

Лекция 2

Выборка. Статистики. Эмпирические функции распределения.

Пусть X — СВ с ф.р. $F(x) = P\{X \leq x\}$,
 $x \in (-\infty, +\infty)$.

Определение 1. Совокупность $\{X_k, k=1, 2, \dots, n\}$ независимых СВ, имеющих одинаковые ф.р. $F_{X_k}(x) = F(x)$, называемая однородной выборкой объема n , соответствующей ф.р. $F(x)$ и обозначаемой Z_n :

$$Z_n = \{X_1, X_2, \dots, X_n\}.$$

СВ X_k ($k=1, 2, \dots, n$) называется k -м элементом выборки.

Определение 2. Реализацией выборки Z_n называется неслучайный вектор $z_n = \{x_1, \dots, x_n\}$, компонентами которого являются реализации СВ X_1, \dots, X_n .

Пусть $z_{(n)} = \{x_{(1)}, \dots, x_{(n)}\}$ — вектор компонент, которого является упорядоченные по возрастанию числа x_1, \dots, x_n , т.е. $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

Определение 3. СВ $X_{(k)}$, реализацией которой для каждой реализации z_n является число $x_{(k)}$, называемое k -й порядковой статистикой, $k=1, \dots, n$. При этом СВ $X_{(1)}$ и $X_{(n)}$ называемые экстремальными статистиками.

Если $\varphi(\dots)$ — некоторая заданная функция от n переменных, то любая случайная величина Y , определяемая как $Y = \varphi(X_1, X_2, \dots, X_n)$, называется статистикой.

Примеры статистик:

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \quad - \text{выборочное среднее}$$

$$\bar{S}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \quad - \text{выборочная дисперсия}$$

$$\bar{V}_r(n) = \frac{1}{n} \sum_{k=1}^n (X_k)^r, \quad r = 1, 2, \dots \quad - \text{выборочный начальный момент } r\text{-го порядка}$$

$$\bar{\mu}_r(n) = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^r, \quad r = 1, 2, \dots \quad - \text{выборочный центрированный момент } r\text{-го порядка}$$

Заметим, что $\bar{X}_n = \bar{V}_1(n)$, $\bar{S}_n^2 = \bar{\mu}_2(n)$.

Теорема 1 (без док-ва). Пусть распределение $F(x)$ таково, что существуют следующие теоретические моменты любого элемента X_k выборки Z_n :

$$m_x = M\{X_k\}, \quad v_r = M\{(X_k)^r\}, \quad \mu_r = M\{(X_k - m_x)^r\},$$

$r = 1, 2, \dots$. Тогда

$$\bar{V}_r(n) \xrightarrow[n \rightarrow \infty]{\text{н.н.}} v_r, \quad \bar{\mu}_r(n) \xrightarrow[n \rightarrow \infty]{\text{н.н.}} \mu_r.$$

Введем функцию-индикатор:

$$I(A) = \begin{cases} 1, & \text{если событие } A \text{ произошло} \\ 0, & \text{если } A \text{ не произошло} \end{cases}$$

Для каждого $x \in (-\infty, +\infty)$ определим случайную величину $F_n^*(x)$ по формуле

$$F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x).$$

Здесь $\sum_{i=1}^n I(X_i \leq x)$ есть случайное число элементов выборки, меньших x . Эту функцию будем называть эмпирической функцией распределения выборки. Каждая реализация выборки дает свою реализацию случайной функции $F_n^*(x)$ (вид этих реализаций см. в 3. об-суждаемое в лекции 1).

Нетрудно убедиться, что при любом фиксированном x случайная величина

$$\sum_{i=1}^n I(X_i \leq x)$$

имеