

UDC 519.23+519.245

Doi: 10.31772/2587-6066-2019-20-1-28-34

For citation: Ushanov S. V., Ogurtsov D. A. [Estimation of the Frocini criteria and omega square criteria statistics by the statistical tests method for a mixture of normal distributions]. *Siberian Journal of Science and Technology*. 2019, Vol. 20, No. 1, P. 28–34. Doi: 10.31772/2587-6066-2019-20-1-28-34

Для цитирования: Ушанов С. В., Огурцов Д. А. Оценка методом статистических испытаний статистики критериев Фроцини и омега-квадрат для смеси нормальных распределений // Сибирский журнал науки и технологий. 2019. Т. 20, № 1. С. 28–34. Doi: 10.31772/2587-6066-2019-20-1-28-34

ESTIMATION OF THE FROCINI CRITERIA AND OMEGA SQUARE CRITERIA STATISTICS BY THE STATISTICAL TESTS METHOD FOR A MIXTURE OF NORMAL DISTRIBUTIONS

S. V. Ushanov*, D. A. Ogurtsov

Reshetnev Siberian State University of Science and Technology
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660037, Russian Federation

*E-mail: ushanov_sv@mail.ru

A lot of sets of subjects and objects in biology, industry, management can be divided into a number of classes, each of which corresponds to a certain distribution component. When analyzing a mixture of distributions, it is necessary to estimate its parameters (task 1) and to assess the correspondence of empirical and theoretical distribution functions (task 2).

To solve the first problem, numerical algorithms that implement the method of moments and the maximum likelihood method are used. In this paper, the problem of estimating the distribution parameters is solved by minimizing the goodness measure by the Quasi-Newton method.

The second problem is solved by comparing the empirical and theoretical distribution functions by one or several statistical goodness measures. Statistics of the distribution of these measures depends on the sample size, the method of forming data and estimating distribution parameters. The paper examines the goodness measure between Frocini and omega-square (Kramer – Mises – Smirnov). The evaluation of the statistics of the goodness measure was carried out by the simulation method based on the results of 50000 statistical tests. In each of the tests, the distribution parameters were estimated by minimizing the calculated value of the corresponding goodness measure. The results of simulation modeling allow estimating the statistics of the parameters of a mixture of distributions.

The results of solving the considered problems for a mixture of two normal distributions of size 240 are presented.

Keywords: Frocini statistics, omega-square statistics, statistical tests, mixture of distributions.

ОЦЕНКА МЕТОДОМ СТАТИСТИЧЕСКИХ ИСПЫТАНИЙ СТАТИСТИКИ КРИТЕРИЕВ ФРОЦИНИ И ОМЕГА-КВАДРАТ ДЛЯ СМЕСИ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

С. В. Ушанов*, Д. А. Огурцов

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31

*E-mail: ushanov_sv@mail.ru

Многие совокупности субъектов и объектов в биологии, промышленности, управлении можно условно разделить на ряд классов, каждому из которых соответствует определенная компонента смеси распределения. При анализе смеси распределений необходима оценка ее параметров (задача 1) и оценка соответствия эмпирической и теоретической функций распределения (задача 2).

Для решения первой задачи обычно применяют численные алгоритмы, реализующие метод моментов и метод максимального правдоподобия. В работе задача оценки параметров распределения решается минимизацией критерия согласия квазиньютоновским методом.

Вторая задача решается сравнением эмпирической и теоретической функций распределения одним или несколькими статистическими критериями согласия. Статистика распределения этих критериев зависит от объема выборки, способа формирования данных и оценки параметров распределения. В работе рассматриваются критерии согласия Фроцини и омега-квадрат (Крамера – Мизеса – Смирнова). Оценка статистики

критериев согласия проводилась методом имитационного моделирования по результатам 50000 статистических испытаний. В каждом из испытаний параметры распределения оценивались минимизацией расчетного значения соответствующего критерия согласия. Результаты имитационного моделирования позволяют оценить статистику параметров смеси распределений.

Представлены результаты решения рассмотренных задач для смеси двух нормальных распределений объемом 240.

