

Результаты экспериментальных исследований системы одно- и двухсторонних сравнений шкал времени // Метрология времени и пространства: Материалы VIII Междунар. симп. 2016. Менделеево: ФГУП «ВНИИФТРИ», 2016. С. 228–230.

21. Droste S., Ozimek F., Udem Th., Predehl K., Hänsch T. W., Schnatz H., Grosche G., Holzwarth R. Optical-Frequency Transfer over a Single-Span 1840 km Fiber Link // Phys. Rev. Lett., 111, 110801. DOI: <https://doi.org/10.1103/PhysRevLett.111.110801>.

22. Фатеев В. Ф., Рыбаков Е. А., Смирнов Ф. Р. Метод релятивистской синхронизации мобильных атомных часов и его экспериментальная проверка // Письма в ЖТФ. 2017. Т. 43. Вып. 10. С. 3–11.

23. Бердасов О. И., Хабарова К. Ю., Стрелкин С. А., Белотелов Г. С., Костин А. С., Грибов А. Ю., Пальчиков В. Г.,

Колачевский Н. Н., Слюсарев С. Н. Оптические стандарты частоты на холодных атомах стронция // Альманах современной метрологии. 2014. № 1. С. 13–36.

24. Чепуров С. В., Луговой А. А., Кузнецов С. Н., Румынин К. М., Охапкин М. В., Тайченачев А. В., Юдин В. И., Багаев С. Н. Оптический стандарт частоты на основе одиночного иона иттербия // Метрология времени и пространства: Материалы VIII Междунар. симп. Менделеево: ФГУП «ВНИИФТРИ», 2016. С. 241–242.

25. Poli N., Schioppo M., Vogt S., Falke St., Sterr U., Lisdat Ch., Tino G. M. A transportable strontium optical lattice clock // Appl. Phys. B. 2014. No. 117 P. 1107–1116. DOI 10.1007/s00340-014-5932-9.

Дата принятия: 07.02.2018.

ОБЩИЕ ВОПРОСЫ МЕТРОЛОГИИ И ИЗМЕРИТЕЛЬНОЙ ТЕХНИКИ

519.24

Оценивание параметров формулы оптимальной дискретизации области значений двумерной случайной величины

А. В. ЛАПКО^{1,2}, В. А. ЛАПКО^{1,2}

¹Институт вычислительного моделирования Сибирского отделения РАН, Красноярск, Россия

²Сибирский государственный университет науки и технологий им. акад. М. Ф. Решетнева, Красноярск, Россия, e-mail: lapko@ict.krasn.ru

Рассмотрены методы оценивания параметров формулы оптимальной дискретизации области определения плотности вероятности двумерной случайной величины. Исследованы свойства предложенных методов и определены условия их компетентности.

Ключевые слова: дискретизация области значений случайных величин, формула оптимальной дискретизации, непараметрическая оценка Розенблатта–Парзена, критерий Пирсона, правило Хайнкольда–Гаеде, правило Скотта.

Methods for estimating the parameters of the formula for the optimal discretization of the range of variation in the probability density of a two-dimensional random variable are considered. The properties of the proposed methods are investigated and the conditions for their competence are defined.

Key words: discretization of the range of values of random variables, optimal sampling formula, choice of histogram bin width, nonparametric estimation Rosenblatt–Parzen, Pearson criterion, Heinhold–Gaede rule, Scott's rule.

