

УДК 519.251.5

МАТЕМАТИКА

Ю. К. БЕЛЯЕВ, Л. В. РЫКОВА

НЕПАРАМЕТРИЧЕСКИЙ КРИТЕРИЙ КОЛМОГОРОВА ДЛЯ ВЫБОРОК ИЗ КОНЕЧНЫХ СОВОКУПНОСТЕЙ

(Представлено академиком А. Н. Колмогоровым 9 X 1972)

Представляет определенный интерес получение аналога непараметрического критерия А. Н. Колмогорова ⁽¹⁻³⁾ для случая выборочного обследования конечных совокупностей в условиях, когда не предполагается взаимная независимость последовательности наблюдаемых значений, а объектом исследования является эмпирическая функция распределения (э.ф.р.) конечной совокупности.

Пусть $\mathfrak{F}$ — конечная совокупность объектов O_i , $i = 1, \dots, N$. Каждому объекту O_i сопоставляется скалярная величина X_i . Предполагается, что $X_i \neq X_j$, $i \neq j$, $i, j = 1, \dots, N$. Совокупность $\mathfrak{F}$ характеризуется э.ф.р. $F_N(x) = l/N$, где l — число $X_i \leq x$. При случайной выборке объема n без возвращения из $\mathfrak{F}$ последовательно отбираются n объектов. При отборе $(k+1)$ -го объекта каждый из $(N-k)$ необследованных объектов может быть отобран с вероятностью $(N-k)^{-1}$. Если $i_1, \dots, i_n$ — номера объектов, попавших в выборку, то наблюдаемые значения величин $x_k = X_{i_k}$, $k = 1, \dots, n$. Результатам выборки сопоставим э.ф.р. $F_n(x) = k/n$, где k — число $x_i \leq x$. Ниже будет показано, что при $N \rightarrow \infty$, $n \rightarrow \infty$, $v = n/N \rightarrow v_0$, $0 \leq v_0 < 1$,

$$P \left\{ \max_x \sqrt{\frac{vN}{1-v}} |F_n(x) - F_N(x)| < y \right\} \rightarrow K(y) = \sum_k (-1)^k e^{-2k^2 y^2}; \quad (1)$$

здесь P соответствует процедуре случайной выборки без возвращения.

Исследуем вероятностные свойства случайной величины

$$\rho_{v,N} = \max_x |F_n(x) - F_N(x)|. \quad (2)$$

Пусть $X_{(1)} < \dots < X_{(N)}$ — вариационный ряд ⁽³⁾, соответствующий всей совокупности, а $x_{(1)} < \dots < x_{(n)}$ — вариационный ряд, построенный по результатам выборки. Так как выборка взята из $\mathfrak{F}$, то $x_{(1)} = X_{(l_1)} < x_2 = X_{(l_2)} < \dots < x_{(n)} = X_{(l_n)}$, $1 \leq l_1 < l_2 < \dots < l_n = N$.

Введем случайный процесс $\kappa(l)$, $l = 1, \dots, N$, равный числу $x_i \leq X_{(l)}$. Соотношение $\{\kappa(l) = k\}$ равносильно $\{x_{(k)} \leq X_{(l)} < x_{(k+1)}\}$.

Теорема 1. *Процесс $\kappa(l)$, $l = 1, \dots, N$, является цепью Маркова с вероятностями переходов*

$$\begin{aligned} P\{\kappa(l+1) = k+1 | \kappa(l) = k\} &= \frac{n-k}{N-l}, \\ P\{\kappa(l+1) = k | \kappa(l) = k\} &= 1 - \frac{n-k}{N-l}. \end{aligned} \quad (3)$$

Введенная выше случайная величина

$$\rho_{v,N} = \max_{1 \leq l \leq N} \left| \frac{\kappa(l)}{n} - \frac{l}{N} \right|. \quad (4)$$

Доказательство (3) основано на совпадении вероятностей отбора в выборку объектов с номерами $l_1, \dots, l_k$ как членов вариационного ряда и вероятностей скачков процесса $\kappa(l)$ в «моменты» $l_1, \dots, l_k$.

Таким образом, как и в классическом случае независимых наблюдений, найдена соответствующая статистика (4), распределение которой не зависит от фактических значений $X_1, \dots, X_N$. Этот результат позволяет строить как доверительные границы для $F_N(x)$, так и критерии значимости для проверки простых гипотез о виде $F_N(x)$. Имея в виду устойчивость результатов вычислений ⁽⁴⁾ при больших N и $v = n/N = \text{const}$, целесообразно искать распределение максимума модуля для процесса

$$\beta_{v,N}(t) = \sqrt{\frac{vN}{1-v}} \left[\frac{\kappa(l)}{n} - t \right], \quad t = \frac{l}{N}, \quad l = 1, \dots, N. \quad (5)$$

Из (4) и (5) имеем

$$\sqrt{\frac{vN}{1-v}} \rho_{v,N} = \max_{1 \leq l \leq N} |\beta_{v,N}(t)| = \bar{\beta}_{v,N}. \quad (6)$$

