

УДК 519.24; 53; 57.017
doi:10.21685/2072-3059-2022-1-5

**Удвоение числа статистических критериев семейства
Крамера – фон Мизеса дифференцированием
малых выборок с нормальным и равномерным
распределением биометрических данных**

А. И. Иванов¹, А. Ю. Малыгин², С. А. Полковникова³

¹Пензенский научно-исследовательский электротехнический институт, Пенза, Россия

^{2,3}Пензенский государственный университет, Пенза, Россия

¹ivan@pnici.penza.ru, ²mal890@yandex.ru, ³1996svetlanaserikova@gmail.com

Аннотация. *Актуальность и цели.* В прошлом веке создано четыре статистических критерия, которые можно объединить в семейство Крамера – фон Мизеса. Целью данной работы является удвоение числа критериев семейства рассматриваемых критериев. *Материалы и методы.* Предложено перед вычислениями выполнять численное дифференцирование данных малой выборки. При синтезе новых статистических критериев по схеме Крамера – фон Мизеса сравниваются производная входных данных с плотностью распределения нормальных данных. *Результаты и выводы.* Показано, что предложенные в работе новые статистические критерии имеют примерно в 10 раз меньшую вероятность ошибок первого и второго рода. Кроме того, они слабо коррелированы с классическими статистическими критериями этого же семейства.

Ключевые слова: статистический критерий Крамера – фон Мизеса, критерий Смирнова – Крамера – фон Мизеса, критерий Андерсона – Дарлинга, критерий Фроцини, искусственные нейроны, синтез новых статистических критериев

Для цитирования: Иванов А. И., Малыгин А. Ю., Полковникова С. А. Удвоение числа статистических критериев семейства Крамера – фон Мизеса дифференцированием малых выборок с нормальным и равномерным распределением биометрических данных // Известия высших учебных заведений. Поволжский регион. Технические науки. 2022. № 1. С. 53–61. doi:10.21685/2072-3059-2022-1-5

**Doubling the number of statistical Cramer – von Mises
criterion by differentiating small samples with normal
and uniform distribution of biometric data**

A.I. Ivanov¹, A.Yu. Malygin², S.A. Polkovnikova³

¹Penza Scientific Research Electrotechnical Institute, Penza, Russia

^{2,3}Penza State University, Penza, Russia

¹ivan@pnici.penza.ru, ²mal890@yandex.ru, ³1996svetlanaserikova@gmail.com

Abstract. *Background.* In the last century, 4 statistical tests were created that can be combined into the Cramer – von Mises criterion. The purpose of this work is to double the number of this criteria, the criteria under consideration. *Materials and methods.* It is proposed to perform numerical differentiation of small sample data before calculations. In the synthesis of new statistical criteria according to the Cramer-von Mises scheme, the derivative of the input data is compared with the density of the distribution of normal data. *Re-*

sults and conclusions. It is shown that the new statistical criteria proposed in the work have about 10 times less probability of errors of the first and second kind. In addition, they are weakly correlated with the classical statistical criteria of the same family.

Keywords: Kramer – von Mises statistical criterion, Smirnov – Kramer – von Mises criterion, Anderson – Darling criterion, Frozini criterion, artificial neurons, synthesis of new statistical criteria

For citation: Ivanov A.I., Malygin A.Yu., Polkovnikova S.A. Doubling the number of statistical Cramer – von Mises criterion by differentiating small samples with normal and uniform distribution of biometric data. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences.* 2022;(1):53–61. (In Russ.). doi:10.21685/2072-3059-2022-1-5

Введение

Прошлый век начался со значимого события: в 1900 г. основателем математической статистики английским ученым Карлом Пирсоном был разработан и предложен критерий согласия хи-квадрат. Сегодня данный критерий является стандартным¹ и широко используемым наряду с другими классическими статистическими критериями². В целом математической общественностью в XX в. было разработано более 21 критерия для проверки гипотезы нормального распределения данных и порядка 24 статистических критериев для проверки гипотезы равномерного распределения данных [1]. К сожалению, большинство ранее разработанных статистических критериев было ориентировано на обработку больших выборок в 200 и более примеров.

