BULLETIN OF NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
ISSN 1991-3494
Volume 1, Number 353 (2015), 17 – 25

DISCRETE CHARACTER OF THE LAW OF CHI-SQUARE DISTRIBUTION CRITERION FOR SMALL TEST SELECTIONS OF VALUES

B. B. Akhmetov[1], A. I. Ivanov[2], N. I. Serikova[3], Ju. V. Funtikova[2]

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

Key words: chi-square distribution of values, fractals, fractional exponent of number of degrees of freedom, histogram, reduction of volumes of test selection, discrete law of distribution of values.

Abstract. It is proved that at small test selections the number of states of the chi-square distribution is finite and generates a discrete range of possible output states. The number of degrees of freedom of chi-square distributions is always fractional. Today practiced usage of the whole values of an indicator of number of degrees of freedom leads to unjustified overestimate of volumes of test selection and decrease in reliability of estimates. Attempts of reduction of volume of test selection inevitably result in need to consider the number of degrees of chi-square distributions as fractal size.

УДК 519.2; 519.66; 57.087.1

ДИСКРЕТНЫЙ ХАРАКТЕР ЗАКОНА РАСПРЕДЕЛЕНИЯ ХИ-КВАДРАТ КРИТЕРИЯ ДЛЯ МАЛЫХ ТЕСТОВЫХ ВЫБОРОК

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

[1]Международный Казахско-Турецкий университет им. Х. А. Ясави, Туркестан, Казахстан,
[2]Пензенский научно-исследовательский электротехнический институт, Россия,
[3]ОАО «Рубин», Пенза, Россия

Ключевые слова: хи-квадрат распределение значений, фракталы, дробный показатель числа степеней свободы, гистограммы, сокращение объемов тестовой выборки, дискретный закон распределения значений.

Аннотация. Доказано, что при малых тестовых выборках число состояний хи-квадрат распределения конечно и порождает дискретный спектр возможных выходных состояний. Число степеней свободы хи-квадрат распределений всегда является дробной величиной. Практикуемое сегодня использование целых значений показателя числа степеней свободы приводит к неоправданному завышению объемов тестовой выборки и снижению достоверности оценок. Попытки сокращения объема тестовой выборки неминуемо приводят к необходимости считать число степеней хи-квадрат распределений фрактальной величиной.

Введение. В настоящее время критерий хи-квадрат проверки статистических гипотез является основой большинства отраслевых методик. Это обусловлено тем, что имеются рекомендации Госстандарта России [1]. В частности, по этим рекомендациям необходимо для проверки гипотезы нормальности закона распределения на N опытах построить гистограмму, содержащую $k \approx N/5$ интервалов, равномерно разбивающих весь динамический диапазон экспериментальных данных. Далее следует рассчитать значение хи-квадрат критерия по следующей формуле:

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

где b_i – число опытов, попавших i-тый интервал гистограммы, $\tilde{p}_i$ – ожидаемая теоретическая вероятность попадания в i-тый интервал гистограммы при нормальном законе распределения значений.

Для последующей проверки гипотезы используется хи-квадрат распределение с m=k-3 степенями свободы. Обоснование такого выбора числа степеней свободы обусловлено тем, что Пирсон доказал сходимость распределения (1) к m=k-1 при большом числе опытов, однако это значение понижают на 2 из-за того, что на той же тестовой выборке вычисляют математическое ожидание и среднеквадратическое отклонение нормального закона распределения.

В итоге возникает тупиковая ситуация при N=10, k=2, тогда m= -1. Аналогично при N=15, k=3, тогда m= 0. Отрицательным и нулевым число степеней свободы быть не может и соответственно хи-квадрат для малых тестовых выборок оказывается не применим.

Численный эксперимент, доказывающий существование дискрентного спектра состояний хи-квадрат распределений при малом числе степеней свободы нормального закона распределения значений. Ряд фундаментальных континуально-квантовых эффектов, присутствующих при привычной всем статистической обработки данных трудно наблюдаемы (не очевидны). Нужно специально организовывать численный эксперимент, который позволит надежно зафиксировать эффект.

