Материалы Российской НТК "Информатика и проблемы телекоммуникаций", Новосибирск. 2005. – Т.1. – С.109-112.

О РАСПРЕДЕЛЕНИЯХ СТАТИСТИК И МОЩНОСТИ КРИТЕРИЕВ ОДНОРОДНОСТИ ДВУХ ВЫБОРОК[1]

Лемешко Б.Ю., Лемешко С.Б.
НГТУ, Новосибирск, E-mail: headrd@fpm.ami.nstu.ru

Задача проверки однородности двух выборок формулируется следующим образом. Пусть имеется две упорядоченные по возрастанию выборки размера m и n: $x_1 < x_2 < ... < x_m$ и $y_1 < y_2 < ... < y_n$.

Проверяется гипотеза о том, что две выборки извлечены из одной и той же генеральной совокупности, т.е. $H_0: F(x) = G(x)$ при любом x.

Критерий однородности двух выборок Смирнова предложен в работе [1]. Предполагается, что функции распределения $F(x)$ и $G(x)$ являются непрерывными. Статистика критерия Смирнова измеряет различие между эмпирическими функциями распределения, построенными по выборкам

$$D_{m,n} = \sup_x \left| G_m(x) - F_n(x) \right|.$$

При практическом использовании критерия значение статистики $D_{m,n}$ рекомендуется вычислять в соответствии с соотношениями [2]

$$D^+_{m,n} = \max_{1 \le r \le m} \left[\frac{r}{m} - F_n(x_r) \right] = \max_{1 \le s \le n} \left[G_m(y_s) - \frac{s-1}{n} \right],$$

$$D^-_{m,n} = \max_{1 \le r \le m} \left[F_n(x_r) - \frac{r-1}{m} \right] = \max_{1 \le s \le n} \left[\frac{s}{n} - G_m(y_s) \right],$$

$$D_{m,n} = \max(D^+_{m,n}, D^-_{m,n}).$$

Если гипотеза H_0 справедлива, то при неограниченном увеличении объемов выборок [2] $\lim\limits_{m \to \infty} P\left\{ \sqrt{\dfrac{mn}{m+n}} D_{m,n} < s \right\} = K(s)$, т.е. статистика

$$S_C = \sqrt{\frac{mn}{m+n}} D_{m,n} \qquad (1)$$

в пределе подчиняется распределению Колмогорова $K(s)$. Однако при ограниченных значениях m и n случайные величины $D^+_{m,n}$ и $D_{m,n}$ являются дискретными, и множество их возможных значений представляет собой решетку с шагом $1/k$, где k наименьшее общее кратное m и n. Для значений $m, n \le 20$ таблицы процентных точек для статистики $D_{m,n}$ приводятся в [2]. Условное распределение $G\left(S_C \middle| H_0\right)$ статистики S_C при справедливости гипотезы H_0

[1] Работа выполнена при поддержке Министерства образования и науки РФ (код проекта 15378)

медленно сходится к $K(s)$ и существенно отличается от него при не очень больших m и n. Асимптотические формулы для распределений $D^{+}_{m,n}$ и $D_{m,n}$ рассматривались в [3, 4, 5].

На рис. 1 показаны построенные условные распределения статистики (1) при справедливости H_0 в зависимости от m и n. Как следует из полученной картины, даже при $m=1000$ и $n=1000$ ступенчатость $G(S_C|H_0)$ сохраняется. Гладкость распределения статистики сильно зависит от k, что затрудняет использование критерия. Исследования распределения статистики показали, что предпочтительнее применять критерий, когда m и n не равны и представляют собой взаимно простые числа. В таких случаях распределение статистики оказывается существенно ближе к предельному $K(s)$ (при меньших m и n).

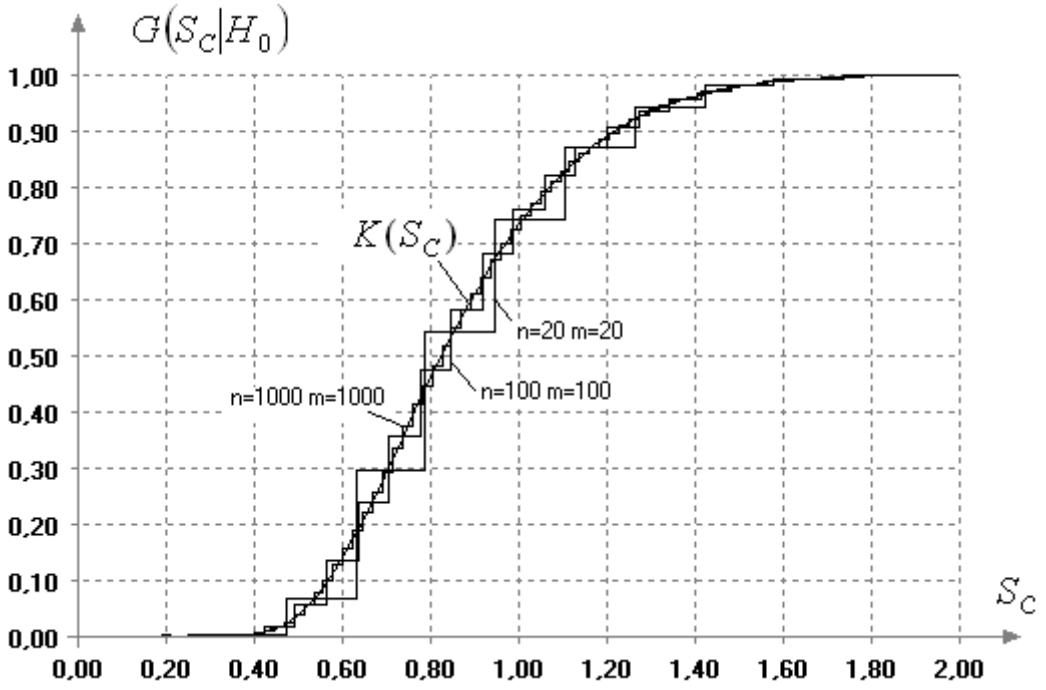

Рис. 1. Распределения статистики (1) при справедливости H_0 в зависимости от m и n

