Доклады Академии Наук СССР
1950. Том LXXIV, № 2

МАТЕМАТИКА

Н. В. СМИРНОВ

О ПОСТРОЕНИИ ДОВЕРИТЕЛЬНОЙ ОБЛАСТИ ДЛЯ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

(Представлено академиком А. Н. Колмогоровым 14 VII 1950)

Пусть произведено n независимых наблюдений случайной величины ξ, распределенной с непрерывной плотностью $f(x)$. Обычно практикуемый прием приближенной оценки неизвестной функции $f(x)$ по данным выборки на сегменте $[a, b]$ заключается в построении „гистограммы" частот $f_n^*(x)$. Подразделяя сегмент $[a, b]$ на s отрезков $\Delta_1, \Delta_2, \ldots, \Delta_s$ равной длины $h = \frac{b-a}{s}$ и подсчитав численности $m_1, m_2, \ldots, m_s$ наблюдений, попавших на последовательные отрезки Δ_k, полагают:

$$f_n^*(x) = \frac{m_k}{nh} \text{ при } x \in \Delta_k. \tag{1}$$

Эту эмпирическую кривую сравнивают с теоретической гистограммой $\overline{f}(x)$:

$$\overline{f}(x) = \frac{p_k}{h} = \frac{1}{h}\int\limits_{\Delta_k} f(x)\,dx. \tag{2}$$

Легко обнаруживается из элементарных соображений, что $f_n^*(x)$ сходится по вероятности к $\overline{f}(x)$ в каждой точке $x\,[a, b]$. Если при возрастании n надлежащим образом увеличивать и число s отрезков подразделения, то можно притти к более точным выводам, впервые полученным В. И. Гливенко ([1]). В настоящей заметке мы формулируем теоремы, которые при известных условиях могут быть использованы для более точной и надежной оценки плотности.

Теорема 1. *Если непрерывная плотность распределения $f(x)$ удовлетворяет на $[a, b]$ условиям:*

$$\operatorname{Min}\{f(x)\} = \mu > 0; \qquad a \leqslant x \leqslant b, \tag{A}$$

$$\int\limits_a^b f(x)\,dx = 1 - \alpha < 1, \tag{B}$$

и при возрастании n и s имеем:

$$\overline{\lim_{n\to\infty}}\ \frac{s^3(\ln s)^3}{n} < \infty, \tag{C}$$

то для любого $\lambda, -\infty < \lambda < \infty$,

$$P_n = P\left\{\operatorname{Max}\frac{|f_n^*(x) - f(x)|}{\sqrt{\overline{f}(x)}} \leqslant \frac{l_s + \dfrac{\lambda}{l_s}}{\sqrt{nh}}\right\} =$$

$$= \left[2\,\Phi\left(l_s + \frac{\lambda}{l_s}\right)\right]^s + O\left(\frac{\sqrt{\ln s}}{s}\right) + O\left(\frac{s^{3/2}}{\sqrt{n}}\right) = e^{-2e^{-\lambda}} + O\left(\frac{1}{\sqrt{\ln s}}\right) \quad (3)$$

и

$$P_n \to e^{-2e^{-\lambda}} \quad (n \to \infty); \quad (3^1)$$

при этом

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-x^2/2}\,dx$$

и l_s ***есть корень уравнения***

$$\frac{1}{2} - \Phi(x) = \frac{1}{s}. \quad (4)$$

В формулировке теоремы 1 теоретическую гистограмму $\overline{f}(x)$ нельзя заменить плотностью $f(x)$, так как (даже при наличии ограниченной производной $f'(x)$) систематическое расхождение между $\overline{f}(x)$ и $f(x)$ значительно превосходит случайную погрешность. Однако в качестве приближающей эмпирической функции берут также „полигон частот" и соответственно заменяют $\overline{f}(x)$ теоретическим полигоном.

Эти функции определяются следующим образом.