Ключевые слова: статистика Фроцини, статистика омега-квадрат, статистические испытания, смесь распределений.

Introduction. One of the tasks of the initial processing of experimental observations is the choice of the distribution law, which adequately describes the random variable for the observed sample. A great number of sets of subjects and objects in biology, industry, management can be divided into a number of classes, each of which corresponds to a specific component of the distribution mix. In biological populations, it is possible to distinguish objects with average values of indicators, objects – indicators which are higher than average (“leaders”) and objects – indicators that are lower than average (“outsiders”) [1]. The dynamics of mass transfer processes of chemical technology depends on the size distribution of the raw materials, which is also determined by a mixture of distributions [2–4].

When analyzing a mixture of distributions, it is necessary to estimate its parameters (task 1) and to evaluate the compliance of empirical and theoretical distribution functions (task 2).

To solve the first problem, usually numerical algorithms are used that implement the method of moments [5] and the maximum likelihood method [6–8]. The peculiarity of this problem solution by the maximum likelihood method for a mixture of distributions is the presence of several local extrema. In this paper, the problem of estimating the distribution parameters is solved by minimizing the agreement criterion by Quasi-Newton methods in MathCad [9] and MATLAB [10] environments.

The second problem is solved by comparing the empirical and theoretical distribution functions by one or several statistical criteria of agreement [5; 11]. Statistics of the distribution of these criteria depends on the sample size, the method of forming data and estimating distribution parameters [12]. The paper examines the criteria of consent Frocini [13; 14]

$$Fr(Xv, a) = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^n \left| F(Xv_i, a) - \frac{i-0.5}{n} \right|,$$

and omega square (Kramer – Mises – Smirnov) [15; 16]

$$KMC(Xv, a) = \frac{1}{12n} + \sum_{i=1}^n \left(F(Xv_i, a) - \frac{i-0.5}{n} \right)^2,$$

where Xv – variational series of random variable X ; n – sample size; i – number of the element of the variation series; a – distribution parameters; $F(Xv_i, a)$ – the value of the integral distribution function for the element of a variational series Xv_i .

The probability density function for a mixture of distributions consisting of K components has the form:

$$f(x, a, \mu) = \sum_{j=1}^k \mu_j \cdot f_j(x, a_j), \quad \sum_{j=1}^k \mu_j = 1,$$

where x – random value; a, μ – distribution parameters; μ_j – the proportion of the j -th component in the mixture.

For a mixture of normal distributions, the probability density of the j -th component is determined by the expression

$$f_j(x, a_j) = \frac{1}{a_{j,0} \cdot \sqrt{2\pi}} \cdot \exp \left(-\frac{1}{2} \left(\frac{x - a_{j,0}}{a_{j,1}} \right)^2 \right),$$

where $a_{j,0}, a_{j,1}$ – estimates of expected value and standard deviation.

The computer approach developed in the works of B. Yu. Lemeshko makes it possible to evaluate the statistics of the compliance criteria when testing various complex hypotheses [10; 16].

When conducting statistical tests, it is necessary to take into account the repetition period of the generated pseudo-random numbers. In the MathCad system, this period for a generator of normally distributed random variables is $784.4 \cdot 10^6$ [17]. For sample size $n = 1000$, this allows to conduct $7 \cdot 10^5$ statistical tests. At the level of significance $\alpha \in [0.001; 0.999]$, the maximum error in estimating the statistics of the criteria under consideration does not exceed 0.0005 [14].

Results of computational experiments. The paper discusses the application of the Frocini criteria [18] and omega-square in estimating the distribution parameters for the analyzed sample by minimizing the calculated value of the corresponding criterion. In each computational experiment for evaluating the statistics of the compliance criteria, 50000 statistical tests were conducted.

In fig. 1 the experimental errors in determining the hydrodynamic quality of the whip beams with a limited buoyancy margin are shown [19] (sample size $n = 240$), in fig. 2 distribution functions that approximate the empirical data with a mixture of two normal distributions are presented; in tab. 1, estimates of distribution parameters obtained by minimizing the Frocini criterion and omega-square are presented.

