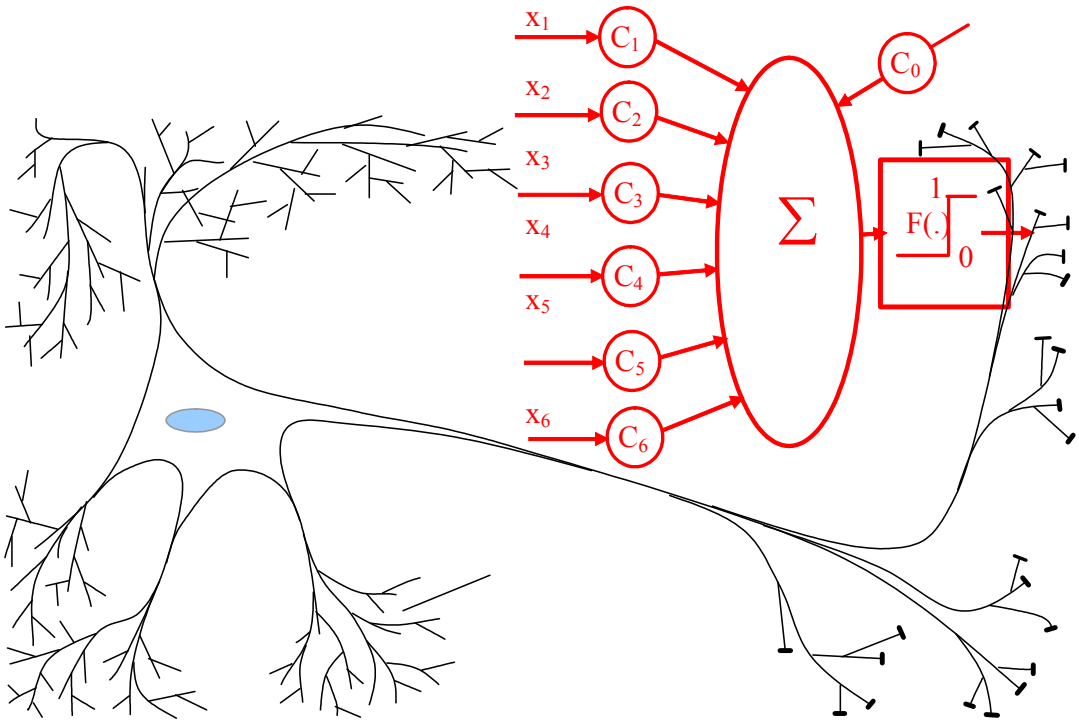
УДК: 519.24; 53; 57.017
И 18
ББК 32.818

Издательство АО «ПНИЭИ»,
электронный вариант пособия размещен на сайте
http://пниэи.рф/activity/science/noc/BOOK18.pdf

Иванов Александр Иванович (Пенза)

ПРОСТЕЙШИЕ ОРАКУЛЫ, ОБУЧЕННЫЕ КОРРЕКТИРОВАТЬ ОШИБКИ ВЫЧИСЛЕНИЯ МЛАДШИХ СТАТИСТИЧЕСКИХ МОМЕНТОВ НА МАЛЫХ ВЫБОРКАХ БИОМЕТРИЧЕСКИХ ДАННЫХ

Учебное пособие



Пенза -2018

УДК: 519.24; 53; 57.017
И 18
ББК 32.818

Рецензенты:
Доктор техн. наук, проф. М.М. Бутаев - ученый секретарь НТС АО НПП «Рубин» 440000, Россия, г. Пенза, ул. Байдукова, 2

Доктор техн. наук, проф. С.И. Геращенко – заведующий кафедрой «Медицинская кибернетика и информатика» ФГБОУ ВО «Пензенский государственный университет» 440000, Россия, г. Пенза ул. Красная, 40.

Иванов А.И.
Простейшие оракулы, обученные корректировать ошибки вычисления младших статистических моментов на малых выборках биометрических данных. Учебное пособие. Пенза – 2018 г. Издательство АО «Пензенский научно-исследовательский электротехнический институт» (ОА «ПНИЭИ») – 35 с. http://пниэи.рф/activity/science/noc/BOOK18.pdf

Изложены теоретические и практические аспекты синтеза, настройки, тестирования простейших оракулов, ориентированных на предсказание интервалов ошибок младших статистических моментов малых выборок (математического ожидания, стандартного отклонения, коэффициентов парной корреляции). Рассматривается аналоговая (континуальная) реализация простейших оракулов, дается критерий оценки качества их работы для нормального закона распределения значений и хи-квадрат распределения Пирсона. Каждый студент сегодня имеет возможность написать программу из пяти строк кода и самостоятельно построить семейство хи-квадрат распределений Пирсона для малой выборки. Такой подход важен тем, что позволяет убедиться в необходимости учета квантовых эффектов, при статистической обработке малых выборок. Даны конструкции, позволяющие это выполнять (математическая хи-квадрат молекула, молекула математического ожидания, молекула стандартного отклонения, корреляционная молекула).

Рассмотренные в пособии процедуры настройки и тестирования простейших оракулов иллюстрируются программными приложениями, написанными в среде инженерных расчетов MathCAD.

Учебное пособие рассчитано на студентов, аспирантов, преподавателей и научных работников, занимающихся обработкой больших объемов «плохих» данных, представленных малыми выборками большого числа контролируемых параметров (например, биометрических данных).

Оглавление стр.

Введение

Информатизация современного общества приводят к необходимости расширения применения криптографии. Обычные люди не могут запоминать длинные пароли доступа и криптографические ключи. Для решения этой проблемы в США и Евросоюзе развиваются технологии «нечетких экстракторов» [1, 2, 3], построенных на корректировке ошибок классическими кодами с обнаружением и исправлением ошибок. При этом выходной код «нечетких экстракторов» является коротким из-за того, что классические коды с приемлемой избыточностью в 50% способны корректировать не более 5% ошибок [4, 5]. Ошибки исходных кодов «нечетких экстракторов» могут составлять от 20% до 30% от длины кода, что заставляет использовать самокорректирующиеся коды с 20-ти и 30-ти кратной избыточностью. То есть длина выходного кода «нечеткого экстрактора» оказывается в 20 или 30 раз меньше, чем число биометрических параметров, из которых «нечеткий экстрактор» восстанавливает код ключа (рисунок 1).

В России развивается технология нейросетевого преобразования биометрии в длинный код доступа или длинный код личного криптографического ключа [5, 6]. Нейросетевые преобразователи биометрия-код, выполненные в соответствии с пакетом стандартов ГОСТ Р 52633.xx [7, 8], обучаются на выборках порядка 20 примеров образа «Свой». При этом вероятность ошибок первого рода (отказ в доступе «Своему») составляет от 0.05 до 0.1. Для снижения вероятности ошибок первого рода необходимо либо увеличивать число примеров в обучающей выборке, либо осуществлять коррекцию ошибок [9, 10]. Для нейросетевых преобразователей биометрия-код нет проблемы коротких выходных кодов, так как за один разряд кода отвечает один нейрон, а число нейронов выбирает разработчик биометрического средства защиты информации (рисунок 1).

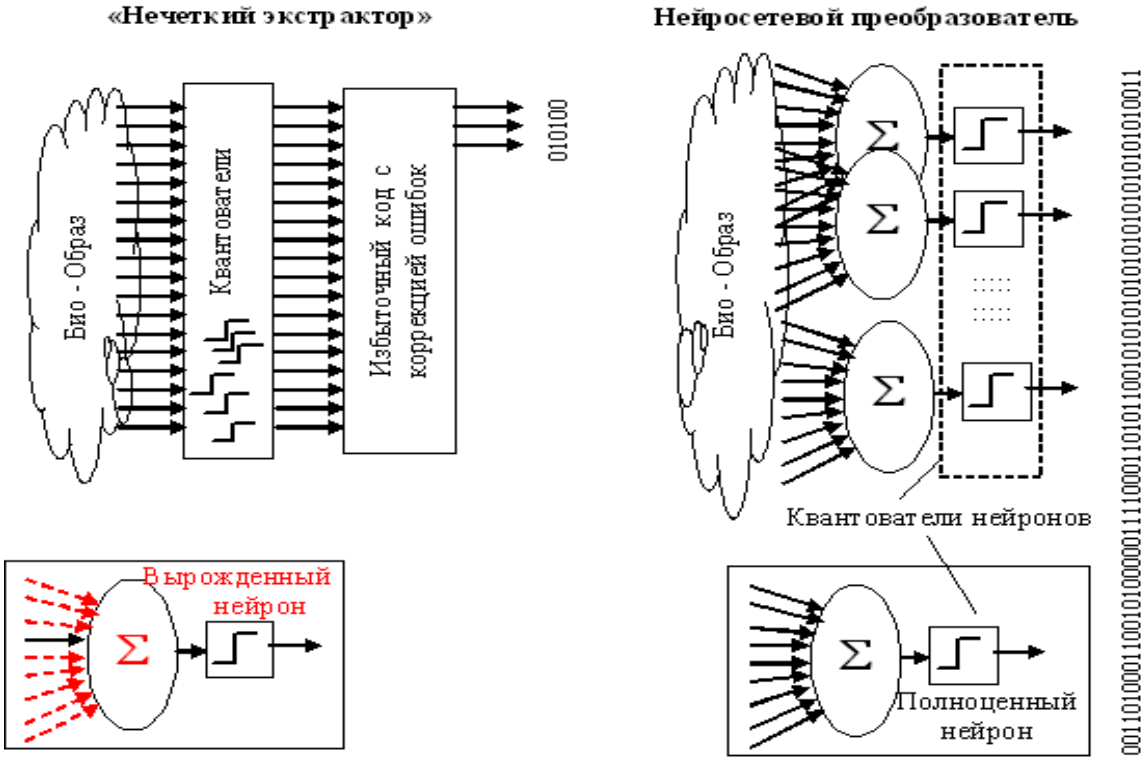


Рис. 1. «Нечеткие экстракторы» как частный случай нейросетевого преобразователя биометрия-код

Получается, что «нечеткие экстракторы» примитивны, но они и не прихотливы к данным, на которых обучаются. Нейросетевые преобразователи биометрия-код гораздо более интеллектуальны, многократно превосходят «нечеткие экстракторы» по всем параметрам, однако они более требовательны к объему и качеству данных в обучающей выборке. Кроме того, нейронные сети более требовательны к качеству предварительной обработки биометрических данных при извлечении из них биометрических параметров.

Еще одним важнейшим условием применения больших искусственных нейронных сетей является то, что они должны быстро автоматически обучаться алгоритмом с низкой вычислительной сложностью. Первый такой алгоритм был создан и стандартизован в России [8], он имеет линейную вычислительную сложность (ниже не бывает). Предельно низкая вычислительная сложность стандартизованного алгоритма обучения ГОСТ Р 52633.5 [8] обусловлена тем, что этот алгоритм не является итерационным. Весовые коэффициенты нейронов вычисляют как простую функцию младших статистических моментов двух переменных:

$$|\mu_i| = f(E(\nu_i), E(\xi_i), \sigma(\nu_i), \sigma(\xi_i)) \qquad (1),$$

где ν_i - биометрический параметр образа «Свой», ξ_i - биометрический параметр образа «Чужой», $E(.)$ - оператор вычисления математического ожидания, $\sigma(.)$ - оператор вычисления стандартного отклонения.

Ожидается, что следующее поколение стандартизованных нейросетевых преобразователей биометрия-код будет автоматически обучаться алгоритмом, имеющим квадратичную вычислительную сложность за счет дополнительного учета одинаковых корреляционных связей [11, 12, 13] биометрических данных одного нейрона:

$$|\mu_i| = f(E(\nu_i), E(\xi_i), \sigma(\nu_i), \sigma(\xi_i), r_i(\nu_i, \nu_j)) \qquad (2),$$

где $r_i(\nu_i, \nu_j) = r_i(\nu_i, \nu_{j+1}) = r_i(\nu_i, \nu_{j+2}) = ...$ одинаково коррелированные биометрические данные одного нейрона, полученные специальной процедурой симметризации корреляционных связей.

На ошибку вычисления весовых коэффициентов нейронов по формуле (1) влияют только ошибки статистических моментов биометрических данных образа «Свой»:

$$\Delta|\mu_i| = f(\Delta E(\nu_i), \Delta\sigma(\nu_i)) \qquad (3).$$

В силу того, что статистические моменты для образов «Чужие» вычисляются заранее по большим выборкам, ошибки вычисления их младших статистических моментов можно считать нулевыми:

$$\begin{cases} \Delta E(\xi_i) = 0 \\ \Delta\sigma(\xi_i) = 0 \end{cases} \qquad (4).$$

Совершенно иная ситуация возникает при оценке математического ожидания - $E(\nu_i)$ и стандартного отклонения - $\sigma(\nu_i)$ биометрических данных образа «Свой». К сожалению, при вычислениях статистических моментов на малых выборках ошибка вычисления может оказаться сопоставимой с самим значением статистических моментов:

$$\begin{cases} |\Delta E(\nu_i)| \approx |E(\nu_i)| \\ |\Delta\sigma(\nu_i)| \approx \sigma(\nu_i) \end{cases} \qquad (5).$$

В связи с этим, возникает необходимость создания оракулов, которые способны предсказывать значения интервалов ошибок $\Delta E(\nu_i)$ и $\Delta\sigma(\nu_i)$. Зная

значения ошибок вычислений, можно скорректировать обучающую выборку, увеличив ее или устранив из нее грубую ошибку.

Подобная постановка задачи уже актуальна сегодня при использовании алгоритма обучения ГОСТ Р 52633.5 [8]. Актуальность задачи только усилится в будущем, когда будут использоваться для обучения больших нейронных сетей алгоритмы с квадратичной вычислительной сложностью, построенные на вычислении коэффициентов парной корреляции. Проблема состоит в том, что при вычислении коэффициентов корреляции накапливаются ошибки младших статистических моментов двух биометрических параметров. То есть функция ошибок коэффициента корреляции четырехмерна:

$$\Delta r(\nu_i, \nu_j) = f(\Delta E(\nu_i), \Delta\sigma(\nu_i), \Delta E(\nu_j), \Delta\sigma(\nu_j)) \qquad (6).$$

Фактически перспективные алгоритмы быстрого обучения больших искусственных нейронных сетей должны иметь в своем составе оракулов, оценивающих ошибки вычислений коэффициентов корреляции (6) на малых обучающих выборках. То есть рассматриваемые в данном учебном пособии оракулы из экзотики сегодняшнего дня в будущем станут важной составляющей алгоритмов обучения.

