statistics for analysis of correspondence of small texts to the Zipf—Mandelbrot law **488**

A. Tyrsin, Ye. Chistova, A. Antonov
A scalar measure of interdependence between random vectors in problems for researching of multidimensional stochastic systems **495**

G. Agarkov, A. Sudakova, A. Tarasyev
Data Mining application features for scientific migration **502**

A. Sherstobitova, T. Emelyanova
On segmentation approach for time series of Arbitrary Nature **510**

D. Rusin, N. Molokova, A. Feldman, N. Nikolaeva
Computer analysis and interpretation of geophysical data **515**

T. Patrusheva, E. Patrushev
Statistical approach to detection of periodic signals under the background noise using the chaotic oscillator Murali-Lakshmanan-Chua **523**

M. Kovalenko, N. Sergeeva
Real-time multiple object tracking algorithm for adaptive traffic control systems **530**

V. Glinskiy, L. Serga, Yu. Ismaiylova, M. Alekseev
Disproportion of Russian Regions development in the sphere of population provision with food of own production **537**

B. Dobronets, O. Popova
A nonparametric approach for estimating the set of solutions of random linear programming **545**

K. Chirikhin, B. Ryabko
Application of artificial intelligence and data compression methods to time series forecasting **553**

N. Galanova
Approaches to customers lifetime value prediction **561**

N. Kononova, D. Zhalnin, O. Chubarova
About the task of leveling the "false" operations of the heat load regulator **566**

Features of testing statistical hypotheses under Big Data analysis

B. Yu. Lemeshko, S. B. Lemeshko and M. A. Semenova
Novosibirsk State Technical University, Novosibirsk, Russia
e-mail: lemeshko@ami.nstu.ru

Abstract

The methods of construction of estimates are considered in the analysis of Big Data. The influence on the results of conclusions according to the Pearson Chi-squared test of choosing the number of intervals and grouping method is demonstrated. It is shows how the limited accuracy of data in large samples effects on the distribution of statistics of non-parametric tests. Recommendations on the application of tests under large samples analysis are given. It is shown that the distribution of statistics of tests for testing laws homogeneity, as well as the tests of homogeneity of the means and tests of homogeneity of the variances, is affected by the non-equilibrium character of the data presented in the compared samples.

Keywords: Big Data; parameter estimation; testing hypotheses; goodness-of-fit tests; homogeneity tests; statistical simulation

Introduction

The questions of application of statistical methods to the analysis of large data arrays (Big Data) are of great interest in recent years. In connection with the rapid accumulation of gigantic volumes of information, there is a need for research the accumulated data, for finding, extracting and using the laws hidden in data, including probabilistic ones. Naturally, one can try to apply methods and tests from the vast arsenal of classical mathematical statistics for the analysis of big data, using popular software systems for statistical analysis. However, application of the classical apparatus of applied mathematical statistics for the analysis of big data, as a rule, leads to specific problems that limit the possibilities of correct application of this apparatus.

In this paper, we will discuss only the methods and tests associated with the analysis of one-dimensional random variables, the real problems of which are most familiar to us. At least three situations can be considered where increasing sample size causes problems in application of methods or tests.

Firstly, due to the "curse of dimension", well-proven methods and algorithms become ineffective. In particular, problems arise under the calculation of estimates of parameters. When using estimation methods that operate on non-grouped data, the computational costs increase cardinally with increasing size of samples analyzed. The convergence of iterative algorithms used in estimation worsens. A significant factor is no robustness of certain types of estimation. The natural way to resolve this situation is the use of estimation methods that involve grouping data [1]. But in this case, the

question arises: how the estimates obtained for grouped data will affect the properties of hypotheses tests in which estimates will be used. For example, how will this affect the statistics distributions of non-parametric goodness-of-fit tests when testing composite hypotheses? In this case, the statistic distributions significantly depend on the method of parameter estimation [2, 3, 4, 5].

Secondly, a lot of popular statistical tests are not adapted even for samples of about thousand observations, since the information on the distributions of statistics of these tests is presented only by brief tables of critical values for some sample sizes n. By rough estimate, the count of such tests is more than 80% of all tests count. It should be noted that the possibility of application such tests with "reasonable" values of sample size is easily resolved by statistical simulation of distributions of statistics for given sample size and validity of the tested hypothesis H_0. This simulation can be carried out interactively during statistical analysis [6, 7]. The empirical distribution $G_N(S_n\,|H_0)$ of statistic S of test constructed as a result of simulation with size N can then be used to estimate the achieved significance level p_{value} by the value of the statistics S^* calculated from the analyzed sample.

Thirdly, the application of tests, for which the limiting (asymptotic) distributions of statistics are known, always leads to rejection of even true tested hypothesis with increasing sample sizes. This is typical, for example, for goodness-of-fit tests, for a lot of special tests for testing hypotheses of normal distribution, uniform distribution or exponential distribution, etc. In [8], it has been shown that this problem is associated not only and not so much with the increasing computational costs, as with the limited accuracy of the analyzed data (with limited measurement accuracy). A similar problem hinders the correctness of application of homogeneity tests (homogeneity of laws, homogeneity of variance, to a lesser degree of homogeneity of means) under large samples. As will be shown, in the case of homogeneity tests, the reason lies in the unevenness of measurements in the analyzed samples.

