DOI 10.37882/2223–2966.2021.04.19

# АНАЛОГИ КОРРЕЛЯЦИОННОГО КРИТЕРИЯ ПИРСОНА

## ANALOGUES OF THE PEARSON CORRELATION TEST

***T. Zolotareva***

*Summary.* This paper considers discrete analogs of the Pearson correlation criterion, presents a numerical experiment that reproduces the conditions for modeling the spectral representation of data, and presents the results of a numerical experiment. The discrete-differential analogs of the Pearson correlation criterion are also considered. The dual representation of randomness using analogs of the Pearson correlation criterion is considered.

Objective: to consider the dual representation of randomness in two different forms.

Methods: discrete analogs of the Pearson correlation test and discrete-differential analogs of the Pearson correlation test.

Results: the classical Pearson correlation coefficient and its differential-discrete analog are almost significantly less correlated (corr (r,"r13")≈ –0.60871) than the Pearson difference correlations. At the same time, the probabilities of errors of the first and second kind for this class of transformations turn out to be applicable at a rate of 0.45 for practice.

Conclusions: it is not the procedure of transition from one form of randomness description to another that is fundamentally important, but the fundamentally different properties of the data presented in two different forms.

*Keywords:* pearson's criterion; discrete representations; continuous representations; neuron; error; continuum.

***Золотарева Татьяна Александровна***
*Старший преподаватель, Липецкий казачий институт технологий и управления (филиал) Московского государственного университета технологий и управления им. К.Г. Разумовского (Первый казачий университет) (Москва), г. Липецк*
*zolotarevatatyana2016@yandex.ru*

*Аннотация.* В данной статье рассмотрены дискретные аналоги корреляционного критерия Пирсона, приведен численный эксперимент, воспроизводящий условия моделирования спектрального представления данных, представлены результаты численного эксперимента. Так же рассмотрены дискретно-дифференциальные аналоги корреляционного критерия Пирсона. Рассмотрено дуальное представление случайности с помощью аналогов корреляционного критерия Пирсона.

Цель: рассмотреть дуальное представление случайности в двух разных формах.

Методы: дискретные аналоги корреляционного критерия Пирсона и дискретно-дифференциальные аналоги корреляционного критерия Пирсона.

Результаты: классический коэффициент корреляции Пирсона и его дифференциально-дискретный аналог практически существенно меньше коррелированы (corr(r,«r13»)≈–0.60871), чем разностные корреляции Пирсона. При этом вероятности ошибок первого и второго рода для этого класса преобразований оказывается применимы PEE≈0.45 на практики.

Выводы: принципиально важна не сама процедура перехода от одной формы описания случайности к другой, а принципиально разные свойства данных представленных в двух разных формах.

*Ключевые слова:* критерий Пирсона; дискретные представления; непрерывные представления; нейрон; ошибка; континуум.

## Введение

По классическим представлениям линейной статистики существуют раздельно непрерывные плотности распределения и дискретные плотности распределения. Дискретные и непрерывные представления данных никак не связаны между собой (у них разная природа, а мы просто наблюдатели, фиксирующие факт нашего наблюдения).

Примерно такая же ситуация наблюдалась в физике в 20-х годах XX-го века, тогда в понимании физиков отсутствовал дуализм частицы материи одновременно, являющейся и волной, и частицей. Однако позднее физики поняли, что обе формы описания частиц одинаковы. Каждый из наблюдателей имеет право пользоваться либо волновым, либо корпускулярным представлением данных (существует корпускулярно-волновой дуализм описания действительности).

## Материалы и методы

### Дискретные аналоги корреляционного критерия Пирсона

Непрерывные распределения состояния частиц (опытов малых выборок) существовало независимо от волнового представления этих же частиц дискрет-