Еще одним фундаментальным моментом нового подхода к классической статистике является наличие у каждого из нас под рукой значительных вычислительных ресурсов. У Пирсона в 1900 году не было такой возможности, и потому он вынужден был аналитически рассматривать предельный случай объема выборки стремящейся к бесконечности. У нас с вами совершенно иная ситуация, сегодня каждый студент имеет возможность самостоятельно построить хи-квадрат распределение для нужной ему выборки малого объема. К сожалению, этому не учат сегодня в высшей школе. Причиной этому является то, что на малых выборках хи-квадрат распределения кардинально меняют свои свойства, из непрерывного распределение становится дискретным [14, 15, 16, 17]. Настоящее учебное пособие является первым по конструктивному статистическому моделированию квантовых эффектов, наблюдаемых на малых тестовых выборках.

1. Оракулы, настроенные на предсказание интервалов ошибок, при вычислении математического ожидания на малых выборках биометрических данных

1.1. Численное определение интервалов ошибок вычисления математического ожидания

Очевидно, что первый статистический момент или математическое ожидание выборки из –n опытов является одной из самых важных статистических характеристик:

$$E(\nu) = \frac{1}{n}\sum_{i=1}^{n} \nu_i \qquad (6).$$

Интуитивно понятно, что ошибка вычисления математического ожидания стремится к нулю $\Delta E(\nu) \to 0$ при монотонном росте объема выборки $n \to \infty$. По классической статистике скорость снижения интервала ошибок должна быть пропорциональна $\sqrt{n}$ для независимых данных. Определим самостоятельно интервал ошибок вычисления математического ожидания для выборки в 11 опытов. Для этой цели необходимо написать и запустить программу из 7 строк (смотри рисунок 2):

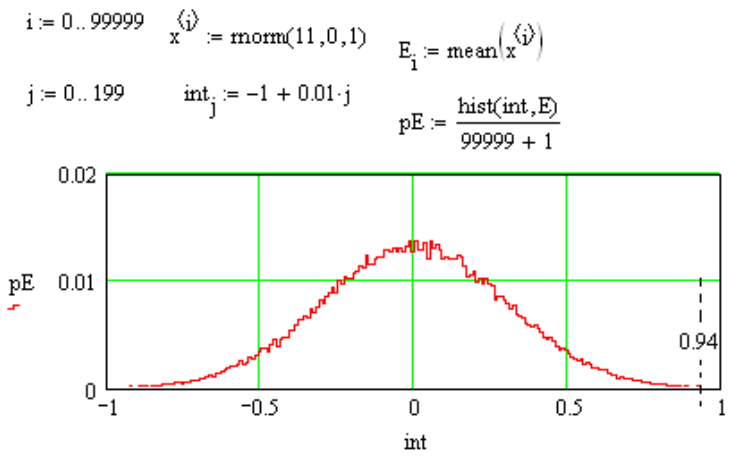


Рис. 2. Гистограмма распределения ошибок ΔE для выборок из 11 опытов

Из рисунка 2 видно, что математическое ожидание математических ожиданий центрированных и нормированных выборок всегда оказывается нулевым $E(E_i)=0$, а интервал ошибок составляет $\Delta E = 0 \mp 0.94$.

Ситуация обучения больших искусственных нейронных сетей на выборке из 11 примеров вполне возможна, например, в среде моделирования «БиоНейроАвтограф» [18] режим обучения запускается для выборок в 8 примеров рукописного образа. Нейросеть хорошо обучается на малых выборках простых рукописных образов, для ее обучения на сложных рукописных образах требуется порядка 21 примера. На рисунке 3 дано распределение для центрированных и нормированных данных 100 000 выборок по 21-му примеру.

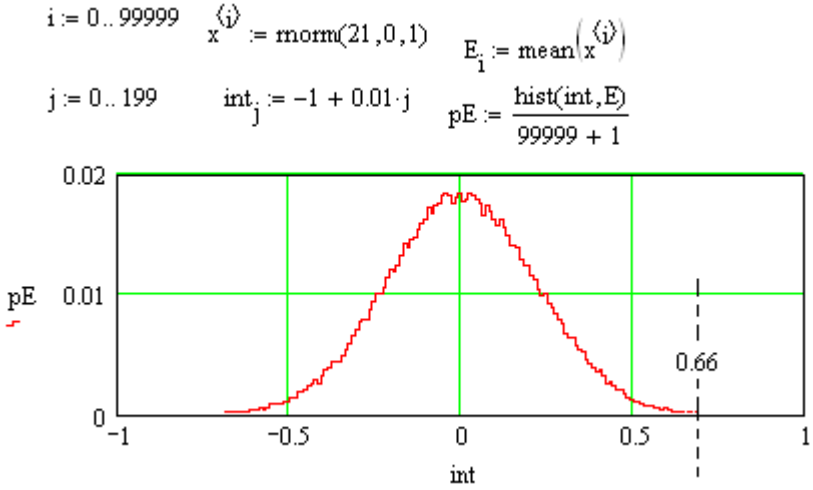


Рис. 3. Гистограмма распределения ошибок ΔE для выборок из 21-го опыта

Из рисунка 3 видно, что почти двукратное увеличение объема тестовой выборки приводит к сужению интервала ошибок примерно в $\sqrt{2} \approx 1.41$ раза. Последнее означает, что для центрированных и нормированных данных ошибка вычисления математического ожидания становится приемлемой только при выборках в сотни примеров. На рисунке 4 приведена гистограмма, построенная для выборок в 500 примеров.

$i := 0..99999 \quad x^{\langle i \rangle} := rnorm(500,0,1) \quad E_i := mean\left(x^{\langle i \rangle}\right)$

$j := 0..199 \quad int_j := -1 + 0.01 \cdot j \quad pE := \frac{hist(int,E)}{99999+1}$

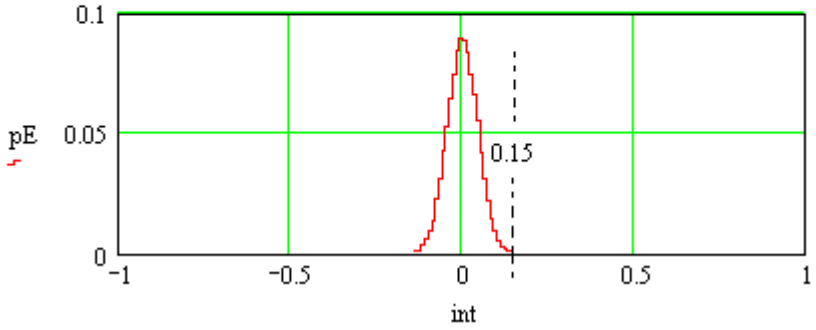


Рис. 4. Гистограмма распределения ошибок ΔE для выборок из 500 опытов

Из рисунка 4 видно, что при выборках в 500 опытов интервал распределения ошибок составляет $\Delta E = 0 \mp 0.15$, мы наблюдаем сжатие интервала $\sqrt{45.5} \approx 6.7$ раза по сравнению с данными рисунка 2.

1.2. Настройка параметров оракула, предсказывающего интервал ошибок вычисления математического ожидания

В связи с тем, что для нормальных независимых данных скорость сжатия интервала ошибок известна - $\sqrt{n}$, настройка параметров оракула примитивна:

$$\Delta E \approx \pm 0.98 \sqrt{\frac{10}{n}} \qquad (7).$$

Корректность работы оракула (7) может быть проверена выполнением соответствующей лабораторной работы в среде математического моделирования MathCAD. Итогом выполнения лабораторной работы является таблица № 1 значений интервалов ошибок, полученная для разных объемов выборки.

Таблица № 1 Значения интервалов -ΔE ошибок вычисления математического ожидания E=0 в зависимости от объема выборки - n

№	0	1	2	3	4	5	6	7	8	9	10	11	12
n	5	6	7	8	9	10	11	12	13	14	15	16	17
ΔE	1.09	1.08	1.08	1.06	1.04	0.98	0.94	0.9	0.83	0.81	0.8	0.78	0.77

№	13	14	15	16	17	18	19	20	21	22	23	24	25
n	18	19	20	21	24	26	28	30	32	36	38	40	42
ΔE	0.73	0.69	0.67	0.66	0.62	0.6	0.56	0.55	0.53	0.52	0.49	0.48	0.47

Если закон распределения данных не является нормальным, а так же если данные в выборке зависимы, то необходимо заранее построить для реальных данных таблицу по аналогии с таблицей №1. Затем следует по данным реальной таблицы найти 2 настраиваемых параметра оракула:

$$\Delta E \approx \pm C_{10} \cdot \left\{ 2^{-Q} \sqrt{\frac{10}{n}} \right\} \qquad (8),$$

где C_{10} полуинтервал ошибок для выборки в 10 примеров, Ω - показатель снижения скорости сжатия интервалов ошибки из-за наличия корреляционных связей в исходных данных, показатель - Ω может находится в интервале от 0 до 2.

В первом приближении показатель снижения скорости пропорционален среднему значению модулей коэффициентов корреляции данных в исследуемой выборке:

$$\Omega \approx 2 \cdot E(|r|) \qquad (9).$$

При отсутствии корреляции $r = 0$ между данными, мы имеем скорость убывания интервала пропорционально $\sqrt{n}$. Если корреляционные связи значительны $|r| \approx 1$, то увеличение тестовой выборки вообще не приводит к сжатию интервала ошибок (скорость сжатия нулевая).

1.3. Учет влияния ошибок стандартного отклонения нормально распределенной выборки биометрических данных

Выше был показан общий принцип синтеза оракулов, предсказывающих ширину интервала допустимых вариаций математического ожидания. При этом мы использовали образцовый программный генератор псевдослучайных чисел. Ради простоты и доходчивости пришлось забыть, что ошибку имеет не только операция вычисления математического ожидания, но и операция вычисления стандартного отклонения. Тем не менее, это неоспоримый факт, при синтезе оракулов необходимо учитывать ошибку вычисления стандартного отклонения $\Delta\sigma$ на малых выборках.

Одним из самых простых способов сделать это является переход к нормированной системе координат, где стандартное отклонение будет всегда единичным. На рисунке 4 приведены программа моделирования данных и графики распределения значений ошибок вычисления математических ожиданий для малых выборок в 45, 20, 10, 5 примеров.

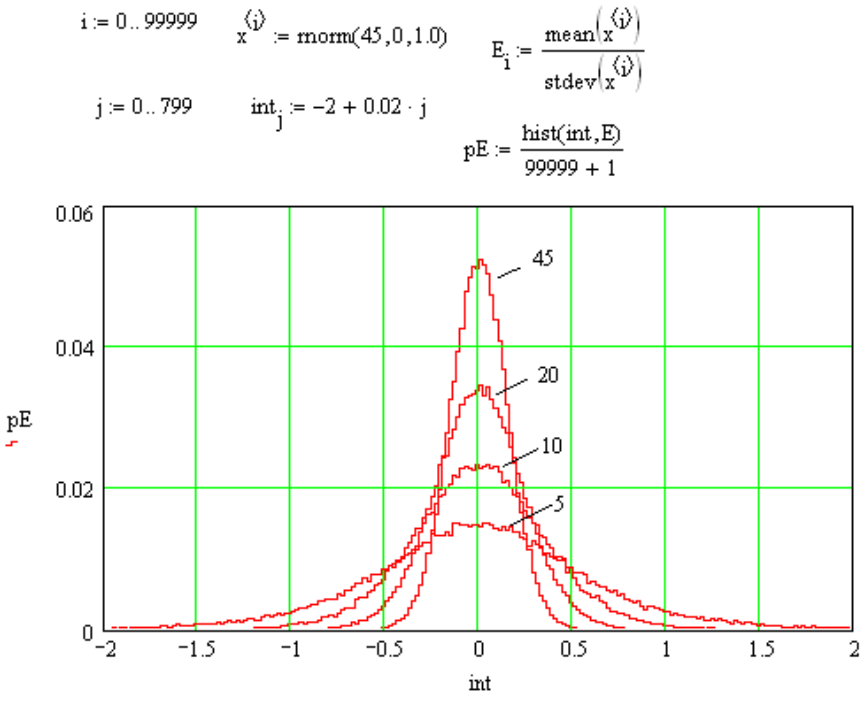


Рис. 5. Распределение нормированных по стандартному отклонению интервалов оценки положения математических ожиданий, вычисленных по малым выборкам 5, 10, 20, 45 примеров

Численные значения интервалов ошибки приведены в таблице 2.

Таблица №2

№	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n	8	9	10	11	12	13	14	15	16	17	18	19	20	25	30	35	40	45
ΔE	1.68	1.6	1.5	1.28	1.24	1.18	1.14	1.06	1.02	0.94	0.92	0.9	0.86	0.74	0.66	0.62	0.65	0.52

Следует подчеркнуть, что данные в таблице 2 примерно в полтора раза хуже, чем данные таблицы 1.

Пользуясь данными таблицы 2, мы можем вычислять интервалы положения математического ожидания для данных с любым стандартным отклонением:

$$\Delta E(v) \approx \pm 1.5 \cdot \sigma(v) \cdot \sqrt{\frac{10}{n}} \qquad (10).$$

Интервал предсказания ошибок расширяется примерно в полтора раза из-за того, что в формуле 10 появляется не точный сомножитель. Мы наблюдаем полуторократное снижение числа обусловленности процедуры вычисления (10) по сравнению с вычислением интервала при точно известном стандартном отклонении (7).

2. Оракулы, настроенные на предсказание интервалов ошибок, при вычислении стандартного отклонения на малых выборках биометрических данных

2.1. Учет методической ошибки вычисления стандартного отклонения на малых выборках биометрических данных

Классическая литература по математической статистики [19, 20] рекомендует вычислять стандартное отклонение по следующей формуле:

$$\sigma(v) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(E(v) - v_i\right)^2} \qquad (11).$$

Необходимо отметить, что формула (11) применима только при больших выборках биометрических данных. На малых выборках она работает некорректно из-за значительной асимметрии распределения значений стандартных отклонений. На рисунке 6 даны примеры распределения значений стандартных отклонений, вычисленных по формуле (11) на малых выборках 3, 5, 7, 11, 15, 20, 30, 40, 60, 80 примеров.

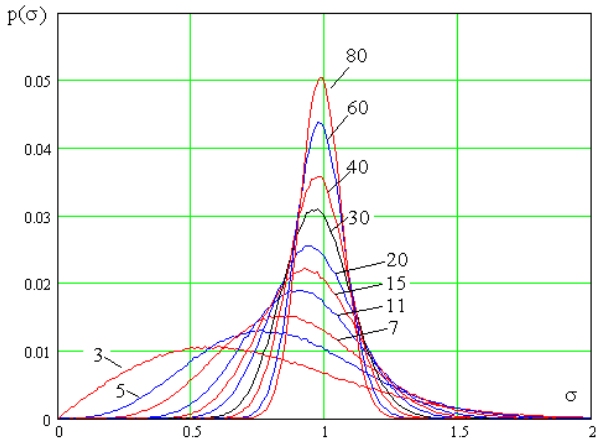


Рис. 6. Примеры асимметричных распределений стандартного отклонения на малых выборках малого объема

Из рисунка 6 видно, что при 60 и 80 опытах распределение значений стандартного отклонения становится почти нормальным, то есть формула (11) для таких выборок и выборок большего размера работает корректно. Иначе обстоит дело с выборками меньшего объема [21]. Для выборок с объемом от 3 до 50 примеров распределения данных оказываются сильно асимметричными. Возникает методическая погрешность вычислений, которую следует учитывать, применяя формулу с корректирующим мультипликативным членом:

$$\sigma(v) = \left\{1.003 + \frac{1}{n}\right\} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(E(v) - v_i\right)^2} \qquad (12).$$

2.2. Оракул, предсказывающий интервалы ошибок вычисления стандартных отклонений на малых выборках биометрических данных

Для того, чтобы построить оракула, предсказывающего интервалы ошибок вычисления стандартного отклонения, необходимо написать и запустить программу из 7 строк, приведенную в верхней части рисунка 7.