1 Estimation of the parameters of distribution

Estimates of the parameters of distributions can be obtained by various methods. The maximum likelihood estimates (MLE) characterized by the best asymptotic properties and calculated by maximizing the likelihood function

$$\hat{\theta} = \arg\max_{\theta} \prod_{j=1}^{n} f(x_j, \theta), \tag{1}$$

or by maximizing the logarithm of this function, where θ is unknown parameter (generally vector), $f(x, \theta)$ is the density function of the distribution law, $x_1, x_2, ..., x_n$ are sample observation. For some laws, the distribution of MLE of parameters is obtained as statistics simply computed from the observations of the samples, but in most cases MLE are the result of using some iterative method.

When calculating MD-estimates (estimates of the minimum distance), some measure of proximity (distance) $\rho(F(x,\ \theta),\ F_n(x))$ between the theoretical $F(x,\ \theta)$ and

empirical $F_n(x)$ distributions is minimized. MD-estimates can be obtain as a result of solving following task

$$\hat{\theta} = \arg \min_{\theta} \rho(F(x,\theta), F_n(x)). \tag{2}$$

For example, the statistics of nonparametric goodness-of-fit tests (Kolmogorov, Cramer-von Mises-Smirnov, Anderson-Darling, Kuiper, Watson, and others [9]) can be used as measures of proximity.

With relatively small sample sizes, L-estimates of parameters can be used. These estimates are some linear combinations of order statistics (elements of variational series $X_{(1)} < X_{(2)} < \ ... < \ X_{(n)}$ constructed from original sample $x_1, x_2, ..., x_n$).

MLE of parameters of distribution, as a rule, are not robust. The presence of anomalies of sample observations or the inaccuracy of the assumption about the form of distribution leads to the construction of models with distribution functions that are unacceptably deviating from empirical distributions. MD-estimations have greater stability.

Obviously, the calculation of estimates (1) and (2) is associated with serious computational difficulties for very large samples. In the case of grouped sample, the sample observations are associated with a set of non-intersecting intervals, which divide the domain of definition of a random variable into k non-intersecting intervals by boundary points

$$x_{(0)} < x_{(1)} < \ldots < x_{(k-1)} < x_{(k)},$$

where $x_{(0)}$ is the lower bound of the domain of definition of random variable X; $x_{(k)}$ is the upper bound of the domain of definition of random variable X.

MLE by grouped sample [1] are calculated by maximizing the likelihood function

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{k} P^{n_i}(\theta), \tag{3}$$

(3) where $P_i(\theta) = \int\limits_{x_{(i-1)}}^{x_{(i)}} f(x,\ \theta)dx$ is the probability of the observation entering in the i-th interval of values, n_i is the number of observations that fell into the i-th interval, $\sum\limits_{i=1}^{k} n_i = n$. Estimates by grouped samples can be obtained by minimizing statistics χ^2

$$\hat{\theta} = \arg\min_{\theta}\ n\sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}, \tag{4}$$

as well as other statistics. In [10], it was shown that all of estimation method for grouped data considered give consistent and asymptotically effective estimates under appropriate regularity conditions. However, the most preferred estimates are MLE. An important advantage of estimates based on grouped data is robustness [11].

In the case of presence of non-grouped data, estimates for grouped data are rarely applied. This is due to the greater computational costs and necessity to numerical integration in the computation $P_i(\theta)$, that requires appropriate software support.

In the case of large sample sizes, the situation changes. Computational costs do not change as computations grow with a fixed number of grouping intervals, but increase only with an increase in the number of intervals k. This means that it is advisable to use MLE by grouped samples in the conditions of Big Data. These are robust and asymptotically efficient estimates. The quality of estimates for small k can be improved using asymptotically optimal grouping (AOG) [1, 12, 13], in which the losses in Fisher information associated with grouping are minimized.

2 Application of χ^2-test under large samples

The statistic of Pearson χ^2 goodness-of-fit test has the following form

$$X_n^2 = n \sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}. \tag{5}$$

In the case of testing simple hypothesis, this statistic obeys χ_r^2-distribution with $r = k - 1$ degrees of freedom if $n \to \infty$ and the null hypothesis is true.

In the case of testing composite hypothesis and estimating m parameters of distribution by sample statistic (4) obeys χ_r^2-distribution with $r = k - m - 1$ degrees of freedom, if the estimates are obtained by minimizing (4) these statistics, or using MLE (3) (or other asymptotically effective estimates for grouped data).

The distribution of statistic (5) does not agree with χ_{k-m-1}^2–distribution when parameter estimations are obtained by non-grouped data. It is recommended to apply the Nikulin-Rao-Robson test when MLE were obtained according to ungrouped data [14, 15].

There are not principal problems that prevent application of Pearson χ^2-test under Big Data. Only computational difficulties are possible.

Let us illustrate the results of application Pearson χ^2-test on the example of testing hypothesis of normal distribution with density

$$f(x, \theta) = \frac{1}{\theta_1 \sqrt{2\pi}} \exp\left\{ -\frac{(x - \theta_0)^2}{2\theta_1^2} \right\}.$$

by sufficiently large sample. The sample of $n = 10^7$ observations was modeled according to the standard normal law $N(0, 1)$ ($\theta_0 = 0$, $\theta_1 = 1$).

In Table 1, there are the results of testing simple hypotheses about standard normal law $N(0, 1)$ with various numbers of intervals in the case of equal-frequency grouping (EFG) and $k = 15$ in the case of asymptotically optimal grouping (AOG).