The maximum deviation between the integral functions of the mixture of distributions, the parameters of which are obtained by minimizing the Frocini criteria and the omega-square is 0.001 for $x = -0.13$, and between the probability density functions is 0.0078 for $x = 0.10$.

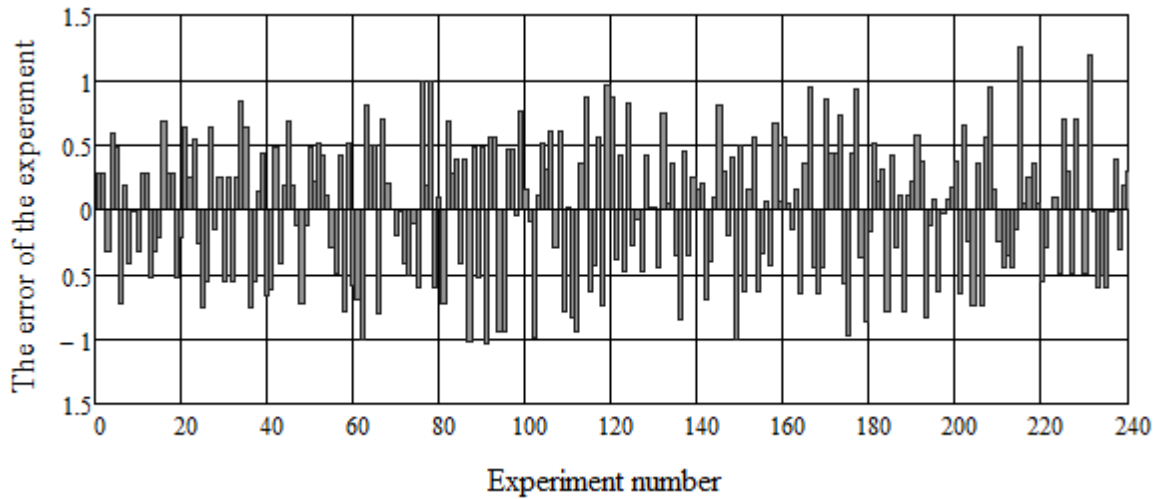


Fig. 1. Experimental errors in determining the hydrodynamic quality of whip beams with a limited buoyancy margin [19]

Рис. 1. Ошибки экспериментов при определении гидродинамического качества хлыстовых пучков с ограниченным запасом плавучести [19]

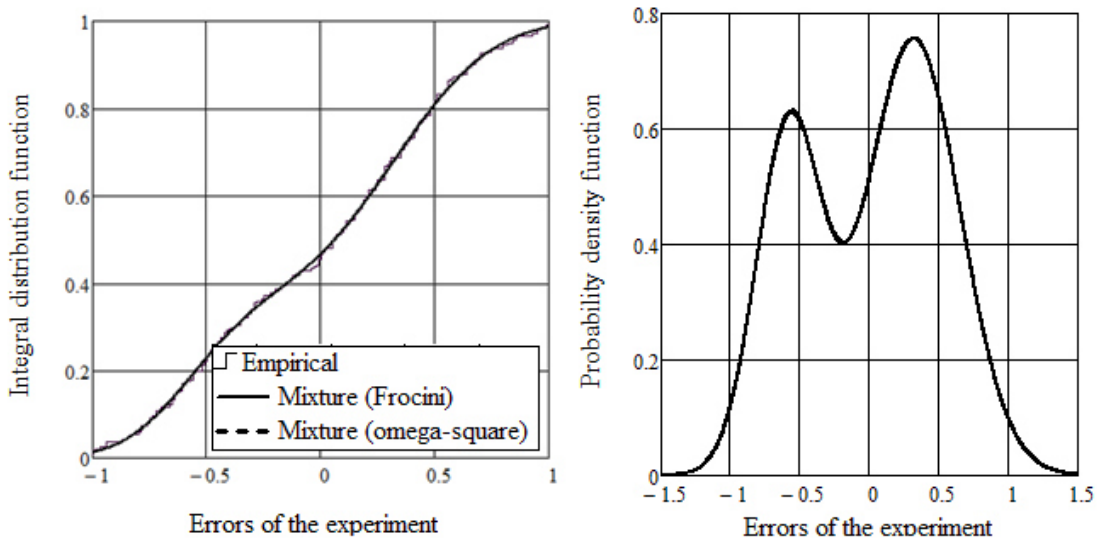


Fig. 2. Empirical and theoretical function of normal distributions mixture

Рис. 2. Эмпирическая и теоретическая функции смеси нормальных распределений

Table 1

The optimal values of the parameters of the mixture of distributions and their estimates obtained by statistical testing ($M = 5000$, $n = 240$) by minimizing the Frocini criterion and omega-square

Parameter	Optimal value	Expected value	Median	Borders of 95 % Confidence Interval		
				lower	upper	
$a_{1.0}$	*	-0.574	-0.569	-0.575	-0.672	-0.437
	**	-0.576	-0.574	-0.580	-0.671	-0.450
$a_{1.1}^2$	*	0.0566	0.0588	0.0556	0.0279	0.112
	**	0.0549	0.0545	0.0510	0.0249	0.105
$a_{2.0}$	*	0.322	0.318	0.320	0.198	0.438
	**	0.318	0.317	0.318	0.199	0.434
$a_{2.1}^2$	*	0.104	0.119	0.116	0.067	0.191
	**	0.103	0.118	0.116	0.068	0.188
μ_1	*	0.361	0.367	0.366	0.243	0.514
	**	0.357	0.353	0.349	0.231	0.483

*Calculations by Frocini criterion; **calculations based on the omega-square test.

Calculated and critical values of the Frocini and omega-square criteria for a mixture of 2 normal distributions with a sample size of $n = 240$

Goodness measure	Calculated values	Critical value at significance level α					
		0.05	0.10	0.15	0.20	0.25	0.30
Frocini	*	0.0776					
	**	0.0785	0.146	0.136	0.130	0.125	0.121
Omega-square	*	0.0104	0.0348	0.0301	0.0277	0.0257	0.0241
	**	0.0102					0.0229

Distribution parameters obtained by minimizing the criteria: * Frocini; **omega-square.

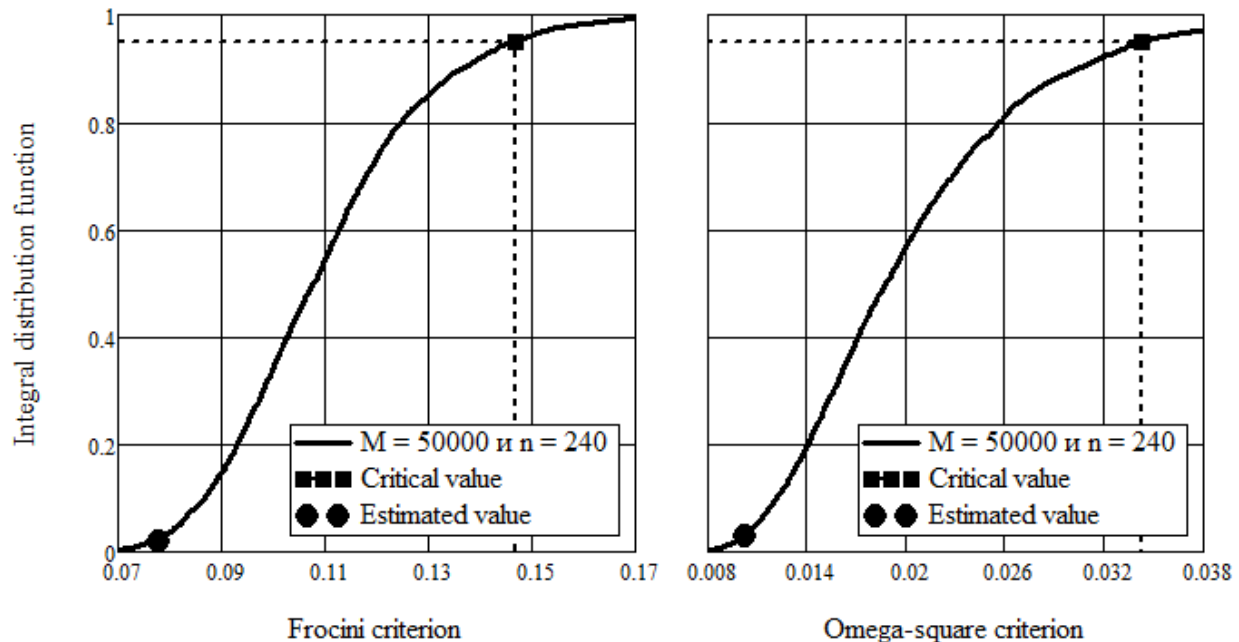


Fig. 3. The results of testing the hypothesis of compliance with the empirical distribution function and the mixture function of two normal distributions by Frocini and omega-square criteria

Рис. 3. Результаты проверки гипотезы соответствия эмпирической функции распределения и функции смеси двух нормальных распределений по критериям Фроцини и омега-квадрат

The calculated and critical values of the Frocini and omega-square criteria for a mixture of 2 normal distributions with a sample size of $n = 240$ are presented in tab. 2.

The visualization of the results of testing the hypothesis of compliance with the empirical distribution function with the mixture function of two normal distributions according to the Frocini and omega-square criteria is presented in fig. 3.

The simulation modeling results allow to evaluate the statistics of the parameters of the distributions mixture. In fig. 4–6 the results of the evaluation of the distribution of the parameters of the first and second components of the mixture, obtained from the results of statistical tests for the Frocini and omega-square agreement criteria, are presented.

Conclusion. The results of computational experiments allow to conclude about the effectiveness of obtaining estimates of distributions mixture parameters, minimizing

the calculated values of the goodness measures. The use of different goodness measures allows improving the quality of the found estimates. The differences in the estimates of the parameters of the mixture of two normal distributions, obtained by minimizing the Frocini and omega-square criteria for experimental samples, did not exceed 1%.

Evaluation of the distribution parameters in combination with the simulation method for evaluating the statistics of the goodness measure allows to test the complex hypothesis of consistency between the empirical and theoretical distribution functions. A related result of this task is an assessment of the statistics of the distribution parameters and confidence intervals of their change.

The choice of the minimum number of components of a distributions mixture is determined by the condition of accepting the hypothesis of compliance with the empirical and theoretical distribution functions.

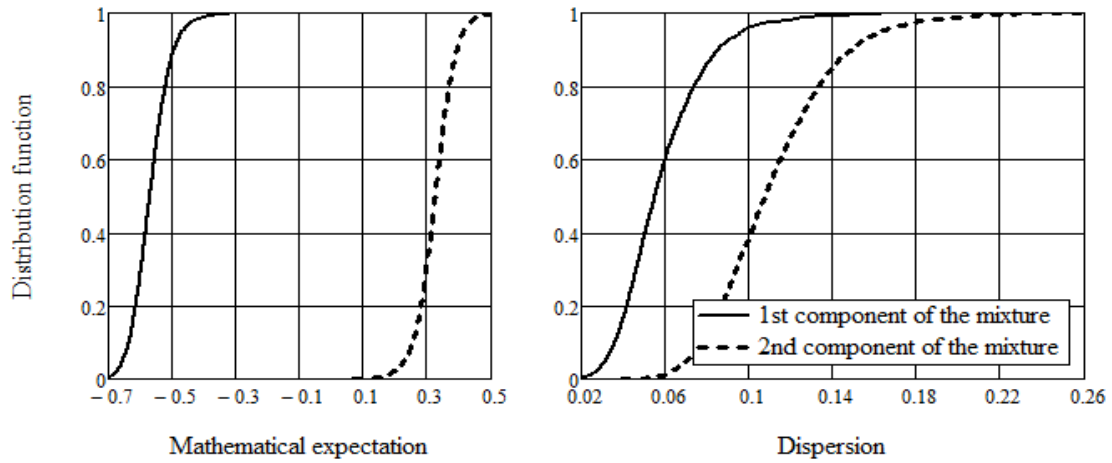


Fig. 4. Estimates of the distribution functions of expected values and dispersions of the mixture components

Рис. 4. Оценки функций распределения математических ожиданий и дисперсий компонентов смеси

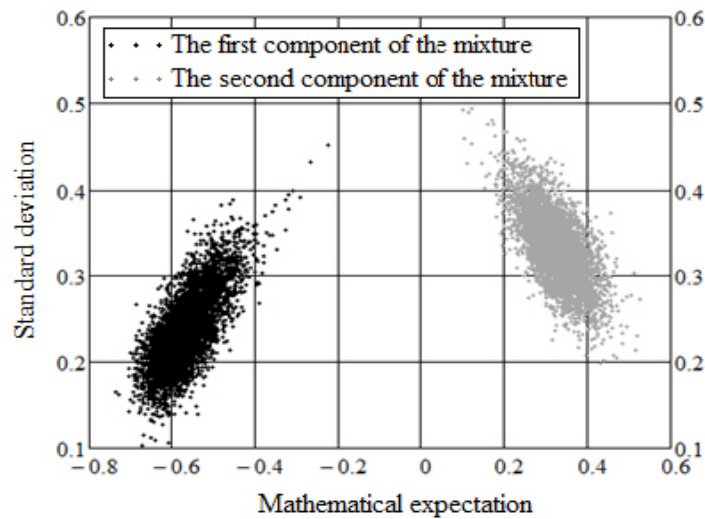


Fig. 5. Estimates of the distribution of the parameters of the first and second components of the mixture

Рис. 5. Оценки распределения параметров первой и второй компоненты смеси

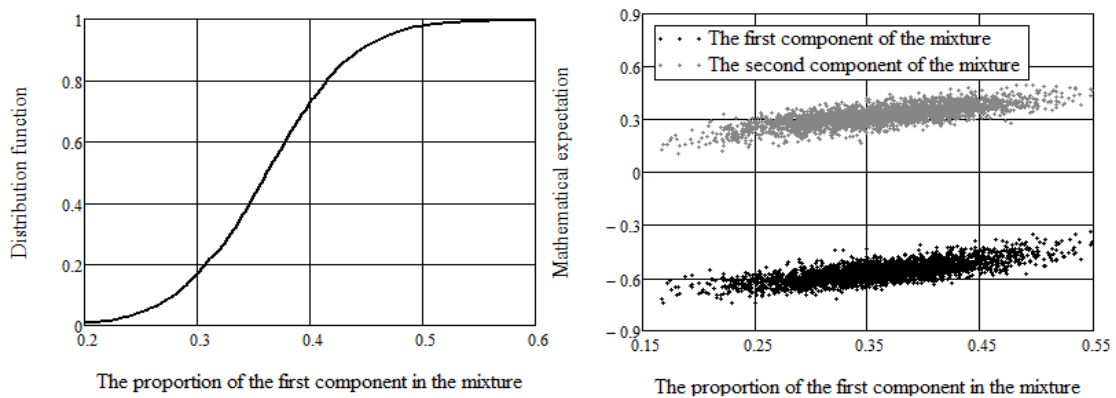


Fig. 6. Estimates of the distribution of the mathematical expectation of the first and the second components and the proportion of the first component in the mixture

Рис. 6. Оценки распределения математических ожиданий первой и второй компоненты и доли первой компоненты смеси

