УДК 519.7

СРАВНЕНИЕ МЕТОДИК ПРОВЕРКИ ГИПОТЕЗЫ О НЕЗАВИСИМОСТИ СЛУЧАЙНЫХ ВЕЛИЧИН, ОСНОВАННЫХ НА НЕПАРАМЕТРИЧЕСКОМ КЛАССИФИКАТОРЕ И КРИТЕРИИ ПИРСОНА

© **А. В. Лапко**[1,2], **В. А. Лапко**[1,2], **А. В. Бахтина**[2]

[1] *Институт вычислительного моделирования СО РАН,*
660036, г. Красноярск, Академгородок, 50, стр. 44
[2] *Сибирский государственный университет науки и технологий*
им. академика М. Ф. Решетнева,
660037, г. Красноярск, просп. им. газеты «Красноярский рабочий», 31
E-mail: lapko@icm.krasn.ru

Методика проверки гипотезы о независимости случайных величин, основанная на непараметрическом алгоритме распознавания образов, используется при анализе неоднозначных зависимостей. Алгоритм распознавания образов соответствует критерию максимального правдоподобия. Оценивание законов распределения в классах осуществляется по исходным статистическим данным в предположении о независимости и зависимости сравниваемых случайных величин. Для оценивания плотностей вероятностей в классах используются непараметрические статистики Розенблатта — Парзена. Коэффициенты размытости ядерных функций в непараметрических оценках плотностей вероятностей в классах определяются из условия минимума их средних квадратических отклонений. В этих условиях вычисляются оценки вероятностей ошибок распознавания образов в классах. По минимальному их значению принимается решение о независимости либо зависимости случайных величин. Проверяется гипотеза о достоверном отличии вероятностей ошибок распознавания образов в классах. Применение предлагаемой методики позволяет обойти проблему декомпозиции области значений случайных величин на интервалы, что свойственно критерию Пирсона. Сравнивается эффективность предлагаемой методики с критерием Пирсона. Приводятся результаты вычислительных экспериментов с применением исследуемых критериев при анализе неоднозначных зависимостей между случайными величинами.

Ключевые слова: проверка гипотезы о независимости случайных величин, двухмерные случайные величины, непараметрический алгоритм распознавания образов, ядерная оценка плотности вероятности, критерий Пирсона, неоднозначные функциональные зависимости.

DOI: 10.15372/AUT20230504

Введение. Проверка гипотезы о независимости случайных величин является одним из основных этапов системного анализа статистических данных. На её результатах осуществляется синтез эффективных алгоритмов принятия решений. Традиционная методика проверки гипотезы о независимости случайных величин основана на использовании критерия Пирсона, которая содержит трудно формализуемый этап разбиения области значений случайных величин на многомерные интервалы [1]. В работах [2, 3] предложена методика проверки гипотезы о независимости случайных величин, основанная на использовании непараметрического алгоритма распознавания образов ядерного типа. Её применение позволяет обойти проблему декомпозиции области значений случайных величин на интервалы. Идея подхода состоит в формировании по исходным статистическим данным обучающей выборки для решения двухальтернативной задачи распознавания образов. Каждый класс определяется в предположении о независимости либо зависимости случайных величин, что проявляется в различии их законов распределения в классах. В этих условиях появляется возможность замены исходной гипотезы задачей проверки достоверности

различий вероятностей ошибок распознавания образов в классах. Эффективность предлагаемого подхода подтверждается результатами исследования однозначных зависимостей между случайными величинами.

Цель данной работы состоит в сравнении эффективности предлагаемой методики проверки гипотезы о независимости случайных величин с традиционным критерием Пирсона при анализе статистических данных, характеризующих неоднозначные зависимости между случайными величинами, и изменении объёма статистических данных.

Непараметрический алгоритм распознавания образов, соответствующий критерию максимального правдоподобия. Имеется выборка $V = (x^i, i = \overline{1,n})$ объёма n, составленная из наблюдений двухмерной случайной величины $x = (x_1, x_2)$. Случайные величины x_1, x_2 характеризуются плотностями вероятности $p(x_1)p(x_2)$ или $p(x_1, x_2)$. Необходимо по статистическим данным V проверить гипотезу

$$H_0: \quad p(x_1, x_2) \equiv p(x_1)p(x_2) \tag{1}$$

о независимости случайных величин x_1, x_2.

Для проверки гипотезы H_0 будем решать двухальтернативную задачу распознавания образов. Под классами Ω_1, Ω_2 понимаются области определения плотностей вероятностей $p(x_1)p(x_2)$ и $p(x_1, x_2)$.

В отличие от традиционной постановки задачи распознавания образов при синтезе решающего правила априори отсутствует обучающая выборка, содержащая сведения о принадлежности элементов выборки V к тому или иному классу. Эти сведения заменяются на предположения о независимости либо зависимости случайных величин в соответствии с гипотезой (1).

