BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 1, Number 353 (2015), 5 – 11

MULTIDIMENSIONAL STATISTICAL ANALYSIS OF BIOMETRIC DATA BY NETWORK OF PRIVATE PEARSON'S CRITERIA

B. B. Akhmetov[1], A. I. Ivanov[2], A. V. Bezyaev[3], Yu. V. Funtikova[2]

[1]International Kazakh-Turkish University named after Kh.A. Yassavi, Turkestan, Kazakhstan,
[2]Penza scientific-research electrotechnical institute, Penza, Russia,
[3]Penza branch of FSUE "STC "Atlas", Russia
E-mail: berik.akhmetov@iktu.kz; ivan@pniei.penza.ru

Key words: multidimensional statistical analysis, network of private Pearson's criteria, symmetrization of a task, accounting of correlation relationships of values of the studied data.

Abstract. It is shown that upon transition to use multidimensional statistical processing it is managed to receive decisions with very high reliability on small test selections. Influence on quality of the solution of correlation relationships of basic biometric data is investigated. It is proved that growth of correlation of data and growth of dimension of their processing compensate each other. Correspondingly the accounting of correlation of data can be carried out through equivalent decrease in dimension. There is a formula allowing to estimate necessary decrease of dimension for data with equal correlation. Values of thresholds of the output quantizer of a network of the private Pearson's criteria, providing equal probabilities of errors of the first and second sort, when checking hypotheses of normal or uniform distribution of values of the studied data, are given.

УДК 681.32 2

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ БИОМЕТРИЧЕСКИХ ДАННЫХ СЕТЬЮ ЧАСТНЫХ КРИТЕРИЕВ ПИРСОНА

Б. Б.Ахметов[1], А. И. Иванов[2], А. В. Безяев[3], Ю. В. Фунтикова[2]

[1]Международный Казахско-Турецкий университет им. Х. А. Ясави, Туркестан, Казахстан,
[2]Пензенский научно-исследовательский электротехнический институт, Россия,
[3]Пензенский филиал ФГУП «НТЦ «Атлас», Россия

Ключевые слова: многомерный статистический анализ, сеть частных критериев Пирсона, симметризация задачи, учет влияния корреляционных связей.

Аннотация. Показано, что при переходе к использованию многомерной статистической обработке удается получать решения с очень высокой достоверностью на малых тестовых выборках. Исследовано влияние на качество решения корреляционных связей исходных биометрических данных. Доказывается, что рост коррелированности данных и рост размерности их обработки компенсируют друг друга. Соответственно учет корреляции данных можно осуществлять через эквивалентное снижение размерности. Дана формула, позволяющая оценить необходимое понижение размерности для данных с равной коррелированностью. Даны значения порогов выходного квантователя сети частных критериев Пирсона, обеспечивающих равные вероятности ошибок первого и второго рода при проверке гипотез о нормальном или равномерном распределении значений исследуемых данных.

Введение. Одним из наиболее популярных при статистическом анализе данных является критерий Пирсона. В частности, только хи-квадрат критерию Пирсона полностью посвящена первая часть рекомендаций Госстандарта [1], тогда как все остальные критерии описаны во второй части рекомендаций [2]. Подробное описание критерия Пирсона в первой части рекомендаций Госстандарта [1], отражает факт высокой востребованности именно этого критерия промышленностью. Большинство методик статистического анализа экспериментальных данных построены на использовании хи-квадрат критерия:

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

где b_i – число опытов, попавших i-тый интервал гистограммы, $\tilde{p}_i$ – ожидаемая теоретическая вероятность попадания в i-тый интервал гистограммы, n – число опытов в тестовой выборке, k – число столбцов гистограммы.

Популярность использования хи-квадрат критерия Пирсона в промышленности во многом обусловлена тем, что при $n \to \infty$ его распределение описывается через гамма функцию с m = k-1 числом степеней свободы:

$$p_{\chi^2}(n = \infty, m = k - 1, x) = \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot x^{\frac{m}{2}-1} \cdot e^{\frac{-x}{2}}. \qquad (2)$$

Аналитическое описание (2) получено Пирсоном в 1904 году и играло крайне важную роль в первой половине 20-го века, когда вычислительные возможности, используемые при статистической обработке данных, были весьма и весьма ограниченными.

К сожалению, традиционное применение хи-квадрат критерия для многомерных зависимых биометрических данных дает неудовлетворительные результаты. В частности, для принятия решений с уровнем доверия 0.99 приходится использовать выборку, состоящую из 400 результатов испытаний. Применение столь больших тестовых выборок недопустимо для биометрии, необходимо добиться их снижения примерно на порядок.

Данная статья посвящена результатам численного моделирования не- классического применения критерия хи-квадрат Пирсона для многомерного анализа выборок, состоящих из малого числа примеров. То, что Пирсон не мог сделать 110 лет назад из-за отсутствия компьютеров, сегодня технически выполнимо. Сегодня повторить эксперимент на компьютере 1 000 000 раз вполне возможно. При этом обнаруживаются устойчивые взаимосвязи, описанию которых и посвящена данная статья.

Оценка мощности одномерного хи-квадрат критерия статистической проверки правдоподобия гипотезы нормального закона распределения. При организации численного эксперимента будем исходить из того, что должны проверяться две статистические гипотезы. Первая гипотеза состоит в том, что данные тестовой выборки имеют нормальный закон распределения значений. Вторая гипотеза состоит в том, что данные этой же выборки могут иметь нормальный закон распределения значений. Как следствие, при организации численного эксперимента необходимо использовать два программных генератора псевдослучайных данных, как это показано на блок-схеме рисунка 1.

Каждый из генераторов случайных данных $Г_1$ (нормальные данные) и $Г_2$ (данные с равномерным законом распределения) случайным образом подаются на вход вычислителя значения хи-квадрат критерия (1). Далее значения хи-квадрат критерия должны сравниваться с некоторым порогом квантователя. Если значение хи-квадрат менее порога, то принимается решение о нормальности исследуемых входных данных. Если значение хи-квадрат критерия (1) оказывается выше порога, то принимается решение о наиболее вероятном равномерном законе распределения значений. На рисунке 2 приведены кривые гистограмм распределения значений хи-квадрат критерия данных, полученных от двух программных генераторов.

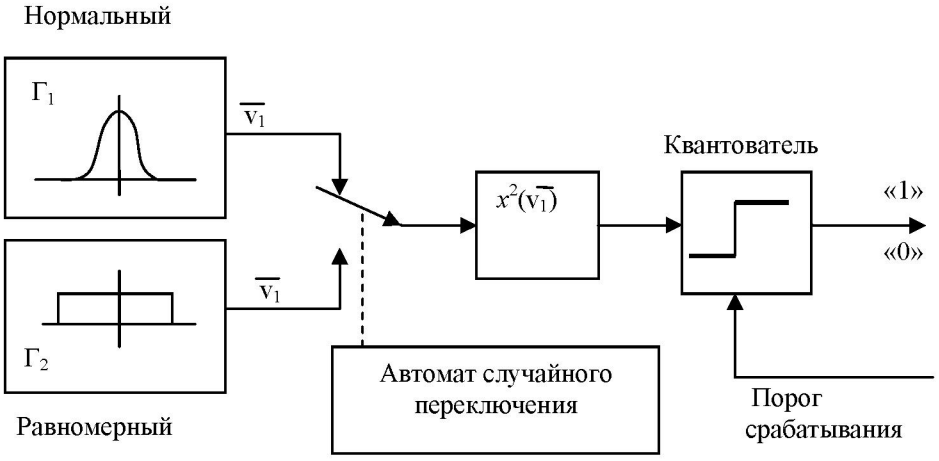

Рисунок 1 – Блок-схема организации численного эксперимента по оценке мощности одномерного критерия хи-квадрат

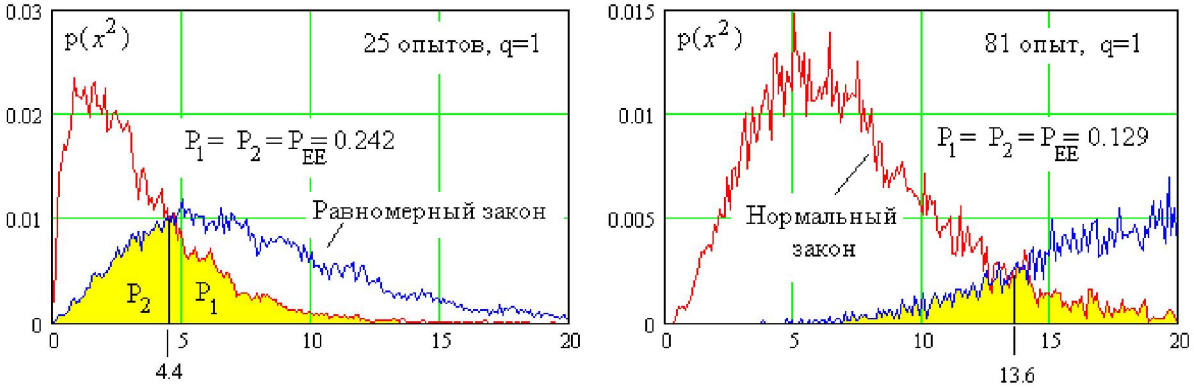

Рисунок 2 – Гистограммы распределения значений одномерного хи-квадрат критерия при проверке гипотезы нормальности и гипотезы равномерности входных данных

Результаты численного моделирования нуждаются в последующем анализе, который усложнен наличием двух видов ошибок. Возникает ошибка первого рода (ложное отклонение верной гипотезы) с вероятностью P_1 и возникает ошибка второго рода (ложное принятие неверной гипотезы) с вероятностью P_2. Анализировать вероятности ошибок первого и второго рода сложно. В связи с этим упростим задачу через ее симметризацию и будем далее рассматривать только равные вероятности ошибок первого и второго рода $P_{EE}=P_1=P_2$. На рисунке 2 отмечены заливкой равные значения вероятностей ошибок первого и второго рода. Из этого рисунка видно, что при увеличении числа опытов, увеличивается число столбцов гистограммы и падает равновероятная ошибка первого и второго рода.

Так, при 25 опытах используется $k=\sqrt{n}=5$ столбцов гистограммы, что обеспечивает равную вероятность ошибок 0.242 (см. левую часть рисунка 2). Однако уже при 81 опыте могут быть использованы 9 столбцов гистограммы, что обеспечивает равные вероятности появления ошибок первого и второго рода на уровне 0.129. Увеличение размеров тестовой выборки в 3 раз приводит к снижению вероятности ошибок в 2 раза. Наблюдается нелинейная зависимость, число опытов в тестовой выборке растет много быстрее в сравнении с падением соответствующей вероятности ошибок P_{EE}. Связь между собой этих двух величин отражена в таблице 1.

Многомерный статистический анализ сложением частных критериев хи-квадрат. Следует отметить, что биометрические данные многомерны. В частности, нейросетевой преобразователь биометрия-код свободно распространяемой среды моделирования «БиоНейроАвтограф» [3], преобразует 416 биометрических параметров в код личного ключа длинной 256 бит. То есть мы

имеем возможность анализировать не один, а 416 биометрических параметров. Если мы имеем выборку из 16 примеров, то у нас появляется возможность анализировать 16×416=6656 отсчетов. Появляется реальная возможность увеличить объем обрабатываемых данных и тем самым поднять достоверность принимаемых решений.

Для многомерной обработки воспользуемся сложением частных критериев Пирсона:

$$\chi^2(\nu_1, \nu_2, ..., \nu_q) = \frac{\chi^2(\nu_1) + \chi^2(\nu_2) + ... + \chi^2(\nu_q))}{q} \quad . \tag{3}$$

Преобразование (3) эквивалентно использованию сети частных критериев Пирсона, структура сети Пирсона приведена на рисунке 3. Сеть частных хи-квадрат критериев Пирсона имеет входные и выходные нелинейные преобразования при линейном суммированием данных между ними (преобразование в соответствии с моделью Гаммерштейна-Винера).