References

1. Pavlov I. N., Ushanov S. V. [Study of the distribution of trees by diameter analysis methods for mixtures of distributions]. *Vestnik SibGTU*. 2005, No. 1, P. 38–46 (In Russ.).
2. Ushanova V. M. *Kompleksnaya pererabotka drevesnoi zeleni i kory pikhty sibirskoi s polucheniem produktov, obladaiushikh biologicheskoi aktivnost'iu. Dokt. Dis.* [Complex processing of wood greens and Siberian fir bark to give products having biological activity. Doct. Dis.]. Krasnoyarsk, 2012, 34 p.
3. Ushanova V. M., Ushanov S. V. [Study of the process of extraction of fir bark by Siberian liquefied carbon dioxide]. *Vestnik Krasgau*. 2009, No. 12 (39), P. 39–44 (In Russ.).
4. Ushanova V. M., Ushanov S. V. *Ekstragirovaniye drevesnoy zeleni i kory pikhty sibirskoy szhizhennym dioksidom ugleroda i vodno-spirtovyimi rastvorami* [Extraction of wood greens and Siberian fir bark with liquefied carbon dioxide and water-alcohol solutions]. Krasnoyarsk, 2009, 191 p.
5. Kobzar A. I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov* [Applied mathematical statistics. For engineers and scientists]. Moscow, Fizmatlit Publ., 2006, 816 p.
6. Vetrov P. P., Kropotov D. A., Osokin A. A. [The automatic determination of the number of components in the mixture of normal distributions]. *Zhurnal vychislitel'noy matematiki i matematicheskoy fiziki*. 2010, Vol. 50, No. 4, P. 770–783 (In Russ.).
7. Korolev V. Yu. *EM-algoritm, yego modifikatsii i ikh primeneniye k zadache razdeleniya smesey veroyatnostnykh raspredeleniy. Teoreticheskiy obzor* [EM-algorithm, its modifications and their application to the problem of separation of probability distributions. Theoretical review]. Moscow, IPI RAN Publ., 2007, 102 p.
8. Celeux G., Chauveau D., Diebolt J. On Stochastic Versions of the EM algorithm: An Experimental study in the Mixture Case. *Journal of Statis. Comput. Simul.* 1996, Vol. 55, P. 287–314.
9. Ohorsin B. A. *Prikladnaya matematika v sisteme MathCad* [Applied Mathematics in the MathCad system]. Moscow, Lan⁷ Publ., 2008, 352 p.
10. Goldstein A. M. *Optimizatsiya v srede MatLAB* [Optimization in MatLAB]. Perm, 2015, 192 p.
11. Lemeshko B. Yu., Lemeshko S. B., Postovalov S. N., Chimitov E. V. *Statisticheskiy analiz dannykh, modelirovaniye, issledovaniye veroyatnostnykh zakonomernostey. Komp'yuternyy podkhod* [Statistical data analysis, modeling, probabilistic regularities research. Computer approach]. Novosibirsk, NGTU Publ., 2011, 888 p.
12. Orlov A. I. [Non-parametric criteria for the agreement of Kolmogorov, Smirnov, omega-square and errors in their application]. *Nauchnyy zhurnal KubGAU*. 2014, No. 97 (03), P. 1–29 (In Russ.).
13. Frozini B. V. A survey of a class of goodness-of-fit statistics. *Metron*. 1978, Vol. 36, No. 1–2, P. 3–49.
14. Ogurtsov D. A., Ushanov S. V. [Evaluation of statistics of the criterion of the normality of the Frozini distribution by the method of statistical tests]. *Aktual'nyye problemy aviatsii i kosmonavtiki*. 2017, Vol. 2, No. 3, P. 290–292 (In Russ.).
15. Martynov G. V. *Kriterii omega-kvadrat* [Criteria omega square]. Moscow, Nauka Publ., 1978, 78 p.
16. Ogurtsov D. A., Ushanov S. V. [Evaluation of statistics on the normality of the distribution of the omega-square method of statistical tests]. *Aktual'nyye problemy aviatsii i kosmonavtiki*. 2017, Vol. 2, No. 3, P. 293–295 (In Russ.).
17. Lemeshko B. Yu. *Neparametricheskiye kriterii soglasiya. Rukovodstvo po primeneniyu* [Non-parametric compliance criteria. Application Guide]. Moscow, INFRA-M Publ., 2014, 163 p.
18. Ushanov S. V., Ogurtsov D. A. [Estimation of statistics of the criterion for the normality of the Frozini distribution using the statistical test method in MATHCAD]. *Reshetnevskiy chteniya*. 2018, Vol. 2, No. 22, P. 171–173 (In Russ.).
19. Zhuk A. Yu. [Hydrodynamic qualities of whip beams made of wood with a limited margin of buoyancy]. *Sistemy. Metody. Tekhnologii*. 2014, No. 4 (24), P. 160–165 (In Russ.).