In the case of AOG, the power of Pearson χ^2-test maximizes for close competing laws [16, 17, 18]. The table shows the values X_n^{2*} of statistics (5), which calculated by the sample, and the corresponding values $p_{value} = P\{X_n^2 \geq X_n^{2*} | H_0\}$ of the achieved significance level. As you can see, the results depend on both the splitting method and the number of intervals. The power of test also depends on these factors [19].

Table 1: Results of testing simple hypothesis about $N(0,1)$

	AOG	EFG						
k	15	15	50	75	100	500	1000	2000
X_n^{2*}	7.75162	9.18380	56.8942	79.4904	96.5701	493.995	1044.57	2099.91
p_{value}	0.90186	0.81910	0.20475	0.31026	0.55038	0.55482	0.15403	0.05702

Table 2 shows the results of testing composite hypotheses. MLE $\hat{\theta}_0$ and $\hat{\theta}_1$ obtained for grouped data with the corresponding number of intervals k, statistics values X_n^{2*} and p_{value} are presented.

MLE of parameters by complete ungrouped sample are $\hat{\theta}_0 = 0.000274$ and $\hat{\theta}_1 = 1.000177$. In [20, 21], models of distributions of statistic (5) were constructed for the case of testing composite hypothesis of normal law using MLE by ungrouped data and AOG. The value of statistic calculated by the sample is $X_n^{2*} = 6.600521$ for $k = 15$, the estimate of p-value obtained in accordance with the limit distribution model given in [20, 21] is $p_{value} = 0.886707$. These values indicate a good agreement between the complete sample and the normal law $N(0.000274, 1.000177)$.

Table 2: Results of testing composite hypothesis

	AOG	EFG						
k	15	15	50	75	100	500	1000	2000
$\hat{\theta}_0$	0.00028	0.00030	0.000244	0.00027	0.00027	0.00028	0.00027	0.00027
$\hat{\theta}_1$	1.00715	1.00263	1.00173	1.00134	1.00112	1.00039	1.00031	1.00024
X_n^{2*}	927.920	99.9963	101.767	104.511	112.151	493.716	1043.47	2098.61
p_{value}	0.0	5.58e-16	6.50e-06	0.00739	0.13938	0.53317	0.14922	0.05572

It should be noted that the MLE by grouped sample for $k = 2000$ and the MLE by ungrouped sample are very close. At the same time, p-value for $k = 2000$ is much lower than 0.886707.

Thus, the result of testing composite hypotheses using Pearson χ^2-test significantly depends on the number of intervals k.

3 Nonparametric goodness-of-fit tests under big samples

If one can omit the growth of computational difficulties, the main reason for possible non-correctness of conclusions by big data using non-parametric goodness-of-fit tests is the limited accuracy of the data in large sample.

As a rule, volumes of samples in Big Data (belonging to some continuous distribution law) are practically unlimited, but the observations itself are presented with

limited accuracy (rounded with some Δ). In essence, there is "violation of assumption" that a continuous random variable is observed.

Suppose, the goodness-of-fit test with statistic S is used to test a simple hypothesis $H_0: \ F_n(x) = F(x)$, where $F_n(x)$ is empirical distribution constructed from sample

$$x_1, x_2, ..., x_n$$

of n observations. Suppose, there is limit distribution of statistic $G(S\,|H_0)$ for this goodness-of-fit test. In the case of trueness of H_0, the empirical distribution $F_n(x)$ corresponding to sample of continuous random variables (without rounding) converges to the distribution function of this random variable $F(x)$ for $n \to \infty$. The empirical distribution of statistics $G_N(S_n\,|H_0)$ based on samples of continuous random variable for $n \to \infty$ (and the number of simulation experiments $N \to \infty$) converges to the limit distribution $G(S\,|H_0)$ of this statistics.

However, the measurement results are rounded off (fixed) with some Δ. Therefore, $\max|F_n(x) - F(x)|$ will cease to decrease starting with certain n, depending on $F(x)$, domain of definition of the random variable and Δ. The distribution $G_N(S_n\,|H_0)$ will deviate from the limiting distribution $G(S\,|H_0)$ with increasing n (the more Δ, that the less n).

The results of studies for demonstrating the effect of accuracy of data on the distribution of statistics will be shown on 3 classical goodness-of-fit tests.

The Kolmogorov test statistics is used with the Bolshev correction[9]

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

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 number of observations; $x_1,\ x_2, \ldots,\ x_n$ are sample values ordered ascending; $F(x,\ \theta)$ is distribution function of law tested. The distribution of S_K under simple hypothesis in the limit obeys the Kolmogorov law with the distribution function $K(S)$ [9].

The Cramer-von Mises-Smirnov test statistic is

$$S_\omega = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i,\ \theta) - \frac{2i-1}{2n}\right\}^2 \tag{7}$$

and under testing simple hypothesis this statistic allows to law with distribution function $a1(s)$ [9]. The Anderson-Darling test statistic has the following form [22]

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\ \theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i,\ \theta))\right\}. \tag{8}$$

In the case of testing simple hypothesis this statistic allows to law with distribution function $a2(s)$ [9].

In [8], the distributions of statistics (6)-(8) of nonparametric goodness-of-fit tests were studied depending on the accuracy of recording the observed values of random

variables. The number of significant decimal places, to which the observed values were rounded, was set. This determined the number of unique values that could be in the generated samples. As a rule, the number of simulation experiments carried out to simulate the empirical distributions of statistics was $N = 10^6$.

