УДК 519.24; 57.017 DOI 10.21685/2307-4205-2016-4-17

УСИЛЕНИЕ МОЩНОСТИ ХИ-КВАДРАТ КРИТЕРИЯ ПРИ ДЕСЯТИКРАТНОМ УВЕЛИЧЕНИИ ЧИСЛА СТЕПЕНЕЙ СВОБОДЫ СТАТИСТИЧЕСКИХ ВЫЧИСЛЕНИЙ НА МАЛЫХ ТЕСТОВЫХ ВЫБОРКАХ

А. И. Иванов, Б. Б. Ахметов, Ю. И. Серикова

Введение

Информационное общество предполагает активное использование интернет-ресурсов. Государственные и частные структуры создают на своих сайтах личные кабинеты пользователей. К сожалению, существующая практика парольной защиты доступа к личным кабинетам обладает существенными уязвимостями. Пользователи не способны запоминать длинные случайные пароли. Владелец информационного ресурса не может быть уверен в том, что к личному электронному кабинету получил доступ именно его хозяин. Пароль может быть перехвачен программной закладкой, также не составляет проблемы подменить IP адрес интернет-пользователя.

Для усиления защиты доступа к электронным кабинетам в настоящее время разрабатываются технологии биометрической аутентификации личности путем преобразования личных биометрических данных человека в его криптографический ключ или длинный случайный пароль доступа. Используются такие биометрические образы, как: рисунок отпечатка пальца [1], рисунок радужной оболочки глаза [2], голосовой пароль [3], рукописный пароль [4], рисунок кровеносных сосудов глазного дна или ладони руки [5]. Естественно, что преобразователи биометрия-код не могут быть идеальными и имеют вероятности ошибок первого и второго рода. Возникает необходимость тестирования ошибок первого и второго рода на реальных биометрических данных. Кроме того, при настройке «нечетких экстракторов» [1–3] и при обучении нейросетевых преобразователей [4, 5] необходимо контролировать отсутствие в биометрических данных грубых ошибок. По сути дела, на небольшом числе примеров биометрического образа необходимо контролировать показатель близости распределения биометрических данных к многомерному нормальному закону [6]. Формально для этой цели может быть использован обычный одномерный хи-квадрат критерий Пирсона [7, 8], однако такой подход далек от оптимального. В данной статье мы попытаемся показать, что наряду с классическим критерием Пирсона можно использовать три варианта критерия Джини. Один из вариантов критерия Джини оказывается лучше, чем классический критерий хи-квадрат.

Проблема применения классического хи-квадрат критерия Пирсона на малых тестовых выборках

Наиболее популярным на сегодня является использование хи-квадрат критерия (созданного Пирсоном в 1900 г.):

$$\chi^2 = n \cdot \sum_{i=1}^{k} \left\{ \frac{\left(\frac{n_i}{n} - p_i \right)^2}{p_i} \right\} = \sum_{i=1}^{k} \left\{ \frac{\left(\tilde{p}_i - p_i \right)^2}{p_i} \right\}, \quad (1)$$

где n_i – число отсчетов, попавших в i-й столбец гистограммы; p_i – вероятность попадания в i-й столбец гистограммы теоретического распределения; k – число столбцов гистограммы.

Широкое распространение применения хи-квадрат критерия обусловлено тем, что для него известно аналитическое описание плотности распределения:

$$p(\chi^2, m) = \left\{ \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \left\{ x^{\frac{m}{2}-1} \cdot \exp\left(\frac{-x}{2}\right) \right\} \right\}, \quad (2)$$

где Γ(.) – гамма функция; m – число степеней свободы.

Число степеней свободы m может быть задано по-разному [8]. Например, оно может быть определено через объем тестовой выборки n:

$$m = \sqrt{n} - 3 = k - 3, \quad (3)$$

если число столбцов гистограммы k выбирается округлением до ближайшего целого величины $\sqrt{n}$:

$$k = \text{round}\left(\sqrt{n}\right). \quad (4)$$

Заметим, что значение числа столбцов гистограммы k и значение числа степеней свободы m для классического хи-квадрат критерия всегда оказывается много меньшими в сравнении с объемом тестовой выборки n. Так для выборки из 16 примеров необходимо интервал наблюдаемых биометрических данных разбить на четыре интервала и построить по ним гистограмму, как это показано на рис. 1.

