УДК 519.24
МРНТИ 81.96.00

ПЕРСПЕКТИВА ИСПОЛЬЗОВАНИЯ МАТЕМАТИЧЕСКОЙ ХИ-КВАДРАТ МОЛЕКУЛЫ ПИРСОНА ДЛЯ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ НА МАЛЫХ ВЫБОРКАХ

В.И. ВОЛЧИХИН[1], Б.Б. АХМЕТОВ[2], Ж.К. АЛИМСЕИТОВА[3], Г.С. БЕКЕТОВА[4], Л.О. ЖАХАН[4]

[1]ФБГОУ ВПО «Пензенский государственный университет»
[2]Есенов университет
[3]Алматинский университет энергетики и связи
[4]Кызылординский государственный университет имени Коркыт Ата

Аннотация: *Целью работы является усиление мощности хи-квадрат критерия на малых выборках. Актуальность обусловлена тем, что данных биометрии и других ряда других практических приложений для классического статистического анализа недостаточно. Предложено воспользоваться средствами имитационного моделирования и численно получать плотность распределения значений хи-квадрат критерия для малых выборок. При синхронизации интервалов гистограммы с математическим ожиданием выборки спектр состояний хи-квадрат становится дискретным. При вычислениях удается поддерживать эффект квантовой суперпозиции анализируемых данных, применяя обычные компьютеры. В отличие от континуально-квантовых уравнений Шредингера, континуально-квантовые уравнения Пирсона оказались гораздо более удобными для реализации программными методами континуально-квантовой суперпозиции. Сделано предположение, что возможно создать усилитель хи-квадрат критерия, позволяющий увеличить его мощность в несколько раз. Дан подсчет числа промежуточных данных такого континуально-квантового вычислителя для выборки из 16 примеров.*

Ключевые слова: *квантовая суперпозиция, хи-квадрат критерий Пирсона, дискретный спектр состояний, статистический анализ малых выборок*

THE PERSPECTIVE OF THE USE OF MATHEMATICAL CHI-SQUARE PEARSON MOLECULE TO CHECK STATISTICAL HYPOTHESIS IN SMALL ELECTIONS

Abstract: *The aim of the work is to increase the power of the chi-square criterion on small samples. The relevance is due to the fact that data from biometrics and other a number of other practical applications for classical statistical analysis is not enough. It is proposed to use simulation tools and numerically obtain the distribution density of the chi-square test for small samples. When synchronizing histogram intervals with the expectation of a sample, the spectrum of chi-square states becomes discrete. In calculations, it is possible to maintain the effect of quantum superposition of the analyzed data, using ordinary computers. In contrast to the continual-quantum Schrödinger equations, the continual-quantum Pearson equations turned out to be much more convenient for the implementation of the continuum-quantum superposition by the program methods. It has been suggested that it is possible to create an amplifier chi-square criterion, which allows to increase its power several times. Given the number of intermediate data of such a continually quantum calculator for a sample of 16 examples.*

Keywords: *quantum superposition, Chi-square Pearson criterion, discrete spectrum of states, statistical analysis of small samples*

ШАҒЫН ІРІКТЕМЕЛЕРДЕ СТАТИСТИКАЛЫҚ ГИПОТЕЗАЛАРДЫ ТЕКСЕРУ ҮШІН ПИРСОН МОЛЕКУЛАСЫНЫҢ МАТЕМАТИКАЛЫҚ ХИ-КВАДРАТЫН ПАЙДАЛАНУ ПЕРСПЕКТИВАСЫ

