Доклады Академии Наук СССР
1952. Том LXXXV, № 3

МАТЕМАТИКА

В. С. МИХАЛЕВИЧ

О ВЗАИМНОМ РАСПОЛОЖЕНИИ ДВУХ ЭМПИРИЧЕСКИХ ФУНКЦИЙ РАСПРЕДЕЛЕНИЯ

(Представлено академиком А. Н. Колмогоровым 29 V 1952)

Н. В. Смирнов [1] доказал следующую теорему:

Пусть ξ — случайная величина, имеющая непрерывную функцию распределения; $x_1, x_2, \ldots, x_{n_1}$ и $y_1, y_2, \ldots, y_{n_2}$ — результаты двух серий независимых испытаний над величиной ξ. Обозначим через $F_1(x)$ и $F_2(x)$ эмпирические функции распределения соответственно для первой и второй серий наблюдений и рассмотрим лестничную кривую

$$T_{n_1 n_2}(x, z) = F_2(x) + \frac{z}{\sqrt{m}} \quad \left(z \geqslant 0, m = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\right).$$

Скажем, что кривая $F_1(x)$ пересекает $T_{n_1 n_2}(x, z)$, в точке x_k, если выполнены неравенства $F_1(x_k - 0) \leqslant T_{n_1 n_2}(x_k, z) < F_1(x_k + 0)$. Обозначим через $v(z, n_1, n_2)$ число точек пересечения в последовательности $x_1, x_2, \ldots, x_{n_1}$. Тогда, при условии, что $n_1 / n_2 = \text{const}$, имеет место соотношение

$$\lim_{n_2 \to \infty} P\{v(z, n_1, n_2) < t\sqrt{m}\} = \begin{cases} 0 & \text{при } t < 0, \\ 1 - e^{-\frac{(t+2z^2)^2}{2}} & \text{при } t \geqslant 0, \end{cases}$$

Отсюда, в качестве простейшего следствия,

$$\lim_{n_2 \to \infty} P\left\{\sup_{-\infty < x < +\infty} [F_1(x) - F_2(x)] < \frac{z}{\sqrt{m}}\right\} = 1 - e^{-2z^2} \quad (z \geqslant 0).$$

Указанные результаты используются для оценки правильности гипотезы, что обе серии наблюдений получены в результате испытаний над случайными величинами с одним и тем же непрерывным распределением вероятностей. Однако, в силу их асимптотического характера, они не улавливают влияния числа наблюдений, которое может быть значительным, если числа n_1 и n_2 невелики. Возникает задача определения соответствующих распределений вероятностей для конечных n_1 и n_2. Б. В. Гнеденко и В. С. Королюк [2] нашли точное распределение для величины $D_n^+(x) = \sup\limits_{-\infty < x < +\infty} [F_1(x) - F_2(x)]$ при условии, что $n_1 = n_2 = n$.

В настоящей заметке определено при условии $n_1 = n_2 = n$ точное распределение вероятностей: 1) для случайной величины $v_n(z) = v(z, n_1, n_2)$; 2) для доли положительных уклонений одной эмпирической кривой распределения относительно границы, определяемой другой.

Введем обозначения $c = [z\sqrt{2n}]$, $k = [t\sqrt{2n}]$.

Теорема 1. *В высказанных предположениях*

$$B_n(z, t) = P\{v_n(z) < t\sqrt{2n}\} = \begin{cases} 0 & \text{при } k \leqslant 0, \\ 1 - \dfrac{C_{2n}^{n-(k+c)}}{C_{2n}^{n}} & \text{при } k > 0,\ c + k \leqslant n, \\ 1 & \text{при } c + k > n. \end{cases}$$

Отсюда легко заключить, что

$$\lim_{n \to \infty} B_n(z, t) = \begin{cases} 0 & \text{при } t < 0, \\ 1 - e^{-2(t+z)^2} & \text{при } t \geqslant 0. \end{cases}$$

В силу результатов, изложенных в работе ([2]), нам достаточно доказать равенство $P\{v_n(z) < t\sqrt{2n}\} = P\{nD_n^+ < c + k\}$ при указанных выше c и k.