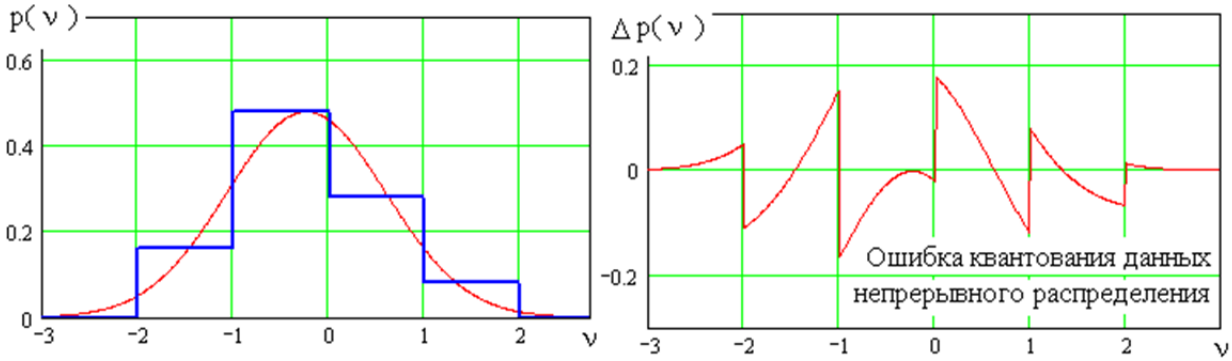

Рис. 1. Приближение непрерывной нормальной плотности распределения значений гистограммы, содержащей четыре интервала $k = 4$ для выборки из 15 примеров

При построении гистограммы мы фактически осуществляем ступенчатое квантование непрерывного закона распределения значений. В связи с этим возникает ошибка квантования, кривая для которой отображена в правой части рис. 1. Получается, что мощность различных статистических критериев во многом определяется тем, как тот или иной критерий подавляет ошибку квантования, возникающую из-за ограниченного объема исходных данных.

В этом отношении критерий хи-квадрат является не самым эффективным. По этой причине в рекомендациях Госстандарта [8] рассматриваются выборки, состоящие из 200 примеров и более. Выборки из 9, 16, 25 примеров считаются слишком малыми для хи-квадрат критерия, так как для них число степеней свободы составит 1, 2, 3. Столь малое число степеней свободы плотности хи-квадрат критерия (2) не дает надежды на приемлемое качество принимаемых решений.

Оценка мощности критерия по равной вероятности ошибок первого и второго рода

Следует отметить, что оценка мощности хи-квадрат критерия во многом остается субъективной. В частности, это связано с тем, что уровень доверительной вероятности принимаемых решений выбирает сам исследователь. Исключим эту неопределенность. Далее будем судить о качестве принимаемых решений по точке равновероятных ошибок первого и второго рода $P_1 = P_2 = P_{EE}$. Этот параметр оказывается работоспособен в ситуации, когда критерий хи-квадрат настроен на нормальный закон распределения значений, а воздействие на него осу-

ществляется как данными с нормальным законом, так и данными с равномерным законом. Подобный численный эксперимент легко реализуем на обычной вычислительной машине. Его результаты отражены на рис. 2.