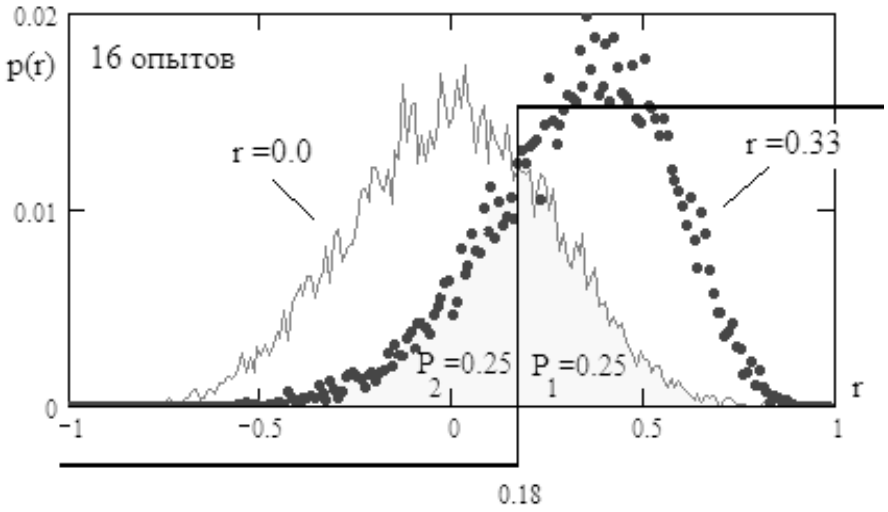



Рис. 1. Вероятности ошибок первого и второго рода корреляционного критерия Пирсона при различении независимых данных малых выборок в 16 опытов и зависимых данных r =0.33

ным спектром. Первый звонок о дуализме статистических наблюдений прозвенел в 1973 году, когда Дахия и Гурланд [1], отказались верить тому, что видят. Классический непрерывный хи-квадрат закон распределения корпускул Пирсона при определенных условиях наблюдения оказался дискретным (волновым) [1]. Авторы этого наблюдения были озадачены своим данными и именно по этой причине поставили знак вопроса в названии своей статьи.

Во второй раз тот же самый эффект дуальности позиции наблюдателя был обнаружен через 40 лет в 2015 году при анализе биометрических данных [2]. Однако в этом случае у наблюдателей никаких сомнений уже не было, так как они знали, как изменить условия наблюдения для кардинального изменения видимого результата. В 2015 году уже были понятны условия, при которых хи-квадрат закон распределения Пирсона выглядит, как непрерывный (корпускулярный) процесс или как дискретный (волновой) процесс [3].

Располагая этими знаниями, следует ожидать, что параллельно с классическим корреляционным непрерывным критерием Пирсона (рис. 1), должен существовать его дискретный (спектральный) аналог.

Из рис. 1 видно, что при вычислениях на малых выборках ошибки первого и второго рода для формулы Пирсона значительны. Если рассматривать критерий Пирсона как некоторый искусственный нейрон с 32 входами (16 входов — для координат отсчетов выборки по одной оси и 16 входов нейрона для координат отсчетов выборки по второй оси), то высока вероятность ошибочных оценок, попадающих в интервал ±0.75 [4].

Численный эксперимент, воспроизводящий условия моделирования спектрального представления данных на языке MathCAD представлен на рис. 2.

## Результаты

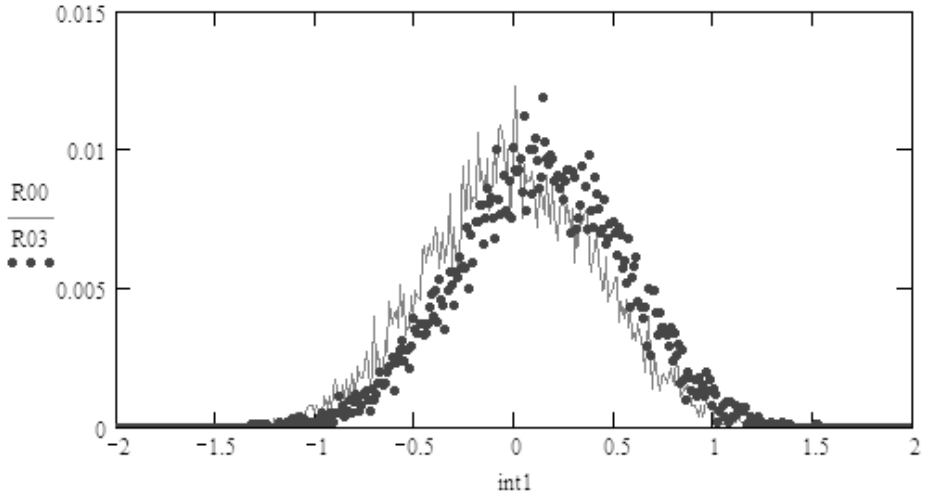
Результаты численного эксперимента представлены на рис. 3.

Из рис. 3 видно, что непрерывный спектр состояний корреляционного критерия Пирсона превратился в дискретный спектр, состоящий из 11 наиболее значимых спектральных линий амплитуд вероятности, сумма которых дает вероятность 0.99.

