УДК 330.4

МОДИФИЦИРОВАННЫЙ КРИТЕРИЙ ПИРСОНА В ЭКОНОМИЧЕСКИХ ИССЛЕДОВАНИЯХ

Рязанский В.П.

НАО «Научно-технический центр», Москва, e-mail: math@intmail.com

Критерий согласия χ2 чаще других статистических критериев используется в экономических и социологических исследованиях. Широкое распространение является причиной некорректности его применения в некоторых случаях. Нами предпринята попытка заострить внимание на особенностях использования критерия χ2. Работа с ним требует использования программного обеспечения, например специализированных пакетов Statgraphics, STATISTICA или MATLAB. Нормальный закон распределения вероятностей получил наибольшее распространение в практических задачах обработки экспериментальных данных. Большинство прикладных методов математической статистики исходит из предположения нормальности распределения вероятностей изучаемых случайных величин. Широкое распространение этого распределения вызвало необходимость разработки специальных критериев согласия эмпирических распределений с нормальным. Существуют как модификации общих критериев согласия, так и критерии, созданные специально для проверки нормальности. Название критерия обусловлено названием непрерывного распределения, к которому сходится статистика критерия по распределению. В случае, когда есть только две взаимоисключающие гипотезы, говорят, что произошла ошибка первого рода, если основная гипотеза отвергнута критерием, тогда как она верна. Вероятность ошибки первого рода называется уровнем значимости критерия.

Ключевые слова: критерий, распределение, Пирсон, гипотеза, вероятность

MODIFIED PEARSON CRITERION IN ECONOMIC RESEARCH

Ryazanskiy V.P.

LOD «Scientific-Technical Center», Moscow, e-mail: math@intmail.com

Criterion consent χ2 statistical criteria most often used in economic and sociological research. Widespread cause incorrectness is its application in some cases. We have attempted to focus on the specifics of the use of χ2 test. Working with them requires the use of software, for example, specialized packages Statgraphics, STATISTICA or MATLAB. The normal law of probability distribution is most prevalent in the practice of data processing tasks. Most of the application of methods of mathematical statistics is based on the assumption of a normal probability distribution of random variables studied. The wide spread of the distribution has necessitated the development of special criteria for the consent of the empirical distributions to normal. There is a modification of the consent of the general criteria and the criteria set up specifically to check normality. The name is due to the name of the criterion of continuous distribution, which converges statistics on the distribution criteria. In the case where there are only two mutually exclusive hypotheses, say that there was an error of the first kind, if the main criterion of the hypothesis is rejected, then it is true. The probability of error of the first kind is called the level of significance criterion.

Keywords: criterion, distribution, Pearson, hypothesis, probability

Критерий согласия Пирсона (χ^2) применяют для проверки гипотезы о соответствии эмпирического распределения предполагаемому теоретическому распределению $F(x)$ при большом объеме выборки ($n \geq 100$). Критерий применим для любых видов функции $F(x)$, даже при неизвестных значениях их параметров, что обычно имеет место при анализе результатов механических испытаний. В этом заключается его универсальность.

Использование критерия χ^2 предусматривает разбиение размаха варьирования выборки на интервалы и определения числа наблюдений (частоты) n_j для каждого из e интервалов. Для удобства оценок параметров распределения интервалы выбирают одинаковой длины.

Число интервалов зависит от объема выборки. Обычно принимают: при $n = 100$ $e = 10–15$, при n = 200 $e = 15–20$, при $n = 400$ $e = 25–30$, при $n = 1000$ $e = 35–40$.

Интервалы, содержащие менее пяти наблюдений, объединяют с соседними. Однако, если число таких интервалов составляет менее 20% от их общего количества, допускаются интервалы с частотой $n_j \geq 2$.

