Доклады Академии наук СССР
1974. Том 214, № 4

УДК 519.251.8 *МАТЕМАТИКА*

Л. В. РЫКОВА

НЕПАРАМЕТРИЧЕСКИЙ КРИТЕРИЙ КОЛМОГОРОВА — СМИРНОВА ДЛЯ ВЫБОРОК ИЗ КОНЕЧНЫХ СОВОКУПНОСТЕЙ

(Представлено академиком А. Н. Колмогоровым 22 X 1973)

Для проверки однородности двух выборок Н. В. Смирновым [1] были введены статистики

$$D^+_{m,n} = \sup_x [F_n(x) - G_m(x)], \qquad D^-_{m,n} = -\inf_x [G_m(x) - F_n(x)],$$
$$D_{m,n} = \sup_x |F_n(x) - G_m(x)|, \tag{1}$$

где $F_n(x)$ и $G_m(x)$ — эмпирические функции распределения (э.ф.р.), построенные по результатам двух выборок $\xi_1, \ldots, \xi_n$ и $\eta_1, \ldots, \eta_m$ объема n и m соответственно. При этом предполагалось, что $\xi_1, \ldots, \xi_n$ $(\eta_1, \ldots, \eta_m)$ — взаимно независимые случайные величины с функцией распределения $F(x)$ $(G(x))$. С помощью (1) можно проверять гипотезу о совпадении функций распределения $F(x)$ и $G(x)$.

В данной работе исследуются свойства аналогичных статистик, которыми можно пользоваться при изучении конечной совокупности без предположения взаимной независимости наблюдений.

Пусть $\mathfrak{P}$ — конечная совокупность объектов O_i, каждому из которых сопоставляется скалярная величина X_i, $i=1, \ldots, N$. Предположим, что $X_i \neq X_j$, $i \neq j$, и объекты занумерованы в порядке возрастания величин X_i, т. е. $X_1 < X_2 < \ldots < X_N$. Совокупность $\mathfrak{P}$ будем характеризовать э.ф.р. $F_N(x) = l/N$, где l — число таких объектов, у которых $X_i < x$.

Из совокупности $\mathfrak{P}$ берутся две случайные выборки без возвращения объема n_1 и n_2. Результатам выборок сопоставим э.ф.р. $F_{n_1}(x) = k_1/n_1$ и $F_{n_2}(x) = k_2/n_2$ соответственно, где k_1 (k_2) равно числу объектов, попавших в первую (вторую) выборку, у которых $X_i \leqslant x$. Ниже будет показано, что при $n_i \to \infty$, $(n_i/N) \to \nu_i$, $0 < \nu_i < 1$, $i=1, 2$,

$$P\left\{\sup_x \left(\frac{N-1}{N} \frac{n_1 n_2}{n_1+n_2}\right)^{1/2} |F_{n_1}(x) - F_{n_2}(x)| < z\right\} \to K(z) = \sum_k (-1)^k e^{-2k^2z^2}. \tag{2}$$

Случай $\nu_2 = 0$ $(\nu_1 = 0)$ идентичен рассмотренному ранее [2].

Определим величину

$$\rho_{n_1, n_2, N} = \sup_x \left(\frac{N-1}{N} \frac{n_1 n_2}{n_1+n_2}\right)^{1/2} |F_{n_1}(x) - F_{n_2}(x)|. \tag{3}$$

Поскольку $F_{n_1}(x)$ и $F_{n_2}(x)$ постоянны на $[X_i, X_{i+1})$, $i=1, \ldots, N-1$, и имеют скачки $1/n_1$ и $1/n_2$ соответственно, в точках X_i, $i=1, \ldots, N$, то отображение X_i в точку i/N отрезка $[0, 1]$ не изменит величины (3). Поэтому распределение величины (3) совпадает с распределением

$$\rho_{n_1, n_2, N} = \max_{1 \leqslant l \leqslant N} \left(\frac{N-1}{N} \frac{n_1 n_2}{n_1+n_2}\right)^{1/2} \left|\frac{\varkappa_1(l)}{n_1} - \frac{\varkappa_2(l)}{n_2}\right|,$$

где $\varkappa_i(l)$ равно числу объектов в i-й выборке, $i=1, 2$, номера которых меньше l.