$i := 0..99999 \qquad x^{\langle i \rangle} := rnorm(5,0,1.0) \qquad \sigma_i := stdev\left(x^{\langle i \rangle}\right)$

$j := 0..799 \qquad int_j := 0 + 0.02 \cdot j \qquad p\sigma := \frac{hist(int,\sigma)}{99999 + 1}$

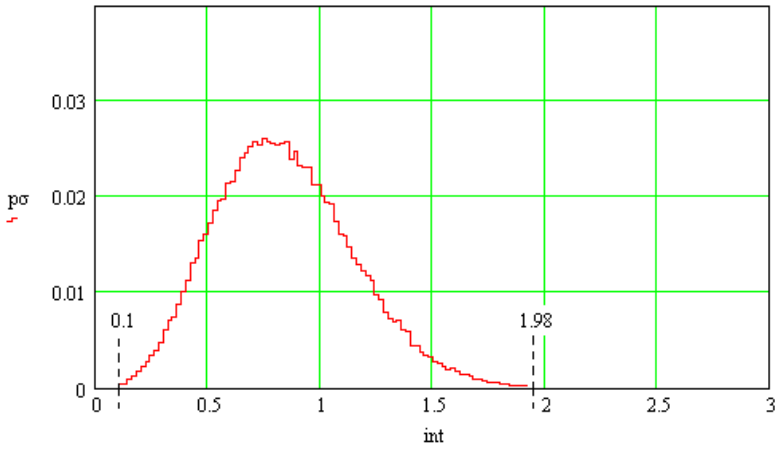


Рис. 7. Интервал неопределенности вычисления стандартного отклонения на малой выборке в 5 опытов

Глядя на нижнюю часть рисунка 7, не трудно определить правую и левую границы интервала неопределенности вычисления стандартного отклонения. Для различных размеров тестовой выборки данные приведены в таблице №3.

Таблица № 3.

№	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
n	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
min(σ)	0.1	0.14	0.2	0.2	0.3	0.28	0.34	0.34	0.38	0.4	0.4	0.44	0.44	0.48	0.48	0.5	0.5	0.5	0.5
max(σ)	1.98	1.9	1.84	1.76	1.74	1.72	1.7	1.68	1.62	1.62	1.58	1.54	1.56	1.56	1.54	1.54	1.52	1.5	1.48

Если сравнивать данные таблицы № 1 и таблицы № 3, то легко заметить значительное расширение интервалов ошибок стандартного отклонения по сравнению с интервалом неопределенности вычисления математического ожидания.

2.3. Учет влияния накапливания ошибок при вычислении стандартного отклонения после приближенного вычисления математического ожидания

В формулу (11) входит значение математического ожидания, вычисляемого с некоторой погрешностью ΔE. В разделе 1.3 (формула (10)) показано, что ошибка вычисления стандартного отклонения Δσ приводит к увеличению ширины интервала ошибок ΔE в полтора раза. Если считать, что ошибка ΔE приводит к похожему сдвигу интервалов неопределенности Δσ, то мы получим следующие корректировки:

$$\begin{cases} \min(\sigma(n), \Delta E) = \dfrac{\min(\sigma(n))}{1.5} \\ \max(\sigma(n), \Delta E) = \max(\sigma(n) \cdot 1.5 \end{cases} \qquad (13).$$

Очевидно, что система преобразований (13) является некоторым приближением, однако этого простого приближения вполне достаточно для инженерных расчетов.

3. Оракулы, настроенные на предсказание интервалов ошибок, при вычислении коэффициентов корреляции на малых выборках биометрических данных

Классическая литература по математической статистики [19, 20] рекомендует вычислять коэффициент корреляции по формуле Пирсона:

$$r(v_i, v_j) = \frac{1}{n} \sum_{k=1}^{n} \frac{(E(v_i) - v_{i,k}) \cdot (E(v_j) - v_{j,k})}{\sigma(v_i) \cdot \sigma(v_j)} \qquad (11).$$

Из-за того, что стандартные отклонения и математические ожидания, входящие в формулу (11) на малых выборках вычисляются с ошибками и эти ошибки накапливаются, коэффициенты корреляции вычисляются с низкой точностью. На рисунке 8 приведены кривые распределения ошибок для выборок объемом 7, 9, 16, 21 опыт.

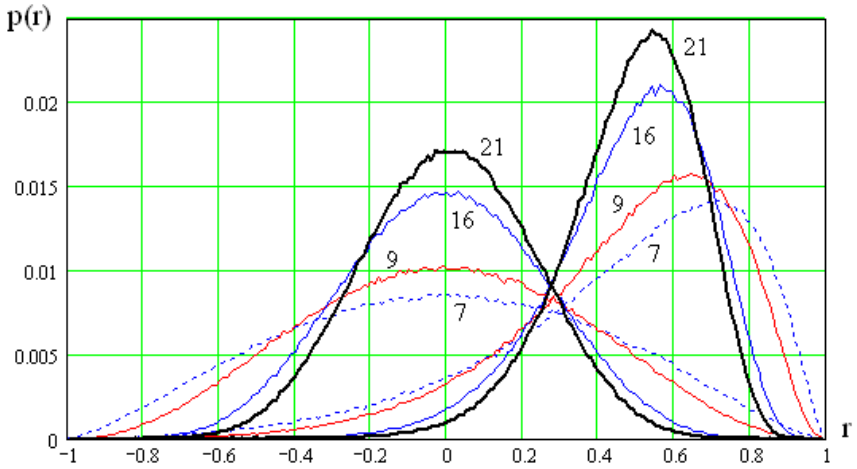


Рис. 8. Распределения значений коэффициентов корреляции, для выборок из 7, 9, 16, 21 примеров при двух заданных значениях коэффициентов корреляции r = 0 и r = 0.5 у программных генераторов случайных чисел

Для того, что бы получить распределения зависимых случайных данных от двух независимых программных генераторов использовалась программа, приведенная в верхней части рисунка 9.

В этой программе следует задавать параметр связанности $a := 10.0$, который однозначно связан со значением коэффициента корреляции независимо от объема выборки. В таблице №4 даны соотношения параметра связывания и итогового среднего значения коэффициентов корреляции.

Таблица № 4.

№	0	1	2	3	4	5	6	7	8	9	10
a	10	2.98	2.01	1.53	1.22	1	0.82	0.65	0.5	0.33	0.13
r	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99

$i := 0..99999$ $n := 10$ $a := 10.0$

$x0^{(i)} := rnorm(n,0,1)$ $x1^{(i)} := rnorm(n,0,1)$ $z^{(i)} := rnorm(n,0,1)$

$xx0^{(i)} := a \cdot x0^{(i)} + z^{(i)}$ $xx1^{(i)} := a \cdot x1^{(i)} + z^{(i)}$

$r_i := corr\left(xx0^{(i)}, xx1^{(i)}\right)$ $mean(r) = 0.01$

$r^T =$

	0	1	2	3	4	5	6	7	8	9
0	-0.607	-0.136	0.554	-0.294	0.258	0.068	-0.109	0.273	-0.035	0.078

$j := 0..799$ $int_j := -1 + 0.02 \cdot j$ $pr := \frac{hist(int, r)}{99999 + 1}$

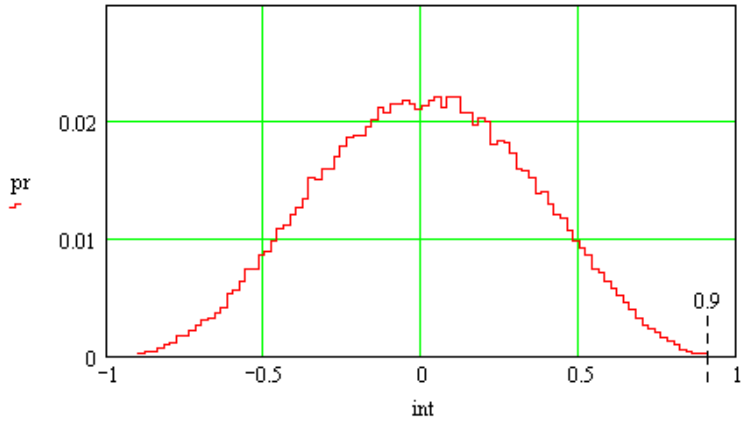


Рис. 9. Моделирование слабо коррелированных данных для выборки из 10 опытов

Для того, что бы получить интервалы ±Δr для различных выборок необходимо менять в программе рисунка 9 параметр связанности $a := 10.0$ и размер выборки $n := 10$. В конечном итоге мы получим симметричные правый и левый интервалы для некоррелированных данных. Эти результаты приведены в таблице №5.

Таблица №5.

№	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
±Δr	0.98	0.96	0.92	0.9	0.88	0.86	0.84	0.82	0.8	0.78	0.76	0.74	0.72	0.7	0.7	0.68	0.68	0.66	0.66

Если данные коррелированны $E(r) = 0.5$, то левая и правая границы престают быть симметричными как это показано на рисунке 10 и в таблице № 6.

Таблица № 6.

№	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
n	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
min(r)	-0.84	-0.74	-0.68	-0.66	-0.56	-0.48	-0.48	-0.46	-0.4	-0.34	-0.32	-0.3	-0.28	-0.26	-0.22	-0.2	-0.18
max(r)	0.98	0.98	0.98	0.98	0.96	0.96	0.96	0.96	0.94	0.94	0.92	0.92	0.92	0.92	0.92	0.9	0.9

$i := 0..99999 \qquad n := 10 \qquad a := 1.0$

$x0^{(i)} := rnorm(n,0,1) \qquad x1^{(i)} := rnorm(n,0,1) \qquad z^{(i)} := rnorm(n,0,1)$

$xx0^{(i)} := a \cdot x0^{(i)} + z^{(i)} \qquad xx1^{(i)} := a \cdot x1^{(i)} + z^{(i)}$

$r_i := corr\left(xx0^{(i)}, xx1^{(i)}\right) \qquad mean(r) = 0.479$

$r^T =$

	0	1	2	3	4	5	6	7	8	9
0	-0.388	0.828	0.72	0.3	0.494	0.918	0.475	0.685	0.279	0.662

$j := 0..799 \qquad int_j := -1 + 0.02 \cdot j \qquad pr := \frac{hist(int, r)}{99999 + 1}$

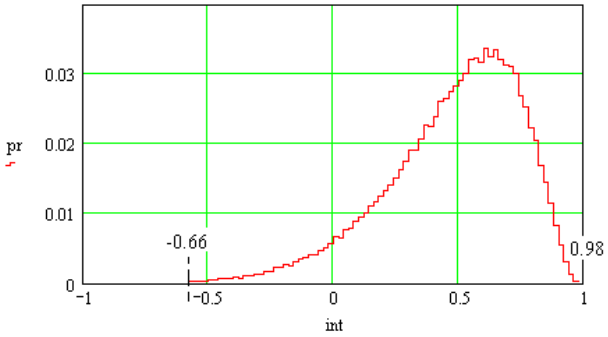


Рис. 10. Моделирование сильно коррелированных данных для определения границ интервала ошибок

4. Оракул, настроенный на вычисление интервалов ошибок хи-квадрат критерия Пирсона для малых выборок

4.1. Аналитическое описание асимптотического хи-квадрат распределения Пирсона

Почему древние обращались за прогнозам к оракулу? Потому, что оракул знал больше, чем обычный человек (у него была возможность накапливать информацию, слушая просьбы других людей и следя за их судьбой). Кроме того, оракул постоянно учился (тренировал свой мозг) учитывая все больше и больше параметров, то есть повышая размерность своих предсказаний. В статистике крайне важно знать вид закона распределения данных. В биометрии данные, как правило, имеют почти нормальное распределение значений с «тяжелыми» хвостами. Оценивая качество биометрических данных целесообразно пользоваться хи-квадрат критерием Пирсона, который анализирует гистограммы реальных данных. Пример такой гистограммы, состоящей из 6 столбцов, построенной для нормальной выборки из 21 опыта приведен на рисунке 11.

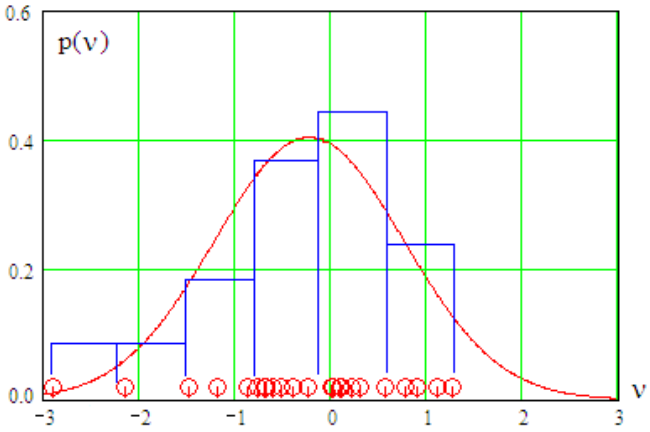


Рис. 11. Гистограмма 21-го примера нормально распределенных данных

Гистограмма строится исходя из заданного числа ее столбцов и наблюдаемого интервала распределения данных. Ширина столбцов гистограммы вычисляется следующим образом:

$$\Delta\nu = \frac{\max(\nu_i) - \min(\nu_i)}{6} \qquad (12).$$

Критерий хи-квадрат для выборки в 21 опыт вычисляется следующим образом:

$$\chi^2 = 21 \cdot \sum_{i=1}^{6} \frac{\left(\frac{n_i}{21} - \widetilde{P}_i\right)^2}{\widetilde{P}_i} \qquad (13),$$

где n_i - число опытов, попавших в i-тый интервал гистограммы, $\widetilde{P}_i$ - вероятность попадания в i-тый интервал гистограммы теоретического распределения значений.

Всеобщая любовь к хи-квадрат критерию обусловлена тем, что для больших выборок его плотность распределения значений имеет аналитическое описание:

$$p(\chi, m) = \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot \chi^{\left(\frac{m}{2}-1\right)} \cdot e^{\frac{-\chi}{2}} \qquad (14),$$

где m – число степеней свободы, зависящее от выбранного числа интервалов гистограммы, Γ(.) - гамма функция Эйлера.

Соотношением (14) можно пользоваться, имея выборку в 400 и более опытов [22]. В этом следует убедиться, написав и запустив программу, приведенную в верхней части рисунка 12.