При вычислениях амплитуды и положения спектральных линий рис. 3, каждый знак элементов суммы коэффициента корреляции Пирсона, рассматривается как положительный или отрицательный ранги. В качестве статистики нового критерия — s рассматривается сумма отрицательных знаков. Компаратор статистики r8, соответствующего нейрона, имеет порог принятия решений k=7.5. При «r8»≥7.5 искусственный нейрон принимает выходное состояние «0», соответствующее решению о подтверждении гипотезы независимости данных малой выборки. Равновероятные ошибки первого и второго рода для нейрона статистики «r8» имеют приемлемые для практического применения значения PEE≈0.45. Важным является то, что классический нейрон Пирсона (рис. 1) и его дискретный аналог «r8»

$j := 0..999$

$$int1_j := -2 + 0.01 \cdot j \quad R00 := \frac{hist(int1, Rr00T^{\langle 1 \rangle})}{9999} \qquad R03 := \frac{hist(int1, Rr03T^{\langle 1 \rangle})}{9999}$$



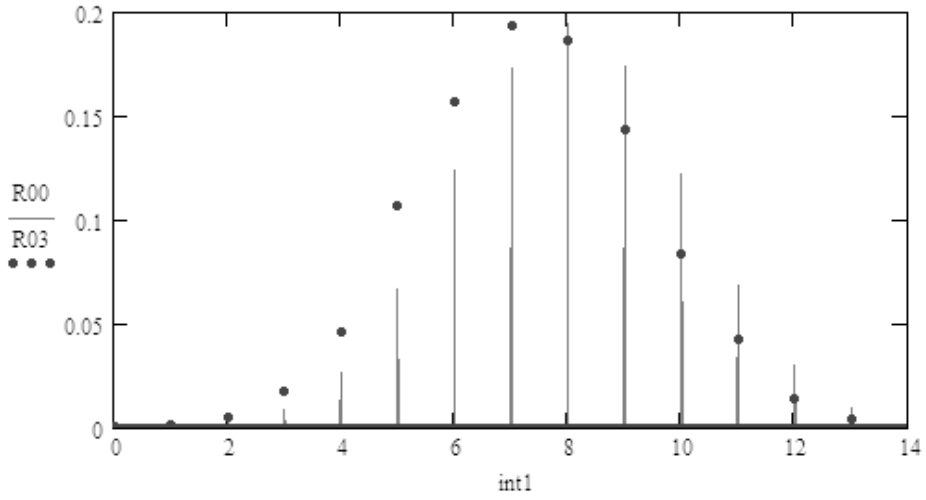

$$\frac{0.427 + 0.423}{2} = 0.425$$

$$1 - \sum_{i=0}^{211} R00_i = 0.384$$

$$\sum_{i=0}^{211} R03_i = 0.487$$

$j := 0..1999$

$$int1_j := -0 + 0.01 \cdot j \quad R00 := \frac{hist(int1, Rr00T^{\langle 2 \rangle})}{9999} \qquad R03 := \frac{hist(int1, Rr03T^{\langle 2 \rangle})}{9999}$$




$$\frac{0.504 + 0.396}{2} = 0.45$$

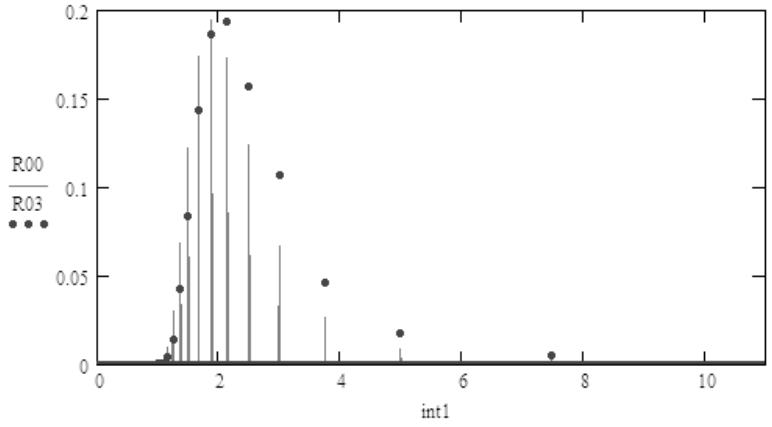
$$\sum_{i=0}^{750} R00_i = 0.401$$

$$1 - \sum_{i=0}^{750} R03_i = 0.474$$

Рис. 2 (часть 1). Дискретные варианты нейронов Пирсона.