Аңдатпа: *Жұмыстың мақсаты шағын іріктемелерде хи-квадрат критерийдің күшін нығайту. Өзектілігі биометрияның және басқа да практикалық қосымшалардың деректері классикалық статистикалық талдау үшін жеткіліксіз болғанынан шығады. Шағын іріктемелер үшін хи-квадрат критерий мәндерінің таралу тығыздығын сан ретінде алуға имитациялық модельдеу құралдарын қолдану ұсынылған. Гистограмма интервалдарын іріктеменің математикалық күтуімен синхронизациялау кезінде хи-квадрат күйлерінің спектрі дискретті болады. Есептеулер кезінде талданатын деректердің кванттық суперпозиция эффекті кәдімгі компьютерлерді қолданамыз. Континуалды-кванттық суперпозицияны бағдарламалық әдістермен жүзеге асыру үшін Шредингердің континуалды-кванттық теңдеулеріне қарағанда Пирсонның континуалды-кванттық теңдеулері әлдеқайда ыңғайлы болып шықты. Хи-квадрат критерийдің күшін бірнеше ретке жоғарылататын күшейткішті жасауға болатыны туралы болжам жасалған. 16 мысалдан тұратын іріктеме үшін сондай континуалды-кванттық есептеуіш аралық деректер санын есептеу берілген.*

Түйінді сөздер: *кванттық суперпозиция, Пирсонның хи-квадрат критерийі, күйлердің дискреттік спектрі, шағын іріктемелердің статистикалық талдауы*

Критерий хи-квадрат был предложен Карлом Пирсоном в 1900 году и по своей сути является фундаментом современной математической статистики [1, 2]. Важнейшим для практики свойством хи-квадрат критерия является простота его вычисления:

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

где N – размер тестовой выборки, k – число интервалов гистограммы, P_i – теоретическая вероятность попадания в i-тый интервал гистограммы, b_i – частота появления данных в i-том интервале гистограммы.

Популярность критерия хи-квадрат обусловлена тем, что для больших выборок Карл Пирсон нашел аналитическое описание закона распределения значений этого критерия:

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

где m – число степеней свободы, $\Gamma\{.\}$ – гамма функция.

Так как мы знаем число опытов – N, то имеем возможность выбирать число интервалов гистограммы – k = round(N/5) и, соответственно, определить число степеней свободы – m=k-3. В итоге, воспользовавшись соотношением (2), можем найти доверительные вероятности принятия того или иного статистического решения [1, 2].

Основной проблемой хи-квадрат критерия является то, что Пирсон нашел асимптотическое решение (2), применимое для больших выборок. Обычно, применяя асимптотическое решение (2), стараются использовать выборки, имеющие от 200 до 400 опытов. В этом случае удается достаточно надежно оценивать вероятности ошибок первого и второго рода $P_1 \approx P_2 \approx P_{EE} \approx 0.01$.

Очевидно, что в жизни возникает множество ситуаций, когда получить несколько сотен опытов технически и/или организационно невозможно. Такие ситуации регулярно возникают в экономике, медицине, биометрии, биологии, химии. Наиболее остро ситуация складывается в биометрии [3, 4], для которой вероятность ошибок первого рода (ошибочный отказ в доступе «Своему») составляет $P_1 \approx 0.05$ при допустимой вероятности ошибок второго рода (ошибочный пропуск «Чужого») $P_2 \approx 0.00001$.

В данной статье авторы пытаются показать, что огромные вычислительные сложности «счастливые» игроки преодолевают, организовывая на подсознательном уровне квантовые вычисления. Естественные и искусственные нейронные сети [5] оказались очень удобны для поддержки эффектов квантовой суперпозиции.

Классические квантовые вычисления [6] и квантовые вычисления нейросетевого подсознания игрока в карты [5] имеют много общего. В связи с этим, по аналогии с планетарной моделью молекулы водорода, введем математическую молекулу Пирсона. Обе эти конструкции иллюстрируются рисунком 1.

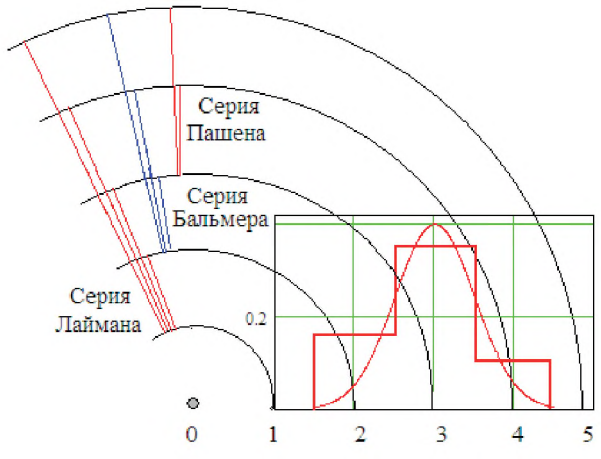

Рис. 1 – Планетарная модель молекулы водорода, построенная исходя из гипотезы нормального закона распределения значений континуума состояний электрона, с квантованием данных по 3 орбитам (по 3 столбцам гистограммы математической молекулы Пирсона)