Задача дискретизации области значений случайных величин является одной из актуальных в математической статистике. Результаты решения поставленной задачи используются при построении доверительных границ плотности вероятности и проверке гипотез о распределениях случайных величин с помощью критерия Пирсона [1–3]. Известен ряд методов дискретизации интервала значений одномерной случайной величины [4–7]. При анализе аппроксимационных свойств непараметрической оценки плотности вероятности $\hat{p}(x)$ парзеновского типа, восстанавливаемой по статистическим данным объема n , получена процедура выбора оптимального количества интервалов дискретизации области Δ

значений одномерной случайной величины x с плотностью вероятности $p(x)$ [8, 9]:

$$N_1^* = \left(\Delta \int_{-\infty}^{+\infty} p^2(x) dx n \right)^{1/2} = k_1 \sqrt{n}.$$

Данная аналитическая зависимость по виду близка к формуле Хайнкольда–Гаеде и совпадает с ней при оценивании плотности вероятности случайной величины с равномерным законом распределения [10]. Установлено, что коэффициент k_1 определяется только видом плотности вероятности и не зависит от её параметров. Дискретизация области значений двумер-

ных независимых случайных величин, основанная на применении формул Старджесса и Скотта, применяется при обработке изображений [11, 12]. Впервые выбор оптимального количества интервалов дискретизации двумерной случайной величины обоснован в работе [13]. Среднее квадратическое отклонение непараметрической оценки плотности вероятности $\hat{p}(x_1, x_2)$ [9] от $p(x_1, x_2)$ можно записать следующим образом:

$$W_2(N) = \left(\frac{1}{4} B^2 (\|\Phi(u)\|^2)^8 (\|p(x_1, x_2)\|^2)^4 \Delta^2 \right)^{1/6} \times \left(\frac{3\Delta}{2N^{4/6}} + \frac{N^{8/6}}{n\Delta \|p(x_1, x_2)\|^2} \right). \quad (1)$$

Из условия минимума (1) определено оптимальное количество N^* двумерных интервалов дискретизации

$$N^* = \left(\frac{3}{4} \Delta^2 \|p(x_1, x_2)\|^2 n \right)^{1/2} = k_2 \sqrt{n}. \quad (2)$$

Ядерная функция $\Phi(u)$ в выражении (1) удовлетворяет условиям

$$\Phi(u) = \Phi(-u); 0 \leq \Phi(u) < \infty;$$

$$\int \Phi(u) du = 1; \int u^2 \Phi(u) du = 1.$$

Здесь и далее бесконечные пределы интегрирования опускаются. Приняты следующие обозначения:

$$\|\Phi(u)\|^2 = \int \Phi^2(u) du;$$

$$\|p(x_1, x_2)\|^2 = \iint p^2(x_1, x_2) dx_1 dx_2;$$

$$B = \iint \left(\sum_{v=1}^2 p_v^{(2)}(x_1, x_2) \right)^2 dx_1 dx_2,$$

где $p_v^{(2)}(x_1, x_2)$ – вторая производная плотности вероятности $p(x_1, x_2)$ по аргументу x_v . Полагается, что интервалы значений случайных величин x_1, x_2 равны, т. е. $\Delta_1 = \Delta_2 = \Delta$.

Количество N_v^* интервалов дискретизации каждой составляющей $x_v, v=1, 2$ соответствуют целому значению $(N^*)^{1/2}$. Показано, что значения коэффициента k_2 для независимых и зависимых случайных величин x_1, x_2 определяются только видом плотности вероятности и не зависят от её параметров. Например, для равномерного, линейного и нормального законов распределения коэффициент k_2 принимает значения 0,866; 1,15; 1,46. Эта закономерность сохраняется при разных значениях Δ_1, Δ_2 .

Цель данной работы – оценить функционал $\|p(x_1, x_2)\|^2$ от плотности вероятности в формуле оптимального количества интервалов дискретизации двумерной случайной величины при оценивании коэффициента k_2 с использованием непараметрической статистики Розенблатта–Парзена и установлении условий компетентности получаемых оценок формулы (2).