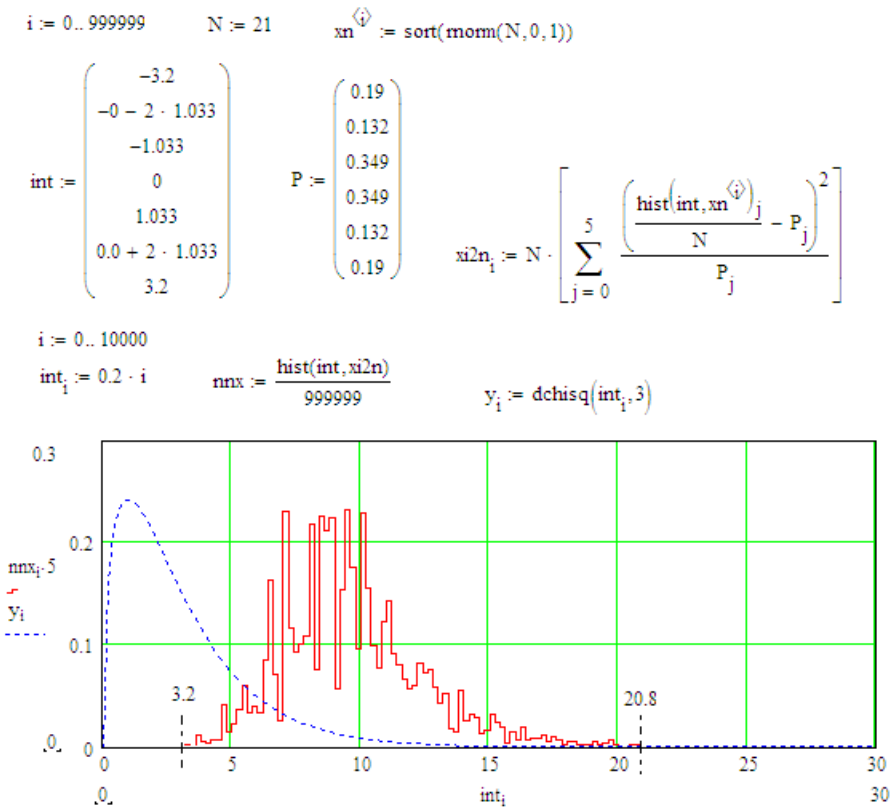


Рис. 12. Распределение значений хи-квадрат критерия при использовании гистограммы с 6 интервалами для выборок объемом в 21 опыт

Из рисунка 11 видно, что значение хи-квадрат критерия с вероятностью близкой к единице будет находится в интервале от 3.2 до 20.8. Значения для нижней и верхней границ при других объемах выборки для гистограмм с 6 столбцами приведены в таблице №7.

Таблица № 7.

№	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
n	8	9	10	11	12	13	14	15	16	17	18	19	20	21	50	100	200	400
$\min(\chi^2)$	0.8	1	1.2	1.6	1.8	1.8	1.8	2.2	2.2	2.2	2.6	2.8	3	3.2	9.6	24.2	52.8	110.8
$\max(\chi^2)$	15.8	17.4	17.4	17.4	18.8	18.8	18.6	21.4	20.4	19.6	20.6	21.6	21.8	20.8	30.4	44.2	77.6	142.6

Методически вычисления интервала неопределенности хи-квадрат критерия из-за малого объема выборки ничем не отличаются от вычисления интервалов неопределенности математического ожидания, стандартного отклонения, коэффициентов корреляции. При этом распределение данных очень сильно отличается от предельного (пунктирная линия в нижней части рисунка 12).

Пирсон в 1900 году не имел вычислительной техники и, соответственно, не мог наблюдать распределения хи-квадрат для малых выборках при малом числе столбцов гистограммы. Сегодня положение изменилось, любой студент может под свои нужды построить частную таблицу квантилей вероятности для любой малой выборки.

4.2. Расчет таблицы квантилей доверительной вероятности хи-квадрат распределения Пирсона для малой выборки в 16 опытов при гистограмме из 6 столбцов

Возможности хи-квадрат критерия могут быть усилены, если под любые имеющиеся экспериментальные данные уметь синтезировать соответствующую

таблицу квантилей доверительной вероятности. В верхней части рисунка 13 дана программа, реализующая вычисления.

Для вычисления квантилей доверительной вероятности необходимо в нижней части программы рисунка необходимо подобрать значения переменной k, близкие к двум ячейкам таблицы доверительной вероятности таблицы № 8.

Таблицы №8. Квантили доверительной вероятности
для выборки 16 опытов, 6 столбцов гистограммы

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
P	0.99	0.98	0.97	0.95	0.9	0.8	0.7	0.5	0.3	0.2	0.1	0.05	0.03	0.02	0.01	$5\cdot10^{-3}$
χ^2	3.3	3.45	3.9	4.35	5.05	5.85	6.45	7.55	8.9	9.95	11.05	12.95	14.1	14.95	16.75	18.85

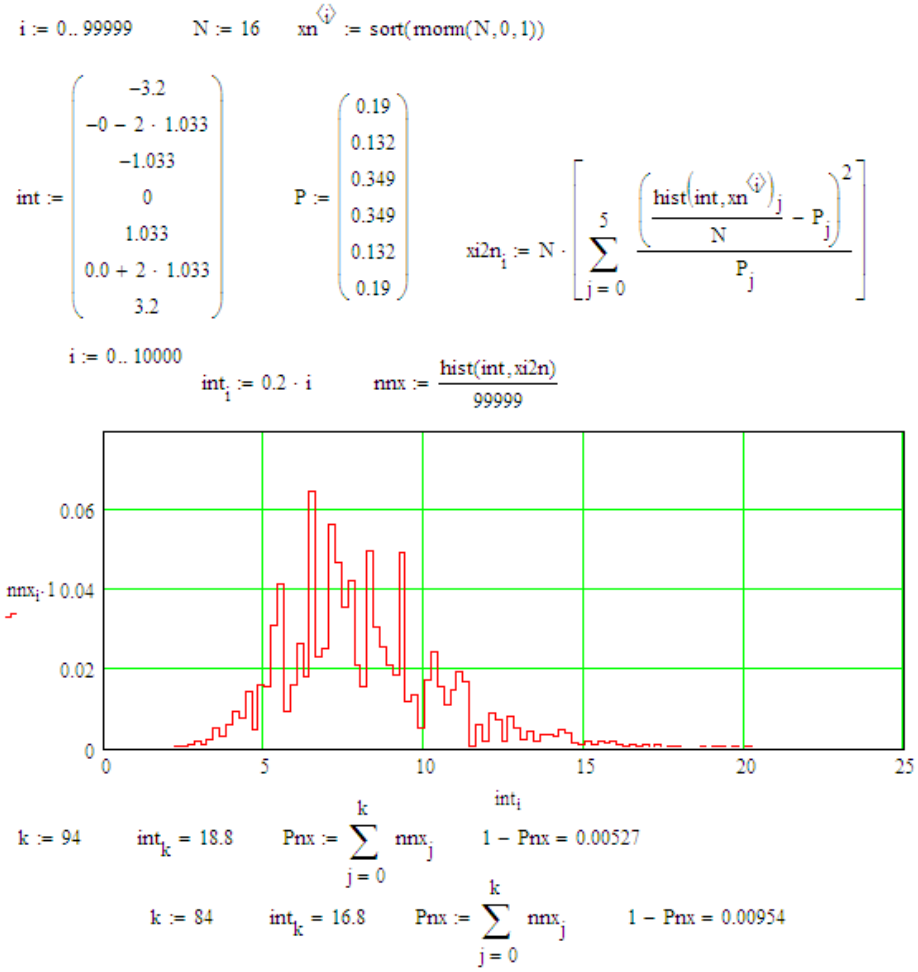


Рис. 13. Программа для вычисления квантилей доверительной вероятности хи-квадрат критерия для выборки в 16 опытов при 6 столбцах гистограммы

При необходимости таблица №8 может быть приближена аналитически по аналогии с формулой (14). Для этой цели необходимо сдвинуть и изменить масштаб хи-квадрат оси. В итоге мы получаем две функции, приведенные на рисунке 14.

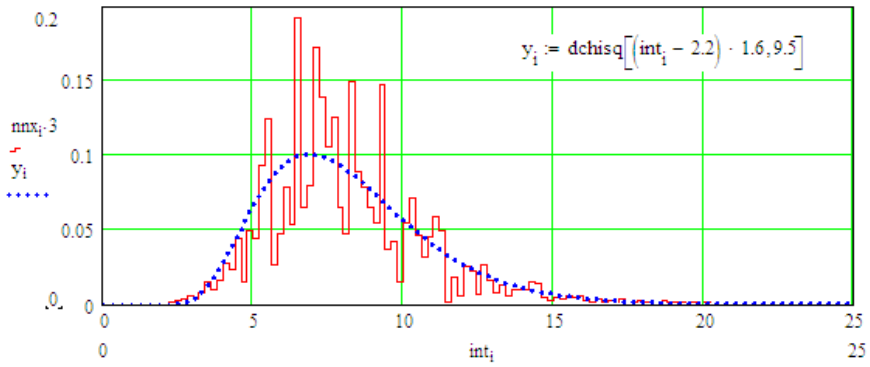


Рис. 14. Приближение распределения хи-квадрат с линейным преобразованием данных по оси абсцисс и дробным числом степеней свободы m=9.5

Существует два пути создания таблиц квантилей доверительной вероятности хи-квадрат распределений малых выборок. Идя по первому пути, мы должны учитывать неровности (мультимодальность) реальных хи-квадрат распределений малых выборок. Иной путь состоит в том, что мы можем сгладить неровности мультимодального распределения хи-квадрат критерия, например, цифровым сглаживающим фильтром. В этом случае, мы получим достаточно гладкое описание хи-квадрат распределения значений, отображенное на рисунке 14 пунктиром.

4.3. Дискретный характер спектра выходных состояний хи-квадрат критерия

Из рисунка 14 видно, что реальное распределение хи-квадрат значений имеет форму «ежика», этот эффект был обнаружен достаточно давно, однако, долгое время он воспринимался исследователями как эффект частичной «случайности» спектра выходных состояний. Только в 2014 году [23] пришло осознание того, что колебательная составляющая хи-квадрат спектра имеет детерминированный характер.

Для того, что бы убедится в этом необходимо многократно выполнить вычисления, по программе, приведенной в верхней части рисунка 13. Затем следует многократно наложить друг на друга полученные рисунки. При этом нестабильная часть данных дает области, достаточно плотно залитые цветом функции рисунка, а стабильная часть будет давать тонкие линии. Этот эффект иллюстрируется рисунком 15.

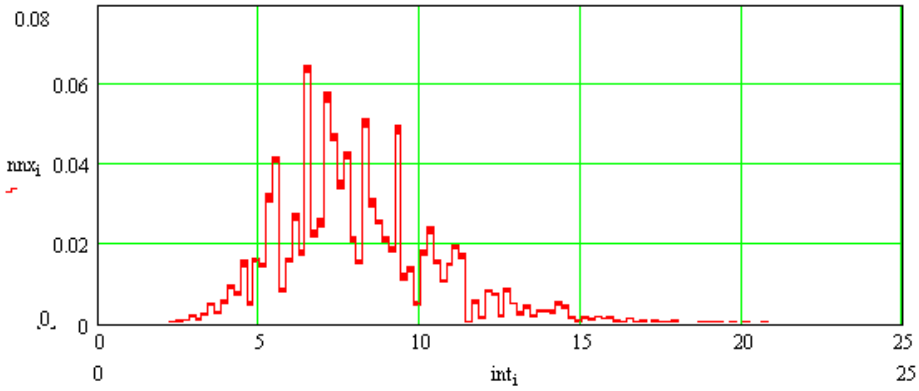


Рис. 15. Двадцать наложений хи-квадрат распределений выборок по 16 опытов, каждая из гистограмм имеет по 6 столбцов, гистограммы строилась на 99999 реализациях

Из рисунка 15 видно, что остаточная неопределенность гистограмм для 99999 реализаций приводит, к примерно, 10-ти кратному утолщению линий функции распределения значений пиков и впадин мульти модальной гистограммы. То есть «иголки на теле ежика» являются практически детерминированными. Однако, если снизить в 100 раз число реализаций, на которых строится каждая гистограмма, что «иголки на теле ежика» увеличат свою нестабильность. Рисунок 16 иллюстрирует это положение.

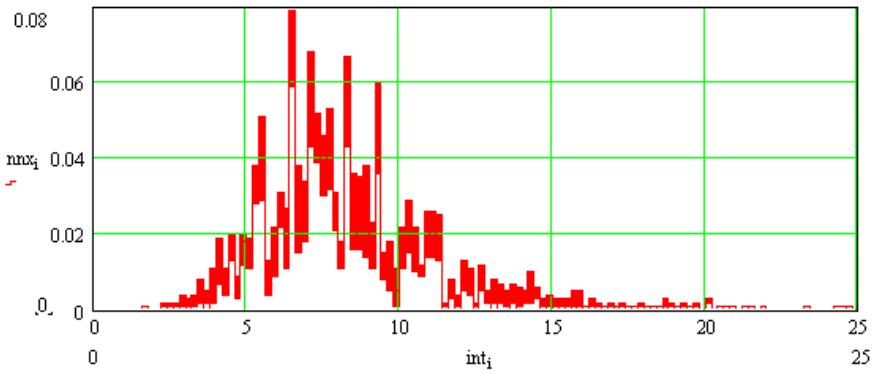


Рис. 16. Двадцать наложений хи-квадрат распределений выборок по 16 опытов, каждая из гистограмм имеет по 6 столбцов, каждая из гистограмм строилась на 999 реализациях

Из рисунка 16 видно, что ширина областей нестабильности по сравнению с областями нестабильности рисунка 15 увеличилась примерно в 10 раз, что хорошо соответствует теоретическим эффектам 100 кратного снижения размеров выборки, на которой строились гистограммы. Общий рисунок чередования максимумов и минимумов колебаний гистограммы сохраняется независимо от размеров выборки. Это означает, что неровности мультимодатности гистограмм значений хи-квадрат распределений не случайны и должны учитываться при вычислениях.

4.5. Борьба с шумами квантования малых выборок через использование сглаживающего данные цифрового фильтра

Эффекты неравномерности плотности распределения значений хи-квадрат критерия обусловлены тем, что мы заменяем континуум плотности распределения значений входных данных на приближение в виде ступенчатой гистограммы. Эта ситуация иллюстрируется рисунком 17. Последний 6-той столбец гистограммы рисунка 17 оказался пустым (имеет нулевое заполнение).