В планетарной модели атома водорода электрон излучает фотон света, перескакивая с одной орбиты на другую. Как следствие, спектр (поглощения, излучения) водорода содержит линии (серии линий). В левой части рисунка 1 даны соответствующие названия серий линий спектра водорода.

Математическая молекула Пирсона не должна точно соответствовать физической реальности. Она соответствует некоторому приближению физических процессов, в соответствии с которыми скорость электронов молекулы имеет нормальное распределение значений, число орбит равно шести и соответствует гистограмме распределения скоростей электронов с шестью интервалами. Всего математическая молекула имеет число электронов, равное числу опытов в тестовой выборке, подвергаемой статистическому анализу.

Очевидно, что таких математических молекул должно существовать множество. Нас будут интересовать только математические молекулы Пирсона, соответствующие малым тестовым выборкам, для которых предельное аналитическое описание (2) очень плохо работает.

Синтезируя математическую молекулу, мы заменили электроны числом опытов и функцию пространственно-волнового квантования уравнения Шредингера на куда более простую функцию пространственного квантования данных, классического построения гистограмм. При этом уравнения, соответствующие этим двум конструкциям молекул должны сохранять некоторое подобие.

Следует отметить то, что при больших объемах выборки хи-квадрат критерий имеет очень много состояний и его спектр действительно можно считать непрерывным. Однако положение меняется, когда объем выборок катастрофически падает [6-8]. Одной из причин, по которой ранее не удавалось наблюдать дискретный характер спектра, состоит в отсутствии синхронизации между столбцами гистограммы и параметрами выборки. Обычно осуществляют нормирование опытных данных и находят интервал между максимальным и минимальным значением в тестовой выборке. Далее вычисляют ширину интервала для нормированной гистограммы. В нашем случае будем вычислять ширину трех столбцов гистограммы по следующей формуле:

$$\Delta = \frac{6 \cdot \sigma(x)}{3} \qquad (3).$$

Синхронизацию данных будем осуществлять через привязку минимального значения выборки к правой границе крайнего правого интервала гистограммы. То есть, границы ин-

тервалов гистограммы зададим следующим образом:

$$\{\min(x),\ \min(x)+\Delta,\ \min(x)+2\Delta\} \quad (4)$$

Тогда молекула Пирсона с 9 электронами, имеющими нормальный закон распределения, будет иметь дискретный спектр с 64 возможными дискретными состояниями, отображенными в верхней части рисунка 2.

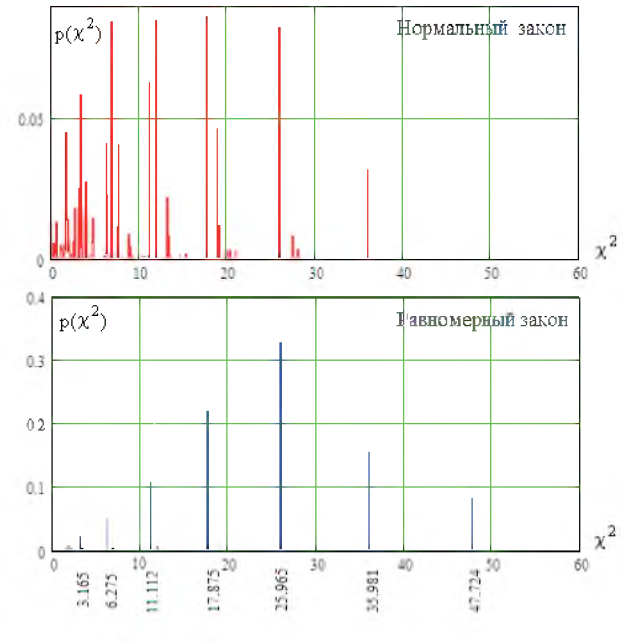

Рис. 2 – Спектры состояний хи-квадрат молекулы с нормальным и равномерным законами распределения континуума 9 состояний по 3 уровням

Как видно из нижней части рисунка 2, спектр состояний хи-квадрат молекулы имеет всего 13 состояний, если континнуум внутренних состояний имеет равномерный закон распределения значений.

