УДК 519.226
ББК 22.172.4
С 37

Симанков Владимир Сергеевич
Профессор, доктор технических наук, профессор кафедры компьютерных технологий и информационной безопасности Кубанского государственного технологического университета, Краснодар, тел. (861) 2980190, e-mail: vs@simankov.ru
Бучацкая Виктория Викторовна
Доцент, кандидат технических наук, доцент кафедры прикладной математики, информационных технологий и информационной безопасности факультета математики и компьютерных наук Адыгейского государственного университета, Майкоп, тел. (8772) 593904, e-mail: strateg_r@adygnet.ru
Теплоухов Семен Васильевич
Аспирант кафедры компьютерных технологий и информационной безопасности Кубанского государственного технологического университета, Краснодар, e-mail: mentory@mail.ru

Определение оптимального сочетания доверительного интервала и доверительной вероятности

(Рецензирована)

Аннотация. *Предложена методика расчета оптимального доверительного интервала и доверительной вероятности для произвольной выборки. Для этого использован способ, основанный на максимизации среднего приращения информации о выборке, который позволяет учесть такой показатель, как неопределенность исходной информации. Проведен численный эксперимент, в ходе которого исследована связь оптимальной доверительной вероятности и доверительного интервала от размера выборки, закона распределения случайной величины и типа неопределенности. Установлены численные значения размера выборки, при достижении которых осуществляется переход от детерминированного типа неопределенности к стохастическому типу, а затем к нечеткому.*

Ключевые слова: *доверительный интервал, доверительная вероятность, закон распределения, размер выборки, неопределенность, приращение информации.*

Simankov Vladimir Sergeevich
Doctor of Technical Sciences, Professor, Professor of Computer Technologies and Information Security Department, Kuban State University of Technology, Krasnodar, ph. (861) 2980190, e-mail: vs@simankov.ru
Buchatskaya Viktoriya Viktorovna
Associate Professor, Candidate of Technical Sciences, Associate Professor of Department of Applied Mathematics, Information Technologies and Information Security of the Faculty of Mathematics and Computer Science, Adyghe State University, Maikop, ph. (8772) 593904, e-mail: strateg_r@adygnet.ru
Teploukhov Semen Vasilyevich
Post-graduate student of Computer Technologies and Information Security Department, Kuban State University of Technology, Krasnodar, e-mail: mentory@mail.ru

Calculation of the optimal ratio of confidence interval and confidence probability

Abstract. *This article proposes a methodology for calculating the optimal confidence interval and confidence probability for a random sample. For this, a method was used based on maximizing the average increment of information about the sample, which allows taking into account such an indicator as the uncertainty of the initial information. A numerical experiment was conducted in which the relationships between the optimal confidence probability and the confidence interval on the sample size, the distribution law of a random variable, and the type of uncertainty were investigated. Numerical values of the sample size are established, upon reaching which the transition is implemented from the deterministic type of uncertainty to the stochastic type, and then to the fuzzy.*

Keywords: *confidence interval, confidence probability, distribution law, sample size, uncertainty, average increment of information.*

Наличие неопределенности в исходных данных при решении различных практических задач приводит к необходимости интервальных оценок исследуемых параметров. Если интервальные оценки будут неоптимальными, то это может привести к неточности и неадекватности полученных результатов и моделей. Поэтому необходимо решить задачу расчета оптимального доверительного интервала и доверительной вероятности.

Как известно, доверительный интервал ξ с доверительной вероятностью α накрывает истинное значение оцениваемого параметра. Величина доверительного интервала зависит от ряда параметров: размера выборки, закона распределения оцениваемой случайной величины, а также априорных сведений, которыми обладает исследователь.

При определении ширины доверительного интервала используются следующие параметры: доверительная вероятность α и уровень значимости $\lambda = 1 - \alpha$. На практике значение α назначается до получения выборки или зависит от предметной области. Однако данный подход имеет ряд недостатков:

- для выбора требуются априорные сведения о процессе или предметной области, которые не всегда имеются или их определение представляет достаточно сложный процесс;
- априорный уровень значимости α принимается за основу на начальном этапе расчета и затем не изменяется, что может привести к высокой погрешности рассчитанного интервала;
- выбор уровня значимости формально не зависит от размера выборки;
- не учитывается степень неопределенности исходной информации, то есть ее неполнота, недостоверность и неточность.

Для устранения указанных недостатков можно использовать метод оптимизации интервальных оценок на основе среднего приращения информации I [1, 2]. Количество полученной информации I возрастает с увеличением объема выборки, а при фиксированном объеме (n=const) зависит от выбора статистической оценки.

