УДК 517.958 : 519.234

МЕТОД ПОДБОРА НАИЛУЧШЕГО ЗАКОНА РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПО ЭКСПЕРИМЕНТАЛЬНЫМ ДАННЫМ

И. А. Клявин[1], **А. Н. Тырсин**[2]

[1] *Челябинский государственный университет,*
454021, г. Челябинск, ул. Бр. Кашириных, 129
[2] *Научно-инженерный центр «Надёжность и ресурс больших систем и машин» УрО РАН,*
620049, г. Екатеринбург, ул. Студенческая, 54а
E-mail: 2ivank@mail.ru
at2001@yandex.ru

Описан новый метод выбора по экспериментальным данным из заданного множества закона распределения, который в наибольшей степени соответствовал бы измеренной случайной величине. Метод основан на сравнении эмпирического распределения, построенного по исходной выборке, с множеством заданных законов с помощью непрерывного отображения функции распределения на отрезок [0; 1]. В результате в качестве наиболее вероятного закона для исходной выборки берётся тот, для которого соответствующее значение функционала будет максимальным. Приведены примеры реализации метода с помощью статистического моделирования методом Монте-Карло.

Ключевые слова: случайная величина, закон распределения, плотность вероятности, случайная выборка, статистическое моделирование методом Монте-Карло, критерий согласия.

Введение. Достаточно часто на практике возникает задача установления по экспериментальным данным закона распределения измеренной случайной величины, например, при контроле и диагностике состояния объектов [1–4], расчёте прочностной надёжности изделий [5, 6], управлении качеством на основе статистических методов [7, 8], в метрологии [9] и т. д.

Такая задача не может быть строго решена, так как имеется конечная выборка и бесконечное количество возможных законов распределения. Известен ряд методов восстановления неизвестной функции плотности, рассчитанных на конкретную ситуацию [5, 10–12]. Эффективное применение этих методов требует достаточно большой выборки данных или наличия априорной информации о форме распределения на малых выборках, что не всегда возможно.

Часто более оправдан иной подход, когда требуется не восстанавливать по конечной выборке неизвестное распределение непрерывной случайной величины, а выбрать модель, достаточно адекватно её описывающую. Это означает, что по экспериментальным данным из заданного множества различных законов распределений необходимо выбрать тот, который в наибольшей степени соответствует измеренной случайной величине.

Одна и та же выборка может принадлежать с различной вероятностью каждому из рассматриваемых законов распределения, поэтому в качестве искомого нужно выбрать наиболее вероятный закон распределения для данной случайной выборки из конечного множества заданных законов.

Использование статистических критериев согласия [13] не решает указанной задачи. Действительно, на основании экспериментальных данных делается предположение о виде закона распределения для выборки. Затем с заданной вероятностью критерий отклоняет или нет гипотезу о том, что имеющаяся выборка не противоречит выбранному закону

распределения. Например, статистическое моделирование методом Монте-Карло показало, что наиболее часто применяющийся критерий согласия χ^2-Пирсона приблизительно в 79 % случаев не отвергает гипотезу о том, что выборка из нормального закона объёмом 100 наблюдений принадлежит закону Лапласа, т. е. он не в состоянии различать нормальный закон и закон Лапласа для малых выборок, поскольку в 79 % случаев критерий согласия указывает, что заданная выборка не противоречит ни одному из этих законов.

Кроме того, критерии согласия имеют разную мощность по отношению к различным альтернативам. Это означает, что у любого из них существуют наиболее близкие альтернативы, для которых он может оказаться несостоятельным, т. е. мощность окажется слишком малой [14]. Таким образом, является актуальной разработка метода, который мог бы различать достаточно близкие законы с приемлемой точностью для малых выборок.

Сформулируем постановку задачи. Пусть имеется случайная выборка $(x_1, \ldots, x_n)$ из генеральной совокупности ξ с некоторой неизвестной непрерывной функцией распределения $F_0(x)$. Также зададим конечный набор непрерывных законов распределения, описываемых с помощью плотностей $\{p_1(x), \ldots, p_m(x)\}$ либо функций распределения $\{F_1(x), \ldots, F_m(x)\}$. Необходимо определить среди них наиболее вероятный закон для данной выборки.

Методика решения. Как известно из метода обратного преобразования [15], выборка $(x_1, \ldots, x_n)$, где $x_i = F^{-1}(u_i)$ и $(u_1, \ldots, u_n) \sim U[0,1]$ ($U[0,1]$ — выборка из непрерывного равномерного распределения от 0 до 1), будет принадлежать распределению, заданному функцией $F(x)$. Отсюда следует, что для любой выборки $(x_1, \ldots, x_n)$ точечная функция $F(x_i) = u_i$, где $(u_1, \ldots, u_n) \sim U[0,1]$, является точечной выборочной функцией распределения. Необходимо доопределить её до непрерывной. Сделать это можно с помощью интерполяции сплайнами. Воспользуемся самым простым, очевидным и наименее ресурсоёмким способом — интерполяцией точечной функции $F(x_i) = u_i$ линейными сплайнами. Получим

$$F(x) = \begin{cases} 0, & x < \hat{x}_1, \\ \dfrac{(x - \hat{x}_{i-1})(F(\hat{x}_i) - F(\hat{x}_{i-1}))}{(\hat{x}_i - \hat{x}_{i-1}) + F(\hat{x}_{i-1})}, & x \in [\hat{x}_i; \hat{x}_{i+1}), \\ 1, & x \geq \hat{x}_n, \end{cases} \tag{1}$$

где $(\hat{x}_1, \ldots, \hat{x}_n)$ — вариационный ряд выборки $(x_1, \ldots, x_n)$.

Генерируя независимые случайные выборки $(u_1^{(j)}, \ldots, u_n^{(j)}) \sim U[0,1]$, $j = 1, \ldots, M$, M раз, будем иметь M возможных выборочных функций распределения для исходной выборки $(x_1, \ldots, x_n)$ (рис. 1). Очевидно, что частота попадания функций распределения в разные области плоскости будет различна.

Пусть даны выборки $(x_1, \ldots, x_n)$, а также $(u_1^{(j)}, \ldots, u_n^{(j)}) \sim U[0,1]$, $j = 1, \ldots, M$. Пусть $\hat{x}_1, \ldots, \hat{x}_n$ — вариационный ряд для выборки $(x_1, \ldots, x_n)$, а $(\hat{u}_1^{(j)}, \ldots, \hat{u}_n^{(j)})$ — вариационные ряды для $(u_1^{(j)}, \ldots, u_n^{(j)})$. Каждой выборке $(u_1^{(j)}, \ldots, u_n^{(j)})$ будет соответствовать функция распределения $F^{(j)}(x)$. Так как $\hat{x}_i$ и $\hat{u}_i^{(j)}$ — независимые случайные величины, то их совместная плотность выражается формулой $F_n(x, y) = p_x(x)p(y, i, n)$, где $p_x(x)$ — плотность распределения для $(x_1, \ldots, x_n)$, а $p(x, i, n)$ — плотность для $(u_i^{(1)}, \ldots, u_i^{(M)})$ при $M \to \infty$. Поскольку $P((x, F^{(j)}(x)) \in O) = \int\limits_O F_n(x, y)dxdy$, где O — некоторая область и $j = 1, \ldots, M$, то эта функция будет отражать вероятность нахождения функций распределения для заданной выборки в некоторой области на плоскости. В качестве $p_x(x)$ можно взять нормализованную гистограмму, основанную на выборке.

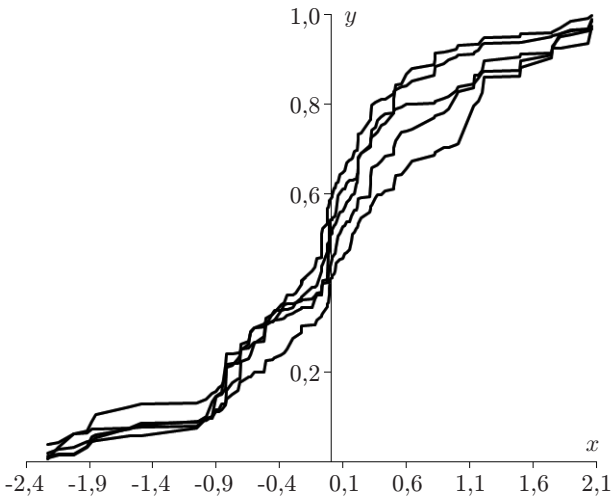

Рис. 1. Различные выборочные функции распределения для выборки объёмом 100 наблюдений, распределение Лапласа, $M = 5$

Построим функцию $p(x, i, n)$. Найдём данную функцию для $i = 1$. Для этого определим вероятность того, что произвольный l-й член выборки ($l = 1, \ldots, n$) является первым в вариационном ряду, т. е. наименьшим. Очевидно, что данная вероятность равна произведению вероятностей того, что l-й член меньше любого другого члена выборки:

$$P(\hat{u}_1 = u_l) = P(u_1 > u_l) \ldots P(u_{l-1} > u_l) P(u_{l+1} > u_l) \ldots P(u_n > u_l) =$$

$$= (1 - F_U(u_l))^{n-1} = (1 - u_l)^{n-1},$$

так как $F_U(x) = x$, $x \in [0; 1]$.

Поскольку любой из n членов выборки с равной вероятностью может быть наименьшим, то, обозначив $u_l = x$, получим

$$p(x, 1, n) = nP(x = \hat{u}_1) = n(1 - x)^{n-1}, \quad x \in [0; 1].$$

Аналогично для $i = 2$ будем иметь

$$P(\hat{u}_2 = u_l) = P(u_1 < u_l) P(u_2 > u_l) P(u_{l-1} > u_l) P(u_{l+1} > u_l) P(u_n > u_l) + \ldots +$$

$$+ P(u_n < u_l) P(u_2 > u_l) P(u_{l-1} > u_l) P(u_{l+1} > u_l) P(u_{n-1} > u_l) = (n - 1) u_l (1 - u_l)^{n-2},$$

$$p(x, 2, n) = n(n - 1) P(x = \hat{u}_2) = n(n - 1) x (1 - x)^{n-2}, \quad x \in [0; 1].$$

Для произвольного i

$$P(\hat{u}_i = u_l) = C_{n-1}^{i-1} \prod_{k \in I_1} P(u_k < u_l) \prod_{s \in I_2} P(u_s > u_l) = \frac{(n-1)!}{(i-1)!(n-i)!} u_l^{i-1} (1 - u_l)^{n-i},$$

где $I_1 = \{k \in \overline{1, n} : k \neq l\}$, $|I_1| = i - 1$; $I_2 = \{s \in \overline{1, n} : s \neq l\}$, $|I_2| = n - i$, $I_1 \cap I_2 = \varnothing$.

Таким образом, функция плотности для $\hat{u}_i^{(j)}$ выражается как

$$p(x,i,n) = \begin{cases} \dfrac{n!}{(i-1)!(n-i)!} x^{i-1}(1-x)^{n-i}, & x \in [0;1], \\ 0, & x \notin [0;\ 1]. \end{cases} \tag{2}$$

Доопределим данную функцию для непрерывного аргумента i следующим образом:

$$p(x,i,n) = \begin{cases} (n-i+1)(1-x)^{n-i}, & x \in [0;1],\ i \in [0;1), \\ \dfrac{1}{B(i,n-i+1)} x^{i-1}(1-x)^{n-i}, & x \in [0;1],\ i \in [1;n], \\ ix^{i-1}, & x \in [0;1],\ i \in (n;n+1], \\ 0, & x \notin [0;1], \end{cases} \tag{3}$$

где $B(x,y)$ — бета-функция.

Функция (3) совпадает с (2) во всех точках $i = 1,\ldots,n$ и непрерывна в остальных точках $i \in [0;n+1]$. Аргумент данной функции $i = (n+1)F(x)$, где $F(x)$ — функция распределения, определяемая по формуле (1).

Следовательно, функция $F_n(x,y)$ принимает вид (рис. 2)

$$F_n(x,y) = p_x(x)p(y,(n+1)F(x),n). \tag{4}$$

Аналогичным образом можно построить функцию $F_n(x,y)$ для любого непрерывного распределения (рис. 3).

Так как $F_n(x,y)$ является плотностью вероятности, то она обладает следующим свойством:

$$\iint\limits_{R\times R} F_n(x,y)dxdy = 1. \tag{5}$$

Пусть дана выборка $(x_1,\ldots,x_n)$ и некоторая непрерывная функция распределения $F^*(x)$. Построим функции вида (4) для выборки $(x_1,\ldots,x_n)$ и для функции распределения $F^*(x)$ соответственно: $F_n(x,y)$ и $F_n^*(x,y)$.

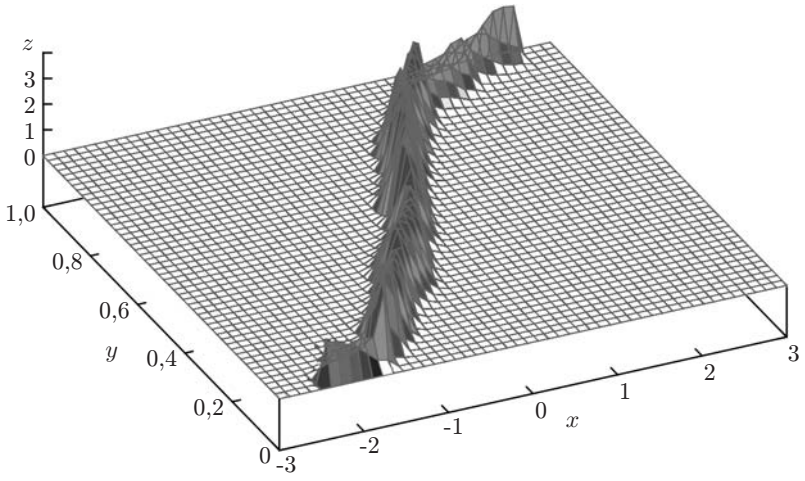

Рис. 2. Функция $F_n(x,y)$ для выборки из нормального распределения, $n = 100$

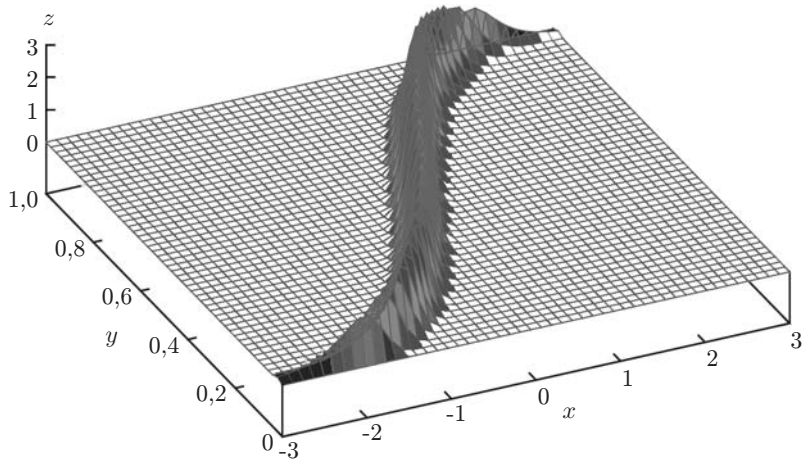

Рис. 3. Функция $F_n(x, y)$ для непрерывного распределения, $n = 100$

Далее найдём объём под функцией $\varphi(x, y) = \min(F_n(x, y), F_n^*(x, y))$:

$$V = \iint\limits_{R \times R} \varphi(x, y) dx dy. \tag{6}$$

Функционал (6) представляет собой непрерывное отображение множества функций распределений на отрезок $[0; 1]$, так как все используемые функции непрерывны. Учитывая свойство (5), видим, что $V \in [0; 1]$ и принимает максимальное значение, если функция распределения $F^*(x)$ совпадает с функцией распределения $F(x)$ исходной выборки, задаваемой в виде (1).

В каждом эксперименте для всех рассматриваемых законов $\{F_1(x), \ldots, F_m(x)\}$ по формуле (6) находим соответствующие оценки V_k, $k \in \{1, 2, \ldots, m\}$. Наиболее вероятный закон для исходной выборки $(x_1, \ldots, x_n)$ будет иметь максимальное значение V_k.

Покажем, что при $n \to \infty$ значение функционала для истинной функции распределения будет больше значения для любой другой функции распределения, не совпадающей с истинной.

Теорема 1. Пусть $(x_1, \ldots, x_n)$ — случайная выборка из генеральной совокупности ξ, имеющей некоторую непрерывную функцию распределения $F_0(x)$. Даны два непрерывных закона распределения, описываемых с помощью функций распределения $\{F_0(x), F_1(x)\}$, причём $\exists O \subseteq \mathbb{R}$: $x \in O \Rightarrow F_0(x) \neq F_1(x)$. Тогда при $n \to \infty$ оценка $V_0 > V_1$.

Доказательство. Пусть $\{p_0(x), p_1(x)\}$ — плотности вероятности, соответствующие $\{F_0(x), F_1(x)\}$, $F^*(x)$ — выборочная функция распределения, а $p^*(x)$ — нормализованная гистограмма для выборки $(x_1, \ldots, x_n)$.

Так как $(x_1, \ldots, x_n)$ из генеральной совокупности ξ, имеющей функцию распределения $F_0(x)$, то

$$F^*(x) \xrightarrow[n \to \infty]{P} F_0(x), \quad p^*(x) \xrightarrow[n \to \infty]{P} p_0(x).$$

Отсюда $F_n^*(x, y) \xrightarrow[n \to \infty]{P} F_n^0(x, y)$ вследствие непрерывности функции $p(x, i, n)$. Тогда

$$\varphi_0(x, y) = \min(F_n^0(x, y), F_n^*(x, y)) \xrightarrow[n \to \infty]{P} F_n^0(x, y),$$

$$\varphi_1(x,y) = \min(F_n^1(x,y), F_n^*(x,y)) \xrightarrow[n\to\infty]{P} \min(F_n^1(x,y), F_n^0(x,y)).$$

Следовательно,

$$V_0 = \iint\limits_{R\times R} \varphi_0(x,y)dxdy \xrightarrow[n\to\infty]{P} \iint\limits_{R\times R} F_n^0(x,y)dxdy = 1. \tag{7}$$

Обозначим $\min(F_n^1(x,y), F_n^0(x,y))$ как $\varphi_1^*(x,y)$. Из условия $\exists O \subseteq \mathbb{R}$: $x \in O \Rightarrow F_0(x) \neq F_1(x)$ вытекает, что $\varphi_1^*(x,y) = F_n^0(x,y)$ при $x \notin O$ и $\exists\Omega \subset \mathbb{R}$: $y \in \Omega \Rightarrow \varphi_1^*(x,y) < F_n^0(x,y)$ при $x \in O$.