The deviation of real (empirical) distribution of statistics from the limit distribution was studied by evaluating median $\tilde{S}_n$ of empirical distribution of statistics obtained as a result of modeling. If real distribution of statistics with sample sizes n does not deviate from the limit distribution, then the probability $P\left\{S > \tilde{S}_n\right\}$ calculated from the corresponding limit distribution is 0.5. If real distribution of statistics shifts to large area of values (to the right of the limit distribution), the estimates $\hat{p}_v = P\left\{S > \tilde{S}_n\right\}$ are decrease. One can judge the correctness of achieved significance level p_{value} calculated from the limit distribution of statistics (in the case of testing simple hypotheses, respectively, by $K(S)$, $a1(S)$ and $a2(S)$) by the value of deviation of estimates $\hat{p}_v$ from 0.5.

When rounding to within 1 in samples belonging to $N(0,1)$, 9 unique values may appear, when rounding to within $\Delta = 0.1$ about 86 unique values, with accuracy $\Delta = 0.01$ – about 956, to within $\Delta = 0.001$ – about 9830.

As the simulation results showed [8], when rounding up observations to integer values, the use of limit distributions of test statistics is **absolutely** excluded.

The distributions of statistic of Kolmogorov test $G(S_n\,|H_0)$ is essentially discrete under $\Delta = 0.1$. The deviation $G(S_n\,|H_0)$ from the limit distribution $K(S)$ for $\Delta = 0.1$ should be taken into account already for $n > 20$, $\Delta = 0.01$ – for $n > 250$, and if $\Delta = 0.001$ the value $n_{\max}$ shifts to value about 10^4. In the case of Cramer-von Mises-Smirnov and Anderson-Darling tests, the deviation $G(S_n\,|H_0)$ from the limit $a1(S)$ and $a2(S)$ for $\Delta = 0.1$ should be taken into account for $n > 30$, $\Delta = 0.01$ – for $n > 1000$, and if $\Delta = 0.001$ – the value $n_{\max}$ shifts to 5×10^5.

Figure 1 shows the dependence of distributions of statistics (7) of Cramer-von Mises-Smirnov test on the degree of rounding Δ at sample size $n = 1000$ for the case of testing simple hypothesis about standard normal law. The limit distribution $a1(S)$, that occurs in the case without rounding, as well as real distributions of statistics $G(S_{1000}\,|H_0)$ at degree of rounding $\Delta = 0.01,\ 0.05,\ 0.1,\ 0.2,\ 0.3$. As you can see if $\Delta = 0.01$ distribution $G(S_{1000}\,|H_0)$ does not practically differ from $a1(S)$, but with increasing Δ deviation $G(S_{1000}\,|H_0)$ from $a1(S)$ rapidly increases.

Consequently, in order to analyze large samples using the appropriate nonparametric goodness-of-fit tests with corresponding limit distributions, statistics should be calculated not over the sample, but according to samples extracted by uniform law from general population (original sample analyzed). The size of extracted sample should take into account the accuracy of the data (the number of possible unique values in the sample) and not exceed certain value $n_{\max}$ at which (for given accuracy) the distribution of test statistics $G(S_{n_{\max}}\,|H_0)$ does not really differ from the limit distribution $G(S\,|H_0)$.

In the case of testing composite hypotheses, the tested hypothesis has the form $H_0 : F(x) \in \{F(x,\,\theta), \theta \in \Theta\}$, where Θ is domain of parameter θ definition. If

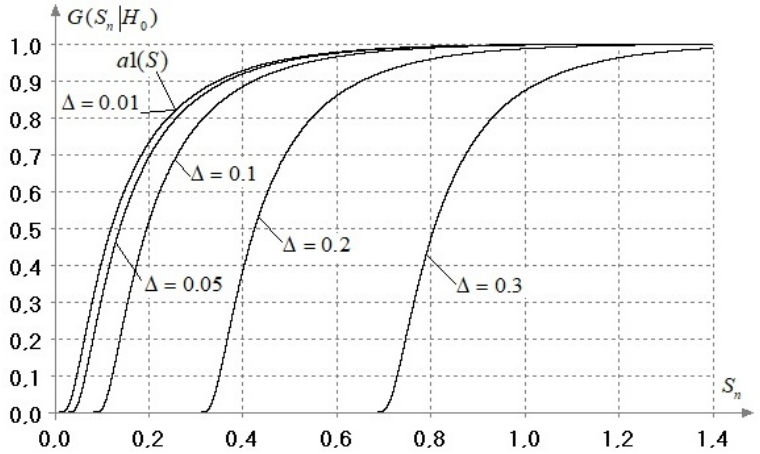

Figure 1: Statistic distributions $G(S_n\,|H_0)$ of Cramer-von Mises-Smirnov test depending on Δ for $n = 1000$

the estimate $\hat{\theta}$ of scalar or vector parameter of law is based on the same sample that the hypothesis is tested on, then the distribution of statistics $G(S\,|H_0)$ for any nonparametric goodness-of-fit test differs significantly from the limit distribution for testing simple hypothesis [23]. If estimates of parameters obtain by the same sample that hypothesis tested, the following factors influence the distribution of statistics $G(S\,|H_0)$ [24]: distribution law $F(x,\,\theta)$ corresponding to the true hypothesis H_0; type of estimated parameter and the number of estimated parameters; in some situations, specific values of parameter (for example, in the case of gamma distribution, etc.); used parameter estimation method.