Семь наиболее мощных спектральных компонент хи-квадрат молекулы с равномерным внутренним спектром, отмечены цифрами значений хи-квадрат в нижней части рисунка 2, они появляются с вероятностью 0.965 (вероятность ошибки первого рода P_1=0.045). Те же спектральные компоненты хи-квадрат молекулы для нормального закона распределений внутреннего состояния появляются с вероятностью 0.329 (вероятность ошибки второго рода P_2=0.671).

Очевидно, что увеличение числа опытов в анализируемой выборке позволяет усложнить спектр состояний хи-квадрат молекулы и одновременно снизить вероятность ошибок проверки статистических гипотез. На рисунке 3 даны спектры хи-квадрат молекул с 16 электронами, случайно размещаемых на 3 орбиталях.

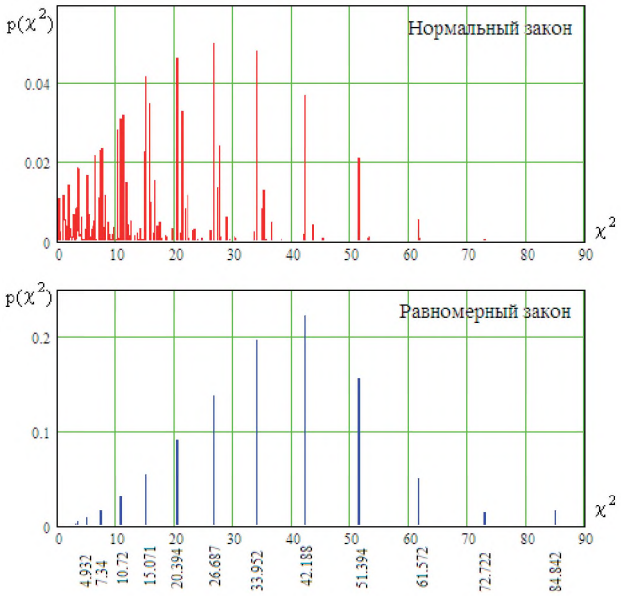

Рис. 3 – Спектры состояний хи-квадрат молекулы с нормальным и равномерным законами распределения континуума 16 состояний по 3 уровням

Увеличение числа примеров в выборке до 16 примеров приводит к незначительному росту мощности спектральных состояний, отмеченных в нижней части рисунка 3. При этом вероятность появления 13 состояний составляет 0.991 (вероятности ошибок первого рода P_1=0.009).

Если появление одного из этих 13 состояний считать обнаружением равномерного закона, то появление любого иного состояния для нормального закона будет возникать с вероятностью 0.321 (вероятность ошибки второго рода P_2=0.679).

Рассмотренное выше решающее правило примитивно, однако оно позволяет в первом приближении найти оценку равновероятных ошибок. Они оказываются сопоставимы:

$$\frac{0.045+0.671}{2} = 0.358 > 0.344 = \frac{0.009+0,679}{2} \quad (5)$$

В первом приближении мы получили подтверждение давно известного факта о низкой эффективности хи-квадрат критерия для одиночных ациклических вычислительных процедур. Почти двукратное увеличение малой тестовой выборки с 9 до 16 опытов приводит к снижению равновероятной ошибки $P_1 \approx P_2 \approx P_{EE}$ примерно на 5%. Для того, чтобы сделать ошибку приемлемой приходится многократно увеличивать размеры тестовой выборки. В этом отношении ациклические квантовые машины имеют возможности, сопоставимые с ациклическими машинами, рассматривающими спектр хи-квадрат критерия как непрерывный [9, 10].