Анализ свойств нелинейного функционала от непараметрической оценки плотности вероятности. Пусть имеется выборка $V = (x_v^i, x_v^j, i=1, n)$ из n независимых наблюдений двумерной случайной величины с неизвестной плотностью веро-

ятности $p(x_1, x_2)$. В качестве приближения по эмпирическим данным V искомым плотности $p(x_1, x_2)$ примем статистику типа Розенблатта–Парзена [14]:

$$\bar{p}(x_1, x_2) = \frac{1}{nc^2} \sum_{i=1}^n \prod_{v=1}^2 \Phi\left(\frac{x_v - x_v^i}{c}\right). \quad (3)$$

Для упрощения дальнейших преобразований будем считать, что интервалы значений случайных величин x_1, x_2 одинаковые. В этих условиях появляется возможность полагать, что и значения коэффициентов размытости $c_1 = c_2 = c$ ядерных функций непараметрической оценки плотности вероятности парзеновского типа, которые убывают с ростом объёма n исходных статистических данных, тоже равны.

При оценивании функционала

$$b = \|p(x_1, x_2)\|^2 = \iint p^2(x_1, x_2) dx_1 dx_2 \quad (4)$$

от плотности вероятности $p(x_1, x_2)$ в формуле дискретизации области значений двумерной случайной величины (2) будем использовать статистику (3). Выбор коэффициентов размытости c ядерных функций осуществляется в соответствии с методиками, предложенными в работе [15].

Сравним эффективность методов оценивания функционала (4) от плотности вероятности $p(x_1, x_2)$. Для этого применим следующие статистики:

$$b_1(c) = \iint \bar{p}^2(x_1, x_2) dx_1 dx_2;$$

$$b_2(c) = \iint \bar{p}(x_1, x_2) p(x_1, x_2) dx_1 dx_2.$$

Опираемся будем на аналитический и вычислительный методы исследования. Используя технологию анализа асимптотических свойств непараметрической оценки плотности вероятности (3), вычислим смещение статистики $b_1(c)$, которое определяется как [16]:

$$M[b_1(c) - b] = \iint [M\bar{p}^2(x_1, x_2) - p^2(x_1, x_2)] dx_1 dx_2, \quad (5)$$

где M – математическое ожидание.

Так как статистически независимые наблюдения $(x_v^i, x_v^j, i=1, n)$ характеризуются одной плотностью вероятности

$p(x_1^i, x_2^i) = p(x_1^j, x_2^j) = \dots = p(x_1^n, x_2^n) = p(t_1, t_2)$, то после несложных преобразований получим

$$\begin{aligned} M \iint \bar{p}^2(x_1, x_2) dx_1 dx_2 &= M \iint \left\{ \frac{1}{n^2 c^4} \left[\sum_{i=1}^n \left(\prod_{v=1}^2 \Phi\left(\frac{x_v - x_v^i}{c}\right) \right)^2 + \right. \right. \\ &\quad \left. \left. + \sum_{\substack{i=1 \\ j \neq i}}^n \sum_{\substack{l=1 \\ l \neq i}}^n \prod_{v=1}^2 \Phi\left(\frac{x_v - x_v^i}{c}\right) \prod_{v=1}^2 \Phi\left(\frac{x_v - x_v^l}{c}\right) \right] \right\} dx_1 dx_2 = \\ &= \frac{1}{n^2 c^4} \iint \left[n \iint \left(\prod_{v=1}^2 \Phi\left(\frac{x_v - t_v}{c}\right) \right)^2 p(t_1, t_2) dt_1 dt_2 + \right. \\ &\quad \left. + n(n-1) \left(\iint \prod_{v=1}^2 \Phi\left(\frac{x_v - t_v}{c}\right) p(t_1, t_2) dt_1 dt_2 \right)^2 \right] dx_1 dx_2. \quad (6) \end{aligned}$$

Проведём замену переменных $u_v = (x_v - t_v)/c$, $v=1,2$ в выражении (6), и разложим функции $p(x_1 - cu_1, x_2 - cu_2)$ в ряд Тейлора в точке (x_1, x_2) . Тогда при достаточно больших значениях n будем иметь

$$M \iint \bar{p}^2(x_1, x_2) dx_1 dx_2 \sim \iint p^2(x_1, x_2) dx_1 dx_2 + \frac{(\int \Phi^2(u) du)^2}{nc^2} + c^2 \iint p(x_1, x_2) [p_1^{(2)}(x_1, x_2) + p_2^{(2)}(x_1, x_2)] dx_1 dx_2, \quad (7)$$

где $\int \Phi^2(u_1) du_1 = \int \Phi^2(u_2) du_2 = \int \Phi^2(u) du$.

При получении данного результата составляющие малости $O(1/n)$, $O(c^2/n)$, $O(c^4)$ не учитывались.

Подставив в правую часть выражения (5) результат (7), запишем

$$M(b_1(c) - b) \sim \frac{(\int \Phi^2(u) du)^2}{nc^2} + c^2 \iint p(x_1, x_2) [p_1^{(2)}(x_1, x_2) + p_2^{(2)}(x_1, x_2)] dx_1 dx_2.$$

Исследуем статистику $b_2(c)$, которая допускает оценивание как

$$\bar{b}_2(c) = \frac{1}{n} \sum_{j=1}^n \bar{p}(x^j) = \frac{1}{n^2 c^2} \sum_{j=1}^n \sum_{i=1}^n \prod_{v=1}^2 \Phi\left(\frac{x_v^j - x_v^i}{c}\right). \quad (8)$$

При выполнении условия $i \neq j$ в выражении (8) статистика $\bar{b}_2(c)$ является несмещённой оценкой $b_2(c)$, т. е. $M(\bar{b}_2(c)) = b_2(c)$ [15].

Найдём смещение статистики $\bar{b}_2(c)$ относительно b . Выполним преобразования по приведённой выше методике, определим

$$\begin{aligned} M(\bar{b}_2(c)) &= \frac{1}{n^2 c^2} \sum_{j=1}^n \sum_{i=1}^n \iiint \prod_{v=1}^2 \Phi\left(\frac{x_v^j - x_v^i}{c}\right) \times \\ &\times p(x_1^j, x_2^j) p(x_1^i, x_2^i) dx_1^j dx_2^j dx_1^i dx_2^i = \frac{n(n-1)}{n^2 c^2} \times \\ &\times \iiint \prod_{v=1}^2 \Phi\left(\frac{t_v - t_v^1}{c}\right) p(t_1, t_2) p(t_1^1, t_2^1) dt_1 dt_2 dt_1^1 dt_2^1 = \frac{n-1}{n} \times \\ &\times \iiint \prod_{v=1}^2 \Phi(u_v) p(cu_1 + t_1^1, cu_2 + t_2^1) p(t_1^1, t_2^1) du_1 du_2 dt_1^1 dt_2^1 \sim \\ &\sim \iint p^2(t_1^1, t_2^1) dt_1^1 dt_2^1 + \\ &+ \frac{c^2}{2} \iint [p_1^{(2)}(t_1^1, t_2^1) + p_2^{(2)}(t_1^1, t_2^1)] p(t_1^1, t_2^1) dt_1 dt_2. \quad (9) \end{aligned}$$

Обозначим асимптотические выражения (7), (9) как $a(7)$, $a(9)$ при оптимальных значениях коэффициента размытости ядерной функции

$$\bar{c} = \left[2(\|\Phi(u)\|^2)^2 / (nB) \right]^{1/6},$$

где $B = \iint [p_1^{(2)}(x_1, x_2) + p_2^{(2)}(x_1, x_2)] dx_1 dx_2$.

Рассмотрим отношения $a(7)/b$, $a(9)/b$ при $b = \iint p^2(x_1, x_2) dx_1 dx_2$. Получим

$$a(1) = \frac{a(7)}{b} = 1 + \frac{1}{b} \left[(\|\Phi(u)\|^2)^4 B / (2n^2) \right]^{1/3} + d;$$

$$a(2) = \frac{a(9)}{b} = 1 + \frac{d}{2},$$

где $d = \frac{\bar{c}^2}{b} B_1$, $B_1 = \iint p(x_1, x_2) \sum_{v=1}^2 p_v^{(2)}(x_1, x_2) dx_1 dx_2$.

Исследуем зависимость введённых отношений от объёма n исходных данных при восстановлении плотности вероятности двумерной случайной величины с нормальным законом распределения. Запишем выражение

$$p(x_1, x_2) = \left[2\pi\sigma_1\sigma_2(1-r^2) \right]^{-1/2} \exp \left\{ -\frac{1}{2(1-r^2)} \times \left[\frac{(x_1 - m_1)^2}{\sigma_1^2} - r \frac{2(x_1 - m_1)(x_2 - m_2)}{\sigma_1\sigma_2} + \frac{(x_2 - m_2)^2}{\sigma_2^2} \right] \right\}, \quad (10)$$

где r – коэффициент корреляции x_1, x_2 .

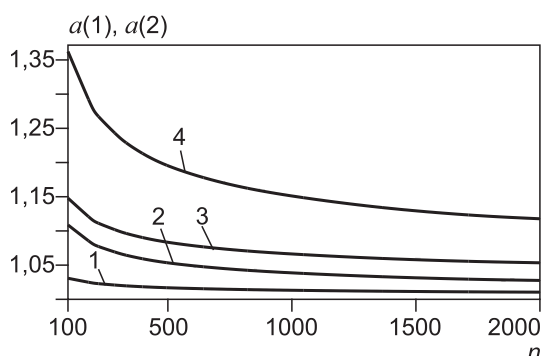
Значения математических ожиданий случайных величин $m_1 = m_2 = 0$, а их средние квадратические отклонения $\sigma_1 = \sigma_2 = 1$. Синтез непараметрической оценки плотности вероятности (3) осуществляется на основе ядерных функций

$$\Phi(u) = \begin{cases} (2\sqrt{3})^{-1} \forall |u| < \sqrt{3}; \\ 0 \forall |u| \geq \sqrt{3}. \end{cases} \quad (11)$$

В этих условиях составляющие выражений (1), (2) принимают значения: $\|\Phi(u)\|^2 = 1/(2\sqrt{3})$; $b = 0,083$, $B = 0,211$ при $r = 0,3$ и $b = 0,133$, $B = 2,702$ при $r = 0,8$, соответственно. Зависимость отношений $a(1)$, $a(2)$ от объёма n исходных статистических данных приведена на рисунке. Полученные зависимости подтверждают преимущество метода оценивания функционала b с использованием статистики $\bar{b}_2(c)$ по сравнению с $b_1(c)$, что следует из анализа соответствующих им асимптотических выражений (9), (7). Эта закономерность сохраняется при разных значениях r . Наблюдаемое снижение аппроксимационных свойств статистик $\bar{b}_2(c)$, $b_1(c)$ с ростом коэффициента корреляции r объясняется увеличением значений производных $p_v^{(2)}(x_1, x_2)$, $v=1,2$ в выражениях B , B_1 . Причём отношение $a(1)$ более чувствительно к изменению значений r по сравнению с $a(2)$. Например, при $n=300$ отношения $a(1)=1,066$ и $a(2)=1,021$ при $\bar{c}=0,372$, $B=0,211$, $d=0,042$ в условиях $r=0,3$. С ростом коэффициента корреляции $r=0,8$ анализируемые отношения увеличиваются: $a(1)=1,238$, $a(2)=1,101$ при $n=300$, $\bar{c}=0,243$, $B=2,702$, $d=0,202$.

Анализ результатов вычислительных экспериментов. Исследуем зависимость количества интервалов дискретизации области значений двумерной случайной величины от методов оценивания функционала $\|p(x_1, x_2)\|^2$ в выражении (2).