Для оценивания плотностей вероятностей $p(x_1, x_2)$ и $p(x_1)p(x_2)$ будем использовать их непараметрические оценки Розенблатта — Парзена. Если случайные величины x_1, x_2 являются зависимыми, то непараметрическая оценка их плотности вероятности запишется в виде [4, 5]

$$\bar{p}(x_1, x_2) = \frac{1}{nc_1c_2} \sum_{i=1}^{n} \Phi\Big(\frac{x_1 - x_1^i}{c_1}\Big)\Phi\Big(\frac{x_2 - x_2^i}{c_2}\Big), \tag{2}$$

где c_v — коэффициент размытости ядерной функции $\Phi(u_v)$. В статистике (2) ядерные функции $\Phi(u_v)$ удовлетворяют условиям:

$$\Phi(u_v) = \Phi(-u_v), \qquad 0 \leqslant \Phi(u_v) < \infty, \qquad \int \Phi(u_v)\, du_v = 1,$$

$$\int u^m \Phi(u_v)\, du_v < \infty, \qquad 0 \leqslant m < \infty, \qquad v = 1, 2.$$

Здесь и далее бесконечные пределы интегрирования опускаются.

При выполнении условий независимости x_1, x_2 непараметрическая оценка $p(x_1)p(x_2)$ представляется как

$$\bar{p}(x_1)\bar{p}(x_2) = \frac{1}{n^2c_1c_2} \sum_{i=1}^{n}\sum_{j=1}^{n} \Phi\Big(\frac{x_1 - x_1^i}{c_1}\Big)\Phi\Big(\frac{x_2 - x_2^j}{c_2}\Big). \tag{3}$$

Тогда непараметрическое решающее правило классификации случайных величин $x = (x_1, x_2)$, соответствующее критерию максимального правдоподобия, запишется в виде

$$\bar{m}(x): \begin{cases} x \in \Omega_1, & \text{если } \bar{p}(x_1, x_2) < \bar{p}(x_1)\bar{p}(x_2); \\ x \in \Omega_2, & \text{если } \bar{p}(x_1, x_2) > \bar{p}(x_1)\bar{p}(x_2). \end{cases} \tag{4}$$

Оптимальный коэффициент размытости c_v ядерных функций непараметрической оценки плотности вероятности $\bar{p}(x_v)$ будем определять из условия минимума статистической оценки критерия

$$W(c_v) = \int (\bar{p}(x_v) - p(x_v))^2 \, dx_v, \tag{5}$$

который характеризует меру близости между $\bar{p}(x_v)$ и $p(x_v)$, $v = 1, 2$.

В работах [6–12] обоснована возможность определения оптимального значения c_v путём минимизации выражения

$$\bar{W}(c_v) = \frac{1}{n^2 c_v^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \int \Phi\Big(\frac{x_v - x_v^j}{c_v}\Big) \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) \, dx_v - \frac{2}{n^2 c_v} \sum_{j=1}^{n} \sum_{i=1, i \neq j}^{n} \Phi\Big(\frac{x_v^j - x_v^i}{c_v}\Big), \tag{6}$$

$$v = 1, 2.$$

Например, при ступенчатой ядерной функции

$$\frac{1}{c_v} \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) = \begin{cases} 0{,}5c_v, & \text{если } |x_v - x_v^i| < c_v; \\ 0, & \text{если } |x_v - x_v^i| \geqslant c_v \end{cases}$$

значения составляющих критерия (6) определяются выражениями

$$\int \Phi\Big(\frac{x_v - x_v^j}{c_v}\Big) \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) \, dx_v = \begin{cases} (2c_v - |x_v^j - x_v^i|)/4, & \text{если } |x_v^j - x_v^i| < 2c_v; \\ 0, & \text{если } |x_v^j - x_v^i| \geqslant 2c_v, \end{cases}$$

$$\Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) = \begin{cases} 0{,}5, & \text{если } |x_v - x_v^i| < c_v; \\ 0, & \text{если } |x_v - x_v^i| \geqslant c_v, \end{cases} \qquad v = 1, 2.$$

По аналогии с выражением (6) нетрудно определить критерий выбора оптимальных коэффициентов размытости непараметрической статистики $\bar{p}(x_1, x_2)$ (2):

$$\bar{W}(c_1, c_2) = \frac{1}{n^2 c_1^2 c_2^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \prod_{v=1}^{2} \int \Phi\Big(\frac{x_v - x_v^j}{c_v}\Big) \Phi\Big(\frac{x_v - x_v^i}{c_v}\Big) \, dx_v - \\ - \frac{2}{n^2 c_1 c_2} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} \prod_{v=1}^{2} \Phi\Big(\frac{x_v^j - x_v^i}{c_v}\Big),$$

минимум которого определяет оптимальные коэффициенты размытости ядерной оценки плотности вероятности $\bar{p}(x_1, x_2)$.

