Тр. V международной конференции "Актуальные проблемы электронного приборостроения" АПЭП-2000. Новосибирск, 2000. - Т. 6. - С. 21-23.

О ВЫБОРЕ ОПТИМАЛЬНОГО ЧИСЛА ИНТЕРВАЛОВ ГРУППИРОВАНИЯ В КРИТЕРИЯХ СОГЛАСИЯ ТИПА χ^2 [1]

Лемешко Б.Ю., Чимитова Е.В.

Новосибирский государственный технический университет
Новосибирск, Россия. E-mail: headrd@fpm.ami.nstu.ru

Аннотация. Выбор числа интервалов в критериях типа χ^2 рассматривается с позиций максимальной мощности при близких альтернативах.

Рекомендуемое в различных источниках количество интервалов группирования, используемое при вычислении оценок параметров, построении гистограмм, а также при проверке статистических гипотез с помощью критерия χ^2 Пирсона, колеблется в очень широких пределах. Большинство рекомендуемых формул для оценки числа интервалов k носит эмпирический характер и обычно дает завышенные величины. Практически все рекомендации по выбору числа интервалов исходят из того, чтобы при данном объеме выборки n как можно лучше приблизить плотность распределения ее непараметрической оценкой (гистограммой). В данной работе выбор оптимального числа интервалов k проводится с позиций построения наиболее мощного критерия согласия при близких альтернативных гипотезах.

При справедливости проверяемой гипотезы H_0 предельным распределением $G(X_n^2|H_0)$ стандартной статистики критерия согласия Пирсона $X_n^2 = \sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)}$, где n - объем выборки, n_i - количество наблюдений, попавших в интервал, $P_i(\theta) = \int_{x_{i-1}}^{x_i} f_0(x,\theta)dx$ - вероятность попадания наблюдения в интервал, является χ^2_{k-1}-распределение. При оценивании в результате минимизации статистики X_n^2 по данной выборке m параметров закона статистика подчиняется χ^2_{k-m-1}-распределению. При справедливой альтернативной гипотезе H_1 предельное распределение $G(X_n^2|H_1)$ представляет собой нецентральное χ^2-распределение с тем же числом степеней свободы и параметром нецентральности $\lambda(\theta) = \sum_{i=1}^{k} \frac{c_i^2(\theta)}{P_i}$, где $c_i(\theta) = \sqrt{n} \int_{x_{i-1}(\theta)}^{x_i(\theta)} \big(f_1(x,\theta) - f_0(x,\theta)\big)dx$.

В [1] и последующих работах Никулиным предложено такое видоизменение стандартной статистики X_n^2, при котором предельное распределение есть обычное распределение χ^2_{k-1} (количество степеней свободы не зависит от числа оцениваемых параметров). Неизвестные параметры распределения $F_0(x,\theta)$ в этом случае должны оцениваться методом максимального правдоподобия по негруппированным данным. При этом вектор вероятностей попадания в интервалы $\bar{p} = (p_1,\ldots, p_k)^{\mathrm{T}}$ предполагается заданным, и граничные точки интервалов определяются соотношениями $x_i(\theta) = F_0^{-1}(p_1 + \ldots + p_i)$, $i = \overline{1,(k-1)}$. Предложенная статистика $Y_n^2(\theta) = X_n^2 + n^{-1}\bar{a}^{\mathrm{T}}(\theta)\Lambda(\theta)\bar{a}(\theta)$ отличается от X_n^2 только при сложных ги-

[1] Работа поддержана Российским фондом фундаментальных исследований (проект № 00-01-00913)

