ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 29 МАТЕМАТИКИ Выпуск 1
2022

В. О. Мironкин, **А. Б. Чухно** (Москва, МИЭМ НИУ ВШЭ). **О применении неравенства Маллоуса при тестировании числовых последовательностей.**

УДК 519.233.33 DOI https://doi.org/10.52513/08698325_2022_29_1_1

Резюме: В работе получены оценки распределений максимума и минимума частот в реализации произвольной полиномиальной схемы, позволяющие строить статистические критерии проверки гипотезы согласия с соответствующим распределением, в том числе в условиях ограниченного объема анализируемых данных.

Ключевые слова: Полиномиальная схема, частота, неравенство Маркова, метод Чернова, дивергенция Кульбака-Лейблера, равновероятное распределение.

При решении задачи проверки гипотезы о непротиворечивости распределения элементов анализируемой выборки (как правило, числовой последовательности произвольного модуля N, где $N \geqslant 2$) свойствам равновероятности и независимости (далее – гипотеза случайности) широко используются статистические критерии, построенные на основе частотных характеристик соответствующей выборки (например, на основе разделимых статистик, статистик меры хи-квадрат и т. д.). Хорошо известно, что для их корректного применения в ряде случаев требуется существенный объем выборки, обеспечивающий выполнение конкретных условий (например, превышение частоты реализации каждого из разбиений в статистике Пирсона величины 5). Как же быть в случае недостаточного объема выборки?

Изложенные ниже рассуждения тесно связаны с указанной проблематикой и описывают один из подходов к построению критериев проверки гипотезы случайности в ситуации, когда анализируемая выборка имеет ограниченный объем и может содержать не все элементы по соответствующему модулю.

Оценки распределений максимума и минимума частот

Для произвольной полиномиальной схемы

$$\mathcal{M}(n, p_1, \ldots, p_N), \qquad n \in \mathbb{N},$$

где $\sum_{i=1}^{N} p_i = 1$, $p_i > 0$, $i = 1, \ldots, N$, известны оценки сверху для совместных распределений частот $\nu_1, \ldots, \nu_N$ элементов схемы [1]:

$$\mathbf{P}\{\nu_1 \leqslant l_1, \ldots, \nu_N \leqslant l_N\} \leqslant \prod_{i=1}^{N} \mathbf{P}\{\nu_i \leqslant l_i\}, \tag{1}$$

$$\mathbf{P}\{\nu_1 \geqslant l_1, \ldots, \nu_N \geqslant l_N\} \leqslant \prod_{i=1}^{N} \mathbf{P}\{\nu_i \geqslant l_i\}, \tag{2}$$

где $\sum\limits_{i=1}^{N} \nu_i = n$.

Введем обозначения

$$\nu_{max}^{(N,n)} = \max\{\nu_j |\ j = 1, \ldots, N\}, \quad \nu_{min}^{(N,n)} = \min\{\nu_j |\ j = 1, \ldots, N\}$$

и в неравенствах (1) и (2) положим $l_1 = l_2 = \ldots = l_N = l \geqslant 0$. В результате получим выражения

$$\mathbf{P}\left\{\nu_{max}^{(N,n)} \leqslant l\right\} = \mathbf{P}\{\nu_1 \leqslant l, \ldots, \nu_N \leqslant l\} \leqslant \prod_{i=1}^{N} \mathbf{P}\{\nu_i \leqslant l\}, \tag{3}$$

$$\mathbf{P}\{\nu_{min}^{(N,n)} \geqslant l\} = \mathbf{P}\{\nu_1 \geqslant l, \ldots, \nu_N \geqslant l\} \leqslant \prod_{i=1}^{N} \mathbf{P}\{\nu_i \geqslant l\}. \tag{4}$$

С учетом того, что для каждого $i \in \{1, \ldots, N\}$ случайная величина ν_i имеет биномиальное распределение $\mathcal{B}i(n, p_i)$, выполняются равенства

$$\mathbf{P}\{\nu_i \leqslant l\} = \sum_{t=0}^{l} C_n^t p_i^t (1-p_i)^{n-t}, \quad \mathbf{P}\{\nu_i \geqslant l\} = \sum_{t=l}^{n} C_n^t p_i^t (1-p_i)^{n-t},$$

позволяющие точно вычислять оценки (3), (4). В свою очередь, выражения для математического ожидания и дисперсии биномиального распределения

$$\mathbf{E}\nu_i = np_i, \quad \mathbf{D}\nu_i = np_i(1-p_i), \tag{5}$$

дают возможность строить на основе классических результатов теории вероятностей оценки для (3), (4), имеющие более простой аналитический вид. Так, например, с использованием неравенства Маркова [2] из (4) и (5) для произвольного $l > 0$ получаем оценку сверху

$$\mathbf{P}\left\{\nu_{min}^{(N,n)} \geqslant l\right\} \leqslant \prod_{i=1}^{N} \frac{np_i}{l} = \left(\frac{n}{l}\right)^N \prod_{i=1}^{N} p_i = \left(\frac{n}{lN^{1-D_N(\overline{p})}}\right)^N, \tag{6}$$