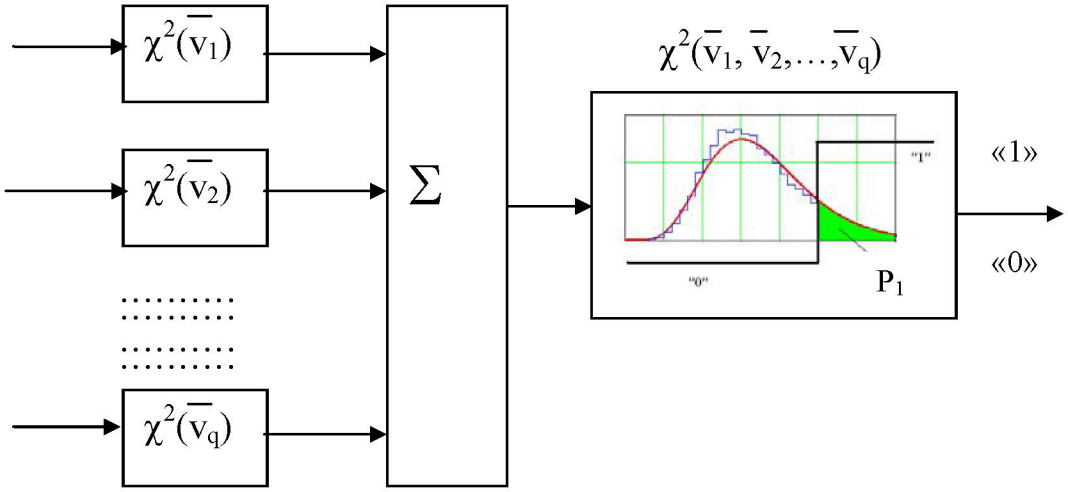

Рисунок 3 – Многомерная статистическая обработка данных сетью Пирсона

При переходе к моделированию сети частных критериев хи-квадрат Пирсона достаточно вместо двух программных генераторов псевдослучайных чисел использовать q пар генераторов. При малой размерности входных данных q ≤ 16 особых сложностей в программировании численного эксперимента не возникает. В таблице приведены значения равных вероятностей ошибок первого и второго рода для разных значений порогов выходного квантователя сетей Пирсона, а также для разной входной размерности.

Вероятности ошибок P_{EE} для разных значений порогов и разных значений показателя входной размерности – q

Число опытов n		9	16	25	36	49	64	81	100	121
Число столбцов гистограммы k		3	4	5	6	7	8	9	10	11
Значения вероятностей $P_{EE} = P_1 = P_2$										
Размерность задачи	q =1	0.42	0.32	0.24	0.22	0.16	0.14	0.13	0.12	0.09
	q =2	0.389	0.262	0.169	0.109	0.08	0.04	0.028	0.023	0.021
	q =3	0.355	0.216	0.119	0.068	0.032	0.024	0.013	0.009	0.006
	q =4	0.332	0.187	0.089	0.054	0.019	0.010	0.006	0.004	0.003
	q =5	0.304	0.154	0.061	0.027	0.012	0.006	0.004	0.002	0.001
Значения порогов для обеспечения вероятностей $P_{EE} = P_1 = P_2$										
Пороги квантования	q =1	2.1	3.1	4.4	5.7	8.3	11.1	13.6	17.1	19.2
	q =2	2.2	3.2	4.8	6.8	9.1	11.5	14.4	17.9	20.1
	q =3	2.2	3.2	5.0	6.9	9.2	11.6	14.3	17.8	20.2
	q =4	2.1	3.2	4.9	6.9	9.2	11.5	14.5	17.8	20.1
	q =5	2.1	3.2	4.9	6.9	9.2	11.6	14.4	17.9	20.1

Следует обратить внимание на то, что значения порога равной вероятности ошибок остается практически одним и тем же для всех показателей размерности таблицы №1. Это крайне интересный факт, свидетельствующий о значительном упрощении задачи из-за корректно выполненной симметризации. Иллюстрацией этой ситуации является рисунок 4, где приведены распределения расстояний на выходе трехмерной сети Пирсона для нормального и равномерного законов распределения значений.