Теорема 1. *Двумерный процесс $(\varkappa_1(l), \varkappa_2(l))$ является двумерным марковским процессом с вероятностями переходов*

$$p_1=P\{\varkappa_1(l+1)=k_1+1,\ \varkappa_2(l+1)=k_2 \mid \varkappa_1(l)=k_1,\ \varkappa_2(l)=k_2\}=(n_1-k_1)/(N-l),$$
$$p_2=P\{\varkappa_1(l+1)=k_1,\ \varkappa_2(l+1)=k_2+1 \mid \varkappa_1(l)=k_1,\ \varkappa_2(l)=k_2\}=(n_2-k_2)/(N-l),$$
$$p=P\{\varkappa_1(l+1)=k_1,\ \varkappa_2(l+1)=k_2 \mid \varkappa_1(l)=k_1,\ \varkappa_2(l)=k_2\}=1-p_1-p_2. \tag{4}$$

Доказательство теоремы состоит в проверке совпадения вероятностей отбора объектов с номерами $l_1, l_2, \ldots, l_{n_1}$ и $m_1, m_2, \ldots, m_{n_2}$ в первую и вторую выборки соответственно и вероятностей скачков процесса с вероятностями переходов (4) в тех же точках.

Поскольку для любых l_1, l_2

$$P\{\varkappa_1(l_1, l_2)=d_1, \varkappa_2(l_1, l_2)=d_2\}=h_{N, l_2-l_1}^{n_1, d_1} h_{N-n_1, l_2-l_1-d_1}^{n_2, d_2}, \tag{5}$$

где $h_{N,M}^{m}=\binom{M}{m}\binom{N-M}{n-m}\Big/\binom{N}{n}$ — гипергеометрические вероятности, $\varkappa_i(l, m)=\varkappa_i(m)-\varkappa_i(l)$, то процесс $(\varkappa_1(l), \varkappa_2(l))$ естественно называть двумерным гипергеометрическим процессом.

Лемма 1. *Пусть m, n, M, N — целые числа, $m\leqslant M\leqslant N$, $m\leqslant n\leqslant N$, $b(n, N)=n(N-n)/N$, $0<\underline{c}<\Delta<\bar{c}<1$,*

$$\nu_+=\overline{\lim_{n\to\infty}}(n/N)<1, \quad |M/N-\Delta|<1/N, \quad \frac{|m-n M/N|^3}{b^2(n, N)}=\alpha_{N,M}^{n,m}\to 0$$

при $n\to\infty$.

Тогда для любого $\varepsilon>0$

$$h_{N,M}^{n,m}=\frac{\exp\{-(m-n\Delta)^2/[2b(n,N)\Delta(1-\Delta)]\}}{[2\pi b(n,N)\Delta(1-\Delta)]^{1/2}}(1+\theta_{N,M}^{n,m}),$$

где

$$|\theta_{N,M}^{n,m}|<\gamma_{N,M}^{n,m}/[\Delta(1-\Delta)(1-\nu_+)(1-\varepsilon)]^2,$$

$$\gamma_{N,M}^{n,m}=\min(\alpha_{N,M}^{n,m}, n^{-1/3}).$$

Рассмотрим процесс $\beta_{n_1,n_2}(l_1/N, l_2/N)=(\xi_{n_1,n_2}^{(1)}(l_1/N), \xi_{n_1,n_2}^{(2)}(l_2/N))$, где

$$\xi_{n_1,n_2}^{(i)}\left(\frac{l}{N}\right)=\left(\frac{N-1}{N}\frac{n_1 n_2}{n_1+n_2}\right)^{1/2}\left[\frac{\varkappa_i(l)}{n_i}-\frac{l}{N}\right].$$