Предлагаемая модификация критерия Пирсона [1, с. 80] позволяет проверять гипотезу о предполагаемом распределении генеральной совокупности [5, с. 51], обладающей функцией распределения

$$F\left(x,\vec{r}\right)=F\left(x\right),$$

где $\vec{r}=\left(r_{1,\ldots,}r_s\right)$ – известный вектор параметров распределения [2, с. 22].

Разобьем носитель случайной величины на m равновероятных интервалов следующим образом:

$$1/m = F(b_j) - F(a_j) = p; \quad F(a_1) = 0;$$

$$F(b_j) = j/m, \quad j = 1, \ldots, m.$$

Имеется выборка $x_1, \ldots, x_n$ из генеральной совокупности, с указанным выше распределением [2, с. 22].

Рассмотрим произвольный интервал (a_j, b_j) на носителе случайной величины [3, с. 68]. Любое наблюдение из выборки с вероятностью $p = 1/m$ попадает в указанный интервал и с дополнительной вероятностью, равной $q = 1 - p$, не попадает в него. Для случайной величины v_i – числа наблюдений из выборки, попавших в указанный интервал, получаем простую схему Бернулли с вероятностью успеха при одном испытании p и числом испытаний n. Таким образом, получаем m простых схем Бернулли при одинаковых вероятностях успеха p и числа испытаний n [4, с. 392].

В силу локальной теоремы Муавра – Лапласа случайные величины $\frac{v_i - np}{\sqrt{npq}}$, $j = 1, \ldots, m$ имеют распределение близкое к стандартному нормальному. Предлагаемая модификация критерия Пирсона заключается в выборе критической статистики в следующем виде:

$$\Lambda(x_1,\ldots,x_n)=\Lambda=\sum_{j=1}^{m}\frac{|v_i-np|}{\sqrt{npq}}.$$

Выбор критической статистики в таком виде обусловлен более устойчивыми ее свойствами. Для применения данного критерия необходимо найти функцию распределения статистики Λ, то есть функцию распределения случайной величины Y_m, где

$$Y_m=\sum_{j=1}^{m}|Z_j|;\ \ Z_j\sim N(0,1),\ \ j=1,\ldots,m.$$

Запишем функцию распределения случайной величины $Z=|Z_j|$ по определению:

$$F_z(x) = P(Z < x) = P(-x < Z_j < x) = 2\Phi(x),$$

где $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-t^2/2}dt.$

После дифференцирования левой и правой частей получим выражение для функции плотности Z:

$$\frac{dF_z}{dx}=p_z(x)=\sqrt{\frac{2}{\pi}}e^{-x^2/2}$$

при $x > 0$ и 0 иначе.

Для дальнейших рассуждений нам потребуется характеристическая функция Z, которая есть по определению:

$$f(t)=\int_0^\infty e^{itx}p_z(x)dx.$$

Продифференцируем левую и правую части равенства по t:

$$\frac{df}{dt}=i\sqrt{\frac{2}{\pi}}\int_0^\infty e^{itx}e^{-x^2/2}dx=i\sqrt{\frac{2}{\pi}}\int_0^\infty e^{itx}de^{-x^2/2}.$$

Далее, интегрируя по частям, приходим к следующей задаче Коши:

$$\frac{df}{dt}+tf(t)=i\sqrt{\frac{2}{\pi}}$$

при начальных условиях $f(0) = 1$.

Как нетрудно проверить, решение этого обыкновенного дифференциального уравнения есть:

$$f(t)=e^{-t^2/2}\left(1+\mathrm{erf}\left(\frac{it}{\sqrt{2}}\right)\right),$$

где $\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$

Интересно отметить, что полученная функция выражается через функцию Фаддеева:

$$f(t)\equiv\omega\left(t/\sqrt{2}\right),$$

где $\omega(x)=e^{-x^2}\left(1+\mathrm{erf}(ix)\right)$ – есть функция Фаддеева. Тогда характеристическая функция случайной величины

$$Y\equiv Y_m=\sum_1^m|Z_j|\ f_Y(t)\equiv f_m(t)=f^m(t)=e^{-mt^2/2}\left(1+\mathrm{erf}\left(it/\sqrt{2}\right)\right)^m.$$

