

**КОРРЕКТИРОВКА МЕТОДИЧЕСКОЙ ПОГРЕШНОСТИ
ВЫЧИСЛЕНИЯ СТАТИСТИЧЕСКИХ МОМЕНТОВ
ЧЕТВЕРТОГО ПОРЯДКА ДЛЯ МАЛЫХ ВЫБОРОК
БИОМЕТРИЧЕСКИХ ДАННЫХ**

А. И. Иванов, Ю. И. Серикова, А. Г. Банных

**ADJUSTMENT OF METHODOLOGICAL ERROR CALCULATING
THE STATISTICAL MOMENTS OF THE FOURTH ORDER
FOR SMALL SAMPLES OF BIOMETRIC DATA**

A. I. Ivanov, Yu. I. Serikova, A. G. Bannyh

Аннотация. *Актуальность и цели.* Поставлена цель по коррекции методических ошибок четвертого статистического момента за счет учета априорной информации о размерах малой выборки реальных биометрических данных. *Материалы и методы.* Используются процедуры численного моделирования малых выборок с нормальным законом распределения значений. Показано, что математическое ожидание ошибки вычисления четвертого момента является функцией стандартного отклонения и числа примеров в тестовой выборке. *Результаты.* Дана таблица поправок, учитывающая конечный размер выборки для нормированных данных. Предложено скомпенсировать методическую погрешность вычисления четвертого момента малых выборок биометрических данных путем гиперболического приближения таблицы поправок. *Выводы.* Подобное приближение позволяет снизить методическую ошибку вычислений до 20 раз при выборках объемом от 4 до 16 примеров. Это открывает возможность начать исследования по снижению случайной составляющей ошибки вычисления четвертого статистического момента на малых тестовых выборках.

Ключевые слова: биометрическая идентификация личности, статистический момент четвертого порядка, объем тестовой выборки, коррекция методической ошибки.

Abstract. *Background.* The aim is for the correction of systematic errors of the fourth statistical moment by taking into account a priori information about the size of a small sample of the actual biometric data. *Materials and methods.* Use the procedure of numerical modeling of small samples with normal values distribution law. It is shown that the expectation of error calculation of the fourth moment is a function of the standard deviation and the number of examples in the test sample. *Results.* Dana table amendments, taking into account the final sample size for the normalized data. It proposed to indemnify for the methodological error of calculation of the fourth moment of the small sample of biometric data by the hyperbolic approximation table amendments. *Conclusions.* Such an approach can reduce the methodological error of calculations up to 20 times while the sample size from 4 to 16 examples. This opens up the potential to begin research on the reduction of the random component of the error calculation of the fourth statistical moment on small test samples.

Key words: biometric identification, statistical moment of the fourth order, the volume of the test sample, the correction of systematic error.

Введение

В настоящее время активно идут процессы информатизации современного общества. Как итог, наша персональная информация постепенно перемещается в интернет-облака. Ярким примером общего вектора развития являются медицинские информационные системы. В 2006 г. в России был введен в действие отечественный стандарт, регламентирующий требования к электронной истории болезни [1], это позволило разработать типовую медицинскую информационную систему [2]. Далее встал вопрос о переходе к использованию типового электронного места врача [3]. В итоге информационная технология уже позволяет создавать интегрированные электронные медицинские карты [4], которые могут размещаться как на локальном сервере медицинской информационной системы, так и на интернет-серверах поставщиков облачных услуг.

Естественно, что эта общая тенденция порождает новые угрозы информационной безопасности, которые должны быть устранены с учетом уже сложившейся технической практики [5] и национального законодательства [6]. За рубежом проблема решается через биометрическую аутентификацию личности человека с использованием так называемых «нечетких экстракторов» [7]. В России для этих же целей используются искусственные нейронные сети [8], применяя которые можно осуществлять обезличивание [9] медицинских электронных документов в случае их размещения в облачных хранилищах.

Очевидным является также то, что массовое использование биометрического обезличивания персональной информации облачных сервисов потребует сертификации криптографических модулей защиты ФСБ России и/или сертификации ФСТЭК России нейросетевых преобразователей биометрии в код пароля доступа. Во втором случае стойкость пароля доступа к атакам подбора со стороны биометрии должна оцениваться статистически по требованиям ГОСТ Р 52633.3–2011 [10].