Obviously, in the case of testing composite hypotheses, we encounter the same problems and must extract sample of size $n < n_{\max}$ from "general population" in order to use when analyzing Big Data with limited accuracy of fixed data. For example, it should be do for application of models of limit distributions of test statistics when testing composite hypotheses [2, 3, 4, 5, 24].

It should be noted, if the estimation $\hat{\theta}$ of parameter is found by one of the above methods by the entire big data array, and then the test is applied to the sample of size $n < n_{\max}$ extracted from the same array, then when testing hypothesis H_0 : $F(x) = F(x, \hat{\theta})$, where $\hat{\theta}$ is previously obtained estimate, the distribution of statistics $G(S\,|H_0)$ will as in the case of testing simple hypothesis.

All of the above fully applies to application of nonparametric Kuiper [25] and Watson [26, 27] goodness-of-fit tests by big samples. The distributions of statistics of third Zhang goodness-of-fit tests [28], which are based on Kolmogorov, Cramer-von Mises-Smirnov and Anderson-Darling tests, depend on sample sizes n. Therefore, there can be no talk about application of limit distributions of statistics. However, distribution of statistics $G(S_n\,|H_0)$ in the same way depends on degree of rounding Δ. Consequently, the critical values of statistics obtained for continuous random variables and n cannot be used with the same n, but with significant degree of rounding Δ. The

problem can be resolved by statistical modeling (including, in the interactive mode [6, 7]) of statistical distributions for given n and Δ with the trueness of the tested hypothesis H_0. The empirical distribution of $G_N(S_n\,|H_0)$ statistics S of corresponding test constructed as a result of N simulation experiments under these conditions can be used to estimate the achieved significance level p_{value}. That is how this problem is solved in the ISW software system being developed [29].

4 Other goodness-of-fit tests under big samples

It should be noted that the degree of rounding of recorded data affects properties of other tests in similar way. In particular, special tests aimed for testing the hypothesis about normal law, uniform law, or exponential law, etc.

It should be noted that in the conditions of large samples (in the presence of repeated observations), a lot of good tests turn out to be inoperable. This is due to the fact that the type of statistics of these tests excludes the presence of repeated observations (or the number of repeated values greater than the size of the "m window" used in statistics). This note concerns tests using entropy estimates (Vacicek [30] and Alizadeh Noughabi [31] normality tests, Dudewics-van der Meulen [32] and Zamanzade [33] uniformity tests), as well as new goodness-of-fit tests using estimates of Kullback-Leibler information [34].

5 Homogeneity tests under big samples

In the case of multi-sample tests, which include homogeneity tests, 2 or more samples are compared. The distributions of statistics of multi-sample tests are influenced by non-uniformity of data presented in the analyzed samples. The two-sample Lehmann-Rosenblatt homogeneity test was proposed in [35] and studied in [36]. Statistic based on two samples $x_{11}, x_{12}, ..., x_{1,n_1}$ and $x_{21}, x_{22}, ..., x_{2,n_2}$:

$$S_{LR} = \frac{1}{n_1 n_2 (n_1 + n_2)} \left[n_1 \sum_{i=1}^{n_1} (r_i - i)^2 + n_2 \sum_{j=1}^{n_2} (s_j - j)^2 \right] - \frac{4 n_1 n_2 - 1}{6(n_1 + n_2)}, \qquad (9)$$

where r_i is serial number (rank) of x_{1i}; s_j is serial number (rank) of x_{2j} in the united variation range.

The limit distribution of statistic (9) under true tested hypothesis $H_0: \; F_1(x) = F_2(x)$ is the same distribution $a1(s)$ [36], which is limit for statistic of Cramer-von Mises-Smirnov goodness-of-fit test.

Let us consider how degree of rounding affects distribution of statistic of homogeneity tests in the case of true H_0 and belonging of analyzed sample observations to the standard normal law.

Figure 2 demonstrates the dependence of distribution of statistic $G(S_{LR}\,|H_0)$ of Lehmann-Rosenblatt homogeneity test on degree of rounding Δ_2 of observations in the second sample when rounding in the first sample $\Delta_1 = 0.01$. The sample sizes are $n_i = 1000$.

The deviation $G(S_{LR}\,|H_0)$ from $a1(S)$ turns out to be significant already for $\Delta_2 = 0.05$. The deviation $G(S_{LR}\,|H_0)$ from $a1(S)$ rapidly increases with increasing sample sizes for fixed Δ_2. The deviation increases with Δ_2 growth for fixed sample size. The distributions of statistic $G(S_{LR}\,|H_0)$ of Lehmann-Rosenblatt homogeneity test depend on the difference between Δ_1 and Δ_2.

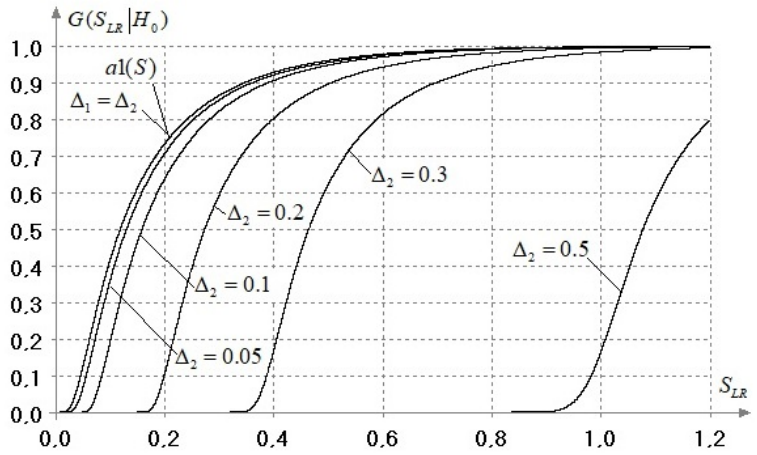

Figure 2: Statistic distributions of Lehmann-Rosenblatt homogeneity test depending on Δ_2 for $\Delta_1 = 0.01$ and $n_i = 1000$