Положим $\bar{\beta}_{v,N}(t) = \max_{1 \leq k \leq l} |\beta_{v,N}(k/N)|$. Тогда для двумерного марковского процесса $(\beta_{v,N}(t), \bar{\beta}_{v,N}(t))$ в соответствии с теоремой 1 из (3) получим

Следствие 1. Вероятности состояний

$$p_l(b_l, B_l) = P \left\{ \beta_{v,N} \left(\frac{l}{N} \right) = b_l, \bar{\beta}_{v,N} \left(\frac{l}{N} \right) = B_l \right\}$$

находятся как решение рекуррентных уравнений

$$\begin{aligned} p_{l+1}(b_{l+1}, B_{l+1}) &= p_l \left\{ \left(b_{l+1} + \sqrt{\frac{v}{N(1-v)}} \right), B_{l+1} \right\} \times \\ &\times \left(1 - v + \frac{\sqrt{Nv(1-v)} b_{l+1} + v}{N(1-t)} \right) + p_l \left\{ \left(b_{l+1} - \sqrt{\frac{1-v}{Nv}} \right), B_{l+1} \right\} \times \\ &\times \left(v + \frac{1-v - \sqrt{Nv(1-v)} b_{l+1}}{N(1-t)} \right) \end{aligned}$$

при $|b_{l+1}| < B_{l+1}$ и аналогичных при $|b_{l+1}| = B_{l+1}$.

Искомое распределение $\bar{\beta}_{v,N}$ получается из соотношений

$$P \{ \bar{\beta}_{v,N} = B \} = p_N(0, B).$$

Разработана программа вычисления квантилей $\beta_v(v, N)$ распределения $\beta_{v,N}$. Пусть $\{B_\alpha\}$ — последовательный набор возможных значений $\bar{\beta}_{v,N}$. $B_{\alpha'} < B_\alpha$, $\alpha' < \alpha$. Тогда $\beta_v(v, N)$ определяется как наименьшее из чисел B_α таких, что $\sum_{\alpha \leq \alpha'} p_N(0, B_{\alpha'}) \geq \gamma$. Таким образом, $P \{ \bar{\beta}_{v,N} > \beta_v(v, N) \} \leq 1 - \gamma$ и, следовательно, с вероятностью, не меньшей γ , э.ф.р. $F_N(x)$ заключена между $\underline{F}_\gamma(x, x)$ и $\bar{F}_\gamma(x, x)$, где

$$\underline{F}_\gamma(x, x) = F_n(x) - \beta_v(v, N) \sqrt{\frac{1-v}{Nv}},$$

$$\bar{F}_\gamma(x, x) = F_n(x) + \beta_v(v, N) \sqrt{\frac{1-v}{Nv}}.$$

Аналогично получаются односторонние верхние и нижние границы, для чего находится распределение максимума и минимума процесса $\beta_{v, N}(t)$.

Заметим, что доверительные границы строятся для $F_N(x)$ и скачки $F_n(x)$ принадлежат множеству скачков $F_N(x)$. Поэтому получаемые доверительные границы будут уже границ, найденных в предположении независимости наблюдений и существовании непрерывной ф.р. ⁽³⁾. Например, при $N = 50$, $n = 3$ критические значения для квантилей $\gamma_1 = 0,1$ и $\gamma_2 = 0,01$ равны $\beta_{0,1} = 1,090$ и $\beta_{0,01} = 1,411$, тогда как по таблице 6.2 ⁽³⁾ $\beta_{0,1}' = 1,136$ и $\beta_{0,01}' = 1,481$.

Пусть $\beta(t)$ — стандартный винеровский процесс, $E\beta(t) = 0$, $E\beta^2(t) = t$, а $\beta_0(t)$ — условный винеровский процесс при условии $\beta(0) = \beta(1) = 0$.

Теорема 2. При $N \rightarrow \infty$, $n \rightarrow \infty$, $v \rightarrow \infty$, $v < N \rightarrow v_0 < 1$ распределение $\beta_{v, N}$ сходится к распределению $\beta_0 = \max_{0 \leq t \leq 1} |\beta_0(t)|$, т. е. справедливо соотношение (1).

Доказательство состоит из двух частей. Сначала для любого целого r и чисел $0 < t_1 < \dots < t_r < 1$ находим, что совместное распределение $\beta_{v, N}(l_i/N)$ при $N \rightarrow \infty$, $v \rightarrow \infty$, $v < N \rightarrow v_0 < 1$, $l_i/N \rightarrow t_i$ является гауссовским и совпадает с $\beta_0(t_i)$, $i = 1, \dots, r$. Для этого используем локальную предельную теорему для гипергеометрического распределения, см. ⁽⁵⁾, стр. 197. Затем показываем, что для любых $\varepsilon > 0$, $\delta > 0$, и $r > 0$ и всех достаточно больших N

$$P \left\{ -u + \varepsilon \leq \beta_{n, N} \left(\frac{l_i}{N} \right) \leq u - \varepsilon, \max_{l_i \leq l \leq l_{i+1}} \left| \beta_{n, N} \left(\frac{l}{N} \right) \right| > u \right\} \leq \frac{\delta}{r+1}. \quad (7)$$

Неравенство (7) получаем заменой на интервалах (l_i, l_{i+1}) процесса $\kappa(l)$ с однородными блужданиями, мажорирующими вероятности выхода за верхнюю и нижнюю границы. После вычитания систематических составляющих используем обобщенное неравенство Колмогорова, см. ⁽⁶⁾, стр. 277.