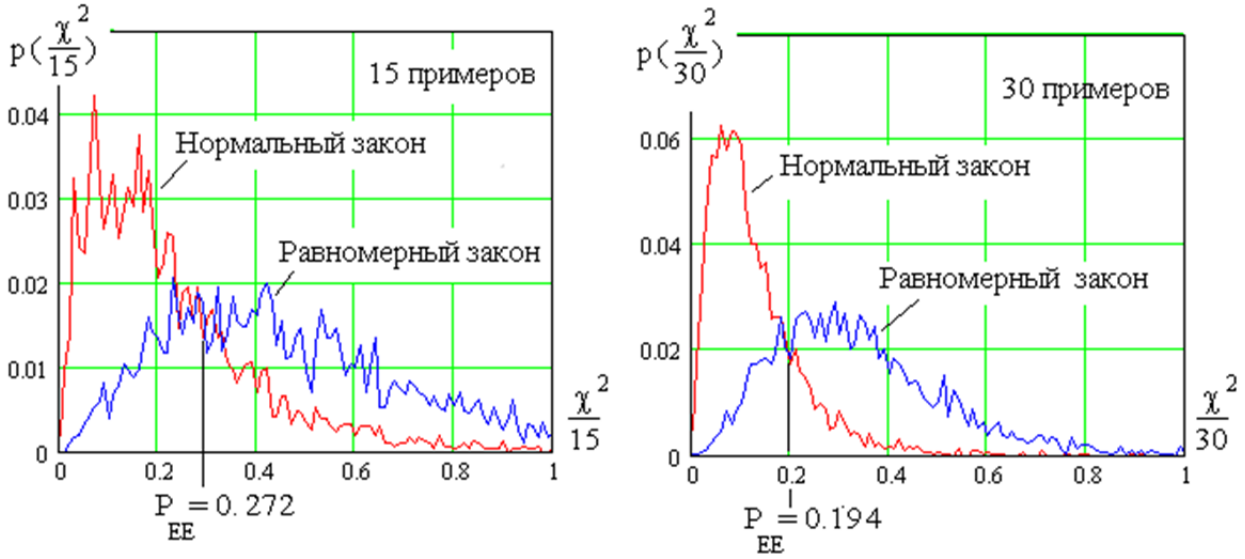

Рис. 2. Результаты численного эксперимента по оценке мощности хи-квадрат критерия для выборок, состоящих из 15 и 30 примеров, при одинаковом числе столбцов гистограммы

Из рис. 2 видно, что для выборок из 15 примеров равновероятная ошибка составляет P_{EE} = 0,272, если же объем тестовой выборки увеличить до 30 примеров, то равновероятная ошибка падает до величины P_{EE} = 0,194. С увеличением объема тестовой выборки в 2 раза происходит снижение примерно в $\sqrt{2}$ раз вероятности появления ошибок.

На практике удобно пользоваться логарифмической шкалой значений равновероятных ошибок. При логарифмическом представлении данных мощность хи-квадрат критерия хорошо описывается ломаными линиями при использовании на каждом участке своего числа примеров в обучающей выборке и своего числа столбцов гистограммы. Для того, что бы уйти от этого эффекта, будем использовать гистограмму, состоящую из шести столбцов для выборки изменяющейся от 5 до 30 примеров. Данные о мощности критерия хи-квадрат отображены в верхней части рис. 3.