Расположим результаты обеих серий наблюдений по величине в одну последовательность $z_1 \leqslant z_2 \leqslant \ldots \leqslant z_{2n}$ и каждому z_k поставим в соответствие случайную величину

$$\xi_k = \begin{cases} +1, \text{ если } z_k \text{ равно одному из } x_i, \\ -1, \text{ если } z_k \text{ равно одному из } y_j. \end{cases}$$

Введем обозначения $s_0 = 0$, $s_k = \xi_1 + \xi_2 + \cdots + \xi_k$ $(k = 1, \ldots, n)$. Нетрудно убедиться, что $v_n(0)$ есть число всех пар $(s_r = 0,\ s_{r+1} = +1)$ в последовательности $s_0, s_1, \ldots, s_{2n}$ точно так же, как $v_n(z)$ есть число всех пар $(s_r = c,\ s_{r+1} = c + 1)$ в этой последовательности. Ясно также, что $nD_n^+ = \sup\limits_{1 \leqslant k \leqslant 2n} s_k$.

Воспользуемся следующей геометрической иллюстрацией: частица, находящаяся в момент $\tau = 0$ в положении $x = 0$, подвержена случайным толчкам в моменты времени $\tau = 1, 2, \ldots, 2n$, в результате каждого из которых она может сдвигаться на $+1$ или -1. В плоскости (τ, x) путь частицы при каждом толчке изобразится перемещением на единицу вправо $(+1)$ или влево (-1). Пусть $c \geqslant 0$, $k > 0$ — целые числа и $k + c \leqslant n$. Тогда: 1) $P\{nD_n^+ = c\}$ и 2) $P\{v_n(z) = k\}$ представляют вероятности того, что движущаяся таким образом частица: 1) достигает прямую $x = c$, но не пересекает ее; 2) за все время движения покинет прямую $x = c$, уйдя вправо, ровно k раз.

Среди $2n$ толчков имеется n «плюсов» и n «минусов», поэтому частица, вышедшая из точки $(0, 0)$, в конце движения придет в точку $(0, 2n)$. Легко подсчитать, что число всех возможных траекторий равно C_{2n}^n, а так как все траектории равновероятны, то вероятность каждой равна $1 / C_{2n}^n$. Поэтому для доказательства теоремы достаточно показать, что число $d_n(c, k)$ траекторий, благоприятствующих событию $v_n(z) = k$, равно числу $\delta_n(c, k)$ траекторий, достигающих прямую $x = c + k$, но не пересекающих ее.

Назовем 0-отрезком (0′-отрезком) часть траектории, состоящую из одинакового числа «минусов» и «плюсов», причем на протяжении всего отрезка число «минусов» не меньше числа «плюсов» (число «плюсов» не меньше числа «минусов»), если двигаться по отрезку снизу вверх. Назовем длиной отрезка количество содержащихся в нем «плюсов» и обозначим через τ_r число всех возможных 0-отрезков длины r $(\tau_0 = 1)$.

Пусть вначале $c = 0$. Из всех траекторий, благоприятствующих событию $v_n(0) = k$, подсчитаем сперва те, которые начинаются «плюсом» (пусть число их есть $\sigma_n(k)$). Каждая из них может быть реализована следующим образом: вначале идет «плюс», за ним 0′-отрезок длины r,