Следовательно,

$$V_1 = \iint\limits_{R\times R} \varphi_1(x,y)dxdy \xrightarrow[n\to\infty]{P} \iint\limits_{R\times R} \varphi_1^*(x,y)dxdy < \iint\limits_{R\times R} F_n^0(x,y)dxdy = 1. \tag{8}$$

Из (7) и (8) получаем $V_0 > V_1$ при $n \to \infty$, что и требовалось доказать.

Таким образом, используя данное отображение, можно сравнивать любые непрерывные функции распределения с функцией распределения $F(x)$, построенной по выборке, т. е. можно выбрать наиболее вероятную функцию распределения для выборки из конечного числа заданных функций.

Отметим, что мы наблюдаем конечную выборку из генеральной совокупности ξ, имеющей некоторую непрерывную функцию распределения $F_0(x)$. Поскольку выборка конечна, то выборочное распределение не может совпадать с теоретическим распределением. Поэтому и сравнивать распределения $F_1(x), \ldots, F_m(x)$ нужно с выборочным распределением $F(x)$. Выше был использован кусочно-линейный вариант (1) доопределения дискретного распределения до непрерывного вида. При наличии априорной информации о виде распределения случайной величины ξ можно применять и любые иные допустимые (с точки зрения свойств функций распределения) варианты интерполяции.

Экспериментальная часть. Рассмотрим семейство плотностей распределения

$$p(x) = \alpha e^{-|x|^\beta/\delta}, \tag{9}$$

где коэффициенты α и δ подбираются для заданного $\beta > 0$ из условий $\int\limits_{-\infty}^{+\infty} p(x)dx = 1$ (свойство плотности) и $\int\limits_{-\infty}^{+\infty} x^2 p(x)dx = 1$ (условие единичной дисперсии). При $\beta = 1$ имеем закон Лапласа, а при $\beta = 2$ — нормальный закон.

Проведём следующий эксперимент, используя метод Монте-Карло [16, 17]. Для L независимых случайных выборок из нормального закона объёмом n будем выбирать наиболее вероятный закон из некоторого набора законов распределения, содержащего нормальный закон, с помощью описанного метода. Для каждого закона определим процент случаев, когда он был выбран наиболее вероятным (табл. 1, столбцы 1). Так же посчитаем процент случаев для каждого закона, когда критерий согласия χ^2-Пирсона не отвергает гипотезу о том, что данная выборка принадлежит к данному закону (табл. 1, столбцы 2).

Из табл. 1 видно, что критерий согласия Пирсона практически не может различить данные законы распределения. Даже для выборки объёмом 800 более чем в половине случаев он не различает законы с $\beta \in \{1{,}5; 2; 2{,}5\}$. Предложенный же метод позволяет определить наиболее вероятный закон: как видно из таблицы нормальный закон ($\beta = 2$) часто

Таблица 1

Результаты эксперимента (в процентах) для пяти законов распределения из семейства $p(x) = \alpha e^{-|x|^\beta/\delta}$, $\beta \in \{1;\ 1{,}5;\ 2;\ 2{,}5;\ 3\}$, $L = 1000$

β	n							
	100		200		400		800	
	1	2	1	2	1	2	1	2
1	0,1	79,3	0	34,7	0	2	0	0
1,5	19,6	98,9	17,3	97	9,2	89,8	2,5	59
2	**41,5**	**99,6**	**53,5**	**99,1**	**69,3**	**98,8**	**84,3**	**98,2**
2,5	25,1	99,3	23,8	96,9	20,3	89,3	13,1	75,9
3	13,7	97,6	5,4	88,3	1,2	60,5	0,1	20,1

Таблица 2

Математическое ожидание, дисперсия и 90 %-ный доверительный интервал для V_k, $\beta \in \{1;\ 1{,}5;\ 2;\ 2{,}5;\ 3\}$, $n = 400$, $L = 1000$

β	Математическое ожидание V_k	Дисперсия V_k	90 %-ный доверительный интервал
1	0,37	0,0024	[0,368; 0,373]
1,5	0,676	0,0045	[0,672; 0,679]
2	0,779	0,002	[0,777; 0,782]
2,5	0,736	0,0037	[0,733; 0,739]
3	0,668	0,0049	[0,664; 0,672]

оказывался самым вероятным законом. Причём достоверность выбора истинного распределения растёт с увеличением объёма выборки.

