ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 18 МАТЕМАТИКИ Выпуск 5
2011

*ЗУБОВ В. Н.**

О СКОРОСТИ СХОДИМОСТИ СТАТИСТИК КРИТЕРИЕВ СОГЛАСИЯ СО СТЕПЕННЫМИ МЕРАМИ РАСХОЖДЕНИЯ К ХИ-КВАДРАТ РАСПРЕДЕЛЕНИЮ

Содержание

Результаты работы позволяют получить оценку скорости слабой сходимости статистик критериев согласия со степенными мерами расхождения $\{T_\lambda(\mathbf{Y}),\ \lambda \in \mathbf{R}\}$, построенных по полиномиальному распределению размерности k, к хи-квадрат распределению. Именно, результаты работы приводят к оценке $\mathbf{P}\{T_\lambda(\mathbf{Y}) < c\} = G_{k-1}(c) + O\left(n^{-1+1/k}\right)$, $n \to \infty$, где $G_r(c)$ — функция распределения случайной величины хи-квадрат с r степенями свободы. Доказательство основано на специальном методе, использующем известный из теории чисел результат о приближении числа точек с целыми координатами, содержащихся в некотором множестве, его объемом.

Ключевые слова: теорема Э. Главки, слабая сходимость, гауссова кривизна, выпуклое многообразие, аппроксимация хи-квадрат распределением, семейство степенных статистик согласия.

---

*Победитель II Общероссийского открытого конкурса «Отмеченная работа молодого исследователя в области прикладной и промышленной математики» (ОРМИ-ОПиПМ'2011). По условиям этого конкурса данная работа была представлена автором на VII Всемирном конгрессе по промышленной и прикладной математике (г. Ванкувер, Канада) и отмечена там согласно протоколу специальной церемонии 21 июля 2011 г. дипломом I степени и вручением нашей золотой медали.

§ 1. Введение и основной результат

1.1. Введение. Рассмотрим вектор $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)^T$, имеющий полиномиальное распределение $M_k(n, \pi)$, т. е. распределение, у которого для любых $n_j = 0, 1, \dots, n$ $(j = 1, 2, \dots, k)$

$$\mathbf{P}\{Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k\} = \begin{cases} \dfrac{n!}{n_1! n_2! \dots n_k!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k} & \text{при } n_1 + n_2 + \dots + n_k = n, \\ 0 & \text{в противном случае,} \end{cases}$$

где $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_k)^T$, $\pi_j > 0$, $\pi_1 + \pi_2 + \ldots + \pi_k = 1$. Далее будем считать выполненной основную гипотезу H_0: $\boldsymbol{\pi} = \mathbf{p} = (p_1, p_2, \dots, p_k)^T$. При этом ковариационная матрица вектора $\mathbf{Y}$, как известно, равна $\Omega = (\delta_i^j\, p_i - p_i p_j)_{ij} \in \mathbf{R}^{(k-1)\times(k-1)}$.

Классический критерий согласия, предложенный К. Пирсоном, использует так называемую статистику χ^2. Эта статистика имеет простой вид и удобна в применении. Вместе с тем, для получения хорошей точности при помощи этого критерия необходимо иметь достаточно большой объем входных данных (в сумме и по отдельности в каждой ячейке). Кроме того, на практике чаще всего приходится заменять распределение статистики асимптотическим. Точность этой аппроксимации зависит от числа ячеек, а величина ошибки чаще всего неизвестна. Непонятно и то, является ли статистика χ^2 оптимальной на малых объемах выборки.

В связи с этим многие ученые исследовали другие подходы к построению критериев согласия с целью найти наиболее эффективный в том или ином статистическом смысле. Особое место в этих исследованиях занимают работы Н. Крисси [6] и Т. Рида [13]. Эти авторы ввели в употребление и произвели первичный анализ семейства статистик

$$t_\lambda(\mathbf{Y}) = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} Y_j \left(\left(\frac{Y_j}{np_j} \right)^{\lambda} - 1 \right), \qquad \lambda \in \mathbf{R},$$

называемых *степенными статистиками согласия*. Это семейство предназначено для построения критериев согласия по сгруппированным данным. Оно параметризовано вещественным параметром λ, при этом как собственно статистика χ^2, так и другие распространенные статистики являются частными случаями.

З а м е ч а н и е 1. Соответствующий англоязычный термин — power divergence family of statistics. Название статистик объясняется тем, что они представляют собой меры несоответствия между эмпирическими частотами $\mathbf{Y}$ и теоретическими вероятностями $\mathbf{p}$. При этом несоответствие измеряется посредством функции степенного вида.

З а м е ч а н и е 2. При $\lambda = 0, -1$ эту запись следует понимать как результат предельного перехода.

З а м е ч а н и е 3. Полагая $\lambda = 1$, $\lambda = -1/2$ и $\lambda = 0$, получаем хи-квадрат статистику, статистику Фримана–Тьюки и логарифмическую статистику отношения правдоподобия соответственно.

Предполагая выполненной основную гипотезу, рассмотрим преобразование

$$X_j = \frac{Y_j - np_j}{\sqrt{n}}, \qquad j=1,2,\ldots,k, \quad r=k-1, \tag{1}$$

и составим вектор $\mathbf{X} = (X_1, X_2, \ldots, X_r)^T$. Компоненты этого вектора сосредоточены на решетке

$$L = \Big\{\mathbf{x} = (x_1, x_2, \ldots, x_r)^T, \qquad \mathbf{x} = \frac{\mathbf{n} - n\,\mathbf{p}}{\sqrt{n}}, \quad \mathbf{n} = (n_1, n_2, \ldots, n_r)^T\Big\},$$

где n_j — неотрицательные целые числа.