за которым следует один «минус», а далее траектория $k-1$ раз покидает ось времени, уходя от нее вправо. Так как r может принимать все целые значения от 0 до $n-k$, а траектории, соответствующие различным r, не могут совпадать, то

$$\sigma_n(k) = \sum_{r=0}^{n-k} \tau_r\, d_{n-r-1}\,(0,\, k-1).$$

Теперь подсчитаем число траекторий, благоприятствующих событию $v_n(0) = k$ и начинающихся «минусом». Каждая из них реализуется следующим образом: вначале идет 0-отрезок длины s, а за ним остальная часть траектории, начинающаяся «плюсом». Так как при различных s траектории не совпадают, а s может меняться от 1 до $n-k$, то число $d_n(0, k) - \sigma_n(k)$ всех таких траекторий равно $d_n(0, k) - \sigma_n(k) = \sum_{s=1}^{n-k} \tau_s \sigma_{n-s}(k)$, т. е. $d_n(0, k) = \sum_{s=0}^{n-k} \tau_s \sigma_{n-s}(k)$, что полезно, меняя порядок суммирования, представить так: $d_n(0, k) = d_{n-1}(0, k-1) + d_{n-2}(0, k-1)[\tau_1 + \tau_1] + \cdots + d_{k-1}(0, k-1)[\tau_{n-k} + \tau_{n-k-1}\tau_1 + \cdots + \tau_1\tau_{n-k-1} + \tau_{n-k}]$.

Покажем, что числа $\delta_n(0, k)$ удовлетворяют такому же соотношению. Для этого заметим, что любая траектория, благоприятствующая событию $nD_n^+ = k$, может быть реализована так: вначале идет 0-отрезок длины r, за ним «плюс», далее траектория от точки $(\tau = 2r+1,\ x = 1)$ до точки $(\tau = 2(r+s)+1, x = 1)$ успевает достигнуть прямую $x = k$, затем следует «минус», после которого траектория не пересекает оси времени, т. е. следует от точки $(\tau = 2(r+s+1),\ x = 0)$ до точки $(\tau = 2n,\ x = 0)$ некоторый 0-отрезок. Учитывая, что s может меняться от $n-1$ до $k-1$, а r (при данном s) от 0 до $n-s-1$, имеем: $\delta_n(0, k) = \delta_{n-1}(0, k-1) + \delta_{n-2}(0, k-1)[\tau_1 + \tau_1] + \cdots + \delta_{k-1}(0, k-1) \times [\tau_{n-k} + \tau_{n-k-1}\tau_1 + \cdots + \tau_1\tau_{n-k-1} + \tau_{n-k}]$.

Так как для $n = 1$ и $n = 2$ наше утверждение легко проверяется непосредственно, то, воспользовавшись принципом полной индукции, получаем $d_n(0, k) = \delta_n(0, k)$ при любых n и $k \leqslant n$. Итак, $P\{v_n(0) = k\} = P\{nD_n^+ = k\}$, т. е. $B_n(0, t) = \Phi_n^+(t)$.

Теперь уже просто доказывается равенство

$$P\{v_n(c) = k\} = P\{nD_n^+ = k + c\} \quad \text{при } c > 0.$$

В самом деле, $d_n(c, k)$ можно подсчитать так: обозначим число всех возможных путей, приводящих из начала координат в точку $(\tau = y,\ x = c)$, но ранее не достигавших прямую $x = c$, через $a(y, c)$, а через $b(y', c, n)$ — число путей, ведущих из точки $(\tau = y',\ x = c)$ в точку $(2n, 0)$, но не достигающих при этом прямой $x = c$. Тогда

$$d_n(c, k) = \sum_{(y, y')} a(y, c)\, d_{\frac{y'-y}{2}}(0, k)\, b(y', c, n),$$

где сумма берется по всем допустимым y и y'.

Подсчитывая $\delta_n(c, k)$ таким же образом, получим аналогичное равенство, только под знаком суммы будет стоять $\delta_{\frac{y'-y}{2}}(0, k)$ вместо $d_{\frac{y'-y}{2}}(0, k)$. Следовательно, $d_n(c, k) = \delta_n(c, k)$ при указанных выше c и k. Теорема доказана.