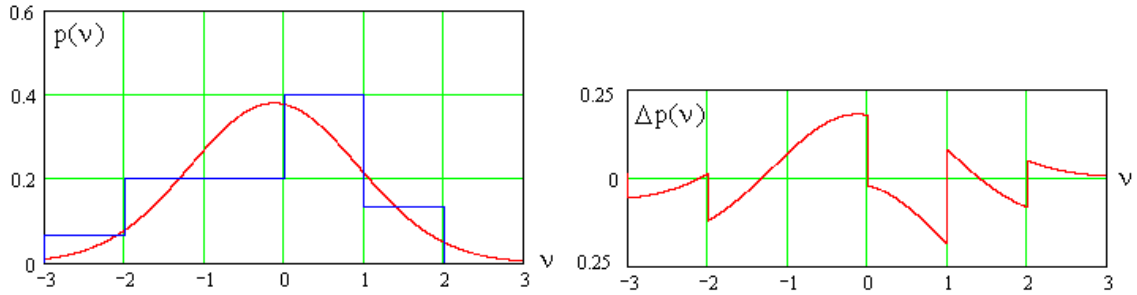


Рис. 17. Появление шумов квантования из-за замены непрерывной плотности распределения значений гистограммой из 6 столбцов, построенной на 16-опытах

Стандартизованные рекомендации [22] требуют выбирать число столбцов гистограммы таким образом, что бы в среднем на один столбец приходилось примерно 5 опытов. Если увеличить число столбцов гистограммы в 10 раз, то

появится много пустых столбцов и вырастет амплитуда ошибок квантования, как это показано на рисунке 18.

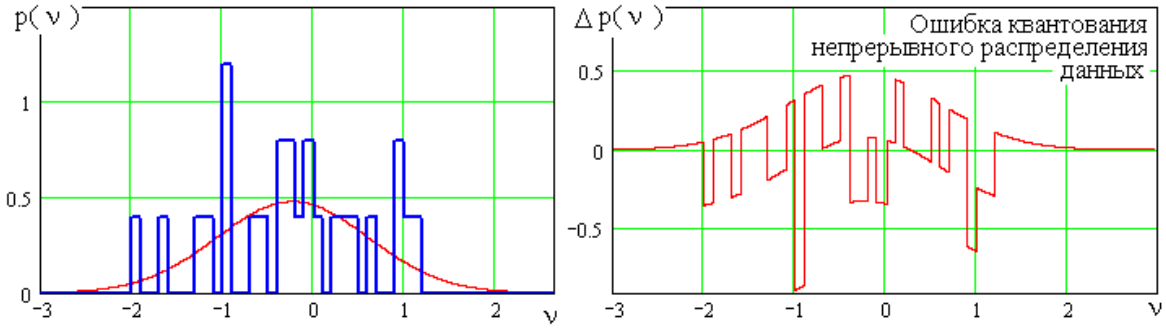


Рис. 18. Попытка увеличить в 10 число столбцов гистограммы приводит к появлению большого числа пустых столбцов гистограммы и росту амплитуды шумов квантования

Уйти от нежелательного эффекта пустых столбцов удается, если воспользоваться усредняющим цифровым фильтром со скользящим окном в 11 отсчетов [24]. В этом случае пустые столбцы гистограммы заполняются, что отображено на рисунке 19.

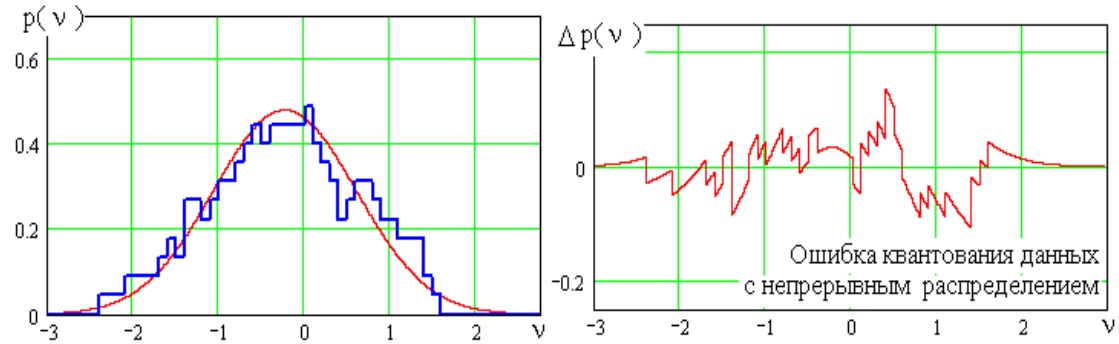


Рис. 19. Восстановленная сглаживанием гистограмма (окно сглаживания 11 отсчетов) с числом столбцов в 10 раз больше чем рекомендуется и соответствующий ей шум квантования

Если сравнивать рисунок 17 и рисунок 19, мы видим, что сглаженная гистограмма из примерно 60 столбцов гораздо больше похожа на нормальное распределение, чем исходная гистограмма, состоящая всего из 6 столбцов.

Программа сглаживающего данные цифрового фильтра со скользящим окном в 11 отсчетов приведена в верхней части рисунка 20.

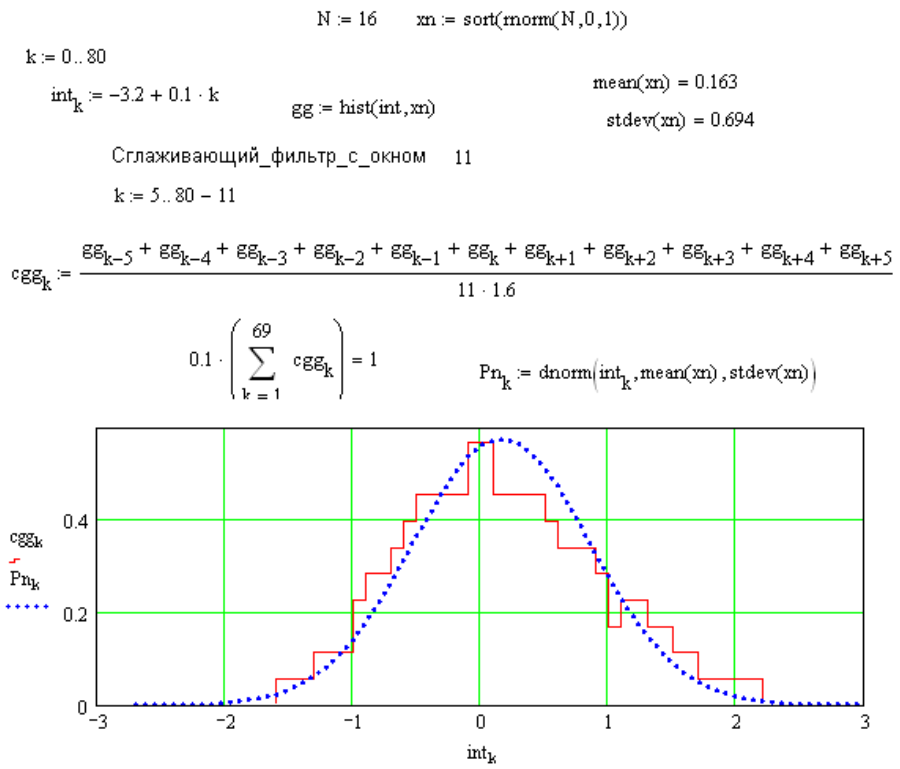


Рис. 20. Программная реализация цифрового фильтра со скользящим окном усреднения в 11 отсчетов

На таких принципах псевдо аналоговой предобработки исходных данных может быть построена целая ветвь оракулов, которые будут более эффективны при вычислении математического ожидания, стандартного отклонения, коэффициента корреляции, хи-квадрат критерия. Для этих вычислительных процедур методы псевдо аналоговой коррекции позволяют увеличить точность вычислений в несколько раз [24, 25] при том же объеме выборки.

В частности, если воспользоваться вычислением обычных значений хи-квадрат критерия для нормальных данных $\mathrm{rnorm}(15,0,1)$ и данных, полученных от программного генератора с равномерным распределением $\mathrm{runif}(15,-3,3)$, то мы получим распределения хи-квадрат критерия, отображенные в левой части рисунка 21.

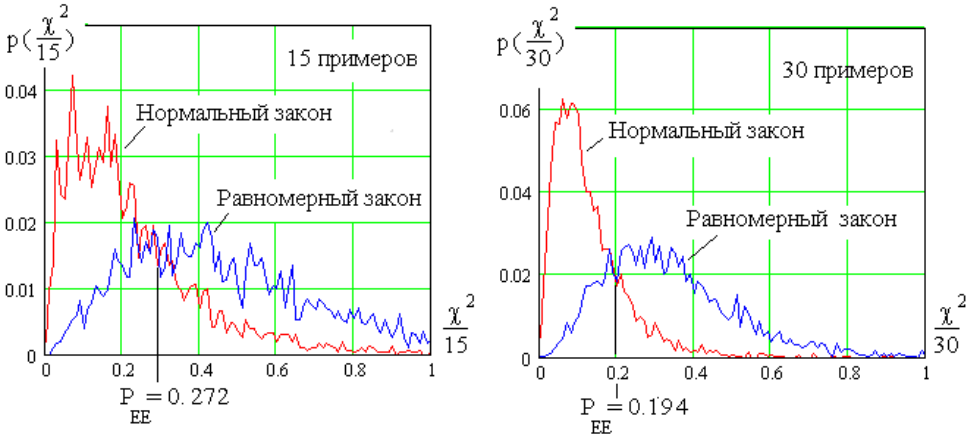


Рис. 21. Соотношения распределений хи-квадрат критерия для данных с нормальным и равномерным распределениями при выборках из 15 и 30 примеров.

Для того, что бы разделять хи-квадрат отклики нормальных и равномерных данных необходимо выставить порог k=0.28 при котором ошибки первого и второго рода оказываются одинаковыми $P_1=P_2=P_{EE}=0.272$.

Если мы увеличим выборку нормальных данных $\mathrm{rnorm}(30,0,1)$ и данных с равномерным законом распределения значений $\mathrm{runif}(30,-3,3)$, то получим распределения значений хи-квадрат критерия, отображенные в правой части рисунка 21. Для этих данных порог равновероятных ошибок $P_1=P_2=P_{EE}=0.194$ составляет k=0.2. При разном числе примеров в тестовой выборке получаются разные данные, которые в логарифмическом масштабе хорошо описываются прямой линией, как это показано на рисунке 22.

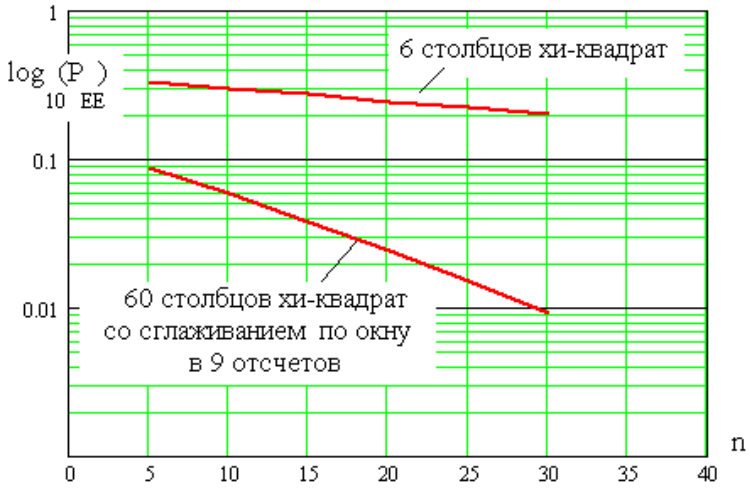


Рис. 22. Мощность хи-квадрат критерия в логарифмической шкале равновероятных ошибок

Если мы выполним процедуру сглаживания данных, увеличив число столбцов гистограммы с 6 до 60, то увеличивается наклон линии (равновероятная ошибка падает с ростом объема выборки в четыре раза быстрее). Прямые рисунка 22, описываются следующими соотношениями:

$$\begin{cases} \log_{10}(P_{EE}(n)) = -0.45 - 0.0086 \cdot n \\ \log_{10}(P_{EE}(n)) = -0.82 - 0.036 \cdot n \end{cases} \qquad (15).$$

В итоге при 30 опытах равновероятную ошибку удается снизить примерно в 20 раз, что показывает наличие существенных резервов повышения точности оценки выходных данных хи-квадрат оракула, построенного на предварительном цифровом сглаживании данных малой выборки.

Следует обратить внимание на то, что при получении данных рисунка 22 использовалось очень простое решающее правило порогового сравнения хи-квадрат преобразований. Реальные статистические оракулы, обученные различать сглаженные гистограммы нормальных и равномерных данных могут быть более интеллектуальными (например, нейросетевыми).

Примеров сглаженных гистограмм малых выборок, может быть создано много. Далее на этих примерах мы можем обучить две нейронные сети, например, стандартизованным для биометрии алгоритмом ГОСТ Р 52633.5 [26] или любым другим алгоритмом обучения [27]. В итоге мы получим две нейронные сети, как это отображено на рисунке 23. Одна нейронная сеть будет обучена распознавать сглаженные гистограммы с 60 столбцами, полученные от фиксированного размера выборки нормальных данных. Вторая нейросеть должна быть обучена распознавать сглаженные гистограммы, полученные от выборки фиксированного объема данных с равномерным законом распределения значений.

На текущий момент нет данных, какой из оракулов будет эффективнее. Работ по созданию нейросетевых оракулов заранее обученных различать формы сглаженных гистограмм в Пензе пока не было.

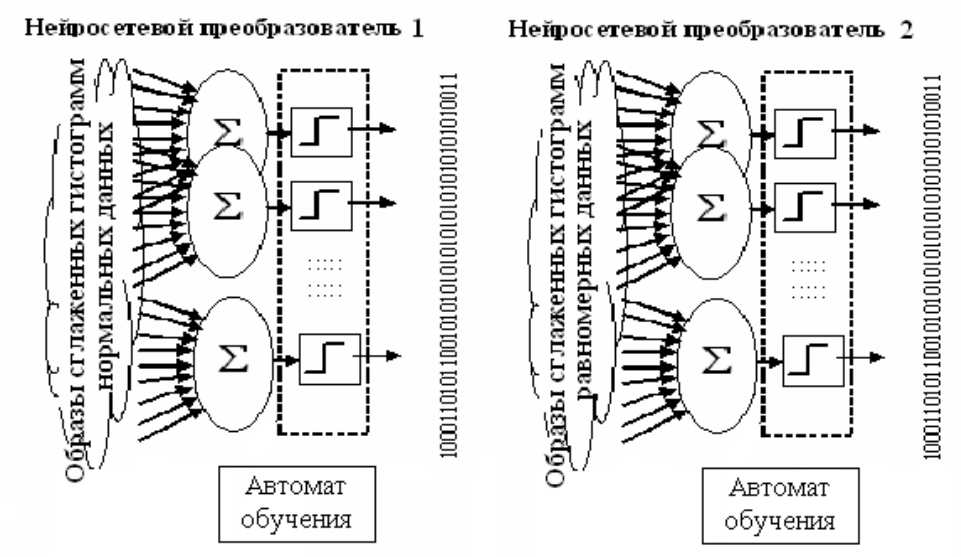


Рис. 23. Использование двух нейронных сетей, заранее обученных различать сглаженные гистограммы малых выборок с нормальным и равномерным законом распределения значений

4.6. Хи-квадрат молекула (подчеркивание квантовых эффектов, возникающих на малых выборках)

Вместо того, что бы бороться с шумами квантования сглаживанием, возможен иной противоположный подход. Мы имеем возможность усилить эффекты [14, 15, 16, 29] квантовой суперпозиции, воспользовавшись дополнительным условием, накладываемым на процедуру формирования гистограммы реальных данных. Из рисунка 11 на странице 13 видно, что обычный статистик выбирает число столбцов гистограммы и их расположение, опираясь на данные в реальной выборке.