Легко проверить, что $E\beta_{n_1, n_2}(l_1/N, l_2/N)=0$,

$$S=\|a_{ij}\|=\begin{pmatrix} \dfrac{(N-n_1)n_2}{N(n_1+n_2)}\dfrac{l_1}{N}\left(1-\dfrac{l_2}{N}\right), & -\dfrac{n_1 n_2}{(n_1+n_2)N}\dfrac{l_1}{N}\left(1-\dfrac{l_2}{N}\right) \\ -\dfrac{n_1 n_2}{(n_1+n_2)N}\dfrac{l_1}{N}\left(1-\dfrac{l_2}{N}\right), & \dfrac{(N-n_2)n_1}{(n_1+n_2)N}\dfrac{l_1}{N}\left(1-\dfrac{l_2}{N}\right) \end{pmatrix},$$

где

$$a_{ij}=E\xi_{n_1,n_2}^{(i)}(l_1/N)\,\xi_{n_1,n_2}^{(j)}(l_2/N), \quad i, j=1, 2.$$

Теорема 2. *При $n_i\to\infty$, $i=1, 2$, $\overline{\lim_{n_i\to\infty}}(n_i/N)=\nu_i<1$ конечномерные распределения процесса $\beta_{n_1, n_2}(l_1/N, l_2/N)$ сходятся к конечномерным распреде-*

лениям двумерного гауссовского процесса $\beta(t, s)=(\beta_1(t), \beta_2(s))$ *с нулевым математическим ожиданием и ковариационной матрицей*

$$Q=\begin{pmatrix} \dfrac{\nu_2(1-\nu_1)t(1-s)}{\nu_1+\nu_2} & -\dfrac{\nu_1\nu_2 t(1-s)}{\nu_1+\nu_2} \\ -\dfrac{\nu_1\nu_2 t(1-s)}{\nu_1+\nu_2} & \dfrac{\nu_1(1-\nu_2)t(1-s)}{\nu_1+\nu_2} \end{pmatrix}, \quad t<s.$$

При доказательстве теоремы надо воспользоваться марковскими свойствами процесса $\beta_{n_1, n_2}(l_1/N, l_2/N)$, равенством (5), леммой 1 и методом интегральных сумм ((3), стр. 188).

Лемма 2. *Для любых* t_1 *и* t_2 *из* $[0, 1]$, *всех* N *и* n_i *таких, что* $0<\underline{\nu}\leq n_i/N\leq\bar{\nu}<1$ *существует такая* $H=H(\underline{\nu}, \bar{\nu})>0$, что

$$\mu_4^{(i)}=E|\overset{(i)}{\xi}_{n_1,n_2}(t_1)-\overset{(i)}{\xi}_{n_1,n_2}(t_2)|^4\leq H|t_1-t_2|^2, \quad i=1, 2. \tag{6}$$

Неравенство (6) получается как следствие гипергеометричности распределения приращений процесса $\varkappa_i(l)$.

Выполнение условий теоремы 2 и леммы 2 дает возможность воспользоваться теоремой (4) о том, что для всех непрерывных на $C_{[0,1]}$ функционалов f распределение $f(\beta_{n_1, n_2}(l_1/N, l_1/N))$ будет сходиться к распределению $f(\beta(t, t))$, $C_{[0,1]}$ — пространство всех непрерывных на $[0, 1]$ функций с метрикой

$$\rho(x, y)=\sup_{0\leq t\leq 1}|x(t)-y(t)|.$$

В частности, распределение

$$f(\beta_{n_1,n_2}(l_1/N, l_1/N))=\max_{1\leq l\leq N}|\overset{(1)}{\xi}_{n_1,n_2}(l/N)-\overset{(2)}{\xi}_{n_1,n_2}(l/N)|=\rho_{n_1,n_2,N}$$

будет сходиться к распределению

$$f(\beta(t,t))=\max_{0\leq t\leq 1}|\beta_1(t)-\beta_2(t)|=\max_{0\leq t\leq 1}|\beta_0(t)|.$$