В XXI в. активно развиваются нейросетевые технологии, более того, пришло понимание, что любой из известных статистических критериев может быть представлен как некоторый эквивалентный ему искусственный нейрон [2–4]. В этой ситуации удастся естественным образом объединять или совместно использовать несколько статистических критериев. Достаточно сформировать однослойную сеть из эквивалентных статистическим критериям нескольких искусственных нейронов.

Каждый из созданных в XX в. статистических критериев плохо работает на малых выборках. Например, на выборке в 16 опытов хи-квадрат критерий Пирсона дает значение появления вероятности ошибок первого и второго рода на уровне $P_1 \approx P_2 \approx P_{EE} \approx 0,33$, что примерно в 10 раз хуже, чем требует практика. Очевидно, что, объединив в одну нейросеть 20 эквивалентных статистическим критериям нейронов, на выходе мы будем получать выходной код с 20-кратной избыточностью. То есть мы всегда можем свернуть избыточный код, скорректировав в его разрядах ошибки [5, 6]. В простейшем случае может быть использован самокорректирующийся код, построенный на равноправном голосовании большинства состояний его разрядов.

Очевидно, что корректирующая способность кодов тем выше, чем больше его избыточность. В связи с этим желательно создавать новые статистические критерии, увеличивающие возможность коррекции ошибок при анализе малых выборок.

¹ Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа χ^2 .

² Р 50.1.037–2002. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии.

Семейство классических статистических критериев Крамера – фон Мизеса, ориентированных на проверку гипотезы нормального распределения малых выборок

Следует отметить, что критерий Крамера – фон Мизеса создан в 1928 г., однако он имеет достаточно скромную мощность на малых выборках. Так, для выборки в 16 опытов этот критерий позволяет различать нормальные и равномерные данные с вероятностями ошибок $P_1 \approx P_2 \approx P_{EE} \approx 0,4$. Это существенно хуже в сравнении с критерием хи-квадрат.

Добиться повышения мощности удалось Смирнову в 1936 г. Новая модификация критерия позволила снизить вероятности ошибок до значений $P_1 \approx P_2 \approx P_{EE} \approx 0,322$, что лучше, чем у критерия хи-квадрат.

Следующая модификация критерия была создана Андерсоном и Дарлинггом в 1952 г., она была получена делением накапливаемых данных на функции вероятности нормально распределенных данных. При этом удалось добиться снижения вероятностей ошибок первого и второго рода $P_1 \approx P_2 \approx P_{EE} \approx 0,272$. То есть такая модификация дает снижение вероятности ошибок на 21 %.

Последняя модификация критерия выполнена Фроцини в 1978 г. Фактически она сводится к замене в критерии Смирнова – Крамера – фон Мизеса операции возведения в квадрат на операцию вычисления модуля. Это позволило на выборках в 16 опытов снизить вероятности ошибок до значений $P_1 \approx P_2 \approx P_{EE} \approx 0,172$ – почти на 58 %.

Таким образом, модификации исходного статистического критерия Крамера – фон Мизеса за 50 лет исследований позволили значительно снизить вероятности ошибок первого и второго рода. Программное обеспечение на языке MathCAD для проведения численного эксперимента для нормальных данных приведено на рис. 1.

```

sx := | x ← sort(morm(16,0,1))
      | m ← mean(x)
      | σ ← stdev(x)
      | KfM ← ∑i=015 (pnorm(xi,m,σ) -  $\frac{i-0.5}{16}$ )2
      | SKfM ← ∑i=015 [ (pnorm(xi,m,σ) -  $\frac{i-0.5}{16}$ )2 · dnorm(xi,m,σ) ]
      | AD ← ∑i=015  $\frac{(pnorm(x_i,m,\sigma) - \frac{i-0.5}{16})^2 \cdot dnorm(x_i,m,\sigma)}{pnorm(x_i,m,\sigma) \cdot (1 - pnorm(x_i,m,\sigma))}$ 
      | F ← ∑i=015 [ | (pnorm(xi,m,σ) -  $\frac{i-0.5}{16}$ ) | · dnorm(xi,m,σ) ]
      | (KfM SKfM AD F)T

```

Рис. 1. Программное обеспечение на языке MathCAD для моделирования классических критериев семейства Крамера – фон Мизеса

Результаты численного моделирования четырех классических статистических критериев на малой выборке в 16 опытов приведены на рис. 2.