Оптимизацию непараметрического решающего правила (4) по коэффициентам размытости ядерных функций можно упростить, если положить в статистиках (2), (3) значения $c_v = c\bar{\sigma}_v$, $v = 1, 2$. Здесь $\bar{\sigma}_v$ — оценка среднего квадратического отклонения случайной величины x_v, вычисляемая по выборке V. Данное утверждение является очевидным, так как большей длине интервала значений x_v соответствует больший коэффициент размытости c_v ядерных функций $\Phi(u_v)$, $v = 1, 2$. Поэтому появляется возможность оптимизацию непараметрического алгоритма распознавания образов (4) проводить лишь по одному параметру c коэффициентов размытости ядерных функций. Подобный подход использовался при построении быстрых процедур оптимизации непараметрических оценок плотностей вероятностей ядерного типа [13–15].

Методика проверки гипотезы о независимости компонент двухмерной случайной величины. Непараметрический алгоритм распознавания образов (4) основан на проверке соотношений между ядерными оценками плотностей вероятностей $\bar{p}(x_1)\bar{p}(x_2)$ и $\bar{p}(x_1, x_2)$. Для ситуаций первого класса Ω_1, в которых вычисляется соотношение $\bar{p}(x_1)\bar{p}(x_2) > \bar{p}(x_1, x_2)$, подтверждается справедливость предположения о независимости случайных величин x_1, x_2. В области определения непараметрической оценки плотности вероятности $\bar{p}(x_1,\ x_2)$ при выполнении соотношения $\bar{p}(x_1)\bar{p}(x_2) < \bar{p}(x_1, x_2)$ следует зависимость случайных величин. Выполнение гипотезы (1) определяет границу в области значений случайных величин x_1, x_2, разделяющую предположения о независимости либо зависимости x_1, x_2. Поэтому методика проверки гипотезы о независимости случайных величин предполагает выполнение ряда действий:

1. Следуя рекомендациям предыдущего раздела, осуществить синтез непараметрического алгоритма распознавания образов (4).

2. Вычислить оценки вероятностей $\bar{\rho}_1$, $\bar{\rho}_2$ ошибок распознавания классов Ω_1, Ω_2 решающего правила (4) по исходным статистическим данным V при оптимальных коэффициентах размытости ядерных статистик $\bar{p}(x_1)\bar{p}(x_2)$, $\bar{p}(x_1, x_2)$.

Значения $\bar{\rho}_t$ вычисляются в режиме «скользящего экзамена» по выборке V в предположении, что её элементы принадлежат к классу Ω_t:

$$\bar{\rho}_t = \frac{1}{n}\sum_{j=1}^{n} 1(\delta(j), \bar{\delta}(j)), \qquad t = 1, 2, \tag{7}$$

где $\delta(j)$ — указания типа $x^j = (x_1^j, x_2^j) \in \Omega_t$, а $\bar{\delta}(j)$ — «решение» алгоритма (4) о принадлежности ситуации x^j к одному из классов Ω_t, $t = 1, 2$.

При вычислении $\bar{\rho}_t$ в соответствии с методикой «скользящего экзамена» ситуация $x^j = (x_1^j, x_2^j)$ из выборки V, которая подаётся на контроль в алгоритм (4), исключается из процесса формирования статистик (2), (3).

Индикаторная функция в формуле (7) определяется выражением

$$1(\delta(j), \bar{\delta}(j)) = \begin{cases} 0, & \text{если } \delta(j) = \bar{\delta}(j); \\ 1, & \text{если } \delta(j) \neq \bar{\delta}(j). \end{cases}$$

3. Сравнить значения $\bar{\rho}_1$, $\bar{\rho}_2$ в предположении, что элементы выборки V принадлежат к классам Ω_1, Ω_2 соответственно. Гипотеза H_0 справедлива, если $\bar{\rho}_1 < \bar{\rho}_2$. В противном случае при $\bar{\rho}_2 < \bar{\rho}_1$ случайные величины x_1 и x_2 являются зависимыми.

При ограниченных объёмах n выборки V возникает задача доверительного оценивания вероятностей ошибок распознавания образов ρ_1, ρ_2. Для её решения используется традиционная методика проверки гипотезы о равенстве вероятностей ρ_1 и ρ_2 [1]. В качестве наблюдаемого значения критерия проверки гипотезы используется статистика

$$U = \frac{|\bar{\rho}_1 - \bar{\rho}_2|}{\sqrt{\dfrac{\bar{\rho}_1 + \bar{\rho}_2}{2}\left(1 - \dfrac{\bar{\rho}_1 + \bar{\rho}_2}{2}\right)\dfrac{2}{n}}}.$$

При уровне значимости 0,05 проверяемая гипотеза выполняется, если $U < 1{,}96$.

