Доклады Академии Наук СССР
1951. Том LXXX, № 4

МАТЕМАТИКА

Действительный член АН УССР Б. В. ГНЕДЕНКО и В. С. КОРОЛЮК

О МАКСИМАЛЬНОМ РАСХОЖДЕНИИ ДВУХ ЭМПИРИЧЕСКИХ РАСПРЕДЕЛЕНИЙ

Н. В. Смирновым была доказана [1] следующая важная теорема математической статистики.

Пусть ξ — случайная величина, имеющая непрерывную функцию распределения $F(x)$. Пусть, далее, $x_1, x_2, \ldots, x_{n_1}$ и $y_1, y_2, \ldots, y_{n_2}$ — результаты двух серий независимых наблюдений над величиной ξ. Если через $F_1(x)$ и $F_2(x)$ обозначить эмпирические функции распределения, соответственно, для первой и второй серий наблюдений, то при условии, что $\lim\limits_{n_1\to\infty} \frac{n_2}{n_1} = \tau$ $(0 < \tau < \infty)$, имеют место соотношения

$$\lim_{n_1\to\infty} P\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup_{-\infty<x<\infty} \{F_1(x) - F_2(x)\} < z\right\} = 1 - e^{-2z^2} \text{ при } z > 0; \quad (1)$$

$$\lim_{n_1\to\infty} P\left\{\sqrt{\frac{n_1 n_2}{n_1+n_2}} \sup |F_1(x) - F_2(x)| < z\right\} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2z^2} \text{ при } z > 0. \quad (2)$$

Эти 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. Для практического применения, однако, указанные теоремы нам представляются недостаточными, так как они носят лишь асимптотический характер и не улавливают влияния числа наблюдений, а это влияние может быть весьма значительным, если только числа n_1 и n_2 невелики. Возникает задача определения распределения соответствующих вероятностей для конечных значений n_1 и n_2.

В настоящей заметке дается исчерпывающее решение этой «конечной» задачи в предположении $n_1 = n_2$, и для этого частного случая отсюда получаются предельные результаты Н. В. Смирнова.

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

$$D_n^+ = \sup_{-\infty<x<\infty} [F_1(x) - F_2(x)], \quad D_n = \sup_{-\infty<x<\infty} |F_1(x) - F_2(x)|.$$

Теорема. *В высказанных ранее предположениях имеют место равенства*

$$\Phi_n^+(z) = P\left\{\sqrt{\frac{n}{2}} D_n^+ < z\right\} = 1 - \frac{C_{2n}^{n-c}}{C_{2n}^n} \qquad \text{при } \frac{1}{\sqrt{2n}} < z \leqslant \sqrt{\frac{n}{2}},$$

$$\Phi_n(z) = P\left\{\sqrt{\frac{n}{2}} D_n < z\right\} = \frac{1}{C_{2n}^n} \sum_{k=-\left[\frac{n}{c}\right]}^{\left[\frac{n}{c}\right]} (-1)^k C_{2n}^{n-kc} \quad \text{при } \frac{1}{\sqrt{2n}} < z \leqslant \sqrt{\frac{n}{2}},$$

$$\Phi_n^+(z) = \Phi_n(z) = \begin{cases} 0 & \text{при } z \leqslant \frac{1}{\sqrt{2n}} \\ 1 & \text{при } z > \sqrt{\frac{n}{2}} \end{cases}.$$

Для доказательства расположим результаты обеих серий наблюдений по величине в одну последовательность $z_1 < z_2 < \ldots < z_{2n}$ и рассмотрим последовательность вспомогательных случайных величин $\xi_1, \xi_2, \ldots, \xi_{2n}$. Каждая величина ξ_k может принимать только одно из двух значений $+1$ или -1 согласно правилу: ξ_k равно $+1$, если z_k принадлежит первой серии наблюдений; ξ_k равно -1, если z_k принадлежит второй серии наблюдений.

Введем обозначение

$$s_k = \xi_1 + \xi_2 + \ldots + \xi_k.$$

Очевидно, что

$$nD_n^+ = \sup_{1 \leqslant k \leqslant 2n} s_k, \quad nD_n = \sup_{1 \leqslant k \leqslant 2n} |s_k|.$$

Прибегнем к следующей иллюстрации: частица, находящаяся в момент $t = 0$ в положении $x = 0$, подвержена случайным толчкам в моменты $t = 1, 2, \ldots, 2n$, в результате каждого из которых она может сдвигаться или на $+1$ или на -1. В плоскости (x, t) путь частицы при каждом толчке изобразится перемещением на единицу «вверх» и на единицу вправо или влево. Пусть c — целое положительное число. В нашей интерпретации вероятности: 1) $P\{nD_n^+ < c\}$ и 2) $P\{nD_n < c\}$ представляют собой вероятности того, что движущаяся указанным способом частица: 1) во все время движения будет оставаться левее прямой $x = c$, 2) во все время движения будет оставаться между прямыми $x = c$ и $x = -c$.

