ОБ ОДНОЙ СТАТИСТИКЕ ДЛЯ ПРОВЕРКИ ОДНОРОДНОСТИ ПОЛИНОМИАЛЬНЫХ ВЫБОРОК

А. М. Зубков, Б. И. Селиванов

Математический институт имени В. А. Стеклова РАН
Москва, Россия
E-mail: zubkov@mi.ras.ru, bor-selivanov@yandex.ru

Предложена новая статистика для проверки гипотезы об однородности нескольких полиномиальных выборок. Показано, что при стремлении к бесконечности объемов выборок предельными распределениями этой статистики являются: распределение хи-квадрат (если выборки однородны), нецентральное распределение хи-квадрат (если распределения выборок сближаются), нормальное (если распределения выборок фиксированы и различны).

Ключевые слова: полиномиальные выборки, критерий однородности, хи-квадрат распределения.

Рассмотрим $M \geq 2$ независимых выборок; k-я ($k = 1,\ldots, M$) выборка является реализацией t_k независимых испытаний, проведенных по полиномиальной схеме с исходами 1,..., N, имеющими вероятности $p_{k,1},\ldots, p_{k,N} > 0$, $p_{k,1}+\ldots+p_{k,N} = 1$. Будем использовать обозначение $t = t_1 +\ldots+ t_k$; через $\nu_{k,i}$ будем обозначать частоту (число появлений) исхода i в k-й выборке.

Будем говорить, что выборки *статистически однородны* и что для них выполняется гипотеза H_0, если

$$p_{1,i} =\ldots= p_{M,i}, \quad i = 1,\ldots, N; \tag{1}$$

в противном случае будем говорить, что выполняется альтернатива H_1. При альтернативе вероятности исходов могут быть постоянными или изменяться с ростом объемов выборок; это будет оговариваться особо.

Для проверки однородности полиномиальных выборок обычно используют статистику хи-квадрат (см., например, [1], § 30.6)

$$\chi^2 = t\sum_{k=1}^{M}\sum_{i=1}^{N}\frac{1}{t_k m_i}\left(\nu_{k,i} - m_i\frac{t_k}{t}\right)^2, \qquad m_i = \sum_{k=1}^{M}\nu_{k,i}. \tag{2}$$

Содержащееся в [1] утверждение о том, что при гипотезе H_0 и $t_0 \to \infty$ распределение статистики (2) сходится к распределению хи-квадрат с $(M–1)(N–1)$ степенями свободы, в работах [2] и [3] было обобщено на случай «почти однородных» полиномиальных выборок.

Статистика (2), по существу, является суммой M статистик Пирсона, построенных по частотам исходов в M выборках в предположении, что вероятности исходов 1,…, N в каждой выборке равны частотам исходов в объединенной выборке. Естественная модификация статистики (2) содержится в следующем утверждении.

Теорема 1. При любом $M \geq 2$ справедливо равенство

$$\zeta^2 \stackrel{\text{def}}{=} \inf_{\substack{q_1,\ldots,q_N>0 \\ q_1+\ldots+q_N=1}} t\sum_{k=1}^{M}\sum_{i=1}^{N}\frac{1}{t_k m_i}\left(\nu_{k,i}-q_i\right)^2 = \left(\sum_{i=1}^{N}\sqrt{\sum\nolimits_{k=1}^{M}\frac{\nu_{k,i}^2}{t_k}}\right)^2.$$

Иначе говоря, ζ^2 – минимум значений суммы статистик Пирсона, вычисленных по частотам всех выборок в предположении, что справедлива гипотеза H_0. Как показывает теорема 1, этот минимум достигается в точке, вообще говоря, отличной от набора частот в объединенной выборке.

Приведем формулировки предельных теорем для статистики ζ^2.