Для нахождения центральных моментов случайной величины Y_m вычислим производные от ее характеристической функции в точке $t = 0$:

$$\frac{df_Y(t)}{dt}=\frac{d\left(e^{-mt^2/2}\left(1+\mathrm{erf}\left(it/\sqrt{2}\right)\right)^m\right)}{dt}=\ im\sqrt{\frac{2}{\pi}}.$$

Отсюда получаем

$$\mu_1=m\sqrt{\frac{2}{\pi}}=M[Y_m].$$

Для нахождения второго центрального момента вычислим вторую производную от характеристической функции в точке $t = 0$:

$$\frac{d^2 f\gamma_m(t)}{dt^2} = i^2\left(m + m(m-1)2/\pi\right),$$

и тогда

$$\mu_2 = \left(m + m(m-1)2/\pi\right).$$

Таким образом, дисперсия

$$D\left[Y_m\right] = \mu_2 - \mu_1^2 = m(1-2/\pi) \approx 0{,}3633m.$$

Дисперсия у Хи-квадрат распределения такова

$$D\left[x^2(m)\right] = 2m.$$

Отношение дисперсий очень красноречиво:

$$\frac{D\left[x^2(m)\right]}{D\left[Y_m\right]} = \frac{2m}{m(1-2/\pi)} \approx 5{,}5.$$

Вернемся теперь к вычислению функции плотности распределения $p_Y(x) = p_m(x)$ через её характеристическую функцию:

$$p_m(x) = \frac{1}{2\pi}\int_0^\infty e^{-itx} f_m(t)dt.$$

Дифференцируя обе части, получаем следующее:

$$\frac{dp_m(x)}{dx} = \frac{1}{2\pi}\int_0^\infty (-it)e^{-itx} f_m(t)dt = \frac{-i}{2\pi}\int_0^\infty te^{-itx}e^{-mt^2/2}\left(1 + erf\left(it/\sqrt{2}\right)\right)^m dt =$$

$$= \frac{i}{2m\pi}\int_0^\infty e^{itx}\left(1 + erf\left(it/\sqrt{2}\right)\right)^m de^{-mt^2/2}.$$

После интегрирования по частям имеем

$$\frac{dp_m(x)}{dx} = -\frac{xp_m}{m} + \sqrt{\frac{2}{\pi}}\frac{1}{2\pi}\int_0^\infty e^{-itx}\left(1 + erf\left(it/\sqrt{2}\right)\right)^{(m-1)} e^{-(m-1)t^2/2}dt\ . \qquad (1)$$

Что приводит к следующей задаче Коши:

$$\frac{dp_m(x)}{dx} = -\frac{xp_m(x)}{m} + \sqrt{\frac{2}{\pi}}p_{m-1}(x). \qquad (2)$$

$$p_m(0) = 0.$$

Положим $m = 2$, получим уравнение для функции плотности распределения такой случайной величины:

$$Y_2 = |Z_1| + |Z_2|; \quad \frac{dp_2(x)}{dx} = -\frac{xp_2(x)}{2} + \sqrt{\frac{2}{\pi}}p_1(x),$$

где $p_1(x) = \sqrt{\frac{2}{\pi}}e^{-x^2/2}$.

Решая эту задачу Коши, получаем

$$p_2(x) = \frac{2}{\sqrt{\pi}}e^{-x^2/4}erf\left(\frac{x}{2}\right), \quad x \ge 0; \quad p_2(x) = 0, \quad x < 0.$$

Решение при $m > 2$ в явном виде найти сложнее. Поэтому удобнее воспользоваться численными методами. Для этого запишем систему обыкновенных дифференциальных уравнений в векторном виде:

$$\frac{d\vec{y}}{dx} = A\vec{y} + \vec{b}; \quad y(\vec{0}) = 0; \quad \vec{y} = \left(p_2(x), \ldots, p_m(x)\right)^T; \quad \vec{b} = \left(\sqrt{\frac{2}{\pi}}p_1(x), 0, \ldots, 0\right)^T,$$

где $p_1(x) = \sqrt{\frac{2}{\pi}}e^{-x^2/2}$.