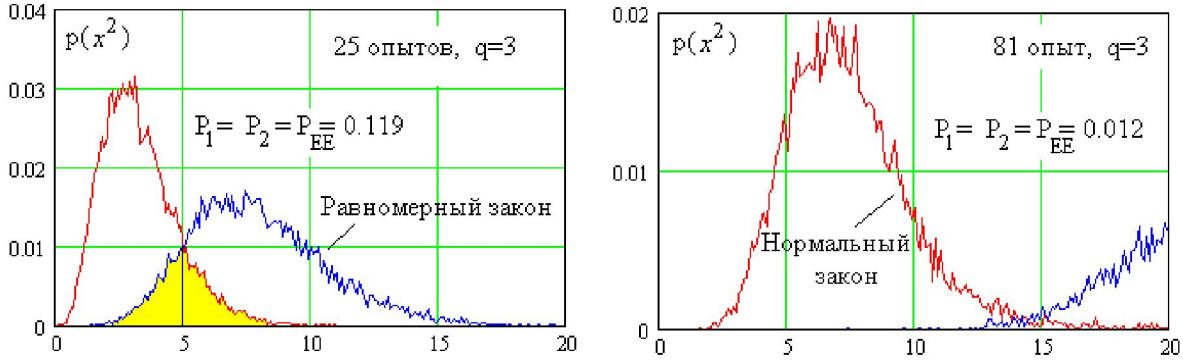

Рисунок 4 – Распределения выходных данных трехмерной сети Пирсона для 25 и 81 опытов

Если сравнивать рисунок 2 и рисунок 4, легко выявить эффект роста линейной разделимости, рассматриваемых распределений значений по мере увеличения числа опытов в обучающей выборке и по мере роста размерности сети частных критериев Пирсона. Это означает, что, увеличивая размерность статистической обработки, мы можем существенно снизить требования к размерам обучающей выборки. Так, при одномерной обработке для получения P_{EE}=0.1 требуется использовать тестовую выборку, состоящую из 112 опытов. Если же мы воспользуемся двухмерной статистической обработкой данных, то для такой же вероятности ошибок P_{EE}=0.1 потребуется выборка из 41 опыта. Наблюдается практически двухкратное снижение требований к размерам обучающей выборки.

Аналитическое описание хи-квадрат распределений для конечных выборок при проверке гипотезы нормального закона распределения значений. Важным свойством хи-квадрат распределений является то, что они имеют точное аналитическое описание не только для выборок бесконечного объема n=∞. Результаты проведения численных экспериментов показали, что для n= 9, 16, 25, 36, 49, … плотность хи-квадрат распределения Пирсона описывается через гамма функцию с целыми показателями числа степеней свободы. В частности, для конечной выборки из 16 опытов (гистограмм из 4 столбцов) плотность распределения будет описываться следующим соотношением:

$$p_{\chi^2}(q=1, n=16, \quad m=3, x) = \frac{1}{2\cdot 2^{\frac{3}{2}}\cdot\Gamma\left(\frac{3}{2}\right)}\cdot(2x)^{\frac{3}{2}-1}\cdot e^{\frac{-2x}{2}} . \tag{4}$$

Для двухмерного хи-квадрат критерия Пирсона плотность распределения будет описываться аналогичным соотношением:

$$p_{\chi^2}(q=2, n=16, \quad m=5, x) = \frac{1}{3\cdot 2^{\frac{5}{2}}\cdot\Gamma\left(\frac{5}{2}\right)}\cdot(3x)^{\frac{5}{2}-1}\cdot e^{\frac{-3x}{2}} . \tag{5}$$