Если попросить человека сравнить спектры в верхней и нижней части рисунка 3, то человек их никогда не перепутает. То есть вероятности ошибок $P_1 \approx P_2 \approx P_{EE}$ оказываются практически нулевыми. То же самое относится и к спектрам рисунка 2, для человека эксперта эти спектральные образы абсолютно разные. То есть, мы можем заранее обучить две нейронные сети распознаванию двух разных образов, используя данные 200 000 опытов [4]. Далее уже при меньшем числе опытов можно проанализировать выходные данные первой и второй нейронной сети и решить, на сколько анализируемое распределение близко к нормальному или равномерному.

При этом будем исходить из размеров тестовой выборки в 16 примеров. В этом случае из полной тестовой выборки можно получить достаточно много частных подвыборок меньшего размера по 9 опытов $C_{16}^{9} = 1144C$. Если использовать и другие подвыборки, то получится:

$$C_{16}^{9} + C_{16}^{8} + C_{16}^{7} + C_{16}^{6} + C_{16}^{5} + C_{16}^{4} + C_{16}^{3} + C_{16}^{2} + C_{16}^{1} = 26332 \quad (6)$$

Если бы все 26 332 были бы независимы, то вероятности ошибок были бы пренебрежимо малы, однако независимость данных, получаемых из одной выборки, невозможна. Тем не менее, воспользовавшись 26 332 опытами, мы можем построить спектры выходных состояний, похожие на спектры, приведенные на рисунках 2 и 3 и обучать на этих образах искусственные нейронные сети [4]. Сильные корреляционные связи в исходных данных не мешают, а помогают распознаванию образов [8].

Заключение

Дискретный спектр состояний хи-квадрат молекулы делает эту математическую конструкцию перспективной для создания квантовых вычислителей, ориентированных на работу с малыми тестовыми выборками. Предположительно использование перспективного квантового усилителя мощности хи-квадрат критерия позволит снизить требования к тестовой выборке с 2000 примеров до 20 примеров при сохранении тех же вероятностей ошибок первого и второго рода.

Еще одним важным моментом рассмотренной темы является то, что трехуровневая молекула Пирсона является некоторым аналогом физически существующей молекулы водорода. То есть уравнения Пирсона (1) и (2) и уравнение Шредингера являются топологическими аналогами. Однако континуально-квантовое уравнение хи-квадрат Пирсона намного удобнее для реализации квантового вычислителя, чем уравнение Шредингена. Моделировать уравнение Шредингера при 9 и 16 электронах намного сложнее, чем моделировать уравнения хи-квадрат Пирсона.

То, что для уравнения Шредингера удалось построить более удобный для реализации математический аналог в виде уравнений Пирсона, является важным шагом в проектировании перспективных квантовых вычислителей.

ЛИТЕРАТУРА

1. Абезгауз Г.Г., Тронь А.П., Копенкин Ю.Н., Коровина И.А. Справочник по вероятностным расчетами. – М.: Воениздат, 1970 г. – 536 с.
2. 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.
3. Болл Руд и др. Руководство по биометрии. / Болл Руд, Коннел Джонатан Х., Панканти Шарат, Ратха Налини К., Сеньор Эндрю У. // Москва: Техносфера, 2007. – 368 с. (перевод с английского).
4. 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.
5. 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
6. Нильсон М., Чанг И. Квантовые вычисления и квантовая информация. – М.: Мир, 2006 г. – 821 с.
7. 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.
8. Иванов А.И. Квантовые компьютеры: прошлое, настоящее, будущее. // «Защита информации. INSAID». – № 2. – 2015 г. – С. 29-32.
9. Ахметов Б.Б., Иванов А.И. Оценка качества малой выборки биометрических данных с использованием более экономичной формы хи-квадрат критерия. // Надежность. №2 (57) . – 2016 г. – С. 43-48. DOI: 10.21683/1729-2640-2016-16-2-43-48.
10. Иванов А.И., Газин А.И., Вятчанин С.Е., Перфилов К.А. Сравнение мощности хи-квадрат критерия и критерия Крамера-фон Мезиса для малых тестовых выборок биометрических данных «Надежность и качество сложных систем». – №2. – 2016. – С. 67-73.