$j := 0..1999$

$int1_j := -0 + 0.01 \cdot j \quad R00 := \frac{hist(int1, Rr00T^{\langle 3 \rangle})}{9999} \quad R03 := \frac{hist(int1, Rr03T^{\langle 3 \rangle})}{9999}$



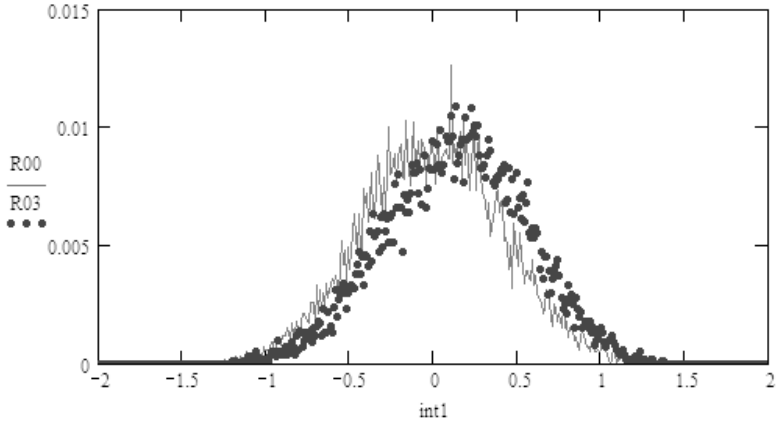

$\frac{0.504 + 0.396}{2} = 0.45$

$1 - \sum_{i=0}^{200} R00_i = 0.4$

$\sum_{i=0}^{200} R03_i = 0.474$

$j := 0..1999$

$int1_j := -2 + 0.01 \cdot j \quad R00 := \frac{hist(int1, Rr00T^{\langle 4 \rangle})}{9999} \quad R03 := \frac{hist(int1, Rr03T^{\langle 4 \rangle})}{9999}$



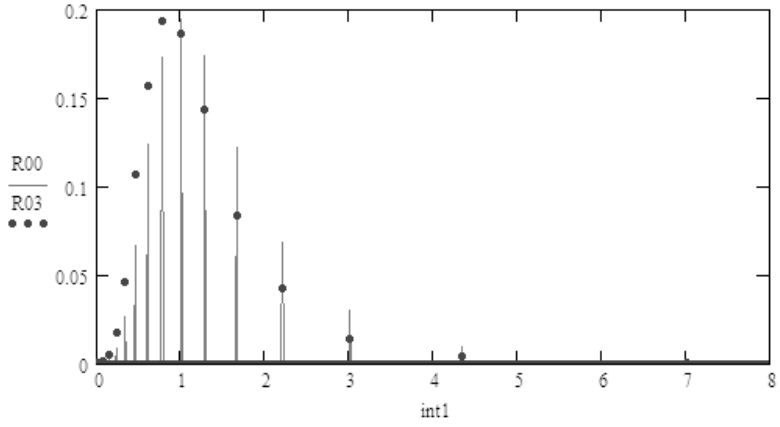

$\frac{0.437 + 0.434}{2} = 0.436$

$1 - \sum_{i=0}^{210} R00_i = 0.399$

$\sum_{i=0}^{210} R03_i = 0.471$

$j := 0..1999$

$int1_j := 0 + 0.01 \cdot j \quad R00 := \frac{hist(int1, Rr00T^{\langle 5 \rangle})}{9999} \quad R03 := \frac{hist(int1, Rr03T^{\langle 5 \rangle})}{9999}$




$\frac{0.398 + 0.481}{2} = 0.439$

$\sum_{i=0}^{95} R00_i = 0.401$

$1 - \sum_{i=0}^{95} R03_i = 0.474$

Рис. 2 (часть 2). Дискретные варианты нейронов Пирсона.

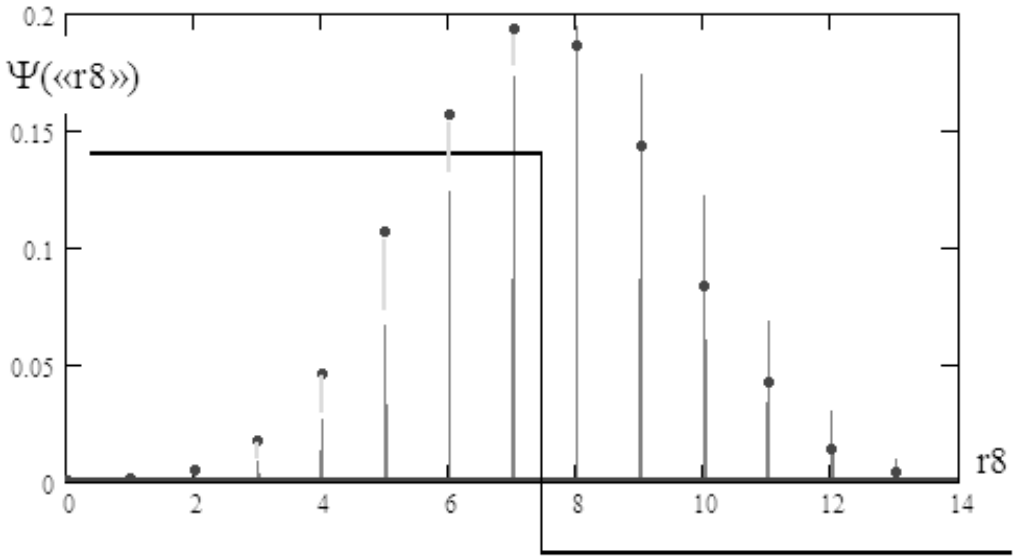



Рис. 3. Дискретный аналог корреляционного критерия Пирсона, содержащий примерно 11 спектральных линий