Пусть имеется выборка $V = (x_1^i, x_2^i, i=\overline{1, n})$ значений случайных величин, характеризующихся плотностью вероятности вида (10). Для оценивания функционала $\|p(x_1, x_2)\|^2$ от плотности



Зависимость отношений $a(1)$, $a(2)$ от объема n статистических данных: 1, 3 – отношение $a(2)$ при $r=0,3$; $0,8$, соответственно; 2, 4 – отношение $a(1)$ в тех же условиях

вероятности $p(x_1, x_2)$ применялись ядерные функции вида (11), а объем n исходных данных выбирался из интервала $[100, 1500]$. При конкретном объеме n значения $b_1(c)$, $b_2(c)$ вычислялись 50 раз и полученные результаты вычислительных экспериментов усреднялись. Результаты применения исследуемых методов оценивания функционала $\|p(x_1, x_2)\|^2$ в формуле (2) сравнивались с правилом выбора количества интервалов дискретизации N_s^* значений двумерной случайной величины $x=(x_1, x_2)$ с нормальным законом распределения [17]:

$$N_s^* = N_1^* N_2^* \quad (12)$$

$$N_v^* = (x_v^{\max} - x_v^{\min}) / h_v^*, \quad v=1, 2,$$

где $h_v^* = 3,504 \sigma_v (1-r^2)^{3/8} / n^{1/4}$ – асимптотически оптимальная ширина интервала дискретизации x_v , $v=1, 2$.

Параметры средних квадратических отклонений σ_v , $v=1, 2$ случайных величин x_1 , x_2 , интервалы Δ_1 , Δ_2 и границы $x_v^{\max}$, $x_v^{\min}$ их изменения в формулах (2), (12) оценивались по статистическим данным вычислительных экспериментов. Количество двумерных интервалов N^* , вычисляемых по формуле (2), может не являться целыми числами. Обозначим через $\bar{N}_v^* = \lfloor N_v^* \rfloor$, $\bar{N}_v^* = \bar{N}_v^* + 1$ – левую и правую границы целого количества интервалов дискретизации случайных величин x_1 , x_2 . Тогда скоррек-

тированное количество интервалов дискретизации $\bar{N}^*$ области значений случайных величин x_1 , x_2 определяется правилом

$$\tilde{N}^* = \begin{cases} (\bar{N}_v^*)^2, & \text{если } W_2((\bar{N}_v^*)^2) < W_2((\bar{N}_v^*)^2); \\ (\bar{N}_v^*)^2, & \text{если } W_2((\bar{N}_v^*)^2) < W_2((\bar{N}_v^*)^2). \end{cases}$$

При этом используется критерий (1), из условия минимума которого находятся значения N^* в соответствии с выражением (2). По аналогии определяется целое количество интервалов дискретизации $\tilde{N}_s^*$, вычисляемых по правилу Скотта (12).

При различных значениях коэффициентов корреляции r количество интервалов дискретизации, вычисляемых по формуле (2) и её оценкам с использованием статистик $b_1(c)$, $b_2(c)$, совпадают (таблица). Эти оценки формулы (2) в дальнейшем будем называть методами 1, 2. Установленное преимущество аппроксимационных свойств $\bar{b}_2(c)$ над $b_1(c)$ компенсируется в процессе определения целых значений $\tilde{N}^*$, $\tilde{N}_s^*$.

В рассмотренных условиях правило Скотта, ориентированное на анализ случайных величин с нормальным законом распределения, даёт завышенные результаты, что особенно проявляется при больших объемах статистических данных и увеличении значений r . Для двумерной случайной величины зависимость количества интервалов дискретизации, вычисляемых по формулам типа (2) и в соответствии с правилом Скотта $N_s^* = k_s \sqrt{n}$, от объема n определяется значением $\sqrt{n}$. Отмеченная выше тенденция объясняется повышенной зависимостью коэффициента

$$k_s = \frac{\Delta_1 \Delta_2}{(3,504)^2 \sigma_1 \sigma_2 (1-r^2)^{3/4}}$$

от интервалов Δ_1 , Δ_2 значений случайных величин по сравнению с правилами дискретизации, основанными на применении выражения (2). В принятых условиях вычислительных экспериментов при больших объемах статистических данных n вероятность появления малых и больших значений случайных величин x_1 , x_2 значительная. Поэтому оценки интервалов Δ_1 , Δ_2 в среднем определяются большими значениями. В условиях выборок меньшего объема зависимость k_s от Δ_1 , Δ_2 ослабевает. Отмеченная тенденция справедлива