Заметим, что для конечных совокупностей возможны различные аналоги критериев Колмогорова — Смирнова ⁽³⁾. Критерий строится аналогичным образом, если конечная совокупность $\mathbb{P}_{N_1+N_2}$ путем случайного отбора разбивается на две $\mathbb{P}_{N_1}$ и $\mathbb{P}_{N_2}$, а затем на основе выборок без возвращения объемов n_1 и n_2 из $\mathbb{P}_{N_1}$ и $\mathbb{P}_{N_2}$ определяются отклонения в эмпирических функциях.

Предложенный подход можно обобщить на выборки из многомерных конечных совокупностей. Для простоты ограничимся случайной выборкой без возвращения из конечной совокупности $\mathbb{P}_N = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$, в которой каждый объект O_i характеризуется вектором (X_i, Y_i) . Вводя вариационные ряды $\{X_{(l_i)}\}$, $\{Y_{(l_i)}\}$, $l_i = 1, \dots, N$, $i = 1, 2$, независимо по каждой из координат X и Y , получаем отображение $\mathbb{P}_N$ в $\mathfrak{R}_N$ — подмножество точек квадратной решетки $K_N = \{(i, j) : i, j = 1, \dots, N\}$. $\mathfrak{R}_N = \{(l_1, l_2)\}$ определяется наборами значений l_1, l_2 пар $(X_{(l_1)}, Y_{(l_2)}) \in \mathbb{P}_N$. Соответственно через $\mathfrak{R}_{n, N} = \{(i, j)\}$ обозначаем множество пар (i, j) значений $(X_{(i)}, Y_{(j)})$ характеристик объектов O_i , попавших в выборку. Число различных $\mathfrak{R}_{n, N} \subset \mathfrak{R}_N$ равно $\binom{N}{n}$, все они равновероятны. Каждому $\mathfrak{R}_{n, N}$ сопоставляем

$$\rho_{n, N}^{(2)} = \max_{(l_1, l_2) \in \mathfrak{R}_N} \left| \frac{\kappa(l_1, l_2)}{n} - \frac{K(l_1, l_2)}{N} \right|, \quad (8)$$

где $\kappa(l_1, l_2)$ ($K(l_1, l_2)$) равно числу $O_i \in \mathfrak{R}_{n, N}$ ($O_i \in \mathfrak{R}_N$), у которых $X_i \leq X_{(l_1)}$, $Y_i \leq Y_{(l_2)}$.

Таким образом, каждому значению $\rho_{n, m}^{(2)} = \rho$ сопоставляется вероятность $m(\rho)$, равная отношению числа наборов $\mathfrak{R}_{n, N}$, для которых $\rho_{n, N}^{(2)} = \rho$, деленному на $\binom{N}{n}$.

Пусть $\rho_\gamma(\mathfrak{R}_{n, N})$ — квантиль уровня γ , а $\rho_\gamma(n, N) = \max_{\mathfrak{R}_{n, N} \subset \mathfrak{R}_N} \rho_\gamma(\mathfrak{R}_{n, N})$. Тогда γ — доверительные границы для $F_N(X, Y) = k/N$, где k — число O_i , для которых $X_i \leq X, Y_i \leq Y$, задаются неравенствами

$$\kappa(X, Y) / n - \rho_\gamma(n, N) \leq F_N(X, Y) \leq \kappa(X, Y) / n + \rho_\gamma(n, N). \quad (9)$$

В (9) $\kappa(X, Y)$ равно числу объектов O_i в выборке, у которых $X_i \leq X, Y_i \leq Y$.

Московский государственный университет
им. М. В. Ломоносова

Поступило
18 IX 1972

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

- ¹ A. N. Kolmogorov, Giorn. dill Inst. Ital. degli Attuari, 4, 83 (1933). ² A. N. Kolmogorov, Ann. Math. Statist., 12, 461 (1941). ³ Л. Н. Б о л ь ш е в, Н. В. С м и р н о в, Таблицы математической статистики, «Наука», 1968. ⁴ И. И. Г и х м а н, А. В. С к о р о х о д, Введение в теорию случайных процессов, «Наука», 1965. ⁵ В. Ф е л л е р, Введение в теорию вероятностей и ее приложения, 1, М., 1967. ⁶ М. Л о э в, Теория вероятностей, ИЛ, 1962.