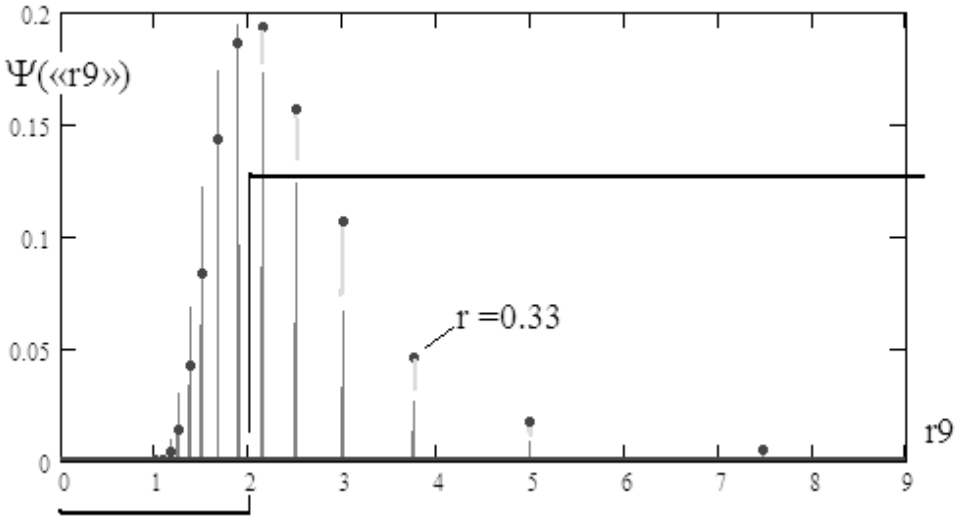



Рис. 4. Нейрон статистики r9, накапливающий (обогащающий) данные в гиперболическом пространстве

имеют значимую независимую компоненту выходных состояний corr(r,«r8»)≈ –0.67346.

Нейрон статистики r8 накапливает (обогащает) данные в линейном пространстве спектральных линий — s, расположенных на одинаковых расстояниях между друг другом. Если мы деформируем пространство накопления данных гиперболой 15/(s+0.01), то спектральные линии утрачивают равномерное расположение спектральных линий, как это показано на рис. 4. Нелинейная деформация пространства накопления данных нейроном статистики r9 позволяет примерно в 10 раз снизить модуль корреляционной сцепленности данных нейронов Пирсона corr(r,«r9»)≈0.0546, corr(«r8»,«r9»)≈-0.0793 при приемлемой вероятности ошибок первого и второго рода PEE≈0.45. В свою очередь среднее значение модулей коэффициентов корреляции для рассматриваемой группы из двух нейронов Пирсона снижается почти в три раза по сравнению с самым большим по модулю

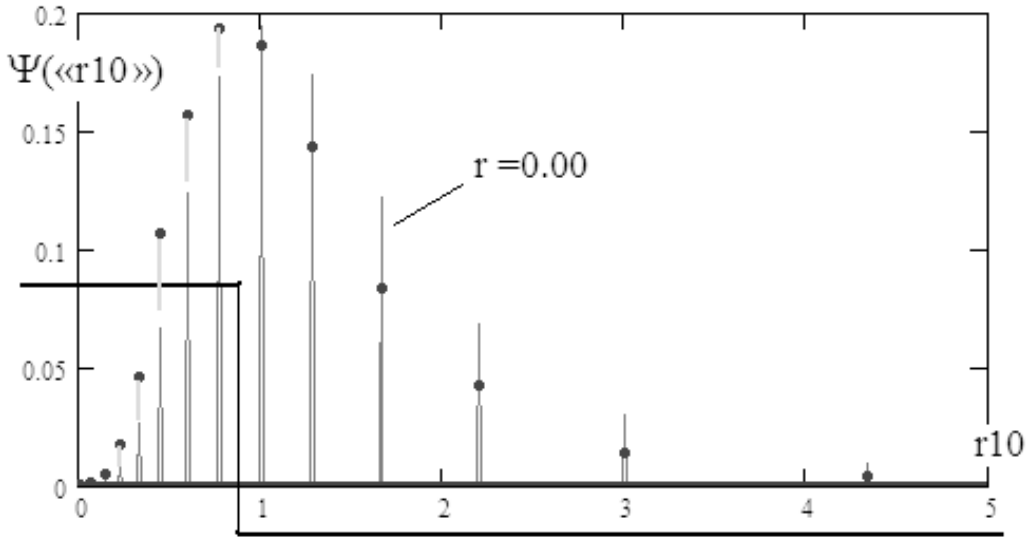



Рис. 5. третий вариант дискретного нейрона Пирсона, полученный нелинейной деформацией s/(16-s) исходного линейного пространства накопления данных первого варианта