Проще всего эффект наблюдается, если использовать N=10, применить гистограмму, состоящую из двух столбиков с разделителем в точке математического ожидания исследуемого распределения. Реализуется этот эксперимент использованием нормального генератора вектора из 10 случайных чисел с нулевым математическим ожиданием и единичной дисперсией. Многократный запуск этого генератора дает следующие примеры векторов: {5, 5}, {4, 6}, {5, 5}, {7, 3}, {6, 4},.... Ожидаемые вероятности попадания в первый и второй интервалы будут всегда одинаковы $\tilde{p}_1 = \tilde{p}_2 = 0.5$. Как следствие, выражение (1) дает дискретный спектр конечного числа состояний: 0.0 – 139 раз, 0.2 – 6 раз, 0.4 – 183 раз, 1.0-6 раз, 1.6 -133 раза, 2.6 – 1 раз, 3.6-53 раза, 6.4- 13 раз, 10.0 – 1 раз из 535 опытов. Пример одной из реализаций спектра состояний выражения (1), отображен на рисунке 1.

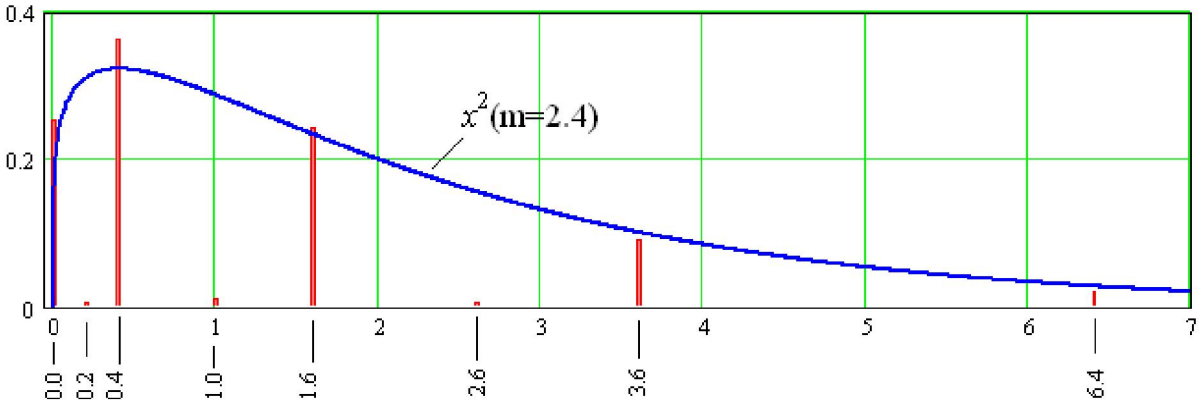

Рисунок 1 – Пример конечного спектра выходных состояний хи-квадрат распределения 10 опытов (минус одна степень числа степеней свободы в теории [1])

Если мы будем использовать выборку из 15 опытов, полученных от генератора случайных чисел с нормальным законом распределения значений и повторим эксперимент 535 раз, то получится хи-квадрат спектр, отображенный на рисунке 2.

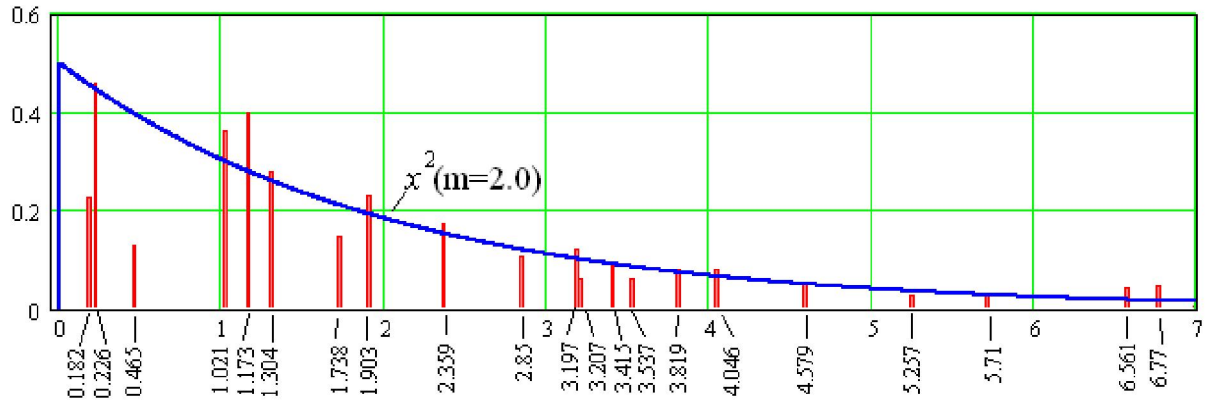

Рисунок 2 – Пример конечного спектра выходных состояний хи-квадрат распределения 15 опытов нормальный закон (нулевой показатель числа степеней свободы в теории [1])