Здесь A – квадратная матрица размером $(m-1)\times(m-1)$

$$A=\begin{pmatrix} -x/2 & 0 & 0 & 0 & \dots & 0 \\ a & -x/3 & 0 & 0 & \dots & 0 \\ 0 & a & -x/4 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ 0 & \cdots & 0 & a & -x/m & \end{pmatrix}.$$

На рис. 1 изображены решения этой системы дифференциальных уравнений, то есть функций плотности распределения случайной величины Y_m при $m = 2, \dots, 9$. Как известно, для суммы независимых случайных величин есть и другой способ найти функцию плотности распределения через свертку. Для случая $m = 2$ оказалось возможным непосредственно найти решение

$$p_2(z)=\frac{2}{\pi}\int_0^z e^{-x^2/2/}e^{-(z-x)^2/2}dx=\frac{2}{\sqrt{\pi}}erf\left(\frac{z}{2}\right)e^{-z^2/4}.$$

В случае $m > 2$ численными методами получено решение полностью совпадающее с решениями системы (2). Необходимо отметить, что численное решение системы (2) многократно эффективнее по времени вычисления по сравнению с нахождением функций плотности через свертку.

Рассмотрим использование данной критической статистики для проверки гипотезы о том, что генеральная совокупность имеет функцию распределения $F(x,\vec{r})$. Практически для всех известных распределений путем моделирования выборки и вычисления критической статистики была построена её (статистики) эмпирическая функция распределения. На том же рис. 2 нанесен график функции распределения, полученный как решение системы (2).

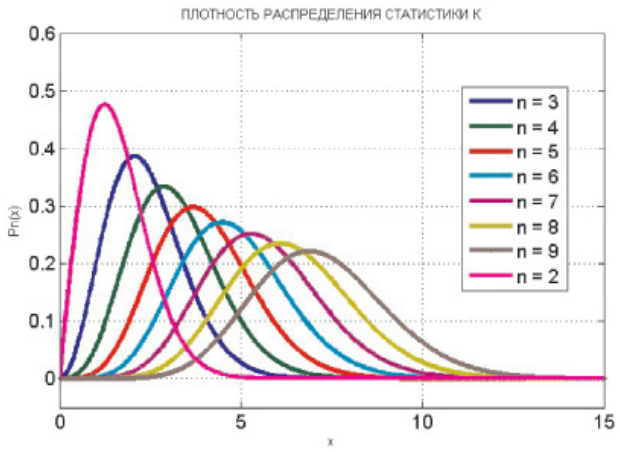

Рис. 1. Плотность распределения статистики Λ

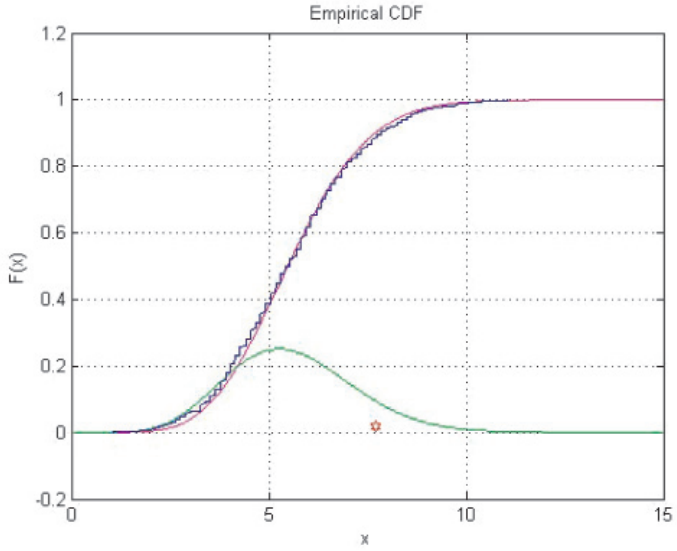

Рис. 2. График функции распределения статистики Λ

Таким образом, для случайной величины Y_m,

$$Y_m = \sum_{j=1}^{m} \left| Z_j \right|; \quad Z_j \sim N(0,1), \quad j = 1, \ldots, m$$