Следует отказаться от этой практики и выбирать ширину интервалов гистограммы исходя из стандартного отклонения реальной выборки, а положение математического ожидания выборки должно быть жестко привязано к положению столбцов гистограммы. Если мы пользуемся гистограммой из 6 интервалов, то для данных с равномерным законом $\mathrm{norm}(11,0,1)$, необходимо строить гистограммы так, что бы математическое ожидание всегда попадало между третьим и четвертым столбцами гистограммы. Программная реализация этих вычислений приведена на рисунке 24.

$$i := 0..9999 \qquad N := 11$$

$$xn^{\langle i \rangle} := sort(rnorm(N,0,1))$$

$$mxn_i := mean\left(xn^{\langle i \rangle}\right) \qquad sxn_i := stdev\left(xn^{\langle i \rangle}\right) \qquad xn^{\langle i \rangle} := \frac{xn^{\langle i \rangle} - mxn_i}{sxn_i}$$

$$int := \begin{pmatrix} -3.2 \\ -2.066 \\ -1.033 \\ 0 \\ 1.033 \\ 2.066 \\ 3.2 \end{pmatrix} \qquad P := \begin{pmatrix} 0.19 \\ 0.132 \\ 0.349 \\ 0.349 \\ 0.132 \\ 0.19 \end{pmatrix}$$

$$xr^{\langle i \rangle} := sort(runif(N,-3.2,3.2))$$

$$mxr_i := mean\left(xr^{\langle i \rangle}\right) \qquad sxr_i := stdev\left(xr^{\langle i \rangle}\right) \qquad xr^{\langle i \rangle} := \frac{xr^{\langle i \rangle} - mxr_i}{sxr_i}$$

$$xi2n_i := N \cdot \left[\sum_{j=0}^{5} \frac{\left(\frac{hist\left(int, xn^{\langle i \rangle}\right)_j}{N} - P_j \right)^2}{P_j} \right] \qquad xi2r_i := N \cdot \left[\sum_{j=0}^{5} \frac{\left(\frac{hist\left(int, xr^{\langle i \rangle}\right)_j}{N} - P_j \right)^2}{P_j} \right]$$

$$i := 0..10000 \qquad int_i := 0.01 \cdot i \qquad nnx := \frac{hist(int, xi2n)}{9999} \qquad rrx := \frac{hist(int, xi2r)}{9999}$$

Рис. 24. Программная реализация вычислений хи-квадрат значений с синхронизацией столбцов гистограммы по математическому ожиданию выборки

При выполнении дополнительных условий синхронизации хи-квадрат отклики становятся полностью дискретными как это показано на рисунке 25.

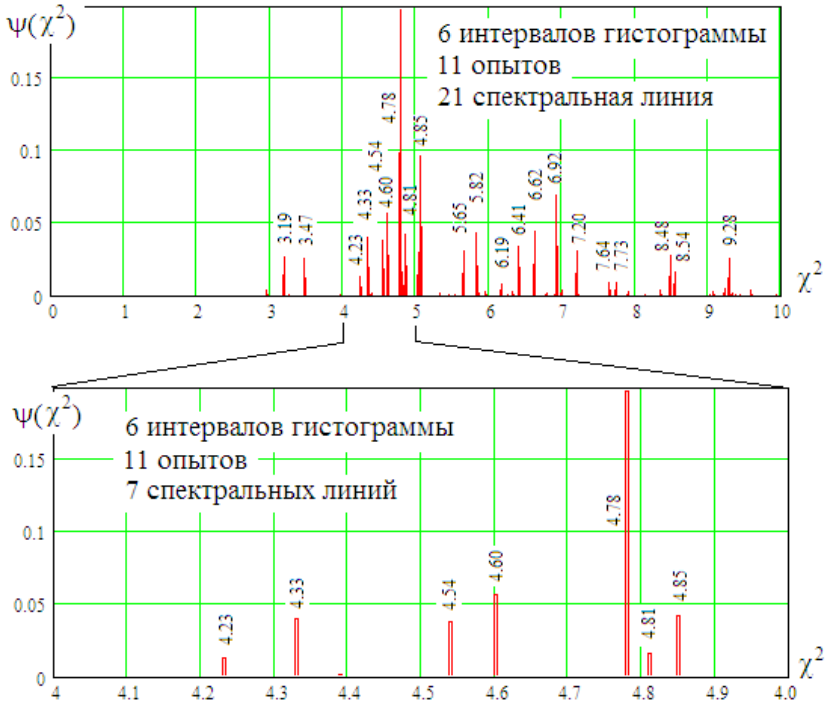


Рис. 25. Положение спектральных линий хи-квадрат молекулы при 11 электронах-опытах и 6 орбиталях-интервалах для нормального распределения данных

Мы видим, что выборки из 11 опытов с нормальным законом распределения значений не могут иметь какие угодно состояния хи-квадрат критерия. В этом контексте нет разницы между молекулой водорода, порождающей дискретный спектр линий водорода и математической хи-квадрат молекулой (ее программная реализация дана на рисунке 24). И в том и в другом случае мы наблюдаем дискретный выходной спектр.

Связь между молекулой водорода и хи-квадрат молекулой иллюстрируется рисунком 26.

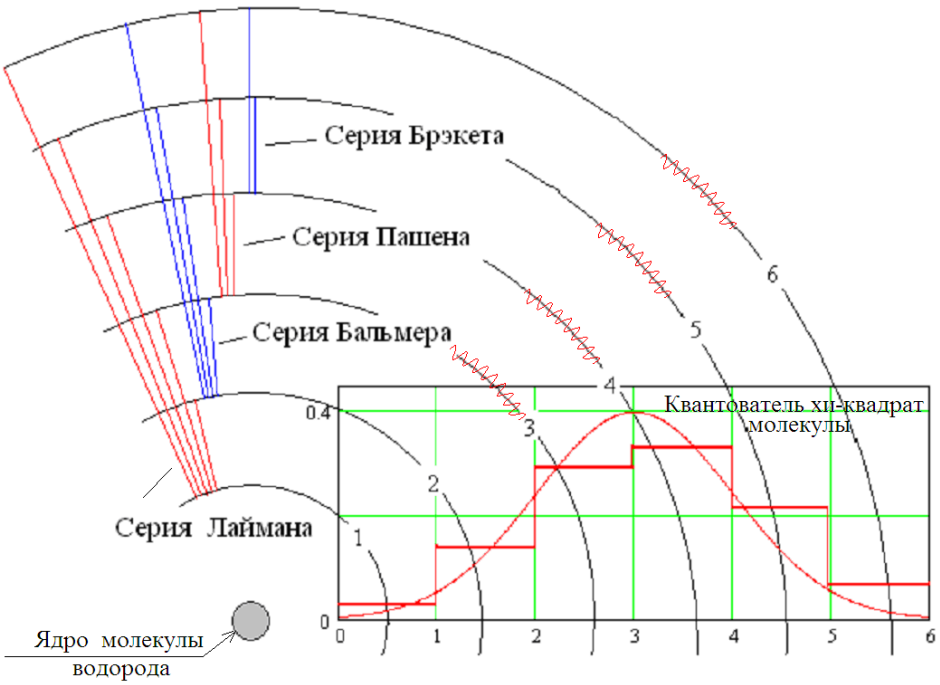


Рис. 26. Планетарная модель молекулы водорода, порождающая серии линий спектра

Планетарная модель атома водорода объясняет дискретный характер ее спектров излучения и поглощения. Под эту модель вычислены серии линий спектра водорода (серия Лаймана, серия Бальмера,…), описывающиеся через целые числа. Такой же эффект возникает при описании непрерывного (континуального) распределения данных гистограммой с конечным числом интервалов. Так как число столбцов гистограммы конечно, а выборка мала, число выходных состояний хи-квадрат молекулы конечно. Всегда возникают пустые интервалы между линиями амплитуды вероятности спектра состояний хи-квадрат молекулы.

Для нас принципиально важным является то, что спектр хи-квадрат молекулы нормальных данных $\mathrm{norm}(11,0,1)$ отличается от спектра данных с равномерным законом распределения значений $\mathrm{runif}(11,-3.2,3.2)$. Положение спектральных линий одинаковое, а вот амплитуда вероятности их появления разные, как это показано на рисунке 27.

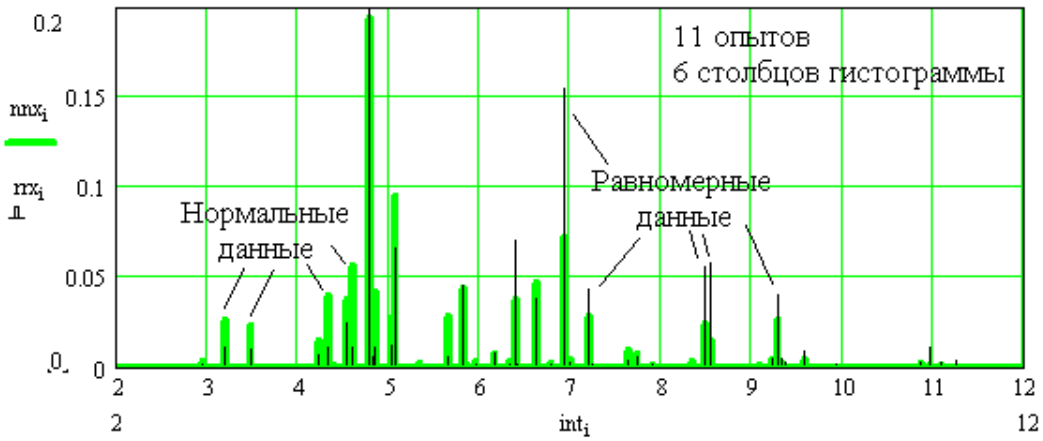


Рис. 27. Разное значение амплитуды вероятности появления спектральных линий с одинаковым положением для нормального и равномерного законов распределения значений

4.7. Оценка потенциальных возможностей квантовых хи-квадрат оракулов, заранее обученных распознавать нормальный и равномерный законы распределения

Программная реализация хи-квадрат молекулы (рисунок 24) дает сразу 10 000 состояний для нормальных выборок и выборок с равномерным распределением. Именно это обстоятельство и позволяет видеть спектр амплитуд вероятностей. Рассуждая формально, мы можем построить две нейронные сети, которые будут заранее обучены распознавать спектры нормальных и равномерных выборок (рисунок 23). Теоретически не имеет значения, какие образы должны распознаваться нейронными сетями сглаженные гистограммы или амплитуды вероятности появления спектральных линий.

В связи с такой интерпретацией попытаемся оценить то, что могут дать нейронные сети при их использовании для распознавания. Оценить это удается через взаимную информативность данных. Очевидно, что спектральная линия для целей различения образов будет бесполезна, если ее амплитуда вероятности будет одинакова для нормальных и равномерных данных. Информативность такой линии будет нулевой. Напротив, информативность линии будет расти по мере увеличения различий амплитуд одной и той же спектральной линии.

Мы можем перейти к двоичным логарифмам амплитуд вероятности и определить информативность каждой из значимых линий спектра. В верхней части рисунка 28 приведена программа, осуществляющая эти вычисления над данными программы рисунка 24.

$$\text{inf} := \left| \begin{array}{l} \text{for } i \in 0..999 \\ \quad \left| \begin{array}{l} \text{in}_i \leftarrow 0 \text{ if } \text{nnx}_i \le 0.0001 \wedge \text{rrx}_i \le 0.0001 \\ \text{in}_i \leftarrow \left| \log(\text{nnx}_i, 2) - \log(\text{rrx}_i, 2) \right| \text{ if } \text{nnx}_i > 0.0001 \wedge \text{rrx}_i > 0.0001 \end{array} \right. \\ \text{in} \end{array} \right. \qquad \sum_{i=0}^{998} \text{inf}_i = 53.277$$

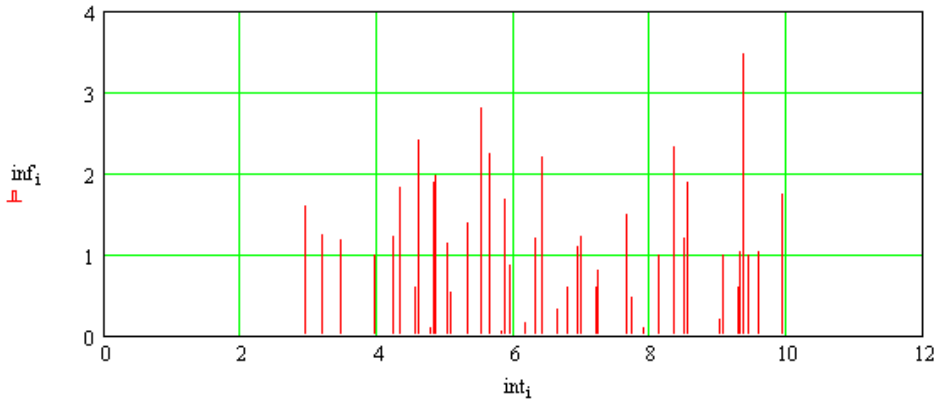


Рис. 28. Определение информативности спектральных линий

Из рисунка 28 мы видим, что максимальная информативность линий спектра составляет 3.5 бита. Сумма информативностей всех спектральных линий составляет 53.3 бита, что является огромной величиной, обеспечивающей равновероятные ошибки на уровне $2^{-53} \approx 0.0000000000000001$ (15 нулей после точки). Очевидно, что такая величина вероятности ошибок практически недостижима, это не более чем приближенная оценка. Однако такая оценка показывает высокий потенциал математических конструкций типа хи-квадрат молекулы.

В программе рисунка 28 при информативности данные сравниваются с порогом вероятности 0.0001. Это вынужденная мера, обусловленная тем, что серия опытов по 11 примеров не может быть как угодно велика. Так, если мы имеем большую выборку в 21 опыт (генеральная выборка), то мы можем получить из нее множество частных не повторяющихся серий по 11 примеров, число которых составит:

$$C_{21}^{11} = \binom{21}{11} = \frac{21!}{11!(21-11)!} = 352716 .$$

Всего мы наблюдаем порядка 30 информативных спектральных линий. То есть на каждую из спектральных линий в среднем будет приходиться 352716/30≈11757 вариантов наблюдения. Такое число опытов обеспечивает вероятность нужных нам событий больше чем величина порога 1/10000 указанного в программной реализации.