Значительный рост числа спектральных линий обусловлен тем, что для 15 опытов строится гистограмма, содержащая три интервала. Каждый из 535 опытов дает соответствующие примеры частот попадания в три разных интервала: {2, 8, 5}, {4, 9, 2}, {5, 8, 3}, … Для нормального закона распределения значений ожидаемые теоретические вероятности составят $\tilde{p}_1 = 0.159$, $\tilde{p}_2 = 0.682$, $\tilde{p}_2 = 0.682$ $\tilde{p}_3 = 0.159$. Подстановка этих данных в выражение (1) приводит к появлению серии из 22 спектральных линий в интервале от 0 до 7 разной интенсивности.

Еще большее усложнение спектра происходит, если использовать тестовые выборки из 20 опытов. В этом случае 535 опытов дает спектр хи-квадрат распределения, отображенный на рисунке 3.

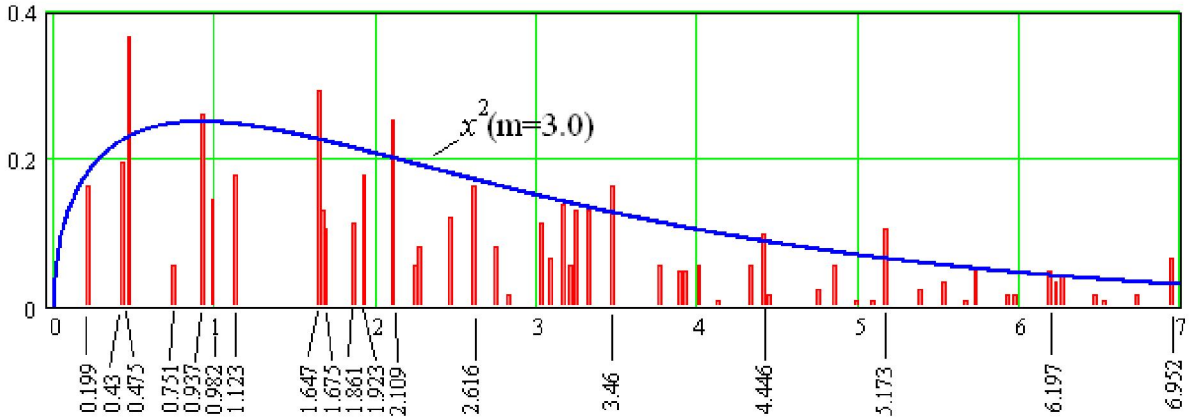

Рисунок 3 – Пример дискретного спектра выходных состояний хи-квадрат распределения для 20 опытов при нормальном законе распределения данных (единичный показатель числа степеней свободы в теории [1])

Всего в интервале от 0 до 7 удается обнаружить 53 спектральные линии с разной степенью интенсивности. Данные об этих спектральных линиях приведены в таблице 1.

Дальнейшее увеличение размеров тестовой выборки приводит к еще большему росту числа спектральных линий. Спектр состояний хи-квадрат для 25 опытов имеет уже более 300 линий (рисунок 3).

Таблица 1 – Положение и интенсивность 85 обнаруженных спектральных линий хи-квадрат распределения для выборки из 20 данных, полученных от генератора псевдослучайных данных с нормальным законом распределения в серии из 631 повторения

χ^2	0.199	0.43	0.475	0.751	0.937	0.982	1.123	1.647	1.675
b_i	21	23	46	6	31	17	21	34	14
χ^2	1.692	1.861	1.923	2.109	2.243	2.278	2.474	2.616	2.750
b_i	12	13	20	32	5	9	14	20	9
χ^2	2.829	3.033	3.095	3.167	3.212	3.246	3.326	3.460	3.770
b_i	2	13	7	16	6	15	15	19	6
χ^2	3.894	3.922	4.018	4.136	4.322	4.418	4.446	4.749	4.846
b_i	5	10	6	1	6	11	2	3	5
χ^2	4.980	5.094	5.173	5.387	5.521	5.673	5.735	5.973	6.197
b_i	1	1	12	3	4	1	6	2	6
χ^2	6.231	6.266	6.473	6.524	6.952	7.121	7.166	7.465	8.079
b_i	4	5	2	1	8	2	2	6	1
χ^2	8.320	8.472	8.506	8.513	8.575	8.799	8.868	8.934	8.934
b_i	2	1	1	2	4	2	1	2	2
χ^2	10.39	10.95	11.35	11.41	11.46	11.52	11.64	12.80	13.09
b_i	1	2	2	1	2	1	1	2	2
χ^2	14.23	14.58	15.03	17.89					
b_i	1	1	1	1					