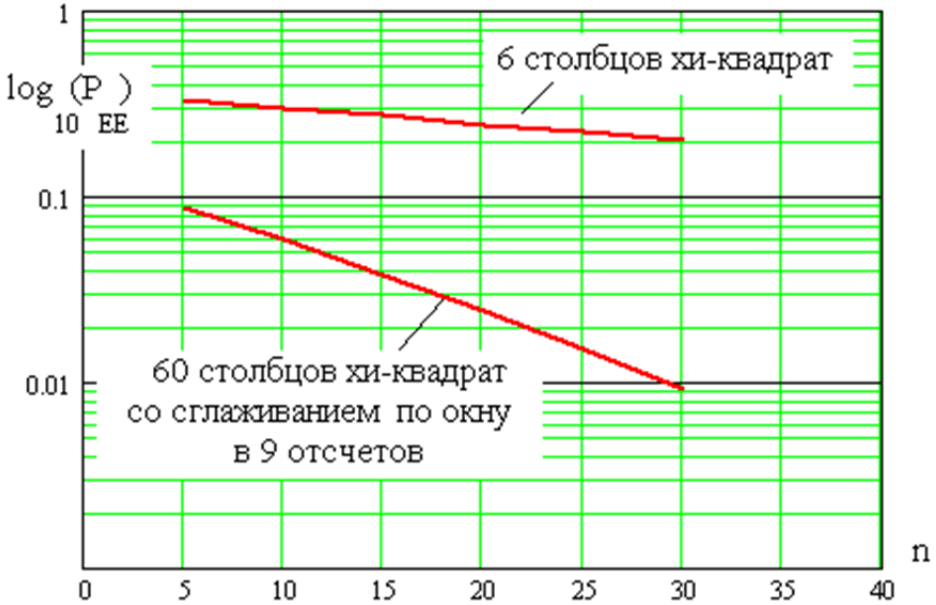

Рис. 3. Мощность хи-квадрат критерия в логарифмической шкале равновероятных ошибок

Из рис. 3 видно, что при одинаковом числе столбцов гистограммы в логарифмическом масштабе происходит линейное уменьшение вероятности ошибок, т.е.

$$\log(P_{EE}(n)) = -0,45 - 0,0086n. \tag{5}$$

При 30 опытах для шести столбцов гистограммы вероятность равновероятной ошибки составляет порядка 0,2, что является слишком большой величиной.

Повышение мощности хи-квадрат критерия Пирсона путем искусственного 10-кратного увеличения столбцов гистограммы

Очевидным является то, что при построении гистограмм реальных данных крайне важным является выбор числа столбцов. Выбор этого параметра во многом субъективен, разные источники дают разные рекомендации. В частности, рекомендации Госстандарта по применению хи-квадрат критерия [8] содержат пять разных правил по выбору числа интервалов гистограммы. Опыт подсказывает, что выбор слишком больших интервалов приводит к большой амплитуде и низкой частоте шумов ошибок квантования. Если же мы примем слишком малые интервалы столбцов гистограммы, то мы получим высокую частоту шума квантования и высокую амплитуду этого шума. На рис. 4 отображена ситуация, когда интервалы гистограммы взяты в 10 раз более узкими, в сравнении с правилом выбора, представленным в выражении (4).

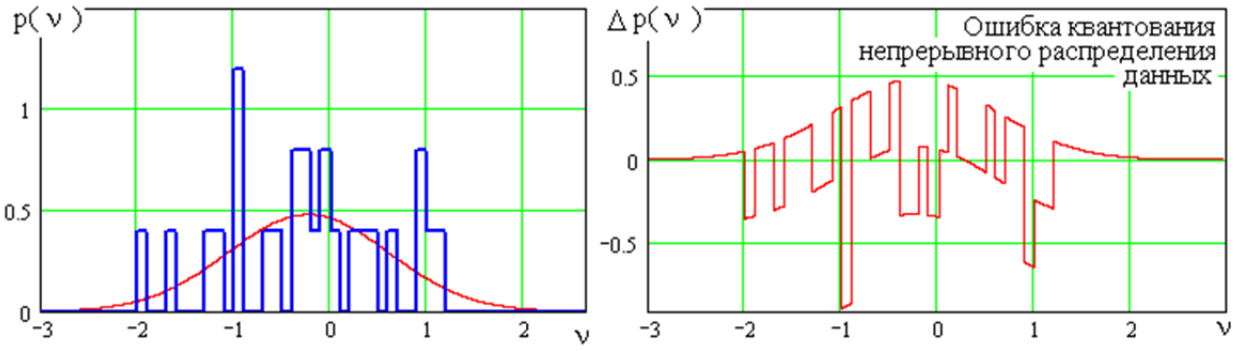

Рис. 4. Гистограмма данных и шум ошибки квантования непрерывного распределения данных при интервалах обработки, взятых в 10 раз уже, чем дает формула (4)

Из левой части рис. 4 видно, что гистограмма имеет очень много пустых столбцов, а ее форма стала не похожа на форму непрерывного распределения данных. Уменьшить амплитуду шума квантования и заполнить пустые интервалы столбцов гистограммы удается, если запустить сглаживающий цифровой фильтр [9–11]. Эта процедура приводит к появлению еще одной модификации хи-кавадрат критерия со сглаживающим шумы квантования фильтром линейным усредняющим фильтром по скальзящему интервалу. Программная реализация такого фильтра для окна сглаживания в девять отсчетов занимает две строки в среде математического моделирования MatCAD:

$$\begin{cases} i := 4,...,(last(g)-4); \\ sg_i := \dfrac{g_{i-4} + g_{i-3} + g_{i-2} + g_{i-1} + g_{i-0} + g_{i+1} + g_{i+2} + g_{i+3} + g_{i+4}}{9}, \end{cases} \tag{6}$$