для всех значений $\tilde{N}_s^*$, отличающихся от $\tilde{N}^*$. Например, в процессе вычислительных экспериментов при $n=300$ произведение $\Delta_1 \Delta_2$ определяется интервалом $(26,2; 38,7)$ со средним значением $\Delta_1 \Delta_2 = 31,3$. При $n=1100$ значения $\Delta_1 \Delta_2 \in (34,5; 45,1)$, а $\Delta_1 \Delta_2 = 39,8$. Увеличение значений r , как следует из анализа зависимости коэффициента k_s от r , дополнительно способствует росту количества интервалов дискретизации $\tilde{N}_s^*$ (см. таблицу). Если при $n=1100$ и $r=0$ среднее значение $k_s = 3,08$, а соответствующее

Количество интервалов дискретизации области значений двумерной случайной величины с нормальным законом распределения, соответствующих объёму n статистических данных, коэффициенту корреляции r между x_1 , x_2 и сравниваемым методам дискретизации

r	Метод дискретизации	n							
		100	300	500	700	900	1100	1300	1500
0	Методы (1), (2), формула (2)	16	25	36	36	49	49	49	64
	Правило Скотта	16	36	64	81	81	100	100	121
0,3	Методы (1), (2), формула (2)	9	16	25	25	36	36	49	49
	Правило Скотта	16	49	64	81	81	100	100	121
0,8	Методы (1), (2), формула (2)	9	25	36	36	49	49	64	64
	Правило Скотта	36	81	121	169	196	225	225	256

ему количество интервалов дискретизации $\tilde{N}_s^* = 100$, то при $r = 0,8$ значения $\bar{k}_s = 7,1$ и $\tilde{N}_s^* = 225$.

Рекомендации по выбору количества интервалов дискретизации $\tilde{N}_s^*$ по правилу Скотта являются менее эффективными, что подтверждается большим значением критерия $W_2(\tilde{N}_s^*)$ по сравнению со значениями $W_2(\tilde{N}^*)$ для методов 1, 2. Данное утверждение справедливо для всех значений $\tilde{N}_s^*$, отличающихся от $\tilde{N}^*$. Например, при $n = 300$: значения $W_2(\tilde{N}_s^*) = 0,088$, $W_2(\tilde{N}_s^*) = 0,094$ в условиях $r = 0$; $W_2(\tilde{N}_s^*) = 0,04$, $W_2(\tilde{N}_s^*) = 0,194$ при $r = 0,8$. При $n = 1100$: значения $W_2(\tilde{N}_s^*) = 0,057$, $W_2(\tilde{N}_s^*) = 0,07$ в условиях $r = 0$; $W_2(\tilde{N}_s^*) = 0,027$, $W_2(\tilde{N}_s^*) = 0,207$ при $r = 0,8$.

Заключение. Оптимальное количество интервалов дискретизации области значений двумерной случайной величины зависит от функционала $b = \|p(x_1, x_2)\|^2$. При использовании статистики Розенблатта–Парзена $\bar{p}(x)$ предлагаются две его оценки, установлены условия их компетентности. Представление функционала b в виде оценки $\bar{b}_2(c)$ математического ожидания $\bar{p}(x_1, x_2)$ обладает преимуществом по сравнению с $b_1(c) = \|\bar{p}(x_1, x_2)\|^2$. Данное утверждение подтверждается выводами асимптотических исследований и результатами вычислительных экспериментов восстановления плотности вероятности зависимых двумерных случайных величин с нормальными законами распределения. Аппроксимационные свойства статистик $b_1(c)$, $\bar{b}_2(c)$ снижаются с ростом коэффициента корреляции между случайными величинами. Количество интервалов дискретизации области значений двумерной случайной величины, вычисляемых по формуле (2) и по её оценкам с использованием статистик $b_1(c)$, $\bar{b}_2(c)$, совпадают. Этот вывод обосновывается тем, что преимущество аппроксимационных свойств $\bar{b}_2(c)$ над $b_1(c)$ компенсируется в процессе определения целых значений $\tilde{N}^*$, $\tilde{N}_s^*$. Правило Скотта для выбора количества интервалов дискретизации менее эффективно по сравнению с анализируемыми методами и даёт завышенные результаты, что особенно характерно при увеличении коэффициента корреляции и объёма статистических данных.