5. Молекула математического ожидания

По аналогии с хи-квадрат молекулой мы можем построить и другие математические молекулы с похожей конструкцией. Например, в верхней части рисунка 29 дана программа, реализующая молекулу математического ожидания для выборок в 11 опытов. Из графической части рисунка 29 виден переход от непрерывного распределения ошибок вычисления математического ожидания к их дискретному спектру, состоящему из 9 линий. Для того, что бы попытаться учесть данные дискретного спектра состояний математической молекулы для улучшения точности вычислений, требуется надстройка нейросетевого обобщения данных. Потребуется создавать квантово-неросетевых оракулов, способных корректировать данные классической процедуры вычисления математического ожидания. Техническая возможность движения по этому пути развития обусловлена тем, что данные по ошибкам вычисления математических ожиданий в непрерывном и дискретном вариантах практически не коррелированны

$\mathrm{corr}(dE, mxn) = 1.396 \times 10^{-3}$

$i := 0..99999 \qquad N := 11$

$xn^{\langle i \rangle} := \mathrm{sort}(\mathrm{rnorm}(N, 0, 1))$

$mxn_i := \mathrm{mean}(xn^{\langle i \rangle}) \qquad sxn_i := \mathrm{stdev}(xn^{\langle i \rangle}) \qquad xn^{\langle i \rangle} := \frac{xn^{\langle i \rangle} - mxn_i}{sxn_i}$

$int := \begin{pmatrix} -3.2 \\ -2.066 \\ -1.033 \\ 0 \\ 1.033 \\ 2.066 \\ 3.2 \end{pmatrix}$

$gE^{\langle i \rangle} := \mathrm{hist}(int, xn^{\langle i \rangle})$

$dE_i := (gE^{\langle i \rangle})_0 \cdot (-0.48) + (gE^{\langle i \rangle})_1 \cdot (-0.32) + (gE^{\langle i \rangle})_2 \cdot (-0.16) + (gE^{\langle i \rangle})_3 \cdot (0.16) + (gE^{\langle i \rangle})_4 \cdot 0.32 + (gE^{\langle i \rangle})_5 \cdot 0.48$

$i := 0..10000 \qquad int_i := -3.01 + 0.02 \cdot i$

$Edx := \frac{\mathrm{hist}(int, dE)}{9999} \qquad Enx := \frac{\mathrm{hist}(int, mxn)}{9999}$

$\mathrm{corr}(dE, mxn) = 1.396 \times 10^{-3}$

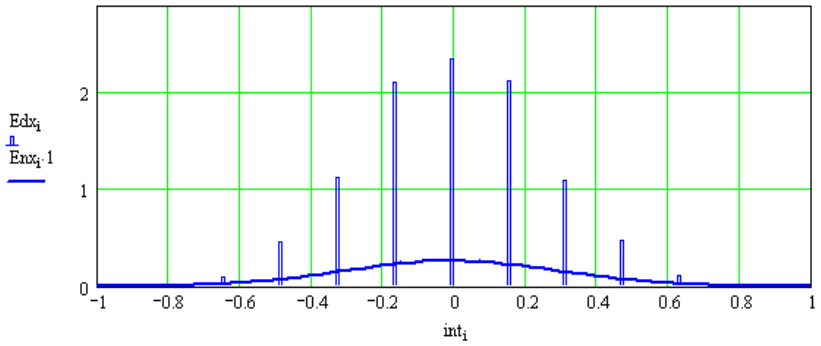
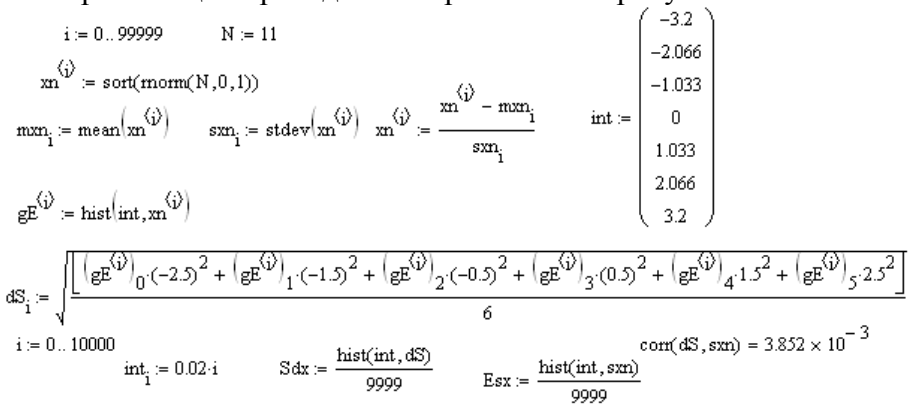


Рис. 29 Программная реализация молекулы математического ожидания

6. Математическая молекула стандартного отклонения

Так же как мы получили молекулу математического ожидания для малых выборок, мы имеем право получить молекулу стандартного отклонения. Ее программная реализация приведена в верхней части рисунка 30.



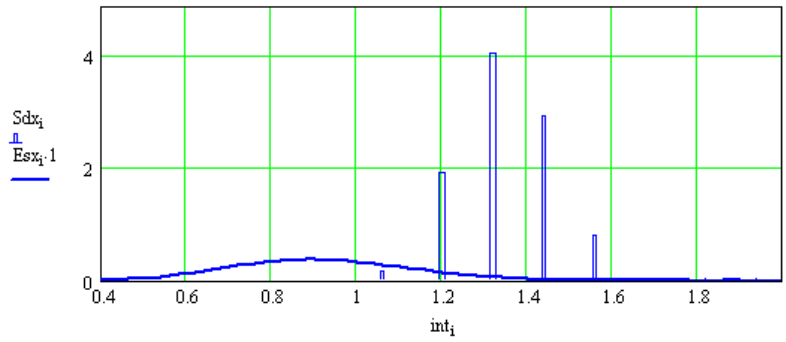


Рис. 30. Программная реализация молекулы стандартного отклонения

Молекула стандартного отклонения не может использоваться самостоятельно. Для нее нужно создавать нейросетевое обобщение. Работоспособность подобных конструкций обусловлена тем, что корреляция ошибок стандартного отклонения в непрерывной форме и в дискретном варианте оказывается мала $\mathrm{corr}(dS, sxn) = 3.852 \times 10^{-3}$.

7. Два варианта корреляционных молекул

7.1. Корреляционная молекула с двумя линейными квантователями

Следует отметить, что при биометрических вычислениях выборок из 20 примеров вполне достаточно для вычисления математического ожидания с нужной точностью. Для вычисления стандартного отклонения таких выборок уже недостаточно, необходимо принимать меры по регуляризации вычислений. Попытки вычислять на столь малых выборках коэффициентов корреляции обычно приводят к неудачам. Предварительные проработки показывают, что модификация алгоритма обучения искусственных нейронных сетей ГОСТ Р 52633.5 [8] с учетом коэффициентов корреляции требует увеличения обучающей выборки до 60 примеров. Трехкратное увеличение объема обучающей выборки крайне нежелательно, так как негативно воспринимается пользователями биометрии. В связи с этим возникает задача трехкратного снижения ошибок при вычислении коэффициентов корреляции.

Из теории статистических оценок [30] известно, что центрированные и нормированные независимые данные, попадают в круг. При этом для больших чисел – N, проводимых опытов вероятности попадания в каждую из четвертей круга, будут одинаковы:

$$P_1 \approx \frac{n_1}{N} \approx P_2 \approx \frac{n_2}{N} \approx P_3 \approx \frac{n_3}{N} \approx P_4 \approx \frac{n_4}{N} \qquad (16),$$

где $n_1, n_2. n_3, n_4$ - число попаданий в первую, вторую, третью и четверную четверти круга.

В случае, если данные коррелированны, то соотношение между вероятностями попадания в разные фрагменты эллипса рассеивания меняются:

$$P_1 \approx \frac{n_1}{N} \approx P_3 \approx \frac{n_3}{N} > P_2 \approx \frac{n_2}{N} \approx P_4 \approx \frac{n_4}{N} \qquad (17).$$

Эта ситуация отображена на рисунке 29.

Можно показать, что для коррелированных данных вероятности попадания в выделенные заливкой сектора эллипса рисунка 29 пропорциональны малому и большому диаметрам эллипса:

$$r = 1 - \frac{d}{D} \approx 1 - \frac{P_2 + P_4}{P_1 + P_3} \approx 1 - \frac{n_2 + n_4}{n_1 + n_3} \qquad (18).$$

Осуществив описанные выше преобразования получим [31] следующую формулу для вычисления коэффициентов корреляции:

$$r(x_1, x_2) \approx 1 - \frac{P_1 + P_3 - P_2 - P_4}{P_1 + P_3 + P_2 + P_4} \approx 1 - \frac{n_1 + n_3 - n_2 - n_4}{n_1 + n_3 + n_2 + n_4} \qquad (19).$$

Получается, что эллипс распределения зависимых данных является двухмерным континуумом корреляционной молекулы, а оси нормированной и центрированной системы координат играют роль двух квантователей, делящих площадь эллипса на четыре части. Получается достаточно простая и понятная математическая конструкция, преобразующая внутренний (не наблюдаемый двухмерный континуум в конечный спектр дискретных выходных состояний). То есть, мы получили желаемую корреляционную молекулу, аналогичную хи-квадрат математической молекуле.

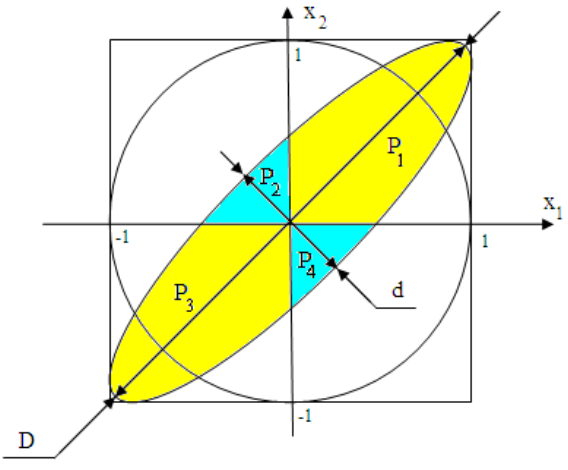


Рис. 31. Описание центрированных и нормированных площадей рассеивания нормальных данных (независимых данных – круг и зависимых данных – эллипс)

На рисунке 32 приведена программная реализация корреляционной молекулы.

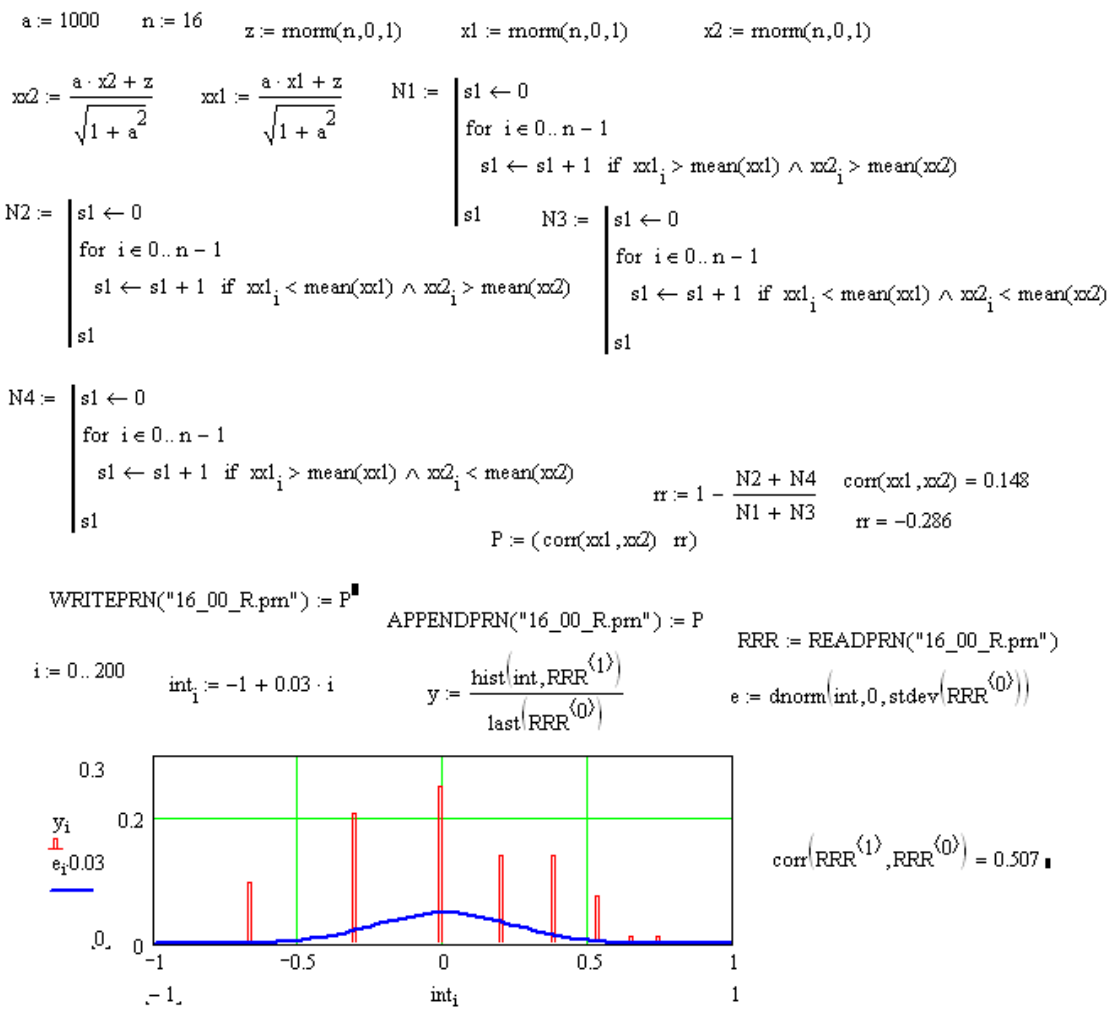


Рис. 32 Корреляционная молекула и спектр ее выходных состояний

Из рисунка видно, что корреляционная молекула дает 8 спектральных линий, ее выходные данные имеют высокий уровень корреляции с данными, полученными по формуле Пирсона $\mathrm{corr}\left(RRR^{\langle 1\rangle}, RRR^{\langle 0\rangle}\right) = 0.507$. Для использования выходных данных корреляционной молекулы с двумя линейными квантователями необходимо создавать нейросетевую обобщающую надстройку.

7.2 Корреляционная молекула с двумя эллиптическими квантователями

Расчет коэффициентов корреляции по формуле (18) дает значение корреляции близкое к нулю для распределения в левой части рисунка 32. Для распределения данных в правой части рисунка корреляция составит r=1/3.