где g_i – отсчеты гистограммы со слишком узкими столбцами; sg_i – выходные отсчеты сглаживающего данные фильтра. Результаты работы сглаживающего фильтра приведены на рис. 5.

Из рис. 5 видно, что форма восстановленной гистограммы хорошо повторяет форму непрерывного распределения данных, шум ошибок квантования существенно снизился. Все это приводит к значительному росту мощности сглаженного дифференциального критерия Джини. На рис. 3 соответствующая прямая понижения вероятности ошибок расположена ниже других, т.е. мощность сглаженного критерия хи-кадрат оказывается много выше мощности обычного хи-квадрат критерия без сглаживания [12–14].

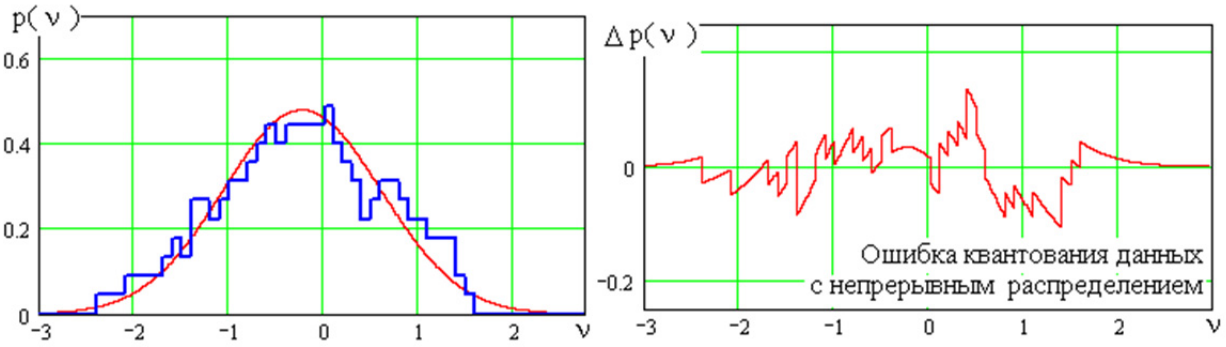

Рис. 5. Восстановленная сглаживанием гистограмма с числом столбцов
в 10 раз больше чем рекомендует формула (4) и соответствующий ей шум квантования

Хорошее приближение мощности сглаженного хи-квадрат критерия дает следующее соотношение:

$$\log(P_{EE}(n)) = -0,82 - 0,039n. \quad (7)$$

Заключение

В силу того, что мы в 10 раз повысили число столбцов гистограммы, как следствие, примерно в 10 раз увеличили и число степеней свободы хи-квадрат критерия. Естественно, что при этом выросла в 10 раз частота шумов квантования, однако при этом многократно выросла и амплитуда шумов квантования. В данной статье показано, что линейного сглаживающего фильтра (6) достаточно для снижения амплитуды шума квантования до приемлемого уровня, обеспечивающего 20-кратный выигрыш по итоговой вероятности принятия ошибочных решений.

Казалась бы, что линейное увеличение частоты квантования и последующее сглаживание линейным фильтром должны друг друга скомпенсировать и не давать ощутимого результата. На самом деле это не так. Каждый статистический критерий является некоторым нелинейным сглаживающим фильтром. Именно по этой причине последовательность выполняемых операций играет важную роль.