Similarly, the distributions of other two-sample homogeneity tests (Smirnov, Anderson-Darling-Pettitt) depend on the difference between Δ_1 and Δ_2. It is natural that the distributions of statistics of all multi-sample tests of homogeneity (set of which is considered in [37]) depend on the non-equivalence of data presentation in the analyzed samples.

The distributions of statistic of parametric tests of homogeneity of means do not suffer from such dependence on degree of rounding of measurements as tests of homogeneity of laws considered above. At the same time, it should be noted that the power of tests decreases with decrease of accuracy of data recorded.

The distributions of statistic of parametric tests of homogeneity of variances, unlike tests of homogeneity of means, are more dependent on degree of rounding. In some ways, this is due to the greater sensitivity of the variance estimates to the accuracy of measurement results.

Parametric tests of Cochran, Bartlett, Fisher, Hartley, Neumann-Pearson and Overall-Woodward Z-test are the most preferable in terms of power among the set of parametric and non-parametric tests of homogeneity of variances. These tests are equivalent in power in the case of two sample and fulfilling the assumption that analyzed samples are normal. But in the case of k sample, the power advantage turns out to be Cochran test has power advantage [38, 39, 40, 41]. Statistic of Cochran test [42] can be written as

$$Q = \frac{S^2_{\max}}{S_1^2 + S_2^2 + \cdots + S_k^2},$$

where $S^2_{\max} = \max(S^2_1,\ S^2_2, \ldots,\ S^2_k)$; k is the number of samples; $S^2_i,\ \ i = \overline{1,k}$, are the estimates of variances obtained by samples. Tested hypothesis $H_0:\ \sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_k$ deviates for large values of statistic. The distributions of statistic $G(Q_n\,|H_0)$ of Cochran test depend on the number of compared samples k and the sizes of these samples n_i.

Figure 3 illustrates the dependence of the distribution of statistics $G(Q_n\,|H_0)$ of Cochran test on degree of rounding of observations in the second sample Δ_2 without rounding in the first sample ($\Delta_1 = 0$). Sample sizes are $n_i = 1000$ and k=2. As can be seen, the dependence of the distribution $G(Q_n\,|H_0)$ on large (different) degrees of rounding Δ_1 and Δ_2 is very significant.

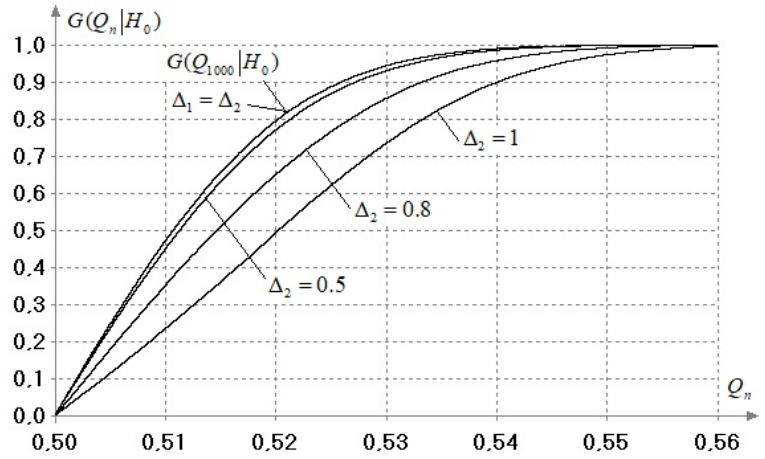

Figure 3: Statistic distributions $G(Q_n\,|H_0)$ of Cochran homogeneity test depending on Δ_2 for $\Delta_1 = 0$ and $n_i = 1000$

The limited accuracy of measurements always leads to decrease of the power of homogeneity tests. The drop in the power of Cochran test with increasing degree of rounding (with equal Δ_i , equal sample sizes $n_1 = n_2 = 100$, and k=2) is shown in Table 3. The competing hypothesis has the form $H_1:\ \ \sigma_2 = 1.2\sigma_1$. Also this table shows power of Klotz nonparametric test [43]. It is interesting that with increasing Δ_i the power of nonparametric test decreases faster than power of parametric one.

Let us emphasize that, similarly, the value of rounding Δ_i affects the distributions of statistics and the power of other tests of homogeneity of variances.

So, the distributions of statistics $G(S\,|H_0)$ of parametric tests of homogeneity of variances with the same degree of rounding Δ_i of measurement in the analyzed samples do not differ from corresponding distributions without rounding ($\Delta_i = 0,\ \ i = \overline{1,k}$). However, the same distributions with different Δ_i differ significantly from distributions without rounding.