Библиографические ссылки

1. Павлов И. Н., Ушанов С. В. Исследование распределения деревьев сосны по диаметру методами анализа смесей распределений // Вестник СибГТУ. 2005. № 1. С. 38–46.
2. Ушанова В. М. Комплексная переработка древесной зелени и коры пихты сибирской с получением продуктов, обладающих биологической активностью : автореф. дисс. ... докт. тех. наук. Красноярск : СибГТУ, 2012. 34 с.
3. Ушанова В. М., Ушанов С. В. Исследование процесса экстрагирования коры пихты сибирской сжиженным диоксидом углерода // Вестник КрасГАУ. 2009. № 12 (39). С. 39–44.
4. Ушанова В. М., Ушанов С. В. Экстрагирование древесной зелени и коры пихты сибирской сжиженным диоксидом углерода и водно-спиртовыми растворами. Красноярск, 2009. 191 с.
5. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006. 816 с.
6. Ветров Д. П., Кротов Д. А., Осокин А. А. Автоматическое определение количества компонент в EM-алгоритме восстановления смеси нормальных распределений // Журнал вычислительной математики и математической физики. 2010. Т. 50, № 4. С. 770–783.
7. Королёв В. Ю. EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор. М. : ИПИРАН, 2007. 102 с.
8. Celeux G., Chauveau D., Diebolt J. On Stochastic Versions of the EM algorithm // An Experimental study in the Mixture Case, Journal of Statis. Comput. Simul. 1996, Vol. 55, P. 287–314.
9. Охорзин В. А. Прикладная математика в системе MathCad. М. : Лань, 2008. 352 с.
10. Гольдштейн А. М. Оптимизация в среде MatLAB. Пермь, 2015. 192 с.

11. Статистический анализ данных, моделирование, исследование вероятностных закономерностей. Компьютерный подход : монография / Б. Ю. Лемешко, С. Б. Лемешко, С. Н. Постовалов [и др.]. Новосибирск : Изд-во НГТУ, 2011. 888 с.

12. Орлов А. И. Непараметрические критерии согласия Колмогорова, Смирнова, омега-квадрат и ошибки при их применении // Научный журнал КубГАУ. 2014. № 97 (03). С. 1–29.

13. Frozini B. V. A survey of a class of goodness-of-fit statistics, *Metron*. 1978. Vol. 36, № 1–2. P. 3–49.

14. Огурцов Д. А., Ушанов С. В. Оценка статистики критерия нормальности распределения Фроцини методом статистических испытаний // Актуальные проблемы авиации и космонавтики. 2017. Т. 2, № 3. С. 290–292.

15. Мартынов Г. В. Критерии омега-квадрат. М. : Наука, 1978. 78 с.

16. Огурцов Д. А., Ушанов С. В. Оценка статистики критерия нормальности распределения омега-квадрат методом статистических испытаний // Актуальные проблемы авиации и космонавтики. 2017. Т. 2, № 3. С. 293–295.

17. Лемешко Б. Ю. Непараметрические критерии согласия. Руководство по применению. М. : Инфра-М, 2014. 163 с.

18. Ушанов С. В., Огурцов Д. А. Оценка статистики критерия нормальности распределения Фроцини методом статистических испытаний в MATHCAD // Решетневские чтения. 2018. Т. 2, № 22. С. 171–173.

19. Жук А. Ю. Гидродинамические качества хлыстовых пучков из древесины с ограниченным запасом плавучести // Системы. Методы. Технологии. 2014. № 4 (24). С. 160–165.

© Ushanov S. V., Ogurtsov D. A., 2019

Ushanov Sergey Viktorovich – Cand. Sc., Associate Professor, Head of the Department of Higher Mathematics and Computer Science; Reshetnev Siberian State University of Science and Technology. E-mail: ushanov_sv@mail.ru.

Ogurtsov Dmitrii Aleksandrovich – student of group 23-7 of the Institute of Informatics and Telecommunications; Reshetnev Siberian State University of Science and Technology. E-mail: dim_ogu@mail.ru.

Ушанов Сергей Викторович – кандидат технических наук, доцент, заведующий кафедрой высшей математики и информатики; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: ushanov_sv@mail.ru.

Огурцов Дмитрий Александрович – студент группы 23-7; Институт информатики и телекоммуникаций; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: dim_ogu@mail.ru.