где $D_N(\overline{p})$ — дивергенция Кульбака-Лейблера [3] распределения $(p_1, \ldots, p_N)$ от $(\frac{1}{N}, \ldots, \frac{1}{N})$ при использовании основания логарифма, равного N. Здесь функционал $D_N(\overline{p})$ выбран не случайно. Несмотря на несимметричность, полуметрика $D_N(\overline{p})$ позволяет отразить зависимость оценки (6) от степени удаленности исходного распределения $(p_1, \ldots, p_N)$ от равновероятного. В частности, при условии равновероятного распределения $\overline{p}$ имеем

$$\mathbf{P}\left\{\nu_{\min}^{(N,n)} \geqslant l\right\} \leqslant \left(\frac{n}{lN}\right)^N.$$

Аналогичным образом может быть получена оценка сверху для (4) с использованием метода Чернова, постулирующего выполнение неравенства

$$\mathbf{P}\{\xi \geqslant \varepsilon\} \leqslant \min\left\{\frac{\mathbf{E}[e^{\lambda\xi}]}{e^{\lambda\varepsilon}} \,\middle|\, \lambda \geqslant 0\right\}$$

для произвольного $\varepsilon > 0$.

Вместе с тем, в условиях ограниченного объема выборки (по тексту это величина $n \in \mathbb{N}$) нас будут интересовать достаточно малые значения величины l, для которых выражение (6) в ряде случаев может стать несодержательным. По этой причине акцентируем внимание на соотношении (3).

Итак, в соответствии с [4] для случайной величины ξ, для которой $\mathbf{E}\xi \geqslant 0$, и любого $\varepsilon \in (0, 1)$ имеет место соотношение

$$\mathbf{P}\{\xi > \varepsilon\mathbf{E}\xi\} \geqslant \frac{(1-\varepsilon)^2(\mathbf{E}\xi)^2}{(1-\varepsilon)^2(\mathbf{E}\xi)^2 + \mathbf{D}\xi}.$$

Тогда для произвольных $i \in \{1, \ldots, n\}$ и $l\colon 0 < l < np_i$ с учетом (5) можем записать

$$\mathbf{P}\{\nu_i \leqslant l\} \leqslant \frac{\mathbf{D}\nu_i}{(1-\frac{l}{\mathbf{E}\nu_i})^2(\mathbf{E}\nu_i)^2 + \mathbf{D}\nu_i} = \frac{\mathbf{D}\nu_i}{(\mathbf{E}\nu_i - l)^2 + \mathbf{D}\nu_i} = \frac{np_i(1-p_i)}{(np_i - l)^2 + np_i(1-p_i)}$$

и соответственно

$$\mathbf{P}\left\{\nu_{max}^{(N,n)} \leqslant l\right\} \leqslant n^N \prod_{i=1}^{N} \frac{p_i(1-p_i)}{(np_i-l)^2 + np_i(1-p_i)}.$$

В частности, при условии равновероятного распределения $\overline{p}$ указанная оценка принимает вид

$$\mathbf{P}\left\{\nu_{max}^{(N,n)} \leqslant l\right\} \leqslant \left(\frac{N-1}{n(1-\frac{lN}{n})^2 + N - 1}\right)^N, \tag{7}$$

где $0 < l < \frac{n}{N}$.

Отметим, что выражения (6), (7) стандартным образом могут быть использованы при построении статистических критериев проверки гипотезы случайности. При этом выражение (7) может применяться в условиях ограниченного объема выборки $n \in \mathbb{N}$.

Список литературы

1. *Mallows C. L.* An Inequality Involving Multinomial Probabilities. — Biometrika, (1968), 55:2, p. 422–424.
2. *Боровков А. А.* Теория вероятностей. М: Эдиториал УРСС, 1999, 472 с.. *Borovkov A. A. //* Probability theory. M: Editorial URSS, 1999, 472 p.
3. *Kullback S.*, *Leibler R. A.* On information and sufficiency. — The Annals of Mathematical Statistics, 1951, 22:1, p. 79–86.
4. *Ширяев А. Н.*, *Эрлих И. Г.*, *Яськов П. А.* Вероятность в теоремах и задачах (с доказательствами и решениями). М: МЦНМО, 2014, 648 с. // *Shiryayev A. N.*, *Erlikh I. G.*, *Yaskov P. A.,* Probability in theorems and problems (with proofs and solutions). M: MCNMO, 2014, 648 p.

UDC 519.233.33 DOI https://doi.org/10.52513/08698325_2022_29_1_1

Mironkin V. O., Chukhno A. B. (Moscow, MIEM, National Research University Higher School of Economics)**. On the application of Mallows inequality in testing numerical sequences.**

Abstract: In this paper estimates of the distributions of maximum and minimum frequencies in the implementation of an arbitrary polynomial scheme, which allow us to build statistical criteria for testing the hypothesis of agreement with the corresponding distribution, including in conditions of a limited amount of analyzed data are obtained.

Keywords: Polynomial scheme, frequency, Markov inequality, Chernov method, Kullback-Leibler divergence, equiprobable distribution.