Описанные в данной статье операции увеличения числа столбцов гистограммы и последующего линейного сглаживания могут быть эффективны только при включении их в состав нелинейного сглаживающего фильтра (некоторого статистического критерия). Для разных статистических критериев выигрыш будет различен. Все известные статистические критерии следует проверить на их возможное увеличение мощности при искусственном увеличении числа степеней свободы или числа столбцов гистограмм. На данный момент проверенными оказываются только дифференциальный критерий Джини [9–11], для которого выигрыш в мощности составляет примерно 2 раза, и хи-квадрат критерий, обеспечивающий 20-кратный выигрыш.

Список литературы

1. Ramírez-Ruiz, J. Cryptographic Keys Generation Using FingerCodes / J. Ramírez-Ruiz, C. Pfeiffer, J. Nolazco-Flores // Advances in Artificial Intelligence – IBERAMIA-SBIA 2006 (LNCS 4140). – 2006. – P. 178–187.
2. Cryptographic key generation from voice / F. Monrose, M. Reiter, Q. Li, S. Wetzel // Proc. IEEE Symp. on Security and Privacy, 2001. – P. 17–19.
3. Feng, Hao. Crypto with Biometrics Effectively / Feng Hao, Ross Anderson and John Daugman // IEEE TRANSACTIONS ON COMPUTERS. – 2006. – Vol. 55, № 9. – September.
4. Нейросетевая защита персональных биометрических данных / Ю. К. Язов, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков, И. Г. Назаров. – М. : Радиотехника, 2012. – 157 с.
5. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа : моногр. / Б. С. Ахметов, А. И. Иванов, В. А.Фунтиков, А. В. Безяев, Е. А. Малыгина. – Казахстан : Изд-во LEM, 2014. – 144 с.
6. Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации / А. Ю. Малыгин, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза : Изд-во Пенз. гос. ун-та, 2006. – 161 с.

7. Кобзарь, А. И. Прикладная математическая статистика для инженеров и научных работников / А. И. Кобзарь. – М. : ФИЗМАТЛИТ, 2006. – 816 с.
8. Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа χ2. – М. : Госстандарт России, 2001. – 140 с.
9. Серикова, Н. И. Линейное сглаживание гистограмм биометрических данных, искусственно увеличивающее число степеней свободы при оценивании статистических гипотез // Труды научно-технической конференции кластера пензенских предприятий, обеспечивающих безопасность информационных технологий. – Пенза, 2014. – Т. 9. – С. 29–31. – URL: http://www.pniei.penza.ru/RV-conf/T9/C29
10. Серикова, Н. И. Биометрическая статистика: сглаживание гистограмм, построенных на малой обучающей выборке / Н. И. Серикова, А. И. Иванов, С. В. Качалин // Вестник Сибирского государственного аэрокосмического университета имени академика М. Ф. Решетнева. – 2014. – № 3 (55). – С. 146–150.
11. Серикова, Н. И. Оценка правдоподобия гипотезы о нормальном распределении по критерию Джини для сглаженных гистограмм, построенных на малых тестовых выборках / Н. И.Серикова, А. И. Иванов, Ю. И. Серикова // Вопросы радиоэлектроники. – М. : ЦНИИ «Электроника», 2015. – Вып. 1. – С. 85–94.
12. Использование среднего геометрического, ожидаемой и наблюдаемой функций вероятности как статистического критерия оценки качества биометрических данных / Б. С. Ахметов, А. И. Иванов, К. А. Перфилов, Е. Д. Проценко, Д. С. Пащенко // Труды Международного симпозиума Надежность и качество. – 2015. – Т. 2. – С. 281–283.
13. Быстрый алгоритм оценки высокоразмерной энтропии биометрических образов на малых выборках / Б. Б. Ахметов, А. И. Иванов, А. Ю. Малыгин, А. В. Безяев, А. И. Газин // Труды Международного симпозиума Надежность и качество. – 2015. – Т. 2. – С. 283–285.
14. Использование множества подобных критериев для случайного выбора контролируемых параметров при многомерном статистическом анализе малой выборки биометрических данных / Б. С. Ахметов, К. Мукапил, Н. И. Серикова, С. Е. Вятчанин, Ю. И. Никитченко // Труды Международного симпозиума Надежность и качество. – 2015. – Т. 2. – С. 285–288.