Повышая размерность входных данных, индукцией удается получить следующее описание хи-квадрат распределения для произвольного значения – q:

$$p_{\chi^2}(q, n=16, x) = \frac{1}{(q+2)\cdot 2^{\frac{(2q+2)}{2}}\cdot\Gamma\left(\frac{2q+2}{2}\right)}\cdot((q+2)\cdot x)^{\frac{2q+2}{2}-1}\cdot e^{\frac{-(q+2)x}{2}} . \quad (5)$$

Получается, что мы можем воспользоваться аналитическими соотношениями вида (5) для того, чтобы построить таблицы квантилей доверительной вероятности для многомерного хи-квадрат распределения Пирсона любой размерности – q. По крайней мере это может быть сделано для конечной выборки, состоящей из 16 опытов. Предположительно, что аналогичные аналитические соотношения могут быть получены и для других тестовых выборок с числом опытов точно совпадающие с квадратом числа столбцов гистограммы. Во всех иных случаях хи-квадрат распределения не могут быть точно описаны целыми показателями числа степеней свободы. Для их описания должны использоваться дробные (фрактальные) показатели числа степеней свободы.

Понижение размерности как способ учесть корреляционные связи биометрических данных. Одной из проблем многомерного обобщения классического хи-квадрат критерия является то, что его таблицы квантилей достоверности легко строятся для независимых (не коррелированных данных). В биометрии все данные обладают существенными зависимостями, эти зависимости следует учитывать.

Если речь идет о действительно многомерной статистической обработке биометрических данных, то вычислять многомерную корреляционную матрицу нет смысла. С ростом размерности решаемой задачи все более и более важным является ее симметризация. В биометрии принято [4, 5] осуществлять симметризацию влияния корреляционных связей через вычисление математического ожидания модулей парных коэффициентов корреляции:

$$R = \frac{1}{N}\sum_{i=1}^{N}\left|r(\nu_k, \nu_j)\right| \text{ при } \begin{cases} k = rnd(N), \\ j = rnd(N), \\ k \neq j \end{cases} . \quad (6)$$

При оценках (6) обычно используется несколько сотен пар случайно выбранных биометрических параметров. При любой размерности задачи удается заменить реальные данные их некоторым эквивалентом, имеющим одинаковые корреляционные связи между всеми учитываемыми параметрами. Такое упрощение задачи позволяет легко оценить эквивалентный показатель размерности:

$$\tilde{q} \approx (q-1)\cdot(1-R^2)+1 . \quad (7)$$

Чем выше корреляция данных, тем меньше оказывается эквивалентная размерность, обработки статистических данных. Рост размерности обработки и рост коррелированности данных работают в противоположных направлениях. Всегда можно понизить размерность задачи, компенсируя тем самым влияние коррелированности данных. То есть, пользуясь преобразованиями вида (7), вполне возможно свести задачу применения многомерных сетей Пирсона к зависимым данным к более простой задаче обработки независимых данных. Соответствующие таблицы пересчета уже созданы для искусственных нейронных сетей [4]. Как следствие, аналогичные таблицы преобразований могут быть построены и для сетей частных критериев Пирсона.

Заключение. Как правило, отраслевые методики статистической оценки гипотез по критерию хи-квадрат предполагают использование тестовых выборок из данных о нескольких сотен опытов. Итоговый результат получается точным, однако во многих случаях получить столь большие объемы данных нельзя. В медицине и биометрии считаются достаточными выборки из 20 примеров. Как правило, все примеры биометрии и медицины многомерны. Переход к многомерной статистической обработке предположительно должен значительно снизить требования к размерам тестовых выборок. В биометрии существует огромный потенциал повышения достоверность, принимаемых статистических решений из-за наличия 400 и более контролируемых биометрических параметров с показателем коррелированности R≈0.3. Необходимо в ближайшее время рассчитать таблицы взаимной компенсации показателя коррелированности данных и их размерности.

ЛИТЕРАТУРА

[1] Р 50.1.037-2002 Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. – Ч. I. Критерии типа χ^2. Госстандарт России. – М., 2001. – 140 с.

[2] Р 50.1.037-2002 Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. – Ч. II. Непараметрические критерии. Госстандарт России. – М., 2002. – 123 с.