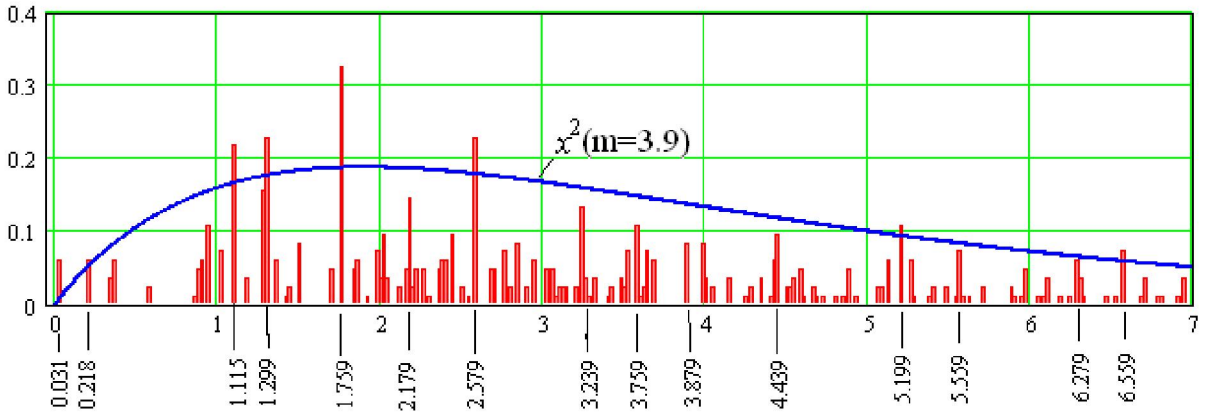

Рисунок 3 – Пример спектра выходных состояний хи-квадрат распределения 25 опытов нормальный закон (степень свободы для хи-квадрат распределения 2 в теории [1])

Численный эксперимент, доказывающий существование дискретного спектра состояний хи-квадрат распределений при малом числе степеней свободы равномерного закона распределения значений. Следует подчеркнуть, что методика проверки статистических гипотез с помощью критерия хи-квадрат [1] универсальна. В связи с этим повторим численный эксперимент для равномерного закона распределения значений при тех же условиях 10, 15, 20, 25 опытов. При этом возникает тот же эффект дискретных спектров возможных состояний хи-квадрат распределений.

В частности, для 10 опытов в тестовой выборке при проведении 535 экспериментов хорошо наблюдаются 5 спектральных линий: 0.0 – 126 раз, 0.4 – 224 раза, 1.6 - 136 раз, 3.6 - 39 раз, 6.4 – 7 раз. Пример распределения спектров данных, полученных для 535 экспериментов, дан на рисунке 4.

Если сравнить рисунок 1 и рисунок 4, станет очевидным различие спектров, которые дают нормальный и равномерный законы распределения значений. Спектр нормального закона богаче, он дает 4 дополнительных линий: 0.2, 1.0, 2.6, 10.0. Появление дополнительных спектральных линий в спектре хи-квадрат распределения нормального закона нельзя списать на ошибку программирования. Это устойчиво повторяющийся факт. Многократное повторение численных экспериментов всегда приводит к появлению дополнительных спектральных линий нормального закона распределения.