получено однопараметрическое семейство распределений со следующими характеристиками:

$$p_1(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}, \quad x \geq 0,$$

$p_2(x), \ldots, p_m(x)$ – решение системы (2);

$\mu_1 = m\sqrt{\frac{2}{\pi}} = M[Y_m]$ – первый центральный момент;

$\mu_2 = \left(m + m(m-1)2/\pi \right)$ – второй центральный момент;

$D[Y_m] = m(1 - 2/\pi) \approx 0{,}3633m$ – дисперсия;

$$f_Y(t) \equiv f_m(t) = \omega\left(\frac{t}{\sqrt{2}} \right)^m =$$

$$= e^{-mt^2/2} \left(1 + erf\left(it/\sqrt{2} \right) \right)^m$$ – характеристическая функция;

$$f_1(t) = \omega\left(\frac{t}{\sqrt{2}} \right),$$

где $\omega(x) = e^{-x^2} \left(1 + erf(ix) \right)$ – функция Фаддеева.

Список литературы

1. Ахряпов О.С. Проверка нормальности распределения эмпирических данных по критерию Пирсона // В мире научных открытий: материалы IV Всероссийской студенческой научной конференции (с международным участием). – 2015. – С. 79–81.

2. Жунисбеков С., Джонсон А., Шевцов А.Н. О некоторых облаках точек хи-квадрат критерия Пирсона // Theoretical & Applied Science. – 2013. – № 8 (4). – С. 1–23.

3. Колгатин А.Г. Информационные технологии в научно-педагогических исследованиях // Управляющие системы и машины. – 2015. – № 1 (255). – С. 66–72.

4. Пилипенко А.Н., Литвиненко Н.И. Влияние институциональной среды на развитие социально-экономических систем // Современные тенденции социального, экономического и правового развития стран Евразии: сборник научных трудов. – 2016. – С. 390–399.

5. Черницына Р.Н. Анализ результатов тестирования с применением методов математической статистики // Вестник Томского государственного педагогического университета. – 2016. – № 4 (169). – С. 46–52.

References

1. Ahrjapov O.S. Proverka normalnosti raspredelenija jempiricheskih dannyh po kriteriju Pirsona // V sbornike: V mire nauchnyh otkrytij Materialy IV Vserossijskoj studencheskoj nauchnoj konferencii (s mezhdunarodnym uchastiem). 2015. pp. 79–81.

2. Zhunisbekov S., Dzhonson A., Shevcov A.N. O nekotoryh oblakah tochek hi-kvadrat kriterija Pirsona // Theoretical & Applied Science. 2013. no. 8 (4). pp. 1–23.

3. Kolgatin A.G. Informacionnye tehnologii v nauchno-pedagogicheskih issledovanijah // Upravljajushhie sistemy i mashiny. 2015. no. 1 (255). pp. 66–72.

4. Pilipenko A.N., Litvinenko N.I. Vlijanie institucionalnoj sredy na razvitie socialno-jekonomicheskih sistem // V sbornike: Sovremennye tendencii socialnogo, jekonomicheskogo i pravovogo razvitija stran Evrazii: sbornik nauchnyh trudov. 2016. pp. 390–399.

5. Chernicyna R.N. Analiz rezultatov testirovanija s primeneniem metodov matematicheskoj statistiki // Vestnik Tomskogo gosudarstvennogo pedagogicheskogo universiteta. 2016. no. 4 (169). pp. 46–52.