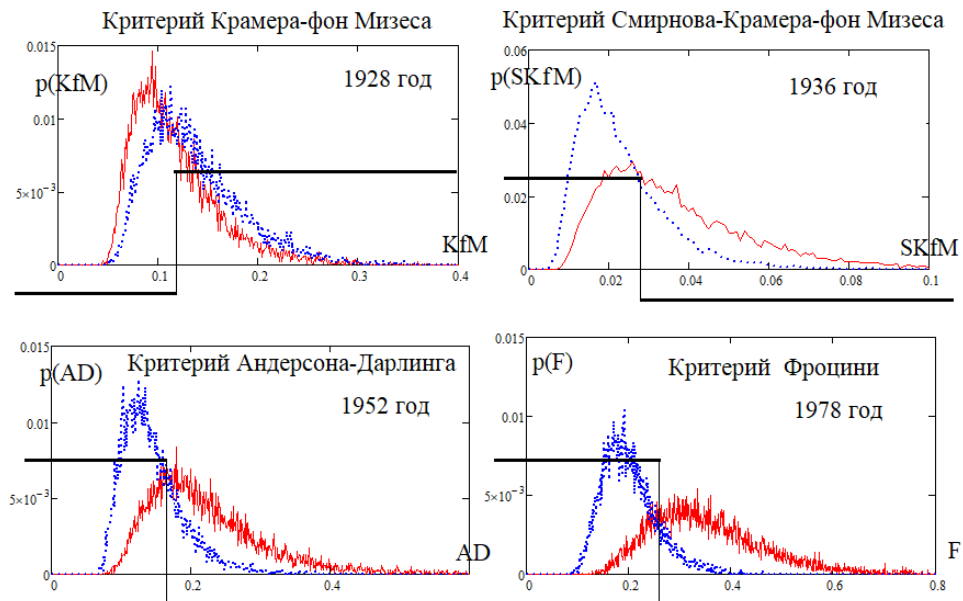


Рис. 2. Распределение данных четырех классических критериев (распределение откликов на равномерные данные отображено точками и дает состояние «0»)

Не случайно эволюция базового критерия Крамера – фон Мизеса шла по пути снижения вероятности ошибок первого и второго рода. Видимо, эволюция семейства рассматриваемых критериев далеко не закончена.

Синтез новых статистических критериев семейства Крамера – фон Мизеса через дифференцирование данных малой выборки

По программному коду (рис. 1) видно, что все критерии рассматриваемого семейства построены на разнице между скачками роста вероятности реальных отсчетов малой выборки и скачков вероятности идеального нормального распределения. Очевидно, что вместо интегральных вероятностей мы можем использовать их производную – плотность распределения вероятности [7]. Для этой цели нужно вычислить производную данных реальной выборки и соответствующую функцию плотности распределения вероятности. На рис. 3 приведен код соответствующего численного эксперимента для равномерно распределенных данных.

Результаты численного эксперимента по имитации четырех новых дифференциальных аналогов приведены на рис. 4.

Из данных рис. 4 видно, что дифференциальный вариант критерия Крамера – фон Мизеса дает снижение вероятности ошибок $P_1 \approx P_2 \approx P_{EE} \approx 0,041$. Этот показатель примерно в 10 раз лучше, чем у базового критерия.