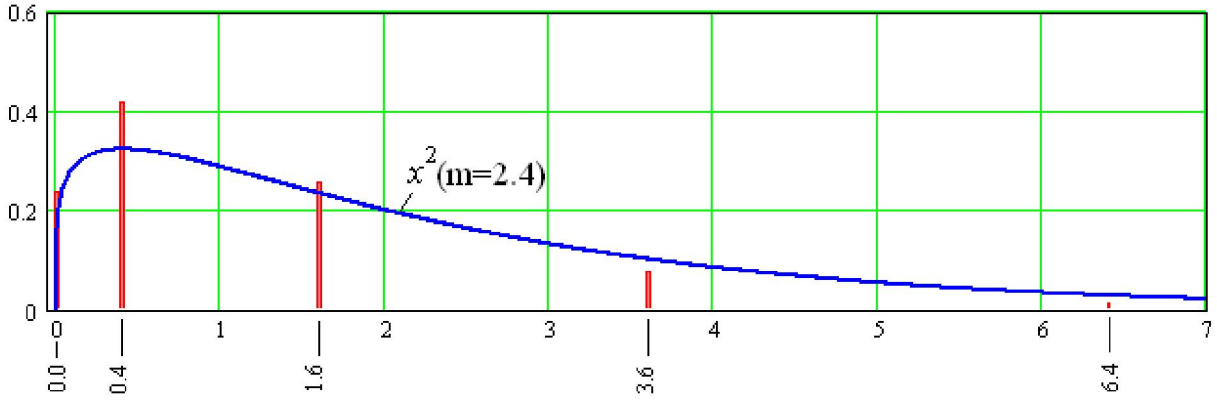

Рисунок 4 – Спектр хи-квадрат распределения (1) для 10 отсчетов равномерного закона распределения

В целом же спектры нормального и равномерных законов имеют очень большие отличия по положению самих спектральных линий и их интенсивности. Для того чтобы убедиться в этом, достаточно сравнить спектр рисунка 2 и спектр рисунка 5, построенные для одной и той же выборки в 15 опытов для нормального и равномерного законов.

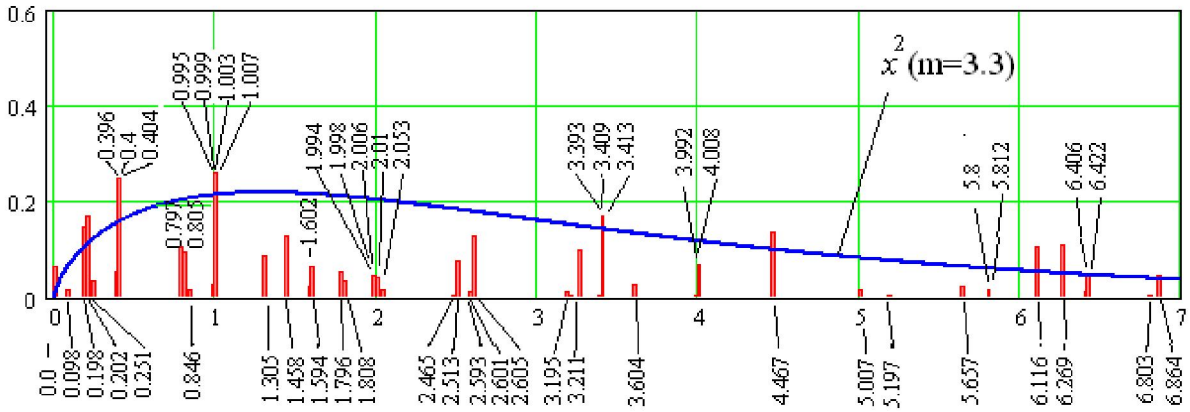

Рисунок 5 – Пример конечного спектра выходных состояний хи-квадрат распределения 15 опытов с равномерным законом (нулевая степень свободы в теории [1])

Участок спектра нормального закона распределения от 0 до 7 (рисунок 2) имеет 22 спектральные линии. На рисунке 5 тот же динамический диапазон хи-квадрат имеет 51 спектральную линию. Чем выше размерность задачи, тем существеннее различия между спектрами. Следует отметить, что сложность спектра (число линий в спектре) является не монотонной функцией. Этот эффект ярко выражен и хорошо наблюдаем. В частности, если перейти к 20 опытам (4 столбцам гистограммы), равномерное распределение дает упрощение спектра (рисунок 6).