Теорема 2. Если существуют такие константы c_0 и c_1, что $0 < c_0 < t_k/t < c_1 < 1$ при всех $k = 1,\ldots,M$, и выполняется гипотеза H_0, то при $t \to \infty$ распределение статистики $\zeta^2 - t$ сходится к распределению хи-квадрат с $(M–1)(N–1)$ степенями свободы.

Теорема 3. Если существуют пределы

$$\alpha_k = \lim_{t\to\infty}\frac{t_k}{t} > 0, \quad k = 1,\ldots,M,$$

и выполняются альтернативы $H_1 = H_1(t)$ с вероятностями исходов $q_{k,i}(t)$, удовлетворяющими при $t \to \infty$ условиям

$$q_{k,i}(t) = p_i + \frac{\lambda_{k,i}}{\sqrt{t_k}} + o\left(\frac{1}{\sqrt{t_k}}\right), i = 1,\ldots,N, \quad \sum_{i=1}^{N} q_{k,i}(t) = 1,$$

то при $t \to \infty$ распределение статистики $\zeta^2 - t$ сходится к нецентральному распределению хи-квадрат с $(M–1)(N–1)$ степенями свободы и параметром нецентральности

$$\delta^2 = \sum_{i=1}^{N}\frac{1}{p_i}\left[\sum_{k=1}^{M}\lambda_{k,i}^2 - \left(\sum_{k=1}^{M}\sqrt{\alpha_k}\lambda_{k,i}\right)^2\right].$$

Доказательства теорем 2 и 3 используют результаты статьи [5].

Предельное распределение статистики ζ^2 в случае альтернативы с фиксированными вероятностями исходов описано в теореме 4.

Теорема 4. Если имеет место альтернатива H_1 с фиксированными вероятностями исходов и существуют пределы

$$\alpha_k = \lim_{t\to\infty}\frac{t_k}{t} > 0, \quad k = 1,\ldots,M,$$

то при $t \to \infty$ статистика ζ^2 асимптотически нормальна с математическим ожиданием μt и дисперсией σt, где μ и σ определяются равенствами

$$\mu = \left(\sum_{i=1}^{N}\sqrt{\sum\nolimits_{k=1}^{M}\alpha_k\, p_{k,i}^2}\right)^2, \quad \sigma = 4\left(\sum_{i=1}^{N}\sqrt{\sum\nolimits_{k=1}^{M}\alpha_k\, p_{k,i}^2}\right)^2 \sum_{k=1}^{M}\left[\sum_{i=1}^{N}p_{k,i}^3 - \left(\sum_{i=1}^{N}p_{k,i}^2\right)^2\right]\frac{\alpha_k}{\sum_{k=1}^{M}\alpha_k p_{k,i}^2}.$$

Доказательство теоремы 4 основано на использовании известных теорем о предельном распределении функции от асимптотически нормальных случайных векторов (см., например, [4]).

Замечание. В силу закона больших чисел при замене всех вероятностей $p_{k,i}$ в формуле для σ выборочными частотами полученная случайная величина будет сходиться к σ при $t \to \infty$.

ЛИТЕРАТУРА

1. *Крамер*, *Г.* Математические методы статистики / Г. Крамер. М.: Мир, 1975. 648 с.

2. *Mitra*, *S. K.* On the limiting power function of the frequency chi-square test / S. K. Mitra // Ann. Math. Statist. 1958. Vol. 29. № 4. P. 1221–1233.

3. *Meng, R. C.* The power of chi square tests for contingency tables / R. C. Meng, D. G. Chapman // J. Amer. Statist. Ass. 1966. Vol. 61. № 316. P. 965–975.

4. *Боровков, А. А.* Математическая статистика / А. А. Боровков. Новосибирск : Наука, 1997. 772 с. (с. 41–45).

5. *Dik, J. J.* The distribution of general quadratic form in normal variables. / J. J. Dick, M. C. M. de Gunst // Statistica Neerlandica. 1985. Vol. 39. № 1. P. 14−26.