Поскольку $E\beta_0(t)\equiv0$, $E\beta_0(t)\beta_0(s)=t(1-s)$, $t<s$, то $\beta_0(t)$ является условным винеровским процессом на $[0, 1]$ с закрепленными концами $\beta_0(0)=\beta_0(1)=1$, а распределение $\max\limits_{0\leq t\leq 1}|\beta_0(t)|$ является известным распределением Колмогорова $K(z)$. Таким образом, верна

Теорема 3. *При* $n_i\to\infty$, $\overline{\lim\limits_{n_i\to\infty}}(n_i/N)=\nu_i$, $0\leq\nu_i<1$, $i=1, 2$, *распределение* $\rho_{n_1, n_2, N}$ *сходится к распределению* $\beta_0=\max\limits_{0\leq x\leq 1}|\beta_0(t)|$, *т. е. справедливо соотношение* (2).

Статистика $\rho_{n_1, n_2, N}$ может быть использована в следующих математических моделях:

а) Объекты совокупности $\mathfrak{P}$ характеризуются скалярными величинами $X_i=X_i(s)$, вообще говоря зависящими от времени (хранения) s, $s\geq0$. Для проверки гипотезы о неизменности характеристик, т. е. гипотезы $H_0=\{X_i(s_1)=X_i(s_2),\ i\in[1, 2, \ldots, N]\}$, в два момента времени $s_1<s_2$ берутся две случайные выборки без возвращения объемов n_1, n_2. Если $n_i\gg1$, $i=1, 2$, $N\gg1$, $n_i/N\leq\bar{\nu}_i<1$, то при тестовой статистике $\rho_{n_1, n_2, N}$ критическое множество с уровнем значимости, близким к α, имеет вид

$$\rho_{n_1, n_2, N}>z_{1-\alpha}, \qquad K(z_{1-\alpha})=1-\alpha. \tag{7}$$

б) Выясняется вопрос о различии влияния двух типов воздействия на характеристики объектов O_i конечной совокупности $\mathfrak{P}$, $i=1, \ldots, N$. Для этого объекты $\mathfrak{P}$ разбиваются на три группы объемом n_1, n_2 и $N-(n_1+n_2)$ на основе случайных выборок без возвращения объема n_1 и n_2. На объекты, составляющие первую выборку, оказывается воздействие первого типа, на объекты, составляющие вторую выборку, оказывается воздействие второго типа. За нулевую гипотезу принимается идентичность воздействия обоих типов или отсутствие его. Тестовая статистика $\rho_{n_1, n_2, N}$ имеет то же критическое множество (7). Отбор объектов для проверки идентичности воздействий можно проводить с постоянной вероятностью. При этом объемы выборок n_1 и n_2 будут случайными величинами.

Заметим, что относительно предельных распределений статистик

$$\rho^{+}_{n_1,n_2,N}=\sup_x[F_{n_1}(x)-F_{n_2}(x)], \quad \rho^{-}_{n_1,n_2,N}=-\inf_x[F_{n_2}(x)-F_{n_1}(x)],$$

$$T=\int_{-\infty}^{\infty}[F_{n_1}(x)-F_{n_2}(x)]^2\,dH_{n_1+n_2}(x),$$

где

$$H_{n_1+n_2}(x)=\varphi_1(n_1, n_2, N)F_{n_1}(x)+\varphi_2(n_1, n_2, N)F_{n_2}(x),$$

будут справедливы утверждения, аналогичные теореме 3, вследствие непрерывности этих статистик, рассматриваемых как функционалы от траекторий гипергеометрических процессов.

Московский государственный университет
им. М. В. Ломоносова

Поступило
4 X 1973

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Н. В. Смирнов, Бюлл. Московск. унив., сер. А, 2, 3 (1939). [2] Ю. К. Беляев, Л. В. Рыкова, ДАН, 210, № 6 (1973). [3] В. Феллер, Введение в теорию вероятностей, 1, М., 1967. [4] И. И. Гихман, А. В. Скороход, Введение в теорию случайных процессов, «Наука», 1965.