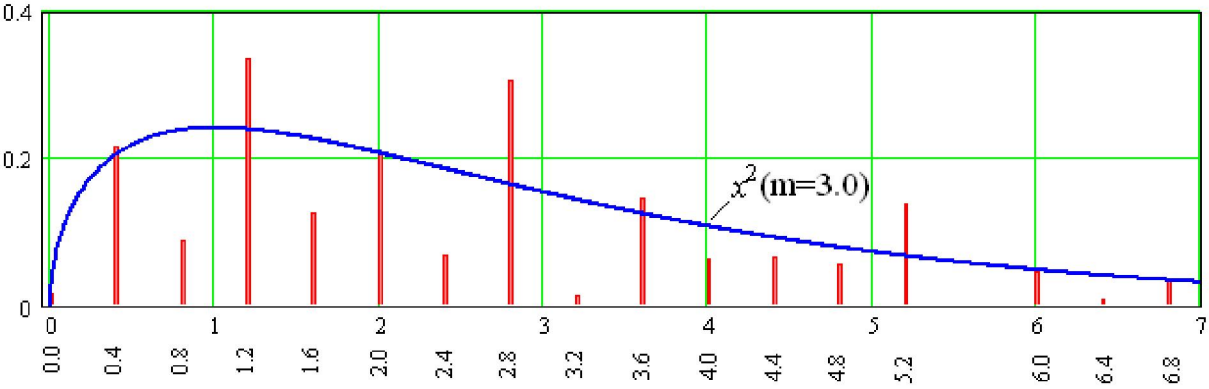

Рисунок 6 – Упрощение спектра при переходе к 20 опытам для равномерного закона распределения значений

Таблица 2 – Положение и интенсивность 29 обнаруженных спектральных линий хи-квадрат распределения выборки из 20 данных, полученных от генератора псевдослучайных данных с нормальным законом распределения значений в серии из 631 повторений

χ^2	0.00	0.40	0.80	1.20	1.60	2.00	2.40	2.80	3.20
b_i	4	5	70	97	67	62	20	91	4
χ^2	3.60	4.00	4.40	4.80	5.2	6.00	6.40	6.80	7.20
b_i	43	18	19	16	41	13	3	10	8
χ^2	7.60	8.00	8.40	8.80	9.20	9.60	10.00	10.80	11.20
b_i	7	1	4	3	4	1	7	4	2
χ^2	12.40	22.00							
b_i	2	1							

В таблице 2 приведены значения положений и интенсивности 29 обнаруженных в спектре линий.

Проверка утверждения Пирсона. Проверим утверждение Пирсона о том, что число степеней свободы хи-квадрат распределения в пределе действительно являются целыми величинами m = k-1. Для этой цели будем считать математическое ожидание данных и среднеквадратическое отклонение известными величинами. Далее будем вычислять хи-квадрат критерий (1) для нормального и равномерного законов распределения значений. При этом будем повторять опыты, пока не убедимся, что результат вычислений стабилизировался. Обычно результат стабилизируется при проведении нескольких миллионов опытов. Данные численных экспериментов для нормального закона приведены в таблице 3. Данные численного эксперимента для равномерного закона приведены в таблице 4.

Таблица 3 – Целые показатели числа степеней свободы - m для плотности распределения значений хи-квадрат при поверке гипотезы нормального закона распределения значений по гистограмме, состоящей из k – столбцов

Число опытов	11	12	13	14	15	16	17	18
Столбцы гистограммы k	2	2	3	3	3	3	3	4
Вычисленное m	1.0	1.0	2.0	2.0	2.0	2.0	2.0	3.0
Теоретическое m	1	1	2	2	2	2	2	3
Число опытов	19	20	21	22	23	24	25	26
Столбцы гистограммы k	4	4	4	4	5	5	5	5
Вычисленное m	3.0	3.0	3.0	3.0	4.0	4.0	4.0	4.0
Теоретическое m	3	3	3	3	5	5	5	5
Число опытов	27	28	29	30	31	32	33	34
Столбцы гистограммы k	5	6	6	6	6	6	7	7
Вычисленное m	4.0	5.0	5.0	5.0	5.0	5.0	6.0	6.0
Теоретическое m	4	5	5	5	5	5	6	6
Число опытов	35	36	37	38	39	40	41	42
Столбцы гистограммы k	7	7	7	8	8	8	8	8
Вычисленное m	6.0	6.0	6.0	7.0	7.0	7.0	7.0	7.0
Теоретическое m	6	6	6	7	7	7	7	7

Сравнивая таблицу 3 и таблицу 4, легко заметить, что утверждение Пирсона верно для 18 опытов и более. В этом случае число степеней свободы в пределе действительно являются целыми величинами m = k-1. Для числа опытов менее 18 показатель числа степеней свободы вполне может являться дробной величиной для всех законов распределения значений, кроме нормального закона.