Пусть $x_k\,(k = 1, 2, \ldots, s)$ — середины промежутков Δ_k и a_k — правые концы этих промежутков; в каждом сегменте $[x_k, x_{k+1}]$ полагаем:

$$\varphi_n^*(x) = \frac{m_k + m_{k-1}}{2\,nh} + (x - a_k)\,\frac{(m_{k+1} - m_k)}{nh^2} \quad (5^1)$$

и

$$\varphi_s(x) = \frac{p_k + p_{k+1}}{2h} + (x - a_k)\,\frac{(p_{k+1} - p_k)}{h^2}, \quad (5^2)$$

так что

$$\varphi_n^*(x_k) = \frac{m_k}{nh};\ \varphi_n^*(x_{k+1}) = \frac{m_{k+1}}{nh};\ \varphi_s(x_k) = \frac{p_k}{h};\ \varphi_s(x_{k+1}) = \frac{p_{k+1}}{h}.$$

Пусть еще

$$\psi_s(x) = \frac{\sqrt{\dfrac{p_k}{h}} + \sqrt{\dfrac{p_{k+1}}{h}}}{2} + (x - a_k)\left(\sqrt{\frac{p_{k+1}}{h}} - \sqrt{\frac{p_k}{h}}\right), \quad (6)$$

$$\psi_s(x_k) = \sqrt{\frac{p_k}{h}};\ \psi_s(x_{k+1}) = \sqrt{\frac{p_{k+1}}{h}}. \quad (6^1)$$

Тогда имеет место

Теорема 2. *При условиях* (A), (B) *и* (C) *теоремы* 1 *для вероятности*

$$P_n' = \underline{P}\left\{\operatorname{Max}\frac{|\varphi_n^*(x) - \varphi_s(x)|}{\psi_s(x)};\ a + \frac{h}{2} \leqslant x \leqslant b - \frac{h}{2}\right\}$$

справедливы утверждения (3) *и* (3¹) *теоремы* 1.

Из теоремы 2 легко получается следующая

Теорема 3. *Если $f(x)$ имеет на $[a, b]$ ограниченную вторую производную и выполняются условия* (A), (B) *и* (C) *теоремы* 1 *и, кроме того,*

$$\frac{n \ln s}{s^5} \to 0 \quad (n \to \infty), \tag{D}$$

то

$$P\left\{ \operatorname{Max} \frac{|\varphi_n^*(x) - f(x)|}{\sqrt{f(x)}} \leqslant \frac{l_s + \frac{\lambda}{l_s}}{\sqrt{nh}} \right\} \xrightarrow[n \to \infty]{} e^{-2e^{-\lambda}}. \tag{7}$$

Условиям (C) и (D) можно удовлетворить, полагая, например, $s = n^\alpha$, $\frac{1}{5} < \alpha < \frac{1}{3}$.

Теорема 3 дает возможность строить доверительные области с заданным коэффициентом довери β для оценки неизвестной плотности $f(x)$ по данным выборки.

Определив при заданном β $(0 < \beta < 1)$ $\lambda = \lambda_\beta$, удовлетворяющее уравнению

$$e^{-2e^{-\lambda}} = \beta$$

(или, более точно, из уравнения $\left[2\Phi\left(l_s + \frac{\lambda}{l_s}\right)\right]^s = \beta$), находим

$$\frac{l_s + \lambda_\beta / l_s}{\sqrt{nh}} = t_\beta.$$

Неравенство

$$\operatorname{Max} \frac{|\varphi_n^*(x) - f(x)|}{\sqrt{f(x)}} \leqslant t_\beta$$

означает, что кривая $f(x)$ на всем сегменте $[a, b]$ не выходит за границы полосы $\mathfrak{A}_\beta$, ограниченной кривыми:

$$y_1(x) = \varphi_n^*(x) + \frac{t_\beta^2}{2} - t_\beta \sqrt{\varphi_n^*(x) + \frac{t_\beta^2}{4}}$$

и

$$y_2(x) = \varphi_n^*(x) + \frac{t_\beta^2}{2} + t_\beta \sqrt{\varphi_n^*(x) + \frac{t_\beta^2}{4}}.$$

Полоса $\mathfrak{A}_\beta$ покрывает кривую плотности $y = f(x)$ при больших n и s при условиях теоремы 3 с вероятностью, сколь угодно близкой к β.

Поступило
4 IV 1950

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] В. И. Гливенко, Курс теории вероятностей, 1939, стр. 180.