К сожалению, стандарт [10] ориентирован только на нормальный закон распределения биометрических данных. Реальные биометрические данные, как правило, описываются распределениями данных, отличающимися от нормального закона. В частности, реальные данные образа «Свой» имеют распределения с «тяжелыми» хвостами [11] и ярко выраженным заострением вершины распределения. Распределения значений коэффициентов парной корреляции биометрических данных [12] имеют явно выраженную плоскую вершину и поджатые к центру хвосты распределения.

Для корректного учета отклонения распределения значений от нормального закона необходимо определить значение нормированного статистического момента четвертого порядка или эксцесса [13]:

$$e_4 = \frac{\sum_{i=1}^n (E(x) - x_i)^4}{n \cdot \sigma^4(x)} - 3, \quad (1)$$

где $E(x)$ – математическое ожидание; $\sigma(x)$ – стандартное отклонение; n – число учитываемых опытов.

Следует подчеркнуть, что формула (1) имеет значительную методическую погрешность на малых выборках. Аналогичная методическая погреш-

ность возникает и при вычислении второго статистического момента (стандартного отклонения). При этом ранее нами было показано, что методическая погрешность стандартного отклонения устранима [14]. Задачей данной статьи является устранение методической погрешности эксцесса и/или четвертого статистического момента при их вычислении на малых выборках.

**Моделирование функции распределения значений
четвертого статистического момента
для нормированного нормального закона распределения значений**

Будем наблюдать методическую погрешность статистических моментов четвертого порядка, обусловленную малым объемом выборки, путем проведения численного эксперимента. Для этой цели осуществим многократное обращение к псевдослучайному программному генератору случайных чисел с нормальным законом распределения значений. Далее для каждой выборки вычислим нормированный статистический момент четвертого порядка

$$\frac{\mu_4(x)}{\sigma^4(x)} = \frac{\sum_{i=1}^n (E(x) - x_i)^4}{n \cdot \sigma^4(x)}. \quad (2)$$

Очевидно, что нормированный статистический момент четвертого порядка (2) будет являться случайной функцией, распределение значений которой зависит от размеров тестовой выборки – n и закона распределения значений тестовой выборки. При нормальном законе распределения значений для больших выборок соотношение (2) становится практически детерминированным и дает значение близкое к трем (1). Для выборок ограниченного объема, содержащих 9, 11, 16, 21, 31, примеры распределения значений нормированного четвертого момента (2) приведены на рис. 1.

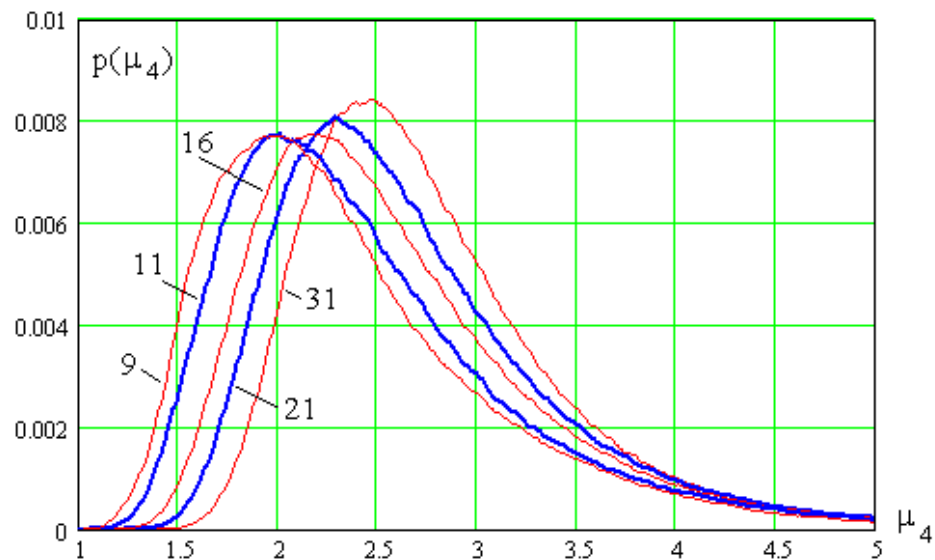


Рис. 1. Плотность распределения значений нормированного четвертого момента для нормального закона распределения значений при разных значениях выборки

Из рис. 1 видно, что с ростом размеров выборки увеличивается математическое ожидание распределений и падает их стандартное отклонение. Данные о взаимодействии этих двух тенденций сведены в табл. 1.