Полученные результаты имеют важное значение при доверительном оценивании плотности вероятности, проверке гипотез о распределениях случайных величин с помощью критерия Пирсона и обработке статистических данных большого объёма.

Л и т е р а т у р а

1. Пугачёв В. С. Теория вероятностей и математическая статистика: Учеб. пособие. М.: ФИЗМАТЛИТ, 2002.
2. Лапко А. В., Лапко В. А. Построение доверительных границ для плотности вероятности на основе её регрессионной оценки // Метрология. 2013. № 12. С. 3–9.
3. Lapko A. V., Lapko V. A. Comparison of empirical and theoretical distribution functions of a random variable on the basis

of a nonparametric classifier // Optoelectronics, Instrumentation and Data Processing. 2012. V. 48. No. 1. P. 37–41.

4. Sturges H. A. The choice of a class interval // J. American Statistical Association. 1926. V. 21. P. 65–66.
5. Wand M. P. Data-based choice of histogram bin width // The American Statistician. 1997. V. 51. No. 1. P. 59–64.
6. Shimazaki H., Shinomoto S. A method for selecting the bin size of a time histogram // Neural Computation. 2007. V. 19. No. 6. P. 1503–1527.
7. Hacine-Gharbi A., Ravier P., Harba R., Mohamadi T. Low bias histogram-based estimation of mutual information for feature selection // Pattern Recognition Letters. 2012. V. 33. No. 10. P. 1302–1308.
8. Лапко А. В., Лапко В. А. Оптимальный выбор количества интервалов дискретизации области изменения одномерной случайной величины при оценивании плотности вероятности // Измерительная техника. 2013. № 7. С. 24–27.
9. Lapko A. V., Lapko V. A. Regression Estimate of the Multidimensional Probability Density and Its Properties // Optoelectronics, Instrumentation and Data Processing. 2014. V. 50. No. 2. P. 148–153.
10. Heinhold I., Gaede K. Ingenieur statistic. München: Wien, Springer Verlag, 1964.
11. Minnotte M. C., Sain S. R., Scott D. W. Multivariate Visualization by Density Estimation // Handbook of Data Visualization. 2008. Pt. 3. P. 389–413.
12. Legg P. A., Rosin P. L., Marshall D., Morgan J. E. Improving accuracy and efficiency of registration by mutual information using Sturges' histogram rule // 11th Annual Conf. in Medical Image Understanding and Analysis. 2007. P. 26–30.
13. Лапко А. В., Лапко В. А. Выбор оптимального количества интервалов дискретизации области значений двумерной случайной величины // Измерительная техника. 2016. № 2. С. 14–17.
14. Parzen E. On estimation of a probability density function and mode // Annals of Mathematics Statistic. 1962. V. 33. No. 3. P. 1065–1076.
15. Лапко А. В., Лапко В. А. Анализ методов оптимизации непараметрической оценки плотности вероятности по коэффициенту размытости ядерных функций // Измерительная техника. 2017. № 6. С. 3–8.
16. Епанечников В. А. Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. Т. 14. № 1. С. 156–161.
17. Scott D. W. Multivariate Density Estimation: Theory, Practice, and Visualization. N. Y.: Wiley, 1992.

Дата принятия: 04.12.2017.