Таблица 4 – Дробные и целые показатели числа степеней свободы - m для плотности распределения значений хи-квадрат при поверке гипотезы равномерного закона распределения значений по гистограмме, состоящей из k – столбцов

Число опытов / Столбцы гистограммы k	11 / 2	12 / 2	13 / 3	14 / 3	15 / 3	16 / 3	17 / 3	18 / 4
Вычисленное m	1.0	1.0	3.5	3.6	3.74	3.86	4.0	3.0
Теоретическое m	1	1	2	2	2	2	2	3
Число опытов / Столбцы гистограммы k	19 / 4	20 / 4	21 / 4	22 / 4	23 / 5	24 / 5	25 / 5	26 / 5
Вычисленное m	3.0	3.0	3.0	3.0	4.0	4.0	4.0	4.0
Теоретическое m	3	3	3	3	4	4	4	4
Число опытов / Столбцы гистограммы k	27 / 5	28 / 6	29 / 6	30 / 6	31 / 6	32 / 6	33 / 7	34 / 7
Вычисленное m	4.0	5.0	5.0	5.0	5.0	5.0	6.0	6.0
Теоретическое m	4	5	5	5	5	5	5	6
Число опытов / Столбцы гистограммы k	35 / 7	36 / 7	37 / 7	38 / 8	39 / 8	40 / 8	41 / 8	42 / 8
Вычисленное m	6.0	6.0	6.0	7.0	7.0	7.0	7.0	7.0
Теоретическое m	6	6	6	7	7	7	7	7

Эффект от вычисления математического ожидания и дисперсии. Как было показано выше, для наблюдения спектров хи-квадрат распределений нужно принимать специальные меры (писать специальное программное обеспечение). Если же мы будем иметь дело с реальными данными, спектральные линии размазываются при вычислениях, спектр хи-квадрат распределения становится непрерывным. Это следствие того, что мы точно не знаем математического ожидания и дисперсии наблюдаемого закона распределения значений.

Тем не менее, опираясь на полученные выше знания, мы можем существенно снизить ошибку проверки статистических гипотез. Для этой цели необходимо отказаться от стандартной практики [1] использования хи-квадрат распределений с целыми положительными степенями свободы. При вычислении вероятностей ошибок первого и второго рода проверяемых статистических гипотез следует применять описание хи-квадрат распределения через гамма- функцию [2]:

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

Для получения более точных оценок следует использовать дробные (фрактальные) значения показателей степеней свободы. В таблице 5 даны значения дробных показателей числа степеней свободы для проверки гипотезы нормального закона распределения значений, вычисленные как математическое ожидание выражения (1) при 1 000 000 повторений. В таблице 6 даны дробные значения показателя числа степеней свободы, рекомендуемые к использованию при проверке гипотезы равномерного закона распределения значений.

Сравнивая между собой таблицу 5 и таблицу 6, мы видим, что показатели числа степеней свободы действительно снижаются примерно $m \approx k-3$, как этого требуют стандартные методики [1, 2]. Однако показатели числа степеней свободы хи-квадрат распределения являются дробными величинами. Наблюдаются существенное расхождение значения показателей размерности вычисленного и заложенного в стандартные методики [1, 2].

Таблица 5 – Дробные показатели числа степеней свободы - m для плотности распределения значений хи-квадрат при поверке гипотезы нормального закона распределения значений по гистограмме, состоящей из k – столбцов

Число опытов	11	12	13	14	15	16	17	18
Столбцы гистограммы k	2	2	3	3	3	3	3	4
Вычисленное m	0.363	0.363	0.727	0.727	0.727	0.727	0.727	1.157
Теоретическое m	-1	-1	0	0	0	0	0	1
Число опытов	19	20	21	22	23	24	25	26
Столбцы гистограммы k	4	4	4	4	5	5	5	5
Вычисленное m	1.157	1.157	1.157	1.157	2.447	2.447	2.447	2.447
Теоретическое m	1	1	1	1	2	2	2	2
Число опытов	27	28	29	30	31	32	33	34
Столбцы гистограммы k	5	6	6	6	6	6	7	7
Вычисленное m	2.447	3.337	3.337	3.337	3.337	3.337	4.264	4.264
Теоретическое m	2	3	3	3	3	3	4	4
Число опытов	35	36	37	38	39	40	41	42
Столбцы гистограммы k	7	7	7	8	8	8	8	8
Вычисленное m	4.264	4.264	4.264	5.205	5.205	5.205	5.205	5.205
Теоретическое m	4	4	4	5	5	5	5	5