In the case of trueness of competing hypotheses, degree of rounding Δ_i (measurement registration accuracy) has significant impact on the distributions of statistics and on the power relative to these competing hypotheses (including under equal Δ_i

Table 3: Estimates of power of Cochran and Klotz tests under H_1

	Cochran test			
α	Without rounding	$\Delta_1 = \Delta_2 = 0.1$	$\Delta_1 = \Delta_2 = 0.2$	$\Delta_1 = \Delta_2 = 0.5$
0.1	0.564	0.562	0.560	0.550
0.05	0.438	0.435	0.434	0.424
	Klotz test			
α	Without rounding	$\Delta_1 = \Delta_2 = 0.1$	$\Delta_1 = \Delta_2 = 0.2$	$\Delta_1 = \Delta_2 = 0.5$
0.1	0.540	0.539	0.535	0.504
0.05	0.413	0.412	0.407	0.378

in samples). Similar conclusions hold for the entire set of parametric tests of homogeneity of variances considered in [37].

Conclusions

It is advisable to use parameter estimation methods involving the grouping of data for constructing probabilistic models by big samples. Such estimates are robust, and computational costs do not depend on sample sizes in contrast to estimates by ungrouped data.

There are no serious objections to application of Pearson χ^2-test for analysis of big samples. This test retains both its positive qualities and its inherent flaws.

The main problem preventing the correct application of nonparametric goodness-of-fit tests for analysis of big samples is limited accuracy of data representation. Due to limited accuracy with increasing sample volumes, the real distributions of statistics deviate from the limit ones that occur under the assumption of continuity of observed random variables. Therefore, the application of classical results for corresponding tests may lead to incorrect conclusions. On the one hand, it is possible to recommend application of these tests to samples extracted from Big data, the size of these samples is limited by accuracy of presenting data analyzed (the number of possible unique values in the sample). On the other hand, it is possible to propose the use of statistical modeling methods to estimate real distributions of test statistics $G_N(S_n\,|H_0)$ (corresponding to degree of rounding Δ of data in sample analyzed) and then use $G_N(S_n\,|H_0)$ to estimate achieved significance level p_{value}.

The reason for possible incorrectness of conclusions when using classical results concerning the distributions of statistics of corresponding homogeneity tests may be the non-equilibrium measurement in the compared samples. Statistical modeling can be proposed to simulate actual distribution of statistics $G_N(S_n\,|H_0)$ of test applied (with appropriate degrees of rounding Δ_i and sizes n_i of compared samples). The distribution $G_N(S_n\,|H_0)$ obtained can then be used to estimate achieved significance level p_{value}.

Similar methodology of analysis of big samples is implemented in ISW software system [29].

Acknowledgements

The research is supported by the Russian Ministry of Education and Science (projects No 1.4574.2017/6.7 and No 1.1009.2017/4.6).

References

[1] Kulldorff G. Contributions to the theory of estimation from grouped and partially grouped samples. Almqvist&Wiksell. 1961.

[2] Lemeshko B.Yu, Lemeshko S.B. Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. P. I // Measurement Techniques. 2009. Vol. 52, No. 6. – P. 555-565.

[3] Lemeshko B.Yu., Lemeshko S.B. Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. P. II // Measurement Techniques. 2009. Vol. 52, No. 8. – P. 799-812.

[4] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N. Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics – Theory and Methods. 2010. Vol. 39, No. 3. – P. 460-471.

[5] Lemeshko B.Yu., Lemeshko S.B. Construction of Statistic Distribution Models for Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses: The Computer Approach // Quality Technology & Quantitative Management. 2011. Vol. 8, No. 4. – P. 359-373.

[6] Lemeshko B.Yu., Lemeshko S.B., Rogozhnikov A.P. Real-Time Studying of Statistic Distributions of Non-Parametric Goodness-of-Fit Tests when Testing Complex Hypotheses // Proceedings of the International Workshop "Applied Methods of Statistical Analysis. Simulations and Statistical Inference" – AMSA'2011, Novosibirsk, Russia, 20-22 September, 2011. – P. 19-27.

[7] Lemeshko B.Yu., Lemeshko S.B., Rogozhnikov A.P. Interactive investigation of statistical regularities in testing composite hypotheses of goodness of fit // Statistical Models and Methods for Reliability and Survival Analysis : monograph. – Wiley-ISTE, 2013. – Chap. 5. – P. 61-76.

[8] Lemeshko B.Yu., Lemeshko S.B., Semenova M.A. To Question of the Statistical Analysis of Big Data // Tomsk State University Journal of Control and Computer Science. 2018. 44. – P. 40-49. DOI: 10.17223/19988605/44/5

[9] Bolshev L.N., Smirnov N.V. Tables for Mathematical Statistics. Moscow: Nauka. 1983. – 416 p.

[10] Rao S.R. Linear statistical methods and their applications. Moscow: Nauka. 1968. – 548 p.

[11] Lemeshko B.Yu. Grouping observations as a way to generate robust estimates // Nadezhnost' i kontrol' kachestva. 1997. 5. – P. 26-35.

[12] Denisov V.I., Lemeshko B.Yu., Tsoi E.B. Optimal grouping, parameter estimation, and planning regression experiments. In 2 parts // Novosibirsk : NSTU Publisher, 1993. – 347 p.

[13] Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N., Chimitova E.D. Statistical Data Analysis, Simulation and Study of Probability Regularities. Computer Approach / B.Yu. Lemeshko, S.B. Lemeshko, S.N. Postovalov, E.V. Chimitova. Novosibirsk : NSTU Publisher, 2011. – 888 p.

[14] Nikulin M.S. On chi-square testing for continuous distributions // Theory of Probability & Its Application. 1973. Vol. 18. No. 3. – P. 675-676.