З а м е ч а н и е 4. Статистика $t_\lambda(\mathbf{Y})$ может быть представлена как функция от $\mathbf{X}$ вида

$$T_\lambda(\mathbf{x}) = \frac{2n}{\lambda(\lambda+1)} \sum_{j=1}^{k} p_j \Big(\Big(1 + \frac{x_j}{p_j\sqrt{n}}\Big)^{\lambda+1} - 1\Big), \tag{2}$$

а затем посредством разложения по Тейлору преобразована к виду

$$T_\lambda(\mathbf{x}) = \sum_{i=1}^{k} \Big(\frac{x_i^2}{p_i} + \frac{(\lambda-1)\,x_i^3}{3\,p_i^2\sqrt{n}} + \frac{(\lambda-1)(\lambda-2)\,x_i^4}{12\,p_i^3\,n} + O\Big(\frac{1}{n^{3/2}}\Big)\Big), \quad n \to \infty. \tag{3}$$

Хорошо известен факт, что распределение всех статистик семейства сходится к распределению хи-квадрат с $k-1$ степенями свободы (см., например, работу [6, с. 443]). Однако для построения эффективных критериев согласия на их основе необходимо оценить качество аппроксимации предельным распределением. В связи с этим большой интерес представляет проблема оценки скорости сходимости степенных статистик согласия к предельному распределению.

Для того чтобы уяснить себе суть используемых в работе результатов, необходимо ввести новое определение. Назовем множество $B \subset \mathbf{R}^r$ *расширенным выпуклым*, если для всех $l = 1, 2, \ldots, r$ оно представимо в виде

$$B = \{\mathbf{x} = (x_1, x_2, \ldots, x_r)^T : \ \lambda_l(x^*) < x_l < \theta_l(x^*) \quad \text{и} \\ x^* = (x_1, x_2, \ldots, x_{l-1}, x_{l+1}, \ldots, x_r)^T \in B_l\},$$

где B_l — некоторое подмножество $\mathbf{R}^{r-1}$, а $\lambda_l(x^*), \theta_l(x^*)$ — некоторые непрерывные функции на $\mathbf{R}^{r-1}$.

Для того чтобы исследовать степенные статистики согласия на слабую сходимость, заметим, что функцию распределения $\mathbf{P}\{T_\lambda(\mathbf{X}) < c\}$ статистик семейства можно записать в виде вероятности попадания случайного вектора $\mathbf{X}$, имеющего решетчатое распределение, в некоторое

СПИСОК ЛИТЕРАТУРЫ

1. *Yarnold J. K.* Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. — Ann. Math. Statist., 1972, v. 43, № 5, p. 1566–1580.
2. *Huxley M. N.* Exponential sums and lattice points II. — Proc. London Math. Soc., 3rd ser. , 1993, v. 66, is. 3, p. 279–301.
3. *Huxley M. N.* Exponential sums and lattice points III. — Proc. London Math. Soc., 3rd ser., 2003, v. 87, is. 3, p. 591–609.
4. *Siotani M.*, *Fujikoshi Y.* Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics. — Hiroshima Math. J., 1984, v. 14, № 1, p. 115–124.
5. *Esseen C. G.* Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law. — Acta Math., 1945, v. 77, p. 1–125.
6. *Cressie N.*, , *Read T. R. C.* Multinomial goodness-of-fit tests. — J. Roy. Statist. Soc., Ser. B, 1984, v. 46, № 3, p. 440–464.
7. *Read T. R. C.* Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics. — Ann. Inst. Statist. Math., 1984, v. 36, № 1, p. 59–69.
8. *Мищенко А. С.*, *Фоменко А. Т.* Курс дифференциальной геометрии и топологии. М.:Факториал Пресс, 2000.
9. *Асылбеков Ж. А.*, *Зубов В. Н.*, *Ульянов В. В.* Об аппроксимации некоторых статистик критериев согласия для случая дискретных трехмерных данных. — Сиб. матем. ж., т. 52, № 4,с. 728–744.
10. *Ivic A.*, *Krätzel E.*, *Kühleitner M.*, *Nowak W. G.* Lattice points in large regions and related arithmetic functions: recent development in a very classic topic. — In: Elementare und Analytische Zahlentheorie. (Tagungsband.) Proceedings ELAZ-Conference. (Mainz, 2004, May 24–28, 2004.)/Ed. by W. Schwarz, J. Steuding. Stuttgart: Franz Steiner Verlag, 2006, p. 89–128.
11. *Hlawka E.* Über integrale auf konvexen körpern. I. — Monatsh. Math., 1950, v. 54, № 1, p. 1–36.
12. *Hlawka E.* Über integrale auf konvexen körpern. II. — Monatsh. Math., 1950, v. 54, № 2, p. 81–99.
13. *Read T. R. C.* Small sample comparisons for the power divergence goodness-of-fit statistics. — J. Amer. Statist. Ass., 1984, v. 79, № 388, p. 929–935.
14. *Bonnesen T.*, *Fenchel W.* Theorie der konvexen körper. Berlin: Springer, 1934.
15. *Ulyanov V. V.*, *Zubov V. N.* Refinement on the convergence of one family of goodness-of-fit statistics to chi-squared distribution. — Hiroshima Math. J., 2009, v. 39, № 1, p. 133–161.

Поступила в редакцию
10.VI.2010