Параметр I рассчитывается по формуле:

$$I = \beta \log \frac{\beta}{\alpha} + (1-\beta) \log \frac{1-\beta}{1-\alpha}, \tag{1}$$

где α – априорная, а β – апостериорная доверительные вероятности того, что неизвестная величина накрыта доверительным интервалом.

Априорная доверительная вероятность α определяется выражением:

$$\alpha = \frac{2n}{k} \left(\frac{1}{\chi^2_{\frac{1-\beta}{2},2n}} - \frac{1}{\chi^2_{\frac{1+\beta}{2},2n+2}} \right), \tag{2}$$

где n – размер исходной выборки;

$X^2_{p,n}$ – квантиль хи-квадрат распределения вероятности p и числа степеней свободы n;

k – характеризует величину априорной информации, которой обладает исследователь, то есть неопределенности исходной информации, и является целой величиной в интервале [1; 6] [1].

Поскольку основной задачей любого исследования является снятие неопределенности, что соответствует росту количества информации, то оптимальная оценка, соответствующая максимуму I, является наилучшей в условиях поставленной задачи. Это соответствует нахождению максимального значения следующей целевой функции:

$$I(\alpha, \beta) \to \max, \quad \alpha \in [0;1], \quad \beta \in [0;1].$$

Для нахождения оптимальных интервальных оценок необходимо найти максимум I при некотором значении β.

Важно отметить, что при отсутствии данных, характеризующих априорное распределение вероятностей значений выборки, целесообразно воспользоваться предположением, что априорные значения равновероятны, то есть можно говорить о равномерном законе распределения случайной величины [2, 3].

Подставив (2) в (1), получим, что среднее приращение информации I зависит от трех параметров: апостериорной доверительной вероятности, размера выборки и вида неопределенности.

Задача нахождения оптимального значения среднего приращения информации I в аналитическом виде является очень сложной, поэтому для оценки данного параметра воспользуемся машинным перебором. Для этого необходимо выполнить ряд действий:

1) пусть целочисленный параметр k примет значение из интервала [1; 6], при этом k=1 соответствует наличию информации об основных закономерностях выборки, а k=6 описывает практически полную неопределенность;

2) для каждого значения параметра k рассмотрим целочисленное значение размера выборки n из интервала [1; 1000];

3) осуществим перебор значения β на интервале [0,8; 1,0] с малым шагом Δ=0,01, что соответствует наиболее применимым на практике доверительным вероятностям;

4) для каждого значения k, n и β рассчитаем параметр I;

5) рассчитаем максимум среднего приращения информации I по формуле (1);

6) вычислим оптимальную доверительную вероятность β для рассчитанного максимального значения среднего приращения информации I для всех значений параметров k и n.

На рисунке 1 приведен график зависимости величины среднего приращения информации I от доверительной вероятности при k=2 и n=100.

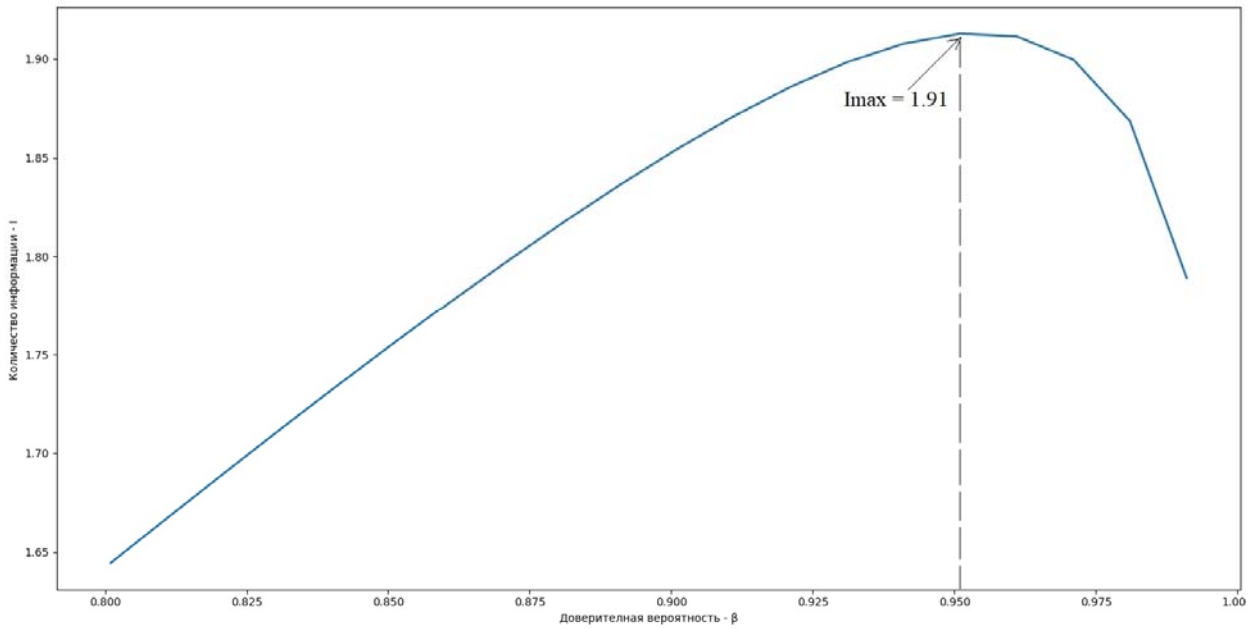

Рис. 1. График среднего приращения информации I в зависимости от значения доверительной вероятности β при k=2 и n=100

Из графика видно, что значение максимума среднего приращения информации равно I_{max}=1,91, и оно достигается при доверительной вероятности β=0,95, что соответствует уровню значимости λ=1–β=0,05. Таким образом, в условиях выборки размера n=100 единиц и при параметре k=2, что соответствует большой малой степени неопределенности исходной информации, оптимальным будет выбор уровня значимости 95%.

На рисунке 2 приведено семейство графиков зависимости величины оптимальной доверительной вероятности β от размера выборки n для каждого значения целочисленного k из интервала [1; 6].

В результате анализа рисунка 2 можно сделать ряд выводов:

1) С ростом количества элементов выборки увеличивается оптимальное значение доверительной вероятности β.

2) При росте значения k оптимальная доверительная вероятность также увеличивается, но, начиная со значения n=200, параметр β становится большим 0,95 для любых k.

3) При n<40 доверительная вероятность сильно отличается для различных значений

k. Это свидетельствует о существенном влиянии неопределенности исходной информации для малых выборок.

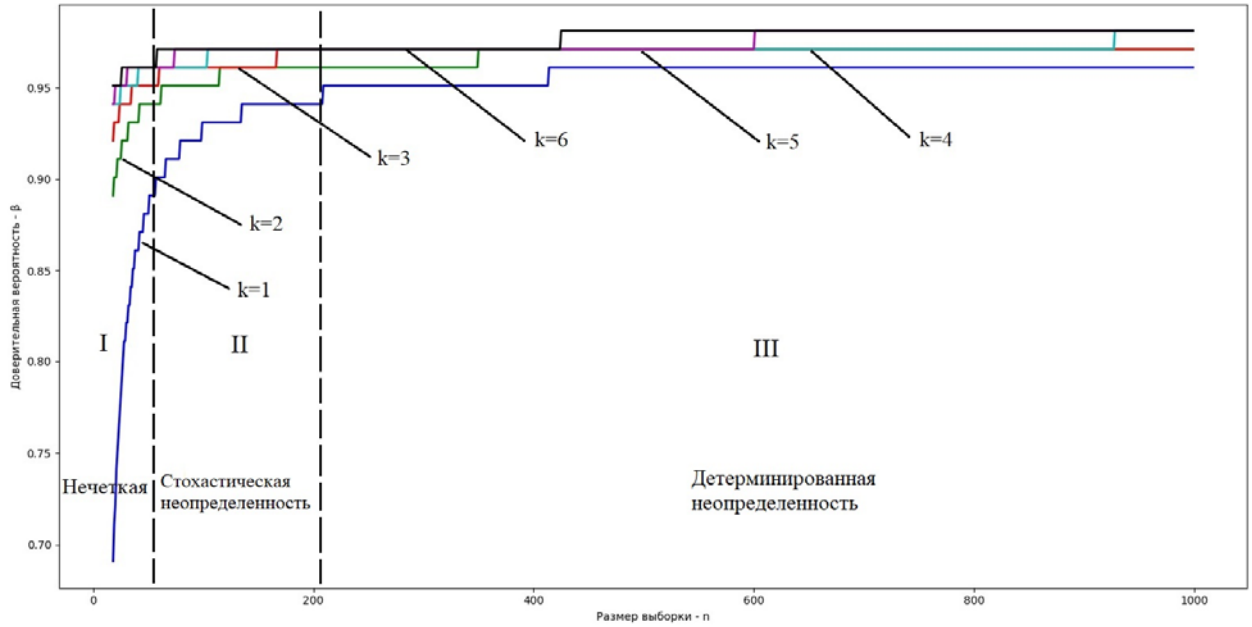

Рис. 2. Оптимальная доверительная вероятность в зависимости от размера выборки при различных значениях $k \notin [1; 6]$

Эти выводы позволяют выделить три основных области на рисунке 2:

• область I характеризуется малым размером выборки $n < 30 - 50$ элементов, что соответствует ситуации, когда у исследователя мало априорной информации. Это соответствует нечеткому типу неопределенности, и в этих условиях целесообразно использовать нечеткие модели [4];

• область II характеризуется средним размером выборки $50 \le n \le 200$ элементов, а также существенной неопределенностью. Это соответствует стохастическому типу неопределенности [4];

• область III соответствует выборке большого размера $n \ge 200$ элементов и большому количеству информации о выборке, то есть малой неопределенности. Это говорит о детерминированном типе неопределенности [4].

Можно заметить, что требуется детально рассмотреть область II, поскольку она является переходной от определенности (область III) к практически полной неопределенности (область I). Исследуем зависимость величины оптимальной доверительной вероятности и доверительного интервала от закона распределения случайной величины в условиях области II.

Рассмотрим следующие законы распределения: нормальный, равномерный и экспоненциальный как наиболее часто встречающиеся при обработке данных [5, 6].

График 3а) был построен при фиксированном значении среднеквадратического отклонения $\sigma = 1$; график 3б) построен для фиксированного значения максимума выборки $\max = 1$; 3в) – для зафиксированного среднего значения элементов исходной выборки $M = 1$ (рис. 3).

Анализ рисунка 3 позволяет сделать вывод о том, что, начиная с некоторого значения n – количества элементов в выборке, ширина доверительного интервала почти перестает изменяться (наклон графиков почти горизонтален). Это происходит при $n \approx 200$ элементов. Следует заметить, что это значение n почти не зависит от формы закона распределения. Также важным является тот факт, что для разных законов распределения при малых n (около 30–50 элементов) ширина доверительного интервала изменяется на значительную величину Δ.

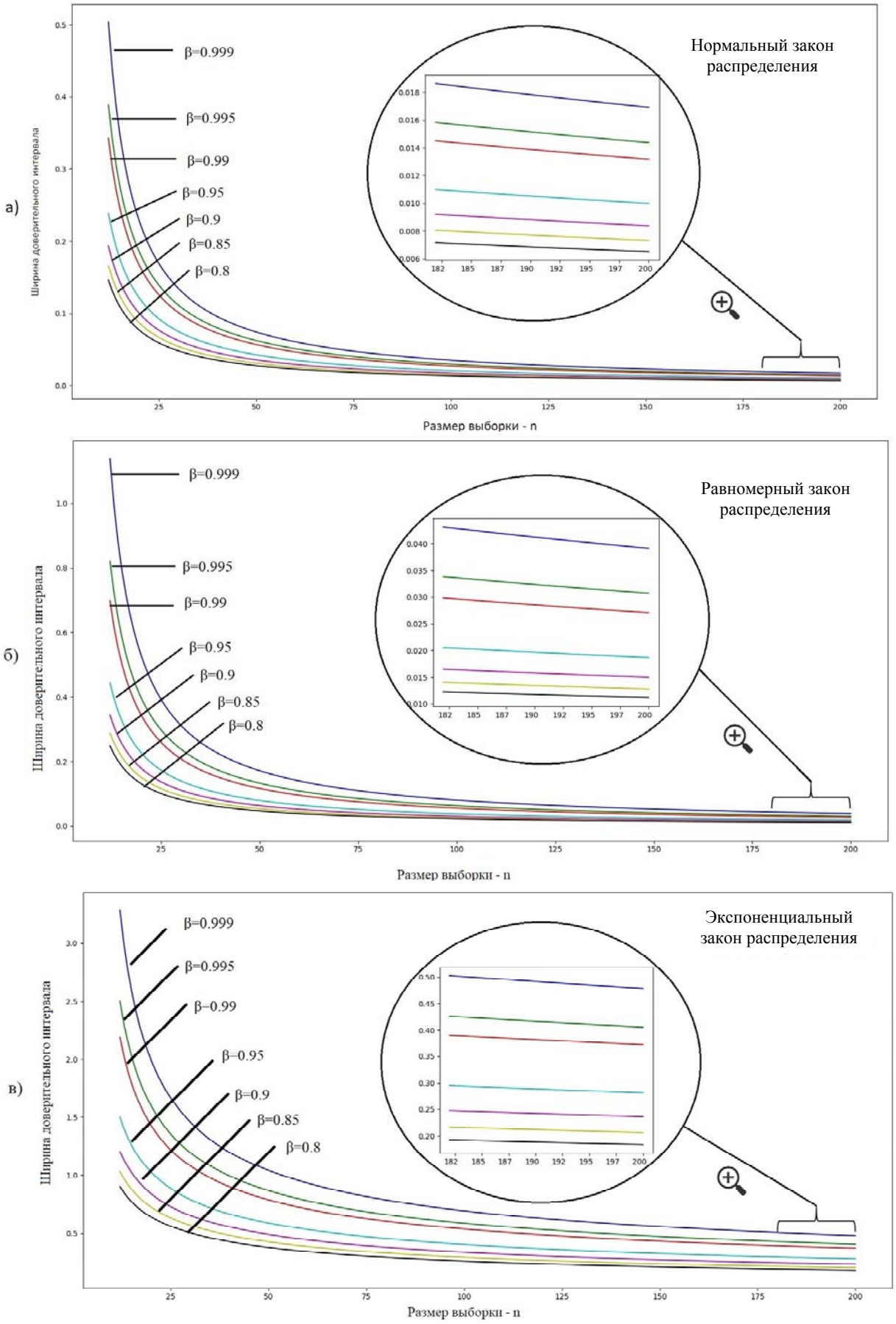

Рис. 3. Зависимость ширины доверительного интервала от размера выборки и доверительной вероятности для различных законов распределения:
а) для нормального закона распределения; б) для равномерного закона распределения;
в) для экспоненциального закона распределения

Таким образом, можно сделать вывод о том, что размер выборки является одной из важнейших характеристик, от которой зависит оптимальная доверительная вероятность и доверительный интервал. В случае если выборка является большой $(n > 200)$, то можно говорить об отсутствии неопределенности, а также о слабой зависимости от вида закона распределения, что соответствует детерминированному типу неопределенности. Если выборка мала $(n < 30-50)$, то существенную роль играет априорная неопределенность, и величина оптимальной доверительной вероятности не может быть однозначно определена, это соответствует нечеткому типу неопределенности. В случае выборки $30-50 \le n \le 200$ можно говорить о существенном влиянии вида закона распределения на доверительный интервал и доверительную вероятность, а также о стохастическом типе неопределенности.

Итак, опираясь на такую важную характеристику выборки, как ее размер, можно определить оптимальные значения доверительной вероятности и доверительного интервала. В результате использования метода максимизации среднего приращения информации установлены численные значения размера выборки, при достижении которых осуществляется переход от детерминированного типа неопределенности к стохастическому типу, а затем к нечеткому типу.

Таким образом, предложенный способ расчета оптимальных интервальных оценок, зависящих от размера выборки, позволит выбирать наиболее адекватные математические модели для решения задачи принятия решений, прогнозирования и планирования.

Примечания:

1. Симанков В.С. Планирование определительных испытаний и оптимизация интервальных оценок при исследовании надежности электрических сетей. Краснодар, 1981, 11 с. Рукопись предоставлена Краснодар. политех. ин-том. Деп. в ВИНИТИ. 1982. № D/987.
2. Рипс Я.А. Информационный аспект статистических оценок надежности // Автоматика и телемеханика. 1967. Вып. 7. С. 140–150.
3. Бонгард М.М. О понятии «полезная информация» // Проблемы кибернетики. 1963. № 9. С. 71–102.
4. Classification of information's uncertainty in system research / V.S. Simankov, V.V. Buchatskaya, P.Yu. Buchatskiy, S.V. Teploukhov // Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM. 2017. P. 167–170.
5. Севастьянов Б.А. Курс теории вероятностей и математической статистики. Москва: Книга по Требованию, 2012. 256 с.
6. Hossein Pishro-Nik. Introduction to Probability, Statistics, and Random Processes. Kappa Research, LLC. 2014. 744 p.

References:

1. Simankov V.S. Planning of definitive tests and optimization of interval estimates in the study of the reliability of electrical networks. Krasnodar, 1981. 11 pp. Manuscript provided by Krasnodar Polytechnic Institute. Dep. of VINITI. 1982. No. D/987.
2. Rips Ya.A. Informational aspect of statistical reliability estimates // Automation and Telemechanics. 1967. Iss. 7. P. 140–150.
3. Bongard M.M. On the concept of “useful information” // Problems of Cybernetics. 1963. No 9. P. 71–102.
4. Classification of information's uncertainty in system research / V.S. Simankov, V.V. Buchatskaya, P.Yu. Buchatskiy, S.V. Teploukhov // Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM. 2017. P. 167–170.
5. Sevastyanov B.A. The course of probability theory and mathematical statistics. Moscow: Book on Demand, 2012. 256 pp.
6. Hossein Pishro-Nik. Introduction to Probability, Statistics, and Random Processes. Kappa Research, LLC. 2014. 744 pp.