Таблица 1

Значения математических ожиданий и стандартных отклонений четвертого статистического момента при разных размерах тестовой выборки – n

n	4	5	6	7	8	9	10	11	12
$E(n)$	1,800	2,000	2,143	2,250	2,332	2,400	2,455	2,500	2,538
$\sigma(n)$	0,349	0,500	0,597	0,597	0,704	0,735	0,754	0,768	0,776
$\Delta E(n) \%$	39,3	32,3	27,5	24,0	21,3	19,2	17,5	16,1	14,9
n	13	14	15	16	17	18	19	20	21
$E(n)$	2,571	2,599	2,625	2,646	2,667	2,685	2,702	2,716	2,727
$\sigma(n)$	0,779	0,779	0,782	0,777	0,776	0,772	0,768	0,764	0,755
$\Delta E(n) \%$	13,9	13,0	12,3	11,6	11,0	10,5	10,99	9,53	9,13
n	22	23	24	25	26	27	28	29	30
$E(n)$	2,738	2,749	2,760	2,768	2,778	2,786	2,794	2,799	2,806
$\sigma(n)$	0,746	0,742	0,736	0,732	0,725	0,720	0,713	0,706	0,699
$\Delta E(n) \%$	8,76	8,42	8,11	7,83	7,50	7,32	7,09	6,87	6,67
n	31	32	33	100	200	300	1000	3000	10000
$E(n)$	2,811	2,819	2,825	2,941	2,971	2,980	2,994	2,997	2,999
$\sigma(n)$	0,694	0,690	0,683	0,455	0,334	0,276	0,154	0,089	0,049
$\Delta E(n) \%$	6,48	6,30	6,13	2,31	1,26	0,88	0,34	0,11	0,032

На рис. 2 дан график монотонного увеличения математического ожидания нормированных значений четвертого момента $E(\mu_4)$ с ростом числа примеров в выборке.

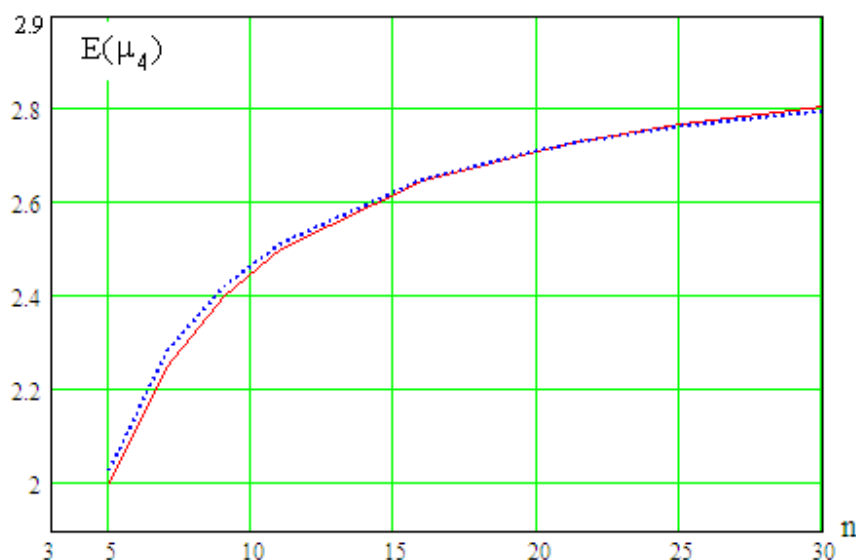


Рис. 2. Приближение математического ожидания четвертого статистического момента как функции числа опытов