Если попытаться воспроизводить дифференциальный аналог критерия Смирнова – Крамера – фон Мизеса, то мы получим снижение вероятности ошибок до величины $P_1 \approx P_2 \approx P_{EE} \approx 0,039$, или в 8,4 раза.

```

sx := x ← sort(runif(16, -3, 3))
m ← mean(x)
σ ← stdev(x)
dKfM ← ∑i=014 [ dnorm(xi, m, σ) -  $\frac{x_{i+1} - x_i}{16 \cdot (x_{15} - x_0)}$  ]2
dSKfM ← ∑i=014 [ [ dnorm(xi, m, σ) -  $\frac{x_{i+1} - x_i}{16 \cdot (x_{15} - x_0)}$  ]2 · dnorm(xi, m, σ) ]
dAD ← ∑i=014  $\frac{[ dnorm(x_i, m, \sigma) - \frac{x_{i+1} - x_i}{16 \cdot (x_{15} - x_0)} ]^2 \cdot dnorm(x_i, m, \sigma)}{pnorm(x_i, m, \sigma) \cdot (1 - pnorm(x_i, m, \sigma))}$ 
dF ← ∑i=014 [ [ [ dnorm(xi, m, σ) -  $\frac{x_{i+1} - x_i}{16 \cdot (x_{15} - x_0)}$  ] ] · dnorm(xi, m, σ) ]
(dKfM dSKfM dAD dF)T

```

Рис. 3. Программное обеспечение на языке MathCAD для моделирования дифференциальных критериев семейства Крамера – фон Мизеса

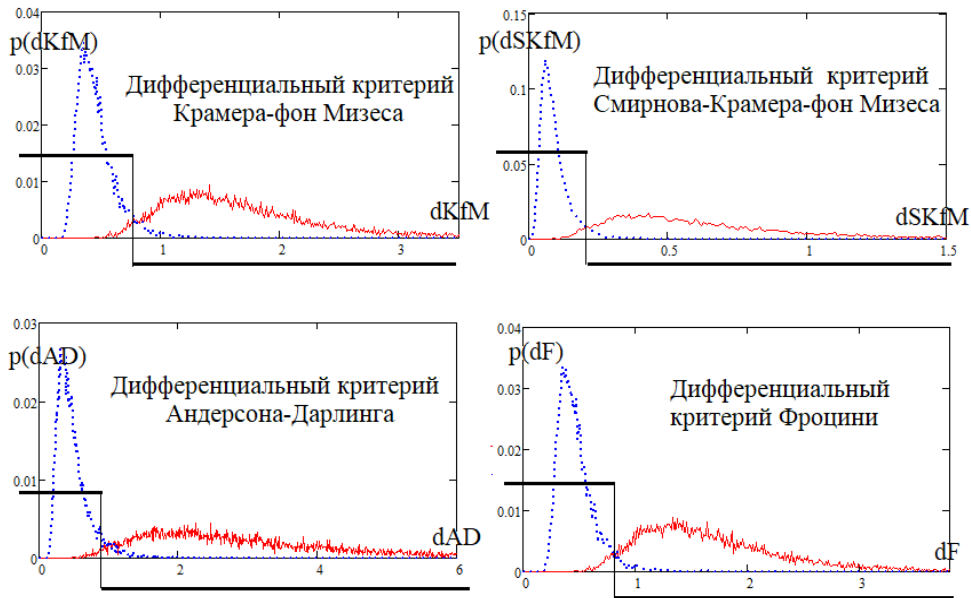


Рис. 4. Распределение данных четырех дифференциальных аналогов классических критериев Крамера – фон Мизеса

Дифференциальный аналог критерия Андерсона – Дарлингга обеспечивает вероятности ошибок $P_1 \approx P_2 \approx P_{EE} \approx 0,039$ или снижение по сравнению с его классическим аналогом в 6,4 раза.

Дифференциальный аналог классического критерия Фрочкини дает тот же уровень вероятности ошибок, что и уже исследованные новые критерии $P_1 \approx P_2 \approx P_{EE} \approx 0,041$. Выигрыш по сравнению с его классическим аналогом – в 4,2 раза.

Корреляционные связи классических интегральных и новых дифференциальных статистических критериев

Одной из негативных особенностей классических статистических критериев является то, что их трудно обобщать [2], заменяя эквивалентными искусственными нейронами. Проблема состоит в сильной корреляционной сцепленности их откликов. Эта ситуация отображена на рис. 5.