коэффициентом коррелированности в этой группе corr(r,«r8»)≈-0.67346.

Следует подчеркнуть, что появление значимой независимой компоненты выходных состояний исходного и модифицированного нейронов обязательно связано с некоторой нелинейной деформацией пространства накопления данных производным искусственным нейроном. Должен существовать континуум возможных нелинейных деформаций, позволяющий получить десятки приемлемых для практики слабо коррелированных нейросетевых решений. Проиллюстрируем этот тезис на примере применения еще одной гиперболической деформации исходно линейного спектрального пространства вида: s/(16-s). Состояния выходного итогового нейрона иллюстрируются рис. 5.

В отличие от первых двух дискретных нейронов, новый третий вариант нейрона, имеет значения вероятностей ошибок примерно на 5% ниже (PEE≈ 0.439). Для практики показатели корреляционной сцепленности являются вполне приемлемыми: {corr(r,«r10»)≈-0.593; corr(«r8»,«r10»)≈-0.409; corr(«r9»,«r10»)≈-0.046}.

## Дискретно-дифференциальные аналоги корреляционного критерия Пирсона

Следует отметить, что варианты критериев проверки статистических гипотез могут быть формально получены заменой функции вероятности на ее производную (плотность вероятности). Например, используя экспериментально полученную функцию вероятности — $\widetilde{P}(x_i)$ и ее гладкий теоретический эквивалент — $P(x_i)$, то мы можем построить статистический критерий Крамера-фон Мизеса (1928 год):

$$KfM = \int_{-\infty}^{\infty} (P(x) - \widetilde{P}(x))^2 \cdot dx \approx$$
$$\approx \sum_{i=1}^{15} \left[P(x_i) - \widetilde{P}(x_i)\right]^2 \cdot (x_{i+1} - x_i) \quad (1)$$

При решении задачи различения нормально распределенных данных и равномерно распределенных данных большей мощностью обладает классический статистический критерий Смирнова-Крамера-фон Мизеса (1936 год):

$$SKfM = \int_{-\infty}^{\infty} (P(x) - \widetilde{P}(x))^2 \cdot dP(x) \approx$$
$$\approx \sum_{i=1}^{15} \left[P(x_i) - \widetilde{P}(x_i)\right]^2 \cdot (P(x_{i+1}) - P(x_i)) \quad (2)$$

На малых выборках в 16 опытов, близкие по программной реализации вычислительные конструкции (1) и (2) дают вероятности ошибок первого и второго рода, отличающиеся в 10 раз (от PEE≈0.403 до PEE≈0.042) [5,6] при очень низком уровне корреляционной сцепленности (corr(KfM, SKfM)≈-0.03).

Понятно, что в первой половине XX века исследователей интересовала прежде всего мощность отдельных синтезируемых статистических критериев. Сегодня в XXI веке исследователей интересует, прежде всего, групповая корреляционная сцепленность системы новых математических конструкций. В связи с этим при синтезе вариантов нейронов Пирсона необходимо проверять похожие друг на друга разносные схемы вычислений.