Назовем «положительными скачками» функции $F_1(x)$ относительно кривой $T_{n_1 n_2}(x, z)$ все те точки x_k, в которых выполняется неравенство $F_1(x_k + 0) > T_{n_1 n_2}(x_k, z)$. Обозначим через $C_{n_1 n_2}(z)$ число всех «положительных скачков» в последовательности $x_1, x_2, \ldots, x_{n_1}$. Мы

найдем распределение вероятностей величины $C_n(z) = C_{n_1 n_2}(z)$ в предположении $n_1 = n_2 = n$.

Положим $\nu = [nt]$ и $\chi_n(z, t) = P\{C_n(z) < nt\}$. В заметке [3] было показано, что случайная величина $C_n(0)$ распределена равномерно, т. е. $\chi_n(0, t) = \frac{\nu}{n+1}$ $(\nu = 0, 1, \ldots, n)$.

Теорема 2. *При* $c > 0$

$$\chi_n(z, t) = \Phi_n^+(z) + \frac{1}{C_{2n}^n}\left[\sum_{r=0}^{\nu-1}(r+1)\frac{C_{2r}^r}{r+1}H(n-r, c) + \right.$$

$$\left. + \nu\sum_{r=\nu}^{n-c}\frac{C_{2r}}{r+1}H(n-r, c)\right]$$

$$\chi_n(z, t) = \begin{cases} & \text{при } 0 \leqslant t \leqslant 1; \\ 0 & \text{при } t < 0, \\ 1 & \text{при } t \geqslant 1, \end{cases}$$

где $H(n, c) = C_{2(n-1)}^{n-c} - C_{2(n-1)}^{n-c-2}$.

Доказательство осуществляется таким же методом с использованием результата Б. В. Гнеденко: вероятность того, что $nD_n^+ = c$ и максимум этот достигается только в одной точке, равна $H(n, c)$.

Замечание. Метод траекторий блуждающей точки может быть применен к критерию Н. В. Смирнова симметрии функции распределения. Помимо результатов, изложенных в работе [4], можно установить такое следствие теоремы 1:

Пусть имеется n результатов независимых испытаний над случайной величиной с непрерывной и симметричной функцией распределения с центром симметрии в точке a. Обозначим через $\mu(x)$ разность между числом наблюдений, попавших в интервал $(a, a+x)$, и числом наблюдений, попавших в интервал $(a-x, a)$ (в работе [4] найдены распределения для величин $\max\limits_{(x>0)}\mu(x)$ и $\max\limits_{(x>0)}|\mu(x)|$). Введем величину, характеризующую поведение функции $\mu(x)$ при изменении x от нуля до бесконечности, а именно, обозначим через $V_n(c)$ число всех таких точек x, для которых $\mu(x) = c$, $\mu(x+0) = c+1$ $(0 \leqslant c \leqslant n)$.

Наше следствие может быть сформулировано так:

$$P\{V_n(c) \geqslant k\} = \frac{1}{2^{n-1}}\sum_{l=0}^{\frac{n-c}{2}-k} C_n^l, \text{ если } n \text{ и } c \text{ одинаковой четности;}$$

$$P\{V_n(c) \geqslant k\} = \frac{1}{2^{n-1}}\sum_{l=0}^{\frac{n-c-1}{2}-k} C_n^l + \frac{1}{2^n}C_n^{\frac{n-c+1}{2}}, \text{ если } n \text{ и } c \text{ разной четности.}$$

В заключение выражаю сердечную благодарность моему учителю Б. В. Гнеденко за постановку задач и за советы, данные при их решении.

Киевский государственный университет

Поступило 29 V 1952

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Н. В. Смирнов, Бюлл. Моск. ун-та, **2**, в. 2 (1939). [2] Б. В. Гнеденко и В. С. Королюк, ДАН, **80**, № 4 (1951). [3] Б. В. Гнеденко и В. С. Михалевич, ДАН, **82**, № 6 (1952). [4] Н. В. Смирнов, ДАН, **56**, № 1 (1947).