Следует отметить, что применение эллипсов для описания распределений эквивалентно использованию симметричных эллиптических квантователей:

$$\begin{cases} y_i^{\,2} = \begin{bmatrix} E(\nu_1) - \nu_{1,i} \\ E(\nu_2) - \nu_{2,i} \end{bmatrix}^{T} \cdot \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} E(\nu_1) - \nu_{1,i} \\ E(\nu_2) - \nu_{2,i} \end{bmatrix} \\ z(y_i^2) = \text{"0"} \quad \text{если} \quad y_i^2 \le k \\ z(y_i^2) = \text{"1"} \quad \text{если} \quad y_i^2 > k \end{cases} \qquad (20),$$

где k –порог квантователя.

Если использовать два квантователя с параметрами r=1/3 и r=-1/3, мы получим корреляционную молекулу, работа которой иллюстрируется рисунком 31.

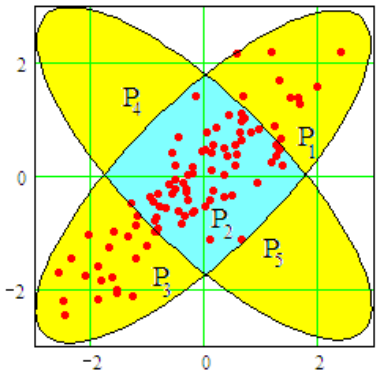


Рис. 32. Работа двух эллиптических квантователей корреляционной молекулы

Работа молекулы построена на том, что число опытов попавших внутрь эллипсов первого и второго квантователя разное:

$$P_1 + P_2 + P_3 > P_2 + P_4 + P_5 \qquad (20)$$

Если исследуемое распределение положительно коррелированно, то число опытов обнаруженное внутри первого квантователя должно быть больше чем число опытов обнаруженное внутри второго квантователя:

$$N_1 > N_2 \qquad (21).$$

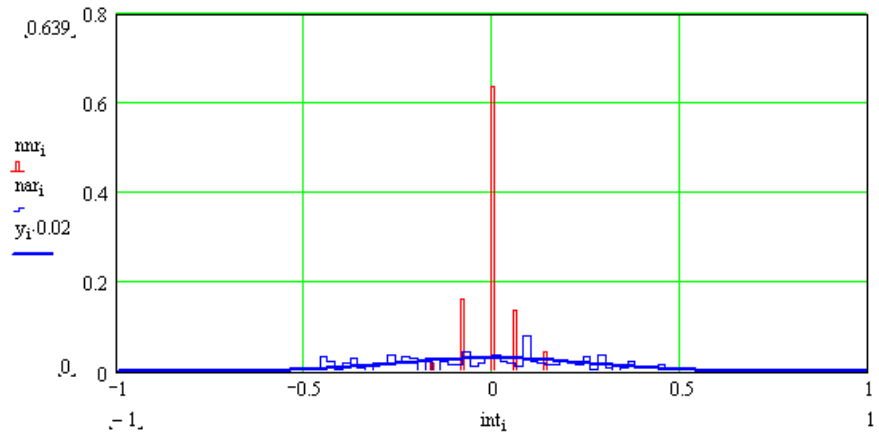
Если корреляция отрицательна, то соотношение (21) меняется на противоположное.

Учитывая это мы получим следующие соотношения для вычисления коэффициентов корреляции:

$$\begin{cases} \tilde{r} \approx 1 - \dfrac{N_1 - N_2}{N_1} \text{ при } N_1 > N_2 \\ \tilde{r} \approx -\left(1 - \dfrac{N_2 - N_1}{N_2}\right) \text{ при } N_2 > N_1 \end{cases} \qquad (22).$$

Очевидно, что вычисления коэффициента корреляции по формуле (22) и по классической формуле Пирсона являются разными вычислительными процедурами. Как следствие мы получаем разные значения ошибок Δr и $\Delta\tilde{r}$. Значения этих ошибок положительно коррелированны $r(\Delta r, \Delta\tilde{r}) = 0.519$. То, что их корреляция не единична является предпосылкой для использования вычислений (22) совместно с результатами, полученными по классической формуле Пирсона.

Результаты реализации корреляционной молекулы с двумя эллиптическими квантователями даны на рисунке 33.



№	0	1	2	3	4	5	6	7	8	9	10	11	12
r	-0.438	0.091	-0.224	-0.238	0.2	0.248	0.455	0.035	$-7.4 \cdot 10^{-3}$	0.365	-0.248	-0.369	-0.288
Mr	-0.067	0	0	-0.067	0	0	0.143	0	0	0.067	0	0	-0.067

Рис. 33. Итоговые результаты работы корреляционной молекулы с двумя эллиптическими квантователями

Из верхней части рисунка 33 видно, что корреляционная молекула имеет 5 спектральных линий, соответственно расположенных в точках r = {-0.143, -0.067, ±0.0, 0.067, 0.143}. Эти пять дискретных значений появляются с вероятностями P = {0.04, 0.18, 0.64, 0.16, 0.05}.

В отличии от корреляционной молекулы с линейными квантователями второй вариант корреляционной молекулы можно использовать самостоятельно без нейросетевого обобщения. При этом у такого корреляционного оракула диапазон ошибок снижается в три раза по сравнению с вычислениями по классической формуле Пирсона.

ЗАКЛЮЧЕНИЕ

Известно, что обычный анализ непрерывных спектров света позволяет определять минимальную концентрацию примесей на уровне 0.1% процента. Если же мы используем спектральный дискретный анализ, то криминалисты обнаруживают присутствие примесей вещества на уровне 0.00000001%. Смочив в спирте вату и протерев весы, криминалисты могут уверенно утверждать, что на этих весах ранее взвешивали золото. Такой результат получается после анализа спектрального состава пламени сжигаемой ваты. Спектральные линии золота не совпадают со спектральными линиями органики (ваты) и кислорода.

Предположительно, похожие результаты можно получить, работая с такими математическими конструкциями, как хи-квадрат молекула, молекула математического ожидания, молекула стандартного отклонения, корреляционная молекула. Предположительно, что использование корреляционных молекул с двумя эллиптическими квантователями позволит использовать обучение сетей высокоразмерных квадратичных форм на выборках из 20 примеров биометрического образа «Свой». Надежный статистический анализ малых выборок, видимо, может быть осуществлен только через анализ спектральных линий математических молекул.

Литература:

1. Juels A., Wattenberg M. A Fuzzy Commitment Scheme // Proc. ACM Conf. Computer and Communications Security, Singapore — November 01 – 04, 1999, p. 28–36.

2. Ramírez-Ruiz J., Pfeiffer C., Nolazco-Flores J. Cryptographic Keys Generation Using FingerCodes. //Advances in Artificial Intelligence - IBERAMIA-SBIA 2006 (LNCS 4140), p. 178-187, 2006

3. Feng Hao, Ross Anderson, and John Daugman. Crypto with Biometrics Effectively, IEEE TRANSACTIONS ON COMPUTERS, VOL. 55, NO. 9, SEPTEMBER 2006.

4. Иванов А.И. Нечеткие экстракторы: проблема использования в биометрии и криптографии. // Первая миля. № 1, 2015 г. с. 40-47.

5. Язов Ю.К. и др. Нейросетевая защита персональных биометрических данных. //Ю.К.Язов (редактор и автор), соавторы В.И. Волчихин, А.И. Иванов, В.А. Фунтиков, И.Г. Назаров // М.: Радиотехника, 2012 г. 157 с. IBSN 978-5-88070-044-8.

6. Ахметов Б.С., Иванов А.И., Фунтиков В.А., Безяев А.В., Малыгина Е.А. Технология использования больших нейронных сетей для преобразования нечетких биометрических данных в код ключа доступа. Монография, Казахстан, г. Алматы, ТОО «Издательство LEM», 2014 г. -144 с., находится в открытом доступе (http://portal.kazntu.kz/files/publicate/2014-06-27-11940.pdf)

7. ГОСТ Р 52633.0-2006 «Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации».

8. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».

9. Безяев А.В. Безкомпроматная индикация качества ввода фрагментов тайного составного биометрического образа «Нейрокомпьютеры: разработка, применение» №6, 2009 с. 59- 62

10. Безяев А.В., Иванов А.И., Фунтикова Ю.В. Оптимизация структуры самокорректирующегося био-кода, хранящего синдромы ошибок в виде фрагментов хеш-функций. «Вестник Уральского федерального округа. Безопасность в информационной сфере» 2014 г. № 3*(13)* с. 4-14.

11. Ложников П.С. Биометрическая защита гибридного документооборота. /Новосибирск. Из-во СО РАН, 2017 г., 130 с.

12. Иванов А.И., Ложников П.С., Качайкин Е.И. Идентификация подлинности рукописных автографов сетями Байеса-Хэмминга и сетями квадратичных форм. «Вопросы защиты информации» №2 2015 г., с. 28-34.

13. Иванов А.И., Ложников П.С., Качайкин Е.И., Сулавко А.Е. Биометрическая идентификация рукописных образов с использованием корреляционного аналога правила Байеса. «Вопросы защиты информации» №3 2015 г., с. 48-54.

14. Ахметов Б.Б., Иванов А.И., Серикова Н.И., Фунтикова Ю.В. Дискретный характер закона распределения хи-квадрат критерия для малых тестовых выборок // Вестник Национальной академии наук Республики Казахстан. – Алматы, 2015. № 1. с. 17-25.

15. Кулагин В.П, Иванов А. И., Газин А.И., Ахметов Б.Б. Циклические континуально- квантовые вычисления: усиление мощности хи-квадрат критерия на малых выборках. /Аналитика, № 5, 2016 (30), с. 22-29. http://www.j-analytics.ru/journal/article/5679

16. Волчихин В.И., Иванов А.И., Пащенко Д.В., Ахметов Б.Б., Вятчанин С.Е. Перспективы создания циклической континуально-квантовой хи-квадрат машины для проверки статистических гипотез на малых выборках биометрических данных и данных иной природы. // Известия высших учебных заведений. Поволжский регион. Технические науки. – Пенза: ПГУ, №1, 2017 с. 3-7. http://izvuz_tn.pnzgu.ru/files/izvuz_tn.pnzgu.ru/ 1117.pdf

17. Волчихин В.И., Иванов А.И., Сериков А.В., Серикова Ю.И. Использование эффектов квантовой суперпозиции при регуляризации вычислений стандартного отклонения на малых выборках биометрических данных. «Измерение. Мониторинг. Управление. Контроль.» №1, 2017, с. 57-63. http://imuk.pnzgu.ru/files/imuk.pnzgu.ru/08117.pdf

18. Иванов А.И., Захаров О.С. Среда моделирования «БиоНейроАвтограф». Программный продукт создан лабораторией биометрических и нейросетевых технологий, размещен с 2009 года на сайте АО «ПНИЭИ», свободный доступ по ссылке http://пниэи.рф/activity/science/noc./bioneuroautugraph.zi

19. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006 г., 816 с.

20. Дерффель К. Статистика в аналитической химии. М.: Мир. 1994 г. 258 с.

21. Волчихин В.И., Иванов А.И., Серикова Ю.И. Компенсация методических погрешностей вычисления стандартных отклонений и коэффициентов корреляции, возникающих из-за малого объема выборок. Известия высших учебных заведений. Поволжский регион. Технические науки. – Пенза: ПГУ, №1, 2016 с. 45-49

22. Р 50.1.037-2002 Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа $\chi 2$. Госстандарт России. Москва-2001 г., 140 с.

23. Ахметов Б.Б., Иванов А.И., Серикова Н.И., Фунтикова Ю.В. Дискретный характер закона распределения хи-квадрат критерия для малых тестовых выборок // Вестник Национальной академии наук Республики Казахстан. – Алматы, 2015. № 1. С. 17-25.

24. Иванов А.И., Ахметов Б.Б., Серикова Ю.И. Усиление мощности хи-квадрат критерия при десяти кратном увеличении числа степеней свободы статистических вычислений на малых тестовых выборках // «Надежность и качество сложных систем» №4 (16)_2016 с. 121-127.

25. Волчихин В.И., Иванов А.И., Ахметов Б.Б., Серикова Ю.И. "Фрактально-корреляционный функционал, используемый при поиске пар слабо зависимых биометрических данных в малых выборках" //«Вестник высших учебных заведений. Поволжский регион. Технические науки» №4, 2016 г., с. 25 – 31.

26. ГОСТ Р 52633.5-2011 «Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа».

27. Саймон Хайкин. Нейронные сети: полный курс. М.: «Вильямс», 2006. — С. 1104.

29. Иванов А.И. Многомерная нейросетевая обработка биометрических данных с программным воспроизведением эффектов квантовой суперпозиции. Издательство АО «ПНИЭИ», Пенза-2016 г., 133 с. Свободный доступ http://пниэи.рф/activity/science/BOOK16.pdf

30. Абезгауз Г.Г., Тронь А.П., Копенкин Ю.Н., Коровина И.А. Справочник по вероятностным расчетами. М.: Воениздат, 1970 г., 536 с.

31. Волчихин В.И., Иванов А.И., Сериков А.В., Серикова Ю.И. Квантовая суперпозиция дискретного спектра состояний математической молекулы

корреляции для малых выборок биометрических данных // Вестник Мордовского университета. Т27. №2, 2017, с 230-243.

СВЕДЕНИЯ ОБ АВТОРЕ:

Иванов Александр Иванович, начальник лаборатории биометрических и нейросетевых технологий (ЛБНТ) АО «ПНИЭИ», 440000, г. Пенза, ул. Советская, 9, телефон: (8412) 59-33-10, e-mail: ivan@pniei.penza.ru. Диссертацию доктора технических наук защитил в 2002 г. по специальности 05.13.01 - Системный анализ, управление и обработка данных. Диплом доцента по специальности 05.13.01 получен в 2009 г.

В период с 2008 г. по 2013 г. являлся экспертом без права голоса от России в двух международных комитетах ISO/IEC JTC1 SC37 (Биометрия) и ISO/IEC JTC1 SC27 (Техника защиты информации) в связи с тем, что был научным руководителем ряда НИР (Исполнитель - АО «ПНИЭИ», Заказчик - ФСТЭК России) по разработке пакета отечественных стандартов: ГОСТ Р52633.0-2006, ГОСТ Р 52633.1-2009, ГОСТ Р 52633.2-2010, ГОСТ Р 52633.3-2011, ГОСТ Р 52633.4-2012, ГОСТ Р 52633.5-2011, ГОСТ Р 52633.6-2013, ГОСТ Р 52633.7-20хх.

В период с 2017 г. по 2018 г. являлся руководителем разработки технической спецификации ТК26 (Криптографическая защита информации) «Защита нейросетевых биометрических контейнеров с использованием криптографических алгоритмов».

Подписано к печати 02.04.2018 формат 60х80 1/16 Усл. печ. л. 2,07
Тираж 300 экз.

Издательство АО «ПНИЭИ», 440000, г. Пенза, ул. Советская, 9

Отпечатано с готового оригинал-макета в Издательстве ФБГОУ ВО «ПГУ»
440026, г. Пенза, ул. Красная, 44