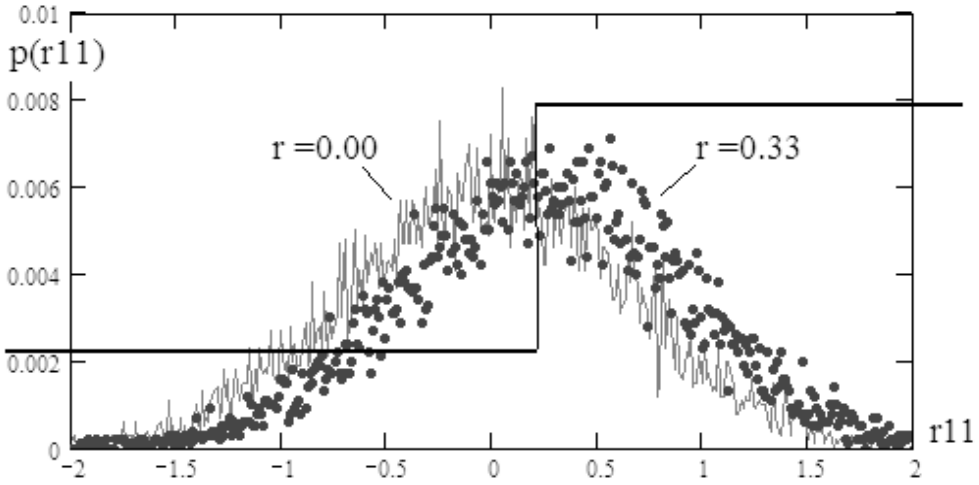



Рис. 6. Разностный аналог классического корреляционного критерия Пирсона

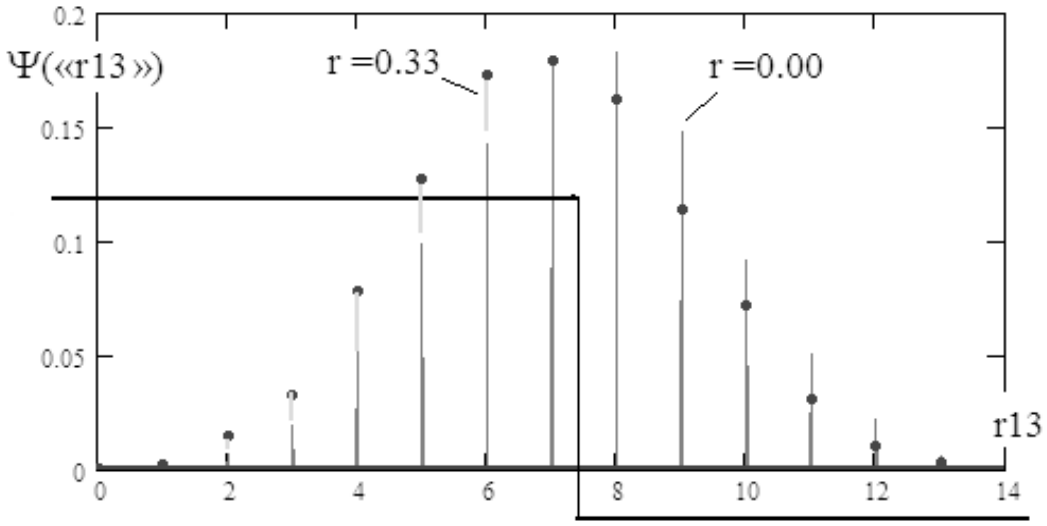



Рис. 7. Дискретный спектр равномерно расположенных 11 наиболее значимых спектральных линий амплитуд вероятности

Новые статистики строятся заменой классической формулы Пирсона

$$r(x,y)=\frac{1}{n}\sum_{i=1}^{n}\frac{(E(x)-x_i)\cdot(E(y)-y_i)}{\sigma(x)\cdot\sigma(y)} \qquad (3)$$

на ее разносный аналог:

$$r11=\sum_{i=0}^{14}\frac{(x_{i+1}-x_i)\cdot(y_{i+1}-y_i)}{\sigma(x)\cdot\sigma(y)\cdot 15} \qquad (4)$$

На рис. 6 отражены результаты численного эксперимента.

Классический критерий Пирсона дает PEE≈0.25, что значительно лучше его разносного аналога PEE≈0.425. Корреляция откликов этих двух нейронов значительна corr(r, r11)≈0.82498. Снизить показатель корреляционной сцепленности удается путем увеличения интервала разностных данных:

$$r12=\sum_{i=0}^{11}\frac{(x_{i+4}-x_i)\cdot(y_{i+4}-y_i)}{\sigma(x)\cdot\sigma(y)\cdot 12} \qquad (5)$$

Этот прием снижает корреляционную сцепленность до величины corr(r, r12)≈0.77155 при практическом сохранении вероятностей ошибок первого и второго рода PEE≈0.436.