Анализ результатов вычислительных экспериментов. Исследуем эффективность предлагаемой методики и критерия Пирсона от объёма n исходных статистических данных при анализе неоднозначных функциональных зависимостей между случайными величинами x_1, x_2.

При формировании выборки $V = (x_1^i, x_2^i,\ i = \overline{1,n})$ использовались следующие функциональные зависимости:

$$x_2' = \varphi_1(x_1) = \pm\sqrt{9 - x_1^2}, \tag{8}$$

$$x_2' = \varphi_2(x_1) = \pm x_1, \tag{9}$$

$$x_2' = \varphi_3(x_1) = \pm 3\sin(x_1), \tag{10}$$

$$x_2' = \varphi_4(x_1) = 0{,}5x_1 \pm 2. \tag{11}$$

Значения x_1 определялись выражением

$$x_1 = -3 + 6\varepsilon_1,$$

где ε_1 — случайная величина с равномерной плотностью вероятности в интервале $[0;1]$.

На значения x_2' накладывалась помеха с нормальным законом распределения $N(0;\sigma)$ при среднем квадратическом отклонении σ. Значения x_2 при использовании функциональных зависимостей (8)–(10) формировались в соответствии с выражением

$$x_2 = \begin{cases} \varphi_j(x_1) + \sigma\Big(\sum\limits_{l=1}^{r}\varepsilon_2^l - 0,5r\Big)\dfrac{6}{\sqrt{3r}}, & \text{если } e \leqslant 0,5; \\ -\varphi_j(x_1) + \sigma\Big(\sum\limits_{l=1}^{r}\varepsilon_2^l - 0,5r\Big)\dfrac{6}{\sqrt{3r}}, & \text{если } e > 0,5, \end{cases} \qquad j = \overline{1,3}. \tag{12}$$

При функциональной зависимости (11) значения x_2 определялись формулой

$$x_2 = \begin{cases} 0,5x_1 + 2 + \sigma\Big(\sum\limits_{l=1}^{r}\varepsilon_2^l - 0,5r\Big)\dfrac{6}{\sqrt{3r}}, & \text{если } e \leqslant 0,5; \\ 0,5x_1 - 2 + \sigma\Big(\sum\limits_{l=1}^{r}\varepsilon_2^l - 0,5r\Big)\dfrac{6}{\sqrt{3r}}, & \text{если } e > 0,5, \end{cases}$$

Здесь ε_2 и e — значения случайных величин с равномерной плотностью вероятности на интервале $[0;1]$, а параметр r принимался равным 12.

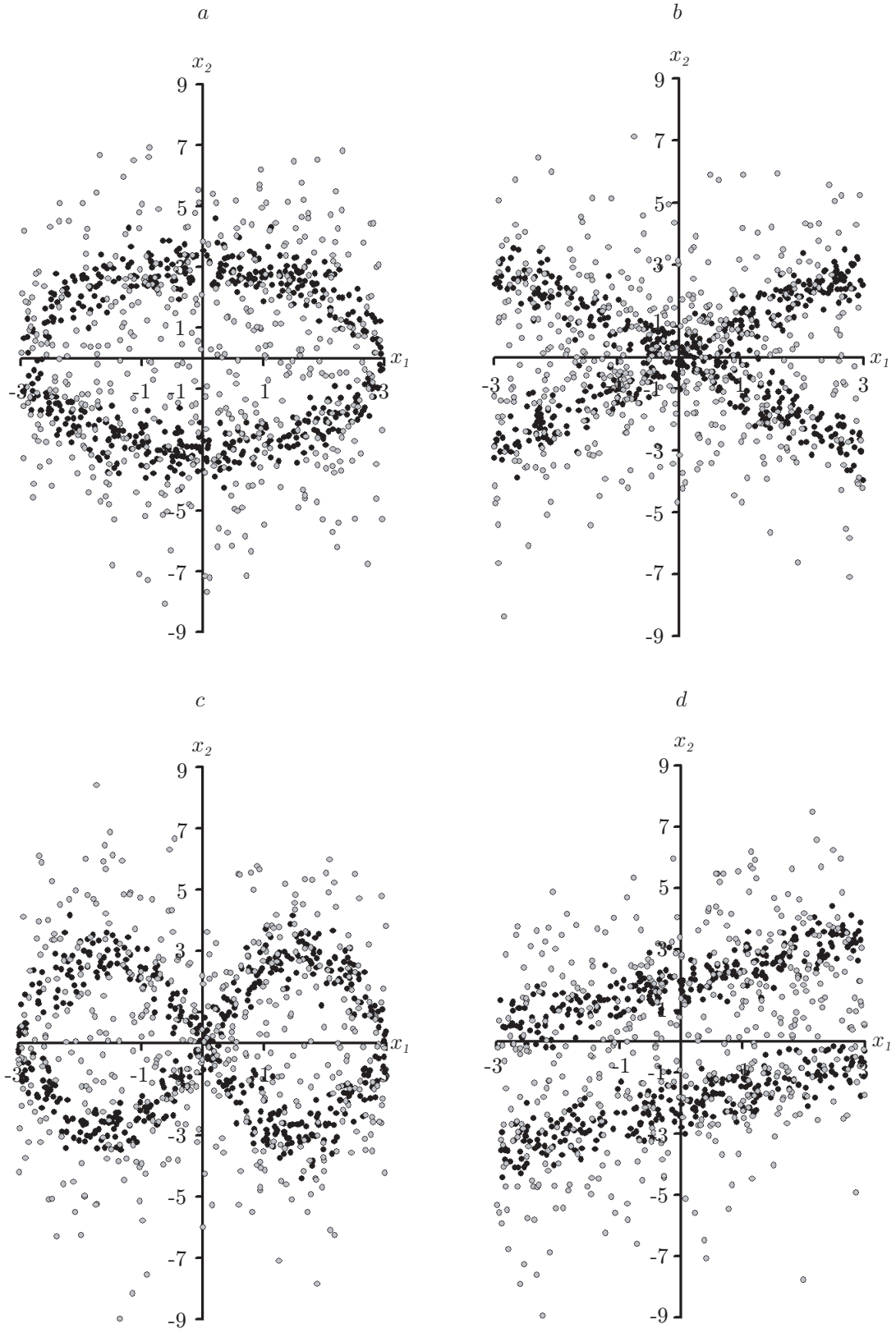
Иллюстрация полученных распределений x_1, x_2 при $n = 500$ для функциональных зависимостей (8)–(11) при различных значениях σ представлена на рисунке a и b.

При проверке гипотезы о независимости компонент двухмерной случайной величины на основе критерия Пирсона используются результаты оптимального выбора количества интервалов дискретизации [16]

$$N^* = ((3/4)\Delta_1\Delta_2\,\|p(x_1,x_2)\|^2\,n)^{\frac{1}{2}}. \tag{13}$$

В (13) значение $\|p(x_1,x_2)\|^2 = \iint p^2(x_1,x_2)\,dx_1\,dx_2$, а Δ_v — интервалы изменения случайной величины x_v, $v = 1,2$.

Результат (13) получен путём исследования асимптотических свойств регрессионной оценки плотности вероятности, синтез которой основывается на дискретизации области её определения для обхода проблемы обработки больших объёмов статистических данных.



Значения случайных величин x_1, x_2 из выборки исходных статистических данных V при $n = 500$ и $\sigma = 0{,}5$ (тёмные точки), а при $\sigma = 2$ (серые точки) с использованием функциональных зависимостей: (8) — a, (9) — b, (10) — c, (11) — d

В формуле (13) оценивание функционала от плотности вероятности $p(x_1, x_2)$ осуществляется по выборке V с использованием статистики

$$\|p(x_1, x_2)\|^2 \approx \iint \bar{p}(x_1, x_2) p(x_1, x_2)\, dx_1\, dx_2 =$$

$$= \frac{1}{n^2 c_1 c_2} \sum_{j=1}^{n} \sum_{\substack{i=1 \\ i \neq j}}^{n} \Phi\Big(\frac{x_1^j - x_1^i}{c_1}\Big) \Phi\Big(\frac{x_2^j - x_2^i}{c_2}\Big). \tag{14}$$

Условие $i \neq j$ в выражении (14) необходимо для обеспечения свойства несмещённости оценки математического ожидания непараметрической оценки плотности вероятности $\bar{p}(x_1, x_2)$.

Реализация теоретического результата (13) предполагает выполнение следующих действий. Вычислить значение $\bar{N} = (N^*)^{1/2}$. Определить количество интервалов дискретизации $\bar{N}_v$ каждой компоненты x_v, $v = 1, 2$, случайной величины $x = (x_1, x_2)$. Здесь $\bar{N}_v$ соответствует целому числу, максимально близкому к значению $\bar{N}$. Общее количество интервалов дискретизации области значений случайной величины x соответствует значению $\bar{\bar{N}} = \bar{N}_1 \bar{N}_2$.

Методика вычислительных экспериментов реализована в среде программирования Delphi-7. Для генерации случайных величин ε_1, ε_2, e с равномерным законом распределения в интервале $(0; 1)$ использовались стандартная функция Random и процедура Randomize, которая учитывает время дня как основу формирования псевдослучайных чисел с равномерным законом распределения.

Объём n исходных статистических данных при организации вычислительных экспериментов определялся значениями 100, 300, 500. При конкретном объёме n исходных данных значения $\bar{\rho}_1$, $\bar{\rho}_2$, U, $\bar{\bar{N}}$, χ^2, порог χ^2 вычисляются 20 раз.

Для удобства анализа результатов вычислительных экспериментов введём ряд обозначений:

— $\bar{P}_1$, $\bar{P}_2$ — частоты достоверного ($|U| > 1{,}96$) принятия решений о независимости либо зависимости случайных величин x_1, x_2 соответственно при использовании предлагаемой методики;

— $\tilde{P}_1$, $\tilde{P}_2$ — частоты выполнения условий $\bar{\rho}_1 \leqslant \bar{\rho}_2$ независимости либо зависимости $\bar{\rho}_1 > \bar{\rho}_2$ случайных величин x_1, x_2, когда значения $\bar{\rho}_1$, $\bar{\rho}_2$ при уровне значимости 0,05 достоверно не отличаются ($|U| < 1{,}96$);

— $\bar{F}_1$, $\bar{F}_2$ — оценки вероятностей принятия решений о независимости либо зависимости случайных величин соответственно на основе критерия Пирсона при уровне значимости 0,05.

В табл. 1–4 в столбце «Количество интервалов $\bar{\bar{N}}$» значения $\bar{\bar{N}}$ определяются в соответствии с формулой (13). Значения элемента столбца, например, 9(1)–16(19) определяют, что в 20 вычислительных экспериментах количество интервалов дискретизации $\bar{\bar{N}} = 9$, $\bar{\bar{N}} = 16$ встречались 1 и 19 раз соответственно.

Предлагаемая методика и критерий Пирсона, который использует процедуру оптимальной дискретизации области значений двухмерной случайной величины, безошибочно определяют зависимость между x_1, x_2 в условиях (8) при $n = 100$ и $\sigma = 0{,}5$ (см. табл. 1).

С увеличением среднего квадратического отклонения σ случайных помех, накладываемых на зависимость (8), эффективность рассматриваемых методик снижается. Например, при $\sigma = 1{,}5$ значение $\bar{P}_1 = 0{,}15$, а $\bar{P}_2 = 0,35$, что подтверждает в основном гипотезу о достоверной зависимости случайных величин при использовании предлагаемой методики по

Таблица 1

Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1, x_2 в условиях (8)

Объём выборки n	Значение σ	Непараметрический подход				Критерий Пирсона		
		$\bar{P}_1$	$\tilde{P}_1$	$\tilde{P}_2$	$\bar{P}_2$	Количество интервалов $\bar{\bar{N}}$	$\bar{F}_1$	$\bar{F}_2$
100	0,5	0	0	0	1	16	0	1
	1	0,15	0	0,15	0,7	9(1)–16(19)	0,2	0,8
	1,5	0,15	0,3	0,2	0,35	9(5)–16(15)	0,6	0,4
	2	0,7	0,15	0,1	0,05	9(9)–16(11)	0,75	0,25
300	0,5	0	0	0	1	25	0	1
	1	0	0	0,05	0,95	25	0	1
	1,5	0,05	0	0,3	0,65	25	0,05	0,95
	2	0,5	0,2	0,2	0,1	25	0,45	0,55
500	0,5	0	0	0	1	36	0	1
	1	0	0	0	1	25(17)–36(3)	0	1
	1,5	0	0,05	0,1	0,85	25(19)–36(1)	0,05	0,95
	2	0,35	0,1	0,35	0,2	25	0,1	0,9

сравнению с критерием Пирсона. В этих условиях критерий Пирсона достоверно принимает решение в пользу независимости случайных величин, так как $\bar{F}_1 = 0,6$, а $\bar{F}_2 = 0,4$. При $\sigma = 2$ оба сравниваемых метода не обнаруживают зависимость между случайными величинами x_1, x_2. С ростом объёма n статистических данных эффективность сравниваемых методов проверки гипотезы о независимости случайных величин повышается. При $n = 300$ и $n = 500$ сравниваемые методы для $\sigma \in [0,5;1]$ достоверно принимают решение о зависимости случайных величин. Для $n = 500$ и $\sigma = 1,5$ они близки по эффективности. Однако при $\sigma = 2$ преимуществом обладает критерий Пирсона. Полученные результаты объясняются особенностями зависимости (8) и большими значениями σ, когда область значений случайных величин скрывает исследуемую зависимость между x_1 и x_2.

Если зависимость между случайными величинами определяется выражением (9), то при $n = 100$, 300, 500 и $\sigma \in [0,5;1]$ сравниваемая методика и критерий Пирсона имеют близкие показатели эффективности (см. табл. 2).

При $n = 500$ критерий Пирсона, который использует предлагаемую процедуру дискретизации области значений двухмерной случайной величины, имеет преимущество. Данный факт можно объяснить проблемой восстановления плотностей вероятностей $p(x_1, x_2)$, $p(x_1)p(x_2)$ по сравнению с оцениванием вероятностей их принадлежности к двухмерным интервалам дискретизации.

В условиях зависимости между случайными величинами (10) при $n = 100$ и $\sigma \in [0,5;2]$ отмечается достоверное преимущество предлагаемой методики по сравнению с критерием Пирсона (см. табл. 3).