	KfM	SKfM	AD	F
KfM	1	0.899	0.848	0.726
SKfM		1	0.949	0.93
AD			1	0.916
F				1

	dKfM	dSKfM	dAD	dF
dKfM	1	0.987	0.984	1
dSKfM		1	0.997	0.987
dAD			1	0.984
dF				1

$$\text{mean}(0.899, 0.848, 0.726, 0.949, 0.93, 0.916) = 0.878 \quad \text{mean}(0.987, 0.984, 1, 0.997, 0.987, 0.984) = 0.99$$

Рис. 5. Демонстрация высокого уровня корреляционной сцепленности для семейства классических критериев и семейства их дифференциальных аналогов

Из данных рис. 5 видно, что классические критерии семейства Крамера – фон Мизеса имеют высокие значения коэффициентов взаимной корреляции. Аналогичная ситуация возникает и для группы дифференциальных аналогов этой группы критериев. Более того, уровень корреляционной сцепленности для дифференциальных критериев оказывается выше уровня корреляционной сцепленности классических статистических критериев.

Из данных на рис. 6 видно, что некоторые корреляционные связи между классическими и новыми статистическими оказываются достаточно низкими, это свидетельствует о взаимном дополнении классических критериев и их дифференциальных аналогов.

Заключение

Таким образом, переход от четырех классических статистических критериев Крамера – фон Мизеса к четырем их дифференциальным аналогам позволяет примерно в 10 раз снизить вероятности ошибок первого и второго рода. При этом далее усилить выигрыш нельзя, обобщая между собой только дифференциальные статистические критерии, из-за их сильной корреляционной сцепленности.

	KfM	SKfM	AD	F	dKfM	dSKfM	dAD	dF
KfM	1	0.899	0.848	0.726	-0.014	-0.012	-0.012	-0.015
SKfM		1	0.949	0.93	0.343	0.331	0.329	0.342
AD			1	0.916	0.462	0.448	0.456	0.462
F				1	0.589	0.569	0.568	0.588
dKfM					1	0.987	0.984	1
dSKfM						1	0.997	0.987
dAD							1	0.984
dF								1

Рис. 6. Данные коэффициентов взаимной корреляции между группой классических критериев и их дифференциальными аналогами

По всей видимости, перспективными будут являться комбинации классических критериев и их дифференциальных аналогов.

На данный момент нет технологий синтеза статистических критериев со значительными отличиями по вероятностям появления ошибок. В этом плане объединение классических статистических критериев и их дифференциальных аналогов является нетривиальной задачей.

Список литературы

1. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников. М. : Физматлит, 2006. 816 с.
2. Иванов А. И., Банных А. Г., Безяев А. В. Искусственные молекулы, собранные из искусственных нейронов, воспроизводящих работу классических статистических критериев // Вестник Пермского университета. Серия: Математика. Механика. Информатика. 2020. № 1 (48). С. 26–32.
3. Иванов А. И. Искусственные математические молекулы: повышение точности статистических оценок на малых выборках (программы на языке MathCAD) : препринт. Пенза : Изд-во ПГУ, 2020. 36 с.
4. Иванов А. И., Банных А. Г., Куприянов Е. Н., Лукин В. С., Перфилов К. А., Савинов К. Н. Коллекция искусственных нейронов эквивалентных статистическим критериям для их совместного применения при проверке гипотезы нормальности малых выборок биометрических данных // Безопасность информационных технологий : сб. науч. ст. по материалам I Всерос. науч.-техн. конф. Пенза : Изд-во ПГУ, 2019. С. 156–164.
5. Морелос-Сарагоса Р. Искусство помехоустойчивого кодирования : пер. с англ. М. : Техносфера, 2007. 320 с.
6. Безяев А. В. Биометрико-нейросетевая аутентификация: обнаружение и исправление ошибок в длинных кодах без накладных расходов на избыточность : препринт. Пенза : Изд-во ПГУ, 2020. 40 с.
7. Иванов А. И., Малыгин А. Ю., Полковникова С. А. Новый статистический критерий большой мощности, полученный дифференцированием случайных данных