Иванов Александр Иванович
доктор технических наук, доцент,
начальник лаборатории биометрических
и нейросетевых технологий,
Пензенский научно-исследовательский
электротехнический институт
(440000, Россия, г. Пенза, ул. Советская, 9)
E-mail: ivan@pniei.penza.ru

Ахметов Берик Бахытжанович
кандидат технических наук, профессор,
вице-президент Международного
Казахско-Турецкого университета
имени А. Ясави
(Казахстан, г. Туркестан, ул. Б. Саттарханов, 29)
E-mail: berik.akhmetov@ayu.edu.kz

Серикова Юлия Игоревна
магистрант,
Пензенский государственный университет
(440026, Россия, г. Пенза, ул. Красная, 40)
E-mail: julia-ska@yandex.ru

Ivanov Aleksandr Ivanovich
doctor of technical sciences, associate professor,
head of laboratory of biometric
and neural-network technologies,
Penza Research Electrotechnical Institute
(440000, 9 Sovetskaya street, Penza, Russia)

Akhmetov Berik Bakhytzhanovich
candidate of technical sciences, professor,
vice president of International Kazakh-Turkish
University named after A. Yasavi
(29 B. Sattarhanov street, Turkestan, Kazakhstan)

Serikova Yuliya Igorevna
master degree student,
Penza State University
(440026, 40 Krasnaya street, Penza, Russia)

Аннотация. *Актуальность и цели*. При статистической обработке реальных данных химии, экономики, биометрии, медицины обычно приходится пользоваться ограниченными тестовыми выборками. Целью данной работы является повышение мощности хи-квадрат критерия за счет искусственного увеличения числа степеней свободы при статистических вычислениях. *Материалы и методы*. Предложено

Abstract. *Background*. Statistical processing of real data of chemistry, economics, biometrics, medicine is usually necessary to use the limited test samples. The aim of this work is to increase the power of the chi-square test due to an artificial increase in the number of degrees of freedom in statistical calculations. *Materials and methods*. It is proposed to increase by 10 times the number of histogram intervals, which leads to the emergence of a

увеличить в 10 раз число интервалов гистограммы, что приводит к появлению большого числа пустых интервалов. Пустые интервалы заполняются линейным сглаживающим фильтром с прямоугольным окном усреднения без фазовых искажений. Ширина окна сглаживающего фильтра выбрана равной девяти отсчетам. *Результаты и выводы*. Показано, что искусственное повышение числа степеней свободы у критерия Пирсона приводит к росту его мощности. Даны оценки повышения мощности критерия для 5, 6, …, 30 примеров в тестовой выборке. Мощность критерия оценивается как отрицательный логарифм равновероятных ошибок первого и второго рода при нормальном и равномерном распределениях тестируемых данных. Отмечено снижение вероятности ошибок до 20 раз на выборке в 30 примеров за счет реализации предложенного в работе алгоритма статистической обработки.

Ключевые слова: хи-квадрат критерий, малые выборки, искусственное увеличение числа степеней свободы.

large number of empty slots. Empty slots are filled with a linear smoothing filter with a rectangular window averaging without phase distortion. The width of the smoothing filter window is selected to be 9 counts. *Results and conclusions*. It is shown that an artificial increase in the number of degrees of freedom at Pearson leads to the growth of its power. Estimations criterion for increasing the power of 5, 6, ..., 30 in a test sample examples. Power criteria is assessed as negative logarithm of the equally probable errors of the first and second kind in the normal and uniform distribution of the test data. A decrease in the probability of error of up to 20 times on a sample of 30 examples by implementing proposed in statistical processing algorithm.

Key words: chi-squared test, a small sample, the artificial increase in the number of degrees of freedom.

УДК 519.24; 57.017

Иванов, А. И.

Усиление мощности хи-квадрат критерия при десяти кратном увеличении числа степеней свободы статистических вычислений на малых тестовых выборках / А. И. Иванов, Б. Б. Ахметов, Ю. И. Серикова // Надежность и качество сложных систем. – 2016. – № 4 (16). – С. 121–127. DOI 10.21685/2307-4205-2016-4-17.