На графике рис. 2 (сплошная линия) мы видим монотонный рост математического ожидания значений четвертого статистического момента для нормального закона распределения значений. Для этой монотонной функции в диапазоне выборок от 5 до 50 примеров хорошую аппроксимацию дает гипербола дробной размерности (пунктирная линия на рис. 2):

$$E(\mu_4, n) = 3,0 - \frac{3,99}{n^{0,88}}. \quad (3)$$

Поведение стандартного отклонения четвертого статистического момента для нормированного нормального закона распределения данных не описывается монотонной функцией. Мы наблюдаем ярко выраженный максимум функции стандартного отклонения в интервале размеров выборки от 14 до 16 опытов, что отображено на рис. 3.

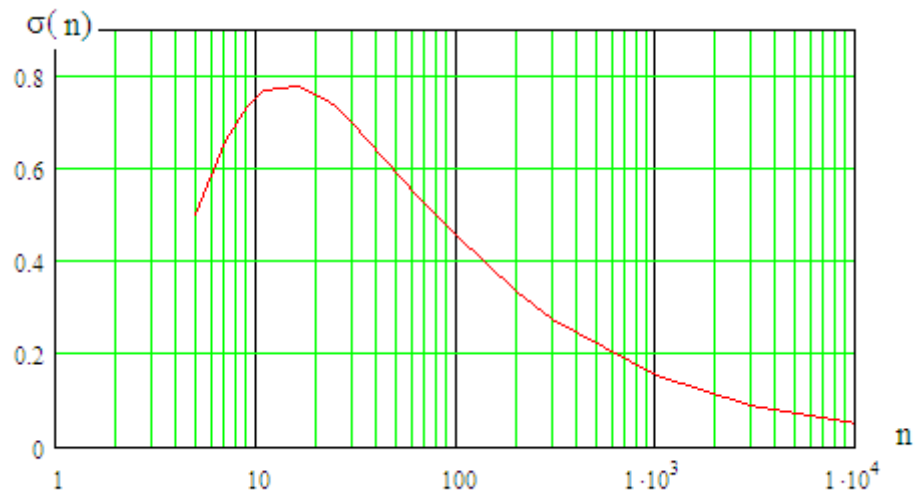


Рис. 3. Влияние размеров выборки на значения стандартного отклонения четвертого статистического момента нормированного нормального закона распределения

Из рис. 3 видно, что для выборок, имеющих более 16 примеров, наблюдается монотонное уменьшение стандартного отклонения, что эквивалентно монотонному снижению случайной составляющей погрешности вычислений при росте объема тестовой выборки. Можно надеяться на то, что только для выборок, имеющих более 16 примеров, можно пользоваться традиционными алгоритмами регуляризации вычислений. Для малых тестовых выборок неустойчивость вычислений (их плохая обусловленность) не может быть уменьшена обычными процедурами регуляризации [15]. Видимо, придется осуществлять регуляризацию вычислений младших статистических моментов на малых выборках на ощупь (в ручном режиме подбора параметров регуляризации [16]). При этом методическая погрешность вычислений случайных статистических моментов должна быть скомпенсирована. В нашем случае компенсация методической погрешности четвертого статистического момента будет описываться следующим соотношением:

$$\mu_4(x) = \frac{\sum_{i=1}^n (E(x) - x_i)^4}{n} - \frac{3,99}{n^{0,88}} \cdot \sigma^4(x). \quad (4)$$

Соотношение (4) оказывается отрицательным, если у распределения биометрических данных имеется плоская вершина и «поджатые» к центру хвосты. Если соотношение (4) дает положительную величину, то скорее всего мы имеем дело с «тяжелыми» хвостами распределения и заостренной вершиной.

Заключение

Проведенные исследования показали, что подавление случайной составляющей погрешности вычисления четвертого статистического момента на малых выборках биометрических данных нельзя выполнять без предварительной компенсации методической погрешности. Такая компенсация может быть выполнена с использованием таблицы поправок № 1 или с использованием аналитического приближения систематической погрешности (4). И в том, и в другом случае методическую погрешность удастся уменьшить более чем в 20 раз, что создает предпосылки для инициации работ по регуляризации вычислительных процедур, снижающих случайную составляющую погрешности.