Таблица 6 – Дробные показатели числа степеней свободы - m для плотности распределения значений хи-квадрат при поверке гипотезы равномерного закона распределения значений по гистограмме, состоящей из k – столбцов

Число опытов	11	12	13	14	15	16	17	18
Столбцы гистограммы k	2	2	3	3	3	3	3	4
Вычисленное m	0.333	0.333	1.351	1.379	1.404	1.426	1.458	1.788
Теоретическое m	-1	-1	0	0	0	0	0	1
Число опытов	19	20	21	22	23	24	25	26
Столбцы гистограммы k	4	4	4	4	5	5	5	5
Вычисленное m	1.787	1.777	1.767	1.767	3.177	3.177	3.177	3.177
Теоретическое m	1	1	1	1	2	2	2	2
Число опытов	27	28	29	30	31	32	33	34
Столбцы гистограммы k	5	6	6	6	6	6	7	7
Вычисленное m	3.177	4.188	4.188	4.188	4.188	4.188	5.415	5.415
Теоретическое m	2	3	3	3	3	3	4	4
Число опытов	35	36	37	38	39	40	41	42
Столбцы гистограммы k	7	7	7	8	8	8	8	8
Вычисленное m	5.415	5.415	5.415	6.488	6.488	6.488	6.488	6.488
Теоретическое m	4	4	4	5	5	5	5	5

Заключение. В литературе по статистике хи-квадрат распределение традиционно рассматривается как непрерывное. Однако в действительности хи-квадрат распределение является дискретным распределением с очень сложным спектром возможных состояний. Считать хи-квадрат распределение непрерывным можно только для большого числа. Когда речь идет о малом числе опытов, необходимо учитывать дискретный характер плотности распределения хи-квадрат

распределения. Хи-квадрат преобразование является континуально-квантовым. Его входные данные являются континуумами, а выходные состояния дискретны. Именно по этой причине возникают дефекты размерности, отраженные в приведенных выше таблицах, которые могут приводить к значительным ошибкам при статистических оценках, строящихся на малых тестовых выборках.

ЛИТЕРАТУРА

[1] Р 50.1.037-2002 Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. – Ч. I. Критерии типа χ2. Госстандарт России. – М., 2001.

[2] Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. – М.: ФИЗМАТЛИТ, 2006. – 816 с.

REFERENCES

[1] R 50.1.037-2002 Recommendations for standardization. Applied statistics. Validation rules of the consent of an experienced distribution with the theoretical. Ch. I. χ2 type criteria. GosStandart of Russia. – M., 2001. (in Russ.)

[2] Kobzar A.I. Applied mathematical statistics. For engineers and scientists. M.: FIZMATLIT, 2006. 816 p. (in Russ.)

КІШІ ТЕСТІЛІК ТАҢДАУЛАР КЕЗІНДЕ ХИ-КВАДРАТ КРИТЕРИЙДІҢ ТАРАТУ ЗАҢЫНЫҢ ДИСКРЕТТІ СИПАТТАМАСЫ

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

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

Тірек сөздер: мәндердің хи-квадрат тарату, фракталдар, бостандық дәрежелер санының бөлшекті көрсеткіші, гистограммалар, тестілік таңдаудың көлемін қысқарту, мәндерді таратудың дискретті заңы.

Аннотация. Кіші тестілік таңдау кезінде хи-квадрат таратудың күйлер саны шектелгені және мүмкінді шығыс күйлердің дискретті спектрін тұдыратыны дәлелденген. Хи-квадрат таратудың бостандық дәрежелер саны әрқашан бөлшекті өлшем болады. Бүгінгі күнде практикада бостандық дәрежелер санның көрсеткіштерін бүтін мән ретінде қолдану тестілік таңдаудың көлемін ақталмаған жоғарлатуына және бағаның шынайлығын төмендетуге алып келеді.

Поступила 15.01.2015 г.