потезах. Элементы и размерность матрицы $\Lambda(\theta)=\left[J(\theta_l,\theta_j)-\sum_{i=1}^{k}\frac{w_{\theta_l i}w_{\theta_j i}}{p_i}\right]^{-1}_{m\times m}$ определяются оцениваемыми компонентами вектора параметров θ, $J(\theta_l,\theta_j)$ - элементы информационной матрицы $\mathbf{J}(\theta)$, $a(\theta_l)=w_{\theta_l 1}n_1/p_1+\ldots+w_{\theta_l k}n_k/p_k$ - элементы вектора $\bar{a}(\theta)$, величины $w_{\theta_l i}$ определяются соотношением $w_{\theta_l i}=-f_0[x_i(\theta),\theta]\frac{\partial x_i(\theta)}{\partial\theta_l}+f_0[x_{i-1}(\theta),\theta]\frac{\partial x_{i-1}(\theta)}{\partial\theta_l}$. При верной альтернативе предельное распределение $G(Y_n^2\,|\,H_1)$ представляет собой нецентральное χ^2_{k-1}-распределение с параметром нецентральности $\lambda(\theta)=\sum_{i=1}^{k}\frac{c_i^2(\theta)}{p_i}+\bar{d}^T(\theta)\Lambda(\theta)\bar{d}(\theta)$, где $d(\theta_l)=w_{\theta_l 1}c_1(\theta)/p_1+\ldots+w_{\theta_l k}c_k(\theta)/p_k$ - элементы вектора $\bar{d}(\theta)$.

Зная предельные распределения статистики $G(S\,|\,H_0)$ и $G(S\,|\,H_1)$, для любого заданного уровня значимости α можно оценить мощность соответствующего критерия, рассматривая её как функцию от числа интервалов k при заданном объеме выборки n. Исследование мощности критериев Пирсона и Никулина как функции n и k проводилось как аналитически, так и методами статистического моделирования. Причем аналитические результаты полностью подтверждаются оценками мощности, полученными на основании моделирования. При моделировании распределений $G(S\,|\,H_0)$ и $G(S\,|\,H_1)$ выборки случайных величин разбивались на интервалы равной вероятности. В качестве примера на рисунках 1-3 приведены функции мощности рассматриваемых критериев в случае близких конкурирующих гипотез (H_0 - нормальный закон, H_1 - логистический) при уровне значимости $\alpha=0.1$. При проверке простой гипотезы функция мощности критерия Пирсона для $n>100$ принимает максимальное значение при $k=4$, и при дальнейшем увеличении объема выборки это оптимальное число интервалов не изменяется (см. рис. 1).

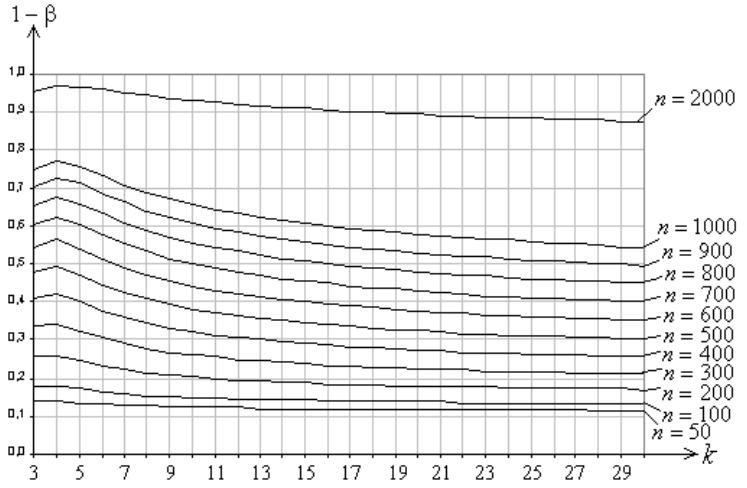

Рис.1. Функции мощности критерия χ^2 Пирсона при проверке простой гипотезы

Статистическое моделирование распределений $G(X_n^2|H_0)$ и $G(X_n^2|H_1)$ подтвердило наличие максимума мощности именно при $k=4$. В случае проверки сложной гипотезы и оценивании по выборке параметров гипотетического распределения функция мощности критерия Пирсона принимает наибольшее значение при минимально возможном числе интервалов $k=4$ и далее монотонно убывает с ростом k (рис. 2). Об этом свидетельствуют как результаты моделирования распределений статистики, так и аналитические расчеты.