Так как среди $2n$ толчков имеются n «правых» и n «левых», то частица, вышедшая из точки $(0, 0)$, в конце движения придет в точку $(0, 2n)$. Легко подсчитать, что число всех путей, приводящих частицу из точки $(0, 0)$ в точку $(0, 2n)$, равно C_{2n}^n, а так как все эти пути равновероятны, то вероятность каждого пути равна $1 / C_{2n}^n$.

Число путей N, благоприятствующих событию $nD_n^+ < c$, определяется методом, подобным тому, который изложен в ((2), § 16). А именно, каждому пути, исходящему из точки $(0, 0)$, достигающему прямой $x = c$ и проходящему в точку $(0, 2n)$, соответствует путь, исходящий из точки $(0,0)$ и симметрично расположенный относительно прямой $x = c$, начиная с момента встречи пути с этой прямой, и приводящий в точку $(2c, 2n)$. Легко видеть, что число таких путей будет C_{2n}^{n-c} (число положительных перемещений равно $n + c$, а число отрицательных $n - c$). Таким образом, число путей, исходящих из точки $(0, 0)$, приходящих в точку $(0, 2n)$ и не достигающих прямой $x = c$, равно $C_{2n}^n - C_{2n}^{n-c}$. Этим, очевидно, доказана часть теоремы, относящаяся к распределению величины D_n^+.

Для определения вероятности события $nD_n < c$ необходимо подсчитать число путей, исходящих из точки $(0, 0)$, приводящих в точку $(0, 2n)$ и не достигающих границ $x = c$ и $x = -c$.

Подсчеты, немногим более сложные, чем предыдущие, показывают, что

$$P\{nD_n < c\} = 1 + 2 \sum_{k=1}^{\left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^n} = \sum_{k=-\left[\frac{n}{c}\right]}^{\left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^n}.$$

Очевидно, что это равенство доказывает теорему.

Мы можем теперь в рассматриваемом нами частном предположении доказать предельные результаты Н. В. Смирнова. Действительно, рассмотрим отношение

$$J_k = \frac{C_{2n}^{n-kc}}{C_{2n}^{n}} = \frac{(n!)^2}{(n-kc)!\,(n+kc)!}$$

при постоянном k. С помощью формулы Стирлинга находим, что

$$J_k = \left(1 - \frac{kc}{n}\right)^{-n+kc} \left(1 + \frac{kc}{n}\right)^{-n-kc} (1 + o\,(1)).$$

Далее, воспользовавшись разложением логарифма в ряд Маклорена и вспомнив равенство $c = [z\sqrt{2n}]$, получаем

$$\lg J_k = -\frac{k^2 z^2}{n} + o\,(1) = -2k^2 z^2 + o\,(1).$$

Таким образом,

$$J_k = e^{-2k^2 z^2} (1 + o\,(1)). \tag{3}$$

Это равенство доказывает соотношение (1).

Для доказательства соотношения (2) рассмотрим $z > 0$, $\varepsilon > 0$ и такое целое N, что

$$e^{-2N^2 z^2} < \frac{\varepsilon}{16}.$$

При этом также

$$\left| \sum_{|k| > N} (-1)^k e^{-2k^2 z^2} \right| < \frac{\varepsilon}{8}.$$

Так как

$$C_{2n}^{n-kc} > C_{2n}^{n-(k+1)c},$$

то

$$\left| \sum_{N < |k| \leqslant \left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^{n}} \right| \text{[redacted]}\, 4 \frac{C_{2n}^{n-Nc}}{C_{2n}^{n}}$$

и, следовательно, при достаточно больших n

$$\left| \sum_{N < |k| \leqslant \left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^{n}} \right| < 4e^{-2N^2 z^2} (1 + o\,(1)) < \frac{\varepsilon}{3}.$$

Таким образом,

$$\left| \sum_{N < |k| \leqslant \left[\frac{n}{c}\right]} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^{n}} - \sum_{N < |k|} (-1)^k e^{-2k^2 z^2} \right| < \frac{\varepsilon}{2}.$$

А так как, согласно (3), при достаточно больших n

$$\left| \sum_{|k| \leqslant N} (-1)^k e^{-2k^2 z^2} - \sum_{|k| \leqslant N} (-1)^k \frac{C_{2n}^{n-kc}}{C_{2n}^{n}} \right| < \frac{\varepsilon}{2},$$

то при каждом $z > 0$ и любом $\varepsilon > 0$ для достаточно больших n

$$\left| P\left\{ \sqrt{\frac{n}{2}} D_n < z \right\} - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2kz^2} \right| < \varepsilon,$$

ч. т. д.

Математический институт
Академии наук УССР и
Артемовский учительский институт

Поступило
2 VIII 1951

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] Н. В. Смирнов, Бюлл. Моск. ун-та, 2, в. 2 (1939). [2] Б. В. Гнеденко, Курс теории вероятностей, 1950.