Критерий однородности Лемана-Розенблатта представляет собой критерий типа ω^2. Критерий был предложен в работе [6] и исследован в [7]. Статистика критерия имеет вид [2]

$$T = \frac{mn}{m+n}\int_{-\infty}^{\infty}\left[G_m(x)-F_n(x)\right]^2 dH_{m+n}(x),$$

где $H_{m+n}(x)=\frac{m}{m+n}G_m(x)+\frac{n}{m+n}F_n(x)$ – эмпирическая функция распределения, построенная по вариационному ряду объединения двух выборок. Статистика T используется в форме [2]

$$T=\frac{1}{mn(m+n)}\left[n\sum_{i=1}^{n}(r_i-i)^2+m\sum_{i=1}^{m}(s_i-j)^2\right]-\frac{mn-1}{6(m+n)}, \qquad (2)$$

где r_i – порядковый номер (ранг) y_i, s_j – порядковый номер (ранг) x_j в объединенном вариационном ряде.

В [7] было показано, что статистика (2) в пределе распределена как $a1(T)$:

$$\lim_{\substack{m\to\infty \\ n\to\infty}} P\{T < s\} = a1(s).$$

В отличие от статистики критерия Смирнова статистика T быстро сходится к предельному $a1(T)$. На рис. 2 условные распределения $G(T|H_0)$ статистики (2) приведены при m и n, равных 20, 100 и 1000. Как видим, уже при $m=100$ и $n=100$ распределение $G(T|H_0)$ визуально совпадает с $a1(T)$.

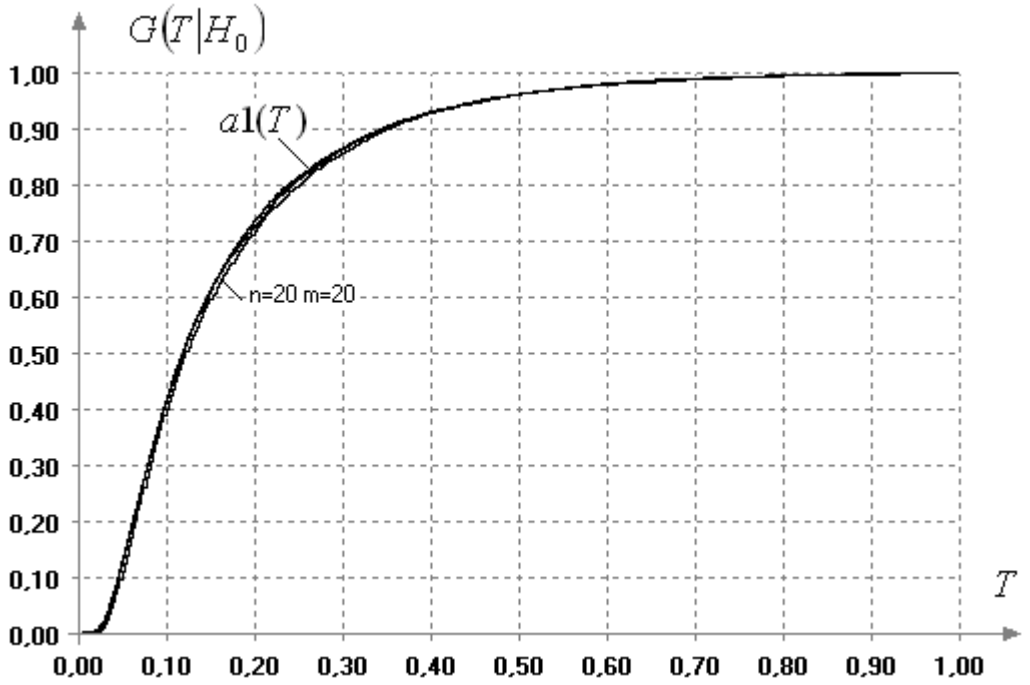

Рис. 2. Распределения статистики (2) при справедливости H_0 в зависимости от m и n

Мощность критериев была исследована для ряда альтернатив и при различных объемах выборок. Исследования показали, что мощность критерия Лемана-Розенблатта, как правило, выше мощности критерия Смирнова. Однако относительно очень близких альтернатив несколько выше оказывается мощность критерия Смирнова.

1. Смирнов Н.В. Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках // Бюллетень МГУ, серия А. – 1939. – Т.2. №2. – С.3-14.
2. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
3. Боровков А.А. К задаче о двух выборках // Изв. АН СССР, серия матем., 1962. Т. 26. – С.605-624.
4. Королюк В.С. Асимптотический анализ распределений максимальных уклонений в схеме Бернулли // Теория вероятностей и ее применения. – 1959. – Т.4. – С. 369-397.

5. Смирнов Н.В. Вероятности больших значений непараметрических односторонних критериев согласия // Труды Матем. ин-та АН СССР. – 1961. – Т.64. – С. 185-210.