При $n = 300$ и $\sigma \in [0,5;1,5]$ эффективность сравниваемых методов сопоставима. Для $\sigma = 2$ оба рассматриваемых метода дают неудовлетворительные результаты. При $n = 500$ и $\sigma \in [0,5;1,5]$ исследуемые методы принимают достоверные решения о зависимости случайных величин x_1, x_2, что является ожидаемым при увеличении объёма n исходных данных. Однако при $\sigma = 2$ преимуществом обладает критерий Пирсона.

В условиях зависимости (11) между x_1 и x_2 при малых значениях $\sigma \in [0,5;1]$ и $n \in [100;500]$ показатели эффективности сравниваемых методов проверки гипотезы о независимости случайных величин сопоставимы (см. табл. 4).

Таблица 2

Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1, x_2 в условиях (9)

Объём выборки n	Значение σ	Непараметрический подход				Критерий Пирсона		
		$\bar{P}_1$	$\tilde{P}_1$	$\tilde{P}_2$	$\bar{P}_2$	Количество интервалов $\bar{\bar{N}}$	$\bar{F}_1$	$\bar{F}_2$
100	0,5	0	0	0	1	16	0	1
	1	0	0	0,05	0,95	16	0,2	0,8
	1,5	0,05	0,25	0,3	0,4	9(1)–16(19)	0,45	0,55
	2	0,15	0,15	0,4	0,3	16	0,65	0,35
300	0,5	0	0	0	1	25	0	1
	1	0	0	0	1	25	0	1
	1,5	0	0	0,1	0,9	25	0	1
	2	0,25	0,15	0,35	0,25	25	0,35	0,65
500	0,5	0	0	0	1	36	0	1
	1	0	0	0	1	25(4)–36(16)	0	1
	1,5	0	0	0	1	25(15)–36(5)	0	1
	2	0,05	0,15	0,2	0,6	25(13)–36(7)	0	1

Таблица 3

Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1, x_2 в условиях (10)

Объём выборки n	Значение σ	Непараметрический подход				Критерий Пирсона		
		$\bar{P}_1$	$\tilde{P}_1$	$\tilde{P}_2$	$\bar{P}_2$	Количество интервалов $\bar{\bar{N}}$	$\bar{F}_1$	$\bar{F}_2$
100	0,5	0	0	0	1	16	0,9	0,1
	1	0,05	0,1	0,1	0,75	9(1)–16(19)	1	0
	1,5	0,35	0,15	0,1	0,4	9(6)–16(14)	0,95	0,05
	2	0,5	0,2	0,2	0,1	9(4)–16(16)	1	0
300	0,5	0	0	0	1	25	0	1
	1	0	0	0	1	25	0	1
	1,5	0	0,15	0,1	0,75	25	0,05	0,95
	2	0,25	0,35	0,35	0,05	25	0,5	0,5
500	0,5	0	0	0	1	36	0	1
	1	0	0	0	1	25(10)–36(10)	0	1
	1,5	0	0,05	0,1	0,85	25(16)–36(4)	0,05	0,95
	2	0,1	0,15	0,5	0,25	25(16)–36(4)	0,2	0,8

Таблица 4

Результаты вычислительных экспериментов при анализе зависимых случайных величин x_1, x_2 в условиях (11)

Объём выборки n	Значение σ	Непараметрический подход				Критерий Пирсона		
		$\bar{P}_1$	$\tilde{P}_1$	$\tilde{P}_2$	$\bar{P}_2$	Количество интервалов $\bar{\bar{N}}$	$\bar{F}_1$	$\bar{F}_2$
100	0,5	0	0	0	1	16	0	1
	1	0	0,1	0,25	0,65	16	0,1	0,9
	1,5	0,2	0,3	0,35	0,15	9(2)–16(18)	0,65	0,35
	2	0,4	0,25	0,2	0,15	9(2)–16(18)	0,6	0,4
300	0,5	0	0	0	1	25	0	1
	1	0	0	0	1	25	0	1
	1,5	0,1	0,2	0,35	0,35	16(1)–25(19)	0,1	0,9
	2	0,2	0,35	0,2	0,25	25	0,05	0,95
500	0,5	0	0	0	1	36	0	1
	1	0	0	0	1	25(18)–36(2)	0	1
	1,5	0	0,1	0,3	0,6	25	0	1
	2	0,05	0,35	0,35	0,25	25(18)–36(2)	0	1

При $\sigma \in [1{,}5; 2]$ преимуществом обладает критерий Пирсона, если при дискретизации области значений x_1, x_2 применяется вышепредложенная процедура.