[3] «БиоНейроАвтограф» – среда моделирования больших искусственных нейронных сетей, преобразующих данные рукописных образов в код пароля. Среда создана лабораторий биометрических и нейросетевых технологий ОАО «Пензенский научно-исследовательский электротехнический институт» в 2009–2014 гг. для свободного использования университетами России, Казахстана и Белоруссии, архивы с исполняемыми файлами размещены в свободном доступе: http://пниэи.рф/activity/science/noc.htm.

[4] Ахметов Б.С., Надеев Д.Н., Фунтиков В.А., Иванов А.И., Малыгин А.Ю. Оценка рисков высоконадежной биометрии. Монография. – Алматы: Из-во КазНТУ им. К.И. Сатпаева, 2014. – 108 с.

[5] Ахметов Б.С., Волчихин В.И., Иванов А.И., Малыгин А.Ю. Алгоритмы тестирования биометрико-нейросетевых механизмов защиты информации Казахстан. – Алматы: КазНТУ им. Сатпаева, 2013. – 152 с. ISBN 978-101-228-586-4, http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf

REFERENCES

[1] R 50.1.037-2002 Recommendations for standardization. Applied statistics. Validation rules of the consent of an experienced distribution with the theoretical. Ch. I. $\chi 2$ type criteria. GosStandart of Russia. – M., 2001.– 140 c. (in Russ.).

[2] P 50.1.037-2002 Applied statistics. Validation rules of the consent of an experienced distribution with the theoretical. – Ch. II. Non-parametric test. GosStandart of Russia. – M., 2002. – 123 c. (in Russ.).

[3] «BioNejroAvtograf» – environment of simulation of large artificial neural networks that convert handwritten image data in a password code. Environment is established by laboratories of biometric and neural network technology of JSC "Penza scientific-research electrotechnical institute" in 2009-2014. for free use by universities of Russia, Kazakhstan and Belarus, archives with executable files are placed in the public domain: http://пниэи.рф/activity/science/noc.htm. (in Russ.).

[4] Akhmetov B.S., Nadeev D.N., Funtikov V.A., Ivanov A.I., Malygin A.Ju. Risk assessment of highly reliable biometrics. Monograph. Almaty: KazNTU named after K.I. Satpayev, 2014. 108 p. (in Russ.).

[5] Akhmetov B.S., Volchihin V.I., Ivanov A.I., Malygin A.Ju. Algorithms of testing of biometric-neural network mechanisms of information protection of Kazakhstan. Almaty: KazNTU named after K.I. Satpayev, 2013. 152 p. ISBN 978-101-228-586-4, http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf (in Russ.).

ПИРСОННЫҢ ЖЕКЕ КРИТЕРИЙЛЕР ЖЕЛІСІМЕН БИОМЕТРИЯЛЫҚ ДЕРЕКТЕРДІҢ КӨПӨЛШЕМДІ СТАТИСТИКАЛЫҚ ТАЛДАУ

Б. Б. Ахметов[1], А. И. Иванов[2], А. В. Безяев[3], Ю. В. Фунтикова[2]

[1]Қ. А. Ясауи атындағы Халықаралық қазақ-түрік университеті, Түркістан, Қазақстан,
[2]Пенза ғылыми-зерттеу электротехникалық институты, Ресей,
[3]«НТЦ «Атлас» ФГУП Пенза филиалы, Ресей

Тірек сөздер: көпөлшемді статистикалық талдау, Пирсонның жеке критерийлер желісі, есепті симметризациялау, корреляциялық байланыстардың әсерін есепке алу.

Аннотация. Көпөлшемді статистикалық өңдеуді қолдануға өткен кезде аз тестілік таңдаулар кезінде өте жоғары шынайлығы бар шешімдерді алуға болатыны көрсетілген. Корреляциялық байланыстарын шешу сапасына алғашқы биометриялық деректердің әсері зерттелген. Деректердің корреляцияланғанының өсуі және оларды өңдеу өлшемінің өсуі бір бірін шегеретіні дәлелденеді.

Поступила 15.01.2015 г.