[15] Rao K.C., Robson D.S. A chi-squared statistic for goodness-of-fit tests within the exponential family // Commun. Statist. 1974. Vol. 3. – P. 1139-1153.

[16] Denisov V.I., Lemeshko B.Yu. Optimal grouping in the processing of experimental data // Measuring Information Systems. Novosibirsk. – P. 5-14.

[17] Lemeshko B.Yu. Asymptotically optimal grouping of observations is to ensure the maximum power of the tests // Nadezhnost' i kontrol' kachestva. 8. – P. 3-14.

[18] Lemeshko B.Yu. Asymptotically optimum grouping of observations in goodness-of-fit tests // Industrial laboratory. Diagnostics of materials. 1998. 64(1). – P. 56-64.

[19] Lemeshko B.Yu., Chimitova E.V. On the choice of the number of intervals in the goodness-of-fit tests of type χ^2 // Industrial laboratory. Diagnostics of materials. 2003. 69(1). – P. 61-67.

[20] Lemeshko B.Yu. Tests for checking the deviation from normal distribution law. Moscow: INFRA–M. 2015. DOI: 10.12737/6086

[21] Lemeshko B.Yu. Chi-Square-Type Tests for Verification of Normality // Measurement Techniques, 2015. Vol. 58, No. 6. – P. 581-591. DOI: 10.1007/s11018-015-0759-2

[22] Anderson T.W., Darling D.A. A test of goodness of fit // J. Amer. Statist. Assoc. 1954. Vol. 29. – P. 765-769.

[23] Kac M., Kiefer J., Wolfowitz J. On tests of normality and other J. tests of goodness of fit based on distance methods // Ann. Math. Stat. 1955. Vol. 26. – P. 189-211.

[24] Lemeshko B.Yu. Nonparametric goodness-of-fit tests. Guide on the application. M: INFRA–M, 2014. – 163 p. DOI: 10.12737/11873

[25] Kuiper N.H. Tests concerning random points on a circle // Proc. Konikl. Nederl. Akad. Van Wettenschappen. 1960. Series A. V.63. – P. 38-47.

[26] Watson G.S. Goodness-of-fit tests on a circle. I // Biometrika. 1961. V. 48. No. 1-2. – P. 109-114.

[27] Watson G.S. Goodness-of-fit tests on a circle. II // Biometrika. 1962. V. 49. No. 1-2. – P. 57- 63.

[28] Zhang J. Powerful goodness-of-fit tests based on the likelihood ratio // Journal of the Royal Statistical Society: Series B. 2002. V.64. No. 2. – P. 281-294.

[29] ISW – Program system of the statistical analysis of one-dimensional random variables. URL: https://ami.nstu.ru/ headrd/ISW.htm. (address date 12.05.2019).

[30] Vacicek O. A test for normality based on sample entropy // Journal of the Royal Statistical Society: Series B. 1976. V. 38, No. 1. – P. 54-59.

[31] Alizadeh Noughabi H. A new estimator of entropy and its application in testing normality // Journal of Statistical Computation and Simulation. 2010. V. 80. No. 10. – P. 1151-1162.

[32] Dudewics E. J., van der Meulen E. C. Entropy-based test of uniformity // J. Amer. Statist. Assoc. 1981. V. 76. No. 376. – P. 967-974.

[33] Zamanzade E. Testing uniformity based on new entropy estimators // Journal of Statistical Computation and Simulation. 2015. V. 85. No. 16. – P. 3191-3205.

[34] Alizadeh Noughabi H. A new estimator of Kullback-Leibler information and its application in goodness of fit tests // Journal of Statistical Computation and Simulation. 2019. V. 89. No. 10. – P. 1914-1934.

[35] Lehmann E. L. Consistency and unbiasedness of certain nonparametric tests // Ann. Math. Statist. – 1951. – Vol. 22, – No. 1. – P. 165-179.

[36] Rosenblatt M. Limit theorems associated with variants of the von Mises statistic // Ann. Math. Statist. 1952. Vol. 23. – P. 617-623.

[37] Lemeshko B.Y. Tests for homogeneity. Guide on the application. M: INFRA–M, 2017. – 207 p. DOI: 10.12737/22368

[38] Lemeshko B.Yu., Lemeshko S.B., Gorbunova A.A. Application and power of criteria for testing the homogeneity of variances. Part I. Parametric criteria // Measurement Techniques. – 2010. – Vol. 53, No. 3. – P. 237-246.

[39] Lemeshko B.Yu., Lemeshko S.B., Gorbunova A.A. Application and power of criteria for testing the homogeneity of variances. Part II. Nonparametric criteria // Measurement Techniques, Vol. 53, No. 5, 2010. – P. 476-486.

[40] Lemeshko B.Y., Sataeva T.S. Application and Power of Parametric Criteria for Testing the Homogeneity of Variances. Part III // Measurement Techniques, 2017. Vol. 60. No. 1. – P. 7-14.

[41] Lemeshko B.Y., Sataeva T.S. Application and Power of Parametric Criteria for Testing the Homogeneity of Variances. Part IV // Measurement Techniques, 2017. Vol. 60. No. 5. – P. 425-431.

[42] Cochran W. G. The distribution of the largest of a set of estimated variances as a fraction of their total // Annals of Eugenics. 1941. Vol. 11. – P. 47-52.

[43] Klotz J. Nonparametric tests for scale // Ann. Math. Stat. 1962. Vol. 33. – P. 498-512.