Заключение. Методика проверки гипотезы о независимости случайных величин основана на использовании непараметрического алгоритма распознавания образов, соответствующего критерию максимального правдоподобия. Её применение позволяет обойти проблему дискретизации области значений случайных величин на многомерные интервалы, что свойственно критерию Пирсона. По результатам вычислительных экспериментов предлагаемая методика и критерий Пирсона при анализе неоднозначных зависимостей между случайными величинами при малых объёмах статистических данных и средних квадратических отклонений σ помех в основном сопоставимы и безошибочно определяют зависимость случайных величин. Данный вывод не соблюдается при зависимости между случайными величинами (10), когда критерий Пирсона не устанавливает зависимость при $n = 100$ и $\sigma \in [0{,}5; 2]$. С увеличением σ эффективность сравниваемых критериев снижается. Этот факт объясняется особенностями неоднозначных зависимостей и большими значениями σ, когда область определения случайных величин скрывает искомую зависимость. С увеличением объёма n исходных данных эффективность сравниваемых критериев проверки гипотезы о независимости случайных величин повышается. Такой вывод является ожидаемым, так как с ростом n повышаются асимптотические свойства непараметрических оценок плотностей вероятностей и частот встречаемости случайных величин в их двухмерных интервалах. Преимущество предлагаемой методики проверки гипотезы о независимости случайных величин наблюдается при малых значениях σ и ограниченных n, а также при больших n и малых значениях σ. При больших n и σ наиболее часто обнаруживается преимущество критерия Пирсона, если соблюдается процедура оптимальной дискретизации области значений двухмерной случайной величины.

Полученные результаты создают основу разработки методики проверки гипотезы о независимости многомерных случайных величин с использованием непараметрического алгоритма распознавания образов. Для этих условий появляется возможность применения при формировании критерия Пирсона процедуры оптимальной дискретизации области значений многомерной случайной величины.

СПИСОК ЛИТЕРАТУРЫ

1. **Пугачёв В. С.** Теория вероятностей и математическая статистика: Учеб. пособие. М.: Физматлит, 2002. 496 с.
2. **Лапко А. В., Лапко В. А.** Проверка гипотезы о независимости двумерных случайных величин с использованием непараметрического алгоритма распознавания образов // Автометрия. 2021. **57**, № 2. С. 41–48. DOI: 10.15372/AUT20210205.
3. **Лапко А. В., Лапко В. А., Бахтина А. В.** Исследование методики проверки гипотезы о независимости двухмерных случайных величин с использованием непараметрического классификатора // Автометрия. 2021. **57**, № 6. С. 90–100. DOI: 10.15372/AUT20210610.
4. **Parzen E.** On estimation of a probability density function and mode // Ann. Math. Statist. 1962. **33**, N 3. P. 1065–1076.
5. **Епанечников В. А.** Непараметрическая оценка многомерной плотности вероятности // Теория вероятности и её применения. 1969. **14**, № 1. С. 156–161.
6. **Rudemo M.** Empirical choice of histogram and kernel density estimators // Scand. Journ. Statist. 1982. N 9. P. 65–78.
7. **Bowman A. W.** A comparative study of some kernel-based non-parametric density estimators // Journ. Statist. Comput. and Simul. 1982. **21**. P. 313–327.
8. **Hall P.** Large-sample optimality of least squares cross-validation in density estimation // Ann. Statist. 1983. **11**. P. 1156–1174.
9. **Jiang M., Provost S. B.** A hybrid bandwidth selection methodology for kernel density estimation // Journ. Statist. Comput. and Simul. 2014. **84**, N 3. P. 614–627. DOI: 10.1080/00949655.2012.721366.
10. **Dutta S.** Cross-validation revisited // Commun. Statist. - Simul. and Comput. 2016. **45**, N 2. P. 472–490. DOI: 10.1080/03610918.2013.862275.
11. **Heidenreich N.-B., Schindler A., Sperlich S.** Bandwidth selection for kernel density estimation: A review of fully automatic selectors // AStA Adv. Statist. Analysis. 2013. **97**, N 4. P. 403–433. DOI 10.1007/s10182-013-0216-y.
12. **Li Q., Racine J. S.** Nonparametric Econometrics: Theory and Practice. Princeton: Princeton University Press, 2007. 768 p.
13. **Scott D. W.** Multivariate Density Estimation: Theory, Practice, and Visualization. New Jersey: John Wiley & Sons, 2015. 384 p.
14. **Sheather S. J.** Density estimation // Statist. Sci. 2004. **19**, N 4. P. 588–597. DOI: 10.1214/088342304000000297.
15. **Silverman B. W.** Density Estimation for Statistics and Data Analysis. London: Chapman and Hall, 1986. 175 p.
16. **Лапко А. В., Лапко В. А.** Оценивание параметров формулы оптимальной дискретизации области значений двумерной случайной величины // Измерительная техника. 2018. № 5. С. 9–13.

Поступила в редакцию 29.11.2022
После доработки 16.01.2023
Принята к публикации 25.05.2023