малой выборки // Известия высших учебных заведений. Поволжский регион. Технические науки. 2021. № 3. С. 67–74. doi:10.21685/2072-3059-2021-3-7

References

1. Kobzar' A.I. *Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov = Applied mathematical statistics. For engineers and scientists*. Moscow: Fizmatlit, 2006:816. (In Russ.)
2. Ivanov A.I., Bannykh A.G., Bezyaev A.V. Artificial molecules assembled from artificial neurons that reproduce the work of classical statistical criteria. *Vestnik Permskogo universiteta. Seriya: Matematika. Mekhanika. Informatika = Bulletin of Perm University. Series: Mathematics. Mechanics. Informatics*. 2020;(1):26–32. (In Russ.)
3. Ivanov A.I. *Iskusstvennye matematicheskie molekuly: povyshenie tochnosti statisticheskikh otsenok na malyykh vyborkakh (programmy na yazyke MathCAD): preprint = Artificial mathematical molecules: improving the accuracy of statistical estimates on small samples (MathCAD programmes): preprint*. Penza: Izd-vo PGU, 2020:36. (In Russ.)
4. Ivanov A.I., Bannykh A.G., Kupriyanov E.N., Lukin V.S., Perfilov K.A., Savinov K.N. Collection of artificial neurons equivalent to statistical criteria for their joint application of the hypothesis of normality of samples of biometric data. *Bezopasnost' informatsionnykh tekhnologiy: sb. nauch. st. po materialam I Vseros. nauch.-tekhn. konf. = Information technology security: proceedings of the 1st All-Russian scientific and engineering conference*. Penza: Izd-vo PGU, 2019:156–164. (In Russ.)
5. Morelos-Saragosa R. *Iskusstvo pomekhoustoychivogo kodirovaniya: per. s angl. = The art of error-correcting coding: translated from English*. Moscow: Tekhnosfera, 2007:320. (In Russ.)
6. Bezyaev A.V. *Biometriko-neyrosetevaya autentifikatsiya: obnaruzhenie i ispravlenie oshibok v dlinnykh kodakh bez nakladnykh raskhodov na izbytochnost': preprint = Biometric-neural network authentication: error detection and correction in long codes without redundancy overhead: preprint*. Penza: Izd-vo PGU, 2020:40. (In Russ.)
7. Ivanov A.I., Malygin A.Yu., Polkovnikova S.A. A new statistical test of high power obtained by differentiating random data from a small sample. *Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskie nauki = University proceedings. Volga region. Engineering sciences*. 2021;(3):67–74. (In Russ.). doi:10.21685/2072-3059-2021-3-7

Информация об авторах / Information about the authors

Александр Иванович Иванов

доктор технических наук, доцент,
научный консультант, Пензенский
научно-исследовательский
электротехнический институт (Россия,
г. Пенза, ул. Советская, 9)

E-mail: ivan@pniei.penza.ru

Aleksandr I. Ivanov

Doctor of engineering sciences, associate
professor, scientific adviser, Penza
Scientific Research Electrotechnical
Institute (9 Sovetskaya street,
Penza, Russia)

Александр Юрьевич Малыгин

доктор технических наук, профессор,
начальник межотраслевой лаборатории
тестирования биометрических устройств
и технологий, Пензенский
государственный университет
(Россия, г. Пенза, ул. Красная, 40)

E-mail: mal890@yandex.ru

Aleksandr Yu. Malygin

Doctor of engineering sciences, professor,
head of the Intersectoral testing laboratory
of biometric devices and technologies,
Penza State University (40 Krasnaya
street, Penza, Russia)

Светлана Андреевна Полковникова
аспирант, Пензенский государственный
университет (Россия, г. Пенза,
ул. Красная, 40)

Svetlana A. Polkovnikova
Postgraduate student, Penza
State University (40 Krasnaya
street, Penza, Russia)

E-mail: 1996svetlanaserikova@gmail.com

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflicts of interests.

Поступила в редакцию / Received 20.01.2022

Поступила после рецензирования и доработки / Revised 03.02.2022

Принята к публикации / Accepted 21.02.2022