Список литературы

1. ГОСТ Р 52636–2006. Электронная история болезни. Общие положения.
2. Федеральная типовая медицинская информационная система (ФТМИС). Разработчик «Крокус Консалдинг» 2008 г., государственный контракт по ФЦП «Электронная Россия (2002–2010 годы)».
3. Электронное рабочее место врача. Руководство пользователя. – М., 2014. – URL: http://miasmaukor.ru/media/ERMV_manual.pdf
4. Интегрированная электронная медицинская карта: задачи и проблемы / Б. В. Зингерман, Н. Е. Шкловский-Корди, В. П. Карп, А. И. Воробьев // Врач и информационные технологии. – 2015. – № 1. – С. 24–27.
5. Костков, Д. Защита облачных вычислений: общие международные подходы / Д. Костков // Первая миля. – 2015. – № 8. – С. 26–29.
6. Федеральный закон «О персональных данных» от 27.07.2006 № 152.
7. Dodis, Y. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy / Y. Dodis, L. Reyzin, A. Smith // Proc. EUROCRYPT. – 2004. – April 13. – P. 523–540.
8. ГОСТ Р 52633.0–2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации».
9. Методические рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 г. № 996 «Об утверждении требований и методов по обезличиванию персональных данных».
10. ГОСТ Р 52633.3–2011 «Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора».
11. Учет «тяжелых» хвостов ненормального закона распределения биометрических параметров все «Чужие» при настройке нелинейного элемента нейрона с несколькими дискретными состояниями / С. В. Куликов, М. В. Секретов, О. С. Захаров,

- А. И. Иванов, А. В. Майоров // Нейрокомпьютеры: разработка, применение. – 2012. – № 3. – С. 56–59.
12. Биометрическая идентификация рукописных образов с использованием корреляционного аналога правила Байеса / А. И. Иванов, П. С. Ложников, Е. И. Качайкин, А. Е. Сулавко // Вопросы защиты информации – 2015. – № 3. – С. 48–54.
13. Кобзарь, А. И. Прикладная математическая статистика. Для инженеров и научных работников / А. И. Кобзарь. – М. : ФИЗМАТЛИТ, 2006. – 816 с.
14. Волчихин, В. И. Компенсация методических погрешностей вычисления стандартных отклонений и коэффициентов корреляции, возникающих из-за малого объема выборок / В. И. Волчихин, А. И. Иванов, Ю. И. Серикова // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2016. – № 1 (37). – С. 103–110
15. Тихонов, А. Н. Методы решения некорректных задач / А. Н. Тихонов, В. Я. Арсенин. – М. : Наука, 1979. – 248 с.
16. Кулагин, В. П. Корректировка методических и случайных составляющих погрешностей вычисления коэффициентов корреляции, возникающих на малых выборках биометрических данных / В. П. Кулагин, А. И. Иванов, Ю. И. Серикова // Информационные технологии. – 2016. – Т. 22, № 9. – С. 705–710.

Иванов Александр Иванович

доктор технических наук, доцент,
начальник лаборатории биометрических
и нейросетевых технологий,
Пензенский научно-исследовательский
электротехнический институт
E-mail: ivan@pniei.penza.ru

Ivanov Alexander Ivanovich

doctor of technical sciences,
associate professor,
head of laboratory of biometric
and neural-network technologies,
Penza Research Electrotechnical Institute

Серикова Юлия Игоревна

студентка,
Пензенский государственный
университет
E-mail: julia-ska@yandex.ru

Serikova Yulia Igorevna

student,
Penza State University

Банных Андрей Григорьевич

аспирант,
Пензенский государственный
университет
E-mail: ibst@pnzgu.ru

Bannyh Andrey Grigoryevich

postgraduate student,
Penza State University

УДК 519.24; 53; 57.017

Иванов, А. И.

Корректировка методической погрешности вычисления статистических моментов четвертого порядка для малых выборок биометрических данных / А. И. Иванов, Ю. И. Серикова, А. Г. Банных // Модели, системы, сети в экономике, технике, природе и обществе. – 2016. – № 4 (20). – С. 108–114.