Г. Б. М а р ш а л к о (Москва, ТВП)**. Вопросы совместного использования нескольких критериев χ^2.**

В современных программных статистических пакетах, таких как NIST-STS [2], TestU01 [3] и др., широкое применение находят варианты критерия χ^2 с различными способами группирования наблюдаемых величин (двоичных векторов длины m). Зачастую получающиеся при одновременном применении критериев результаты будут зависимы между собой, при этом эта зависимость будет определяться не только исходными данными, но и видом функций группирования исходов. В работе, представленной данным докладом, описывается экспериментальный подход к исследованию независимости статистических критериев подобного вида.

Пусть $x^n = x_0, x_1, \ldots, x_{n-1}$ — исследуемая последовательность одинаково распределенных случайных величин, $x_i \in V_k$, $i = 0, 1, \ldots, n-1$. Нулевой гипотезой H_0 является предположение о том, что случайные величины x_i независимы и $\mathbf{P}\{x_i = t\} = 1/2^k$, $t = 0, 1, \ldots, 2^k - 1$. Рассмотрим два критерия χ^2, заданных на элементах последовательности (векторах длины m) и отличающихся различным способом построения ячеек. Введем в рассмотрение две функции $f\colon V_m \to U$ и $g\colon V_m \to W$, где $U = \{0, 1, \ldots, s-1\}$ и $W = \{0, 1, \ldots, t-1\}$, которые в случае нулевой гипотезы определяют распределения на множествах U и W. Обозначим $\mathbf{P}\{f(x) = i\} = p_i$, $i = 0, 1, \ldots, s-1$ и $\mathbf{P}\{g(x) = j\} = q_j$, $j = 0, 1, \ldots, t-1$, ν_i $(i = 0, 1, \ldots, s-1)$ и μ_j $(j = 0, 1, \ldots, t-1)$ — частоты встречаемости соответствующих значений при маркировке исследуемой последовательности соответственно с использованием каждой из функций.

Нас будет интересовать вопрос о коррелированности частот ν_i, $i = 0, 1, \ldots, s-1$, и μ_j, $j = 0, 1, \ldots, t-1$, который очевидным образом влияет на коррелированность значений соответствующих статистик χ^2: $\chi^{2'} = \sum_{i=0}^{s-1} \nu_i^2/(np_i) - n$, $\chi^{2''} = \sum_{j=0}^{t-1} \mu_j^2/(nq_j) - n$.

Одним из подходов является оценка выборочной средней квадратичной сопряженности признаков [1], выбор именно этой характеристики вызван, в первую очередь, тем, что стандартный подход, основанный на анализе коэффициента корреляции, может быть использован только для нормального распределения. В большинстве же используемых на практике критериев χ^2 число степеней свободы не позволяет использовать нормальное приближение.

В соответствии с описанием функций f и g, мы можем построить таблицу сопряженности признаков размеров $s \times t$ для частот попадания в соовтетствующие интервалы группирования. Если обозначить ν_{ij}, $i = 0, 1, \ldots, s-1$, $j = 0, 1, \ldots, t-1$, совместную частоту встречаемости пар значений при группировани элементов исследуемой последовательности с использованием обеих функций, то статистика будет иметь вид $\Psi^2 = \sum_{i=0}^{s-1} \sum_{j=0}^{t-1} \nu_{ij}^2/(\nu_i \mu_j) - 1$. При этом ее значение заключено в интервале $[0, \min\{t, s\} - 1]$ и достигает максимального значения в случае, когда исследуемые признаки однозначно зависимы (в таблице в каждом столбце (при $t \geqslant s$) или каждой строке (при $t \leqslant s$) только один ненулевой элемент) (см. [1]).

Исследования критериев из [1, 2] показывают, что для некоторых пар критериев функция $\Psi^2(m)$ (как функция от длины m векторов исходной последовательности, которые используются при расчете статистики) для небольших значений m $(m \leqslant 30)$ является монотонной. В данном случае с использованием метода наименьших квадратов можно построить аппроксимацию функции $\Psi^2(m)$ с целью оценки корреляции критериев для больших значений m.

В целом можно сказать, что предложенный подход может быть использован для первичной оценки независимости при анализе частотных критериев, с тем, чтобы изначально исключить заведомо зависимые критерии.

С точки зрения исключения критериев, зависимость которых определяется видом их статистик в смысле предложенной характеристики, важно, чтобы функции, в соответствии с которыми происходит заполнение классов, по которым вычисляются

статистики критериев, были по возможности близки к равновероятным функциям. Кроме этого, вероятности попадания в классы таблицы сопряженности для значений этих функций также должны, по возможности, быть одинаковыми (т. е. в приведенных выше обозначениях равняться $1/(ts)$).

Работа выполнена при поддержке гранта Президента РФ НШ-4.2008.10.

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А.*, *Мхитарян В. С* Прикладная статистика и основы эконометрики, М.: ЮНИТИ, 1998.
2. *Rukhin A.*, *Soto J.*, *Nechvatal J.*, *Smid M.*, *Barker E.*, *Leigh S.*, *Levenson M.*, *Vangel M.*, *Banks D.*, *Heckert A.*, *Dray J.*, *Vo S.* A statistical test suite for random and pseudorandom number generators for cryptographic applications. NIST Special Publication 800-22 (Revision 1, August 2008). http://csrc.nist.gov/rng/.
3. *P.L. L'Ecuyer P. L.*, *Simard R.* TestU01. A software library in ANSI C for empirical testing of random number generators. http://www.iro.umontreal.ca/ simardr/.