Функция мощности критерия типа χ^2 Никулина, как видно из рис. 3, на области значений k, содержащей максимальное значение мощности, является выпуклой вверх функцией.

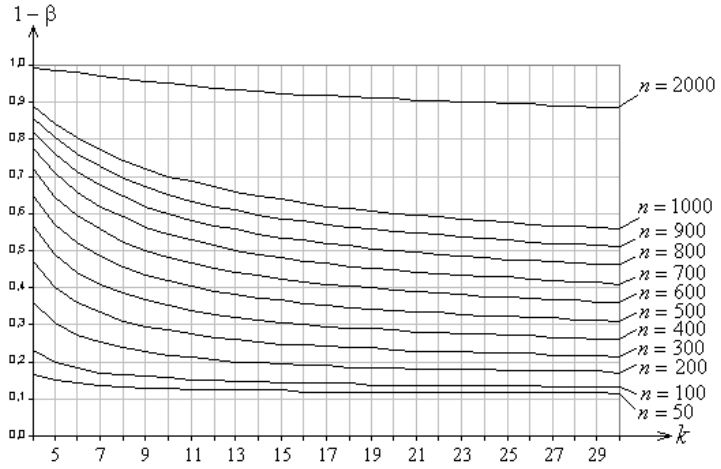

Рис.2. Функции мощности критерия χ^2 Пирсона при проверке сложной гипотезы

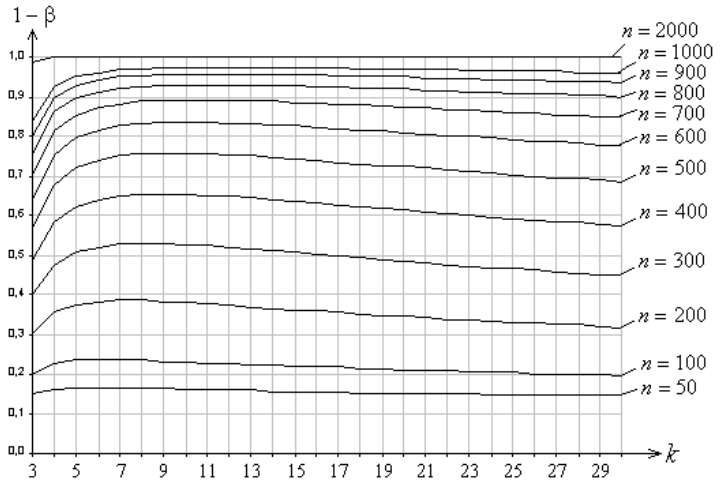

Рис.3. Функции мощности критерия типа χ^2 Никулина при проверке сложной гипотезы

Таким образом, при проверке гипотез с помощью критерия χ^2 Пирсона мощность критерия будет максимальной против близких конкурирующих гипотез, если выборку разбивать на минимально возможное число интервалов группирования. С ростом числа интервалов мощность критерия падает (в полном соответствии с работами [2,3]). Этот факт ускользает от большинства исследователей, использующих данный критерий, и не упоминается в рекомендациях различного уровня. В случае же критерия Никулина существует оптимальное число интервалов, которое также существенно меньше значений, рекомендуемых любыми действующими регламентирующими документами и справочными источниками.

ЛИТЕРАТУРА

1. Никулин М.С. О критерии хи-квадрат для непрерывных распределений // Теория вероятностей и её применение. 1973. Т.XVIII. № 3. – С.675-676.
2. Чибисов Д.М., Гванцеладзе Л.Г. О критериях согласия, основанных на группированных данных // III советско-японский симпозиум по теории вероятностей. Ташкент: изд-во “Фан”, 1975. – С. 183-185.
3. Боровков А.А. О мощности критерия χ^2 при увеличении числа групп // Теория вероятностей и ее применение. 1977. Т. XXII. № 2. – С.375-378.