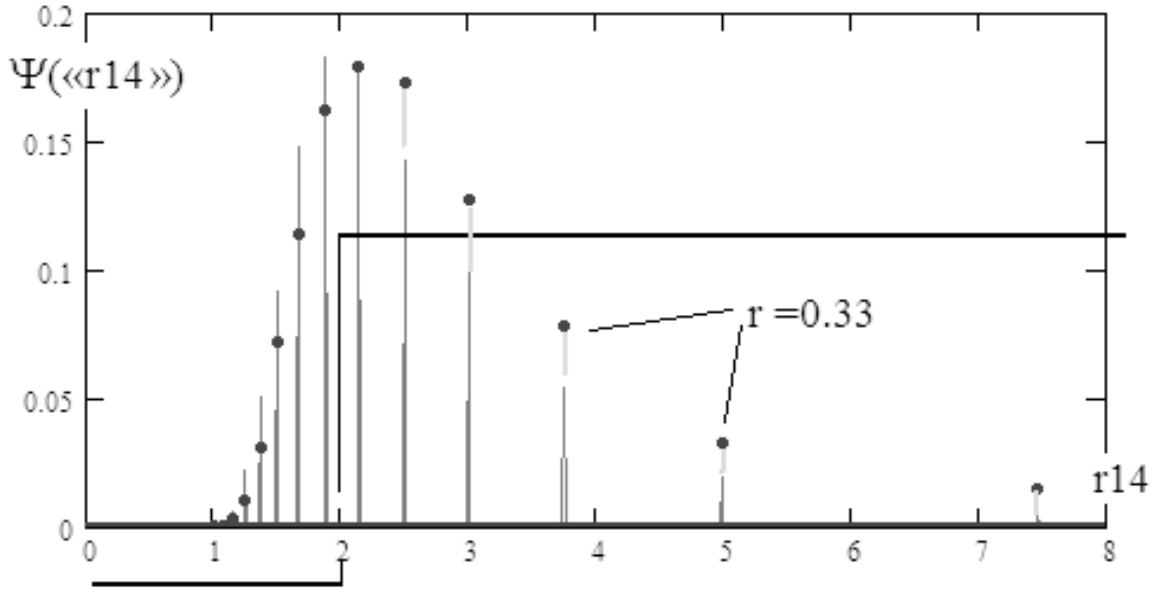



Рис. 8. Дискретный спектр 11 наиболее значимых спектральных линий амплитуд вероятности с гиперболическим расширением расстояний между спектральными линиями

Значительного снижения показателей коррелированности удается добиться переходом к дискретным аналогам разностных преобразований (4) и (5). Для этой цели знаки суммируемых данных формулы (2), рассматриваются как положительный и отрицательный ранги. Если мы будем рассматривать в качестве статистики сумму отрицательных рангов, то мы получим спектральные линии, отраженные на рис. 7.

Классический коэффициент корреляции Пирсона и его дифференциально-дискретный аналог практически существенно меньше коррелированы (corr(r,«r13»)≈-0.60871), чем разностные корреляции Пирсона. При этом вероятности ошибок первого и второго рода для этого класса преобразований оказывается применимы PEE≈0.45 на практики.

Снизить показатель корреляционной сцепленности примерно в 10 раз до величины corr(r,«r14»)≈0.05774, если данные линейного пространства спектральных линий «r13» деформировать гиперболой 15/(s+0.01). Соответствующий этой деформации спектр приведен на рис. 8.

Состояние спектров состояний метрики «r9» (рис. 4) и спектров состояний метрики «r14» (рис. 8) похожи, но не совпадают.

## Заключение

Хи-квадрат критерий Пирсона в 1900 создавался как континуальный критерий, однако уже в 1973 году [1], стала очевидной возможность перехода к дискретной форме его описания для малых выборок. Сегодня дискретно-континуальные формы описания $(x) \Leftrightarrow \Psi(\text{"x"})$ случайностей построены более чем для десятка статистических критериев. Принципиально важна не сама процедура перехода от одной формы описания случайности к другой, а принципиально разные свойства данных представленных в двух разных формах. Необходимо научиться извлекать пользу из возможности дуального представления случайности в двух разных, но дополняющих друг друга формах.

ЛИТЕРАТУРА

1. Dahiya R.C., Gurland J. How many class in the Pearson hi-square test? //Journal of the American Statistical Association. 1973. V.68, № 303, p.p. 707–712.
2. Ахметов Б.Б., Иванов А.И., Фунтикова Ю.В. Дискретный характер закона распределения хи-квадрат критерия для малых тестовых выборок //Вестник Национальной академии наук Республики Казахстан. 2015. № 1. С. 17–25.
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. Zolotareva T.A. Statistical Characteristics of Decisions Made by a Neural Network Molecule with Quadrant Quantization and a Molecule with Data Quantization by Two Ellipses / T.A. Zolotareva, A.I. Ivanov // Software Engineering Perspectives in Intelligent Systems Proceedings of 4th Computational Methods in Systems and Software 2020, Vol.1 — Q3.— ISSN: 2194–5357 (electronic), p. 829–835.
5. 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. /Сборник научных статей по материалам I Всероссийской научно-технической конференции «Безопасность информационных технологий», 24 апреля, Пенза 2019, с. 156–164.
6. 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.


Журнал «Современная наука: актуальные проблемы теории и практики»

Г. Липецк