Из табл. 2 видно, что дисперсия значений V_k невысока.

Заключение. Предложен новый метод выбора закона распределения случайной величины по экспериментальным данным. Он позволяет из заданного конечного множества выбрать наиболее вероятный непрерывный закон распределения.

Качественный выигрыш рассматриваемого метода по сравнению с известными критериями согласия достигается за счёт двух факторов:

— в основе метода лежит непрерывное отображение множества функций распределений на отрезок [0; 1];

— при расчёте V используются как плотность распределения (гистограмма), так и функция распределения, вследствие чего повышается чувствительность к малым отклонениям в исходных данных.

Проведённое исследование метода на основе статистического моделирования показало его работоспособность. Он может применяться для более точного определения закона распределения по выборке в совокупности с различными критериями согласия.

СПИСОК ЛИТЕРАТУРЫ

1. **Биргер И. А.** Техническая диагностика. М.: Машиностроение, 1978. 241 с.

2. **Генкин М. Д., Соколова А. Г.** Виброакустическая диагностика машин и механизмов. М.: Машиностроение, 1987. 288 с.
3. **Cempel C.** Vibroacoustic Condition Monitoring. N. Y.: Ellis Horwood, 1991. 212 p.
4. **Вибродиагностика** /Под ред. Г. Ш. Розенберга, Е. З. Мадорского, Е. С. Голуба и др. С.-Пб.: ПЭИПК, 2003. 284 с.
5. **Сызранцев В. Н., Невелев Я. П., Голофаст С. Л.** Расчет прочностной надежности изделий на основе методов непараметрической статистики. Новосибирск: Наука, 2008. 218 с.
6. **Иванов В. А., Лысяный К. К.** Надежность и работоспособность конструкций магистральных нефтепроводов. С.-Пб.: Наука, 2003. 317 с.
7. **Ноулер Л.** Статистические методы контроля качества продукции. М.: Изд-во стандартов, 1989. 96 с.
8. **Миттаг Х.-Й., Ринне Х.** Статистические методы обеспечения качества. М.: Машиностроение, 1995. 615 с.
9. **Новицкий П. В., Зограф И. А.** Оценка погрешностей результатов измерений. Л.: Энергоатомиздат, 1985. 248 с.
10. **Тарасенко Ф. П.** Непараметрическая статистика. Томск: Изд-во ТГУ, 1976. 294 с.
11. **Деврой Л., Дьерфи Л.** Непараметрическое оценивание плотности. L_1-подход. М.: Мир, 1988. 408 с.
12. **Орлов А. И.** Прикладная статистика. М.: Экзамен, 2004. 656 с.
13. **Кендалл М., Стьюарт А.** Статистические выводы и связи. М.: Наука, 1973. 900 с.
14. **Ивченко Г. И., Медведев Ю. И.** Введение в математическую статистику. М.: Изд-во ЛКИ, 2010. 600 с.
15. **Вадзинский Р. Н.** Справочник по вероятностным распределениям. С.-Пб.: Наука, 2001. 295 с.
16. **Ермаков С. М.** Метод Монте-Карло и смежные вопросы. М.: Наука, 1975. 472 с.
17. **Лапко А. В., Лапко В. А.** Сравнение эмпирической и предлагаемой функций распределения случайной величины на основе непараметрического классификатора // Автометрия. 2012. **48**, № 1. С. 45–49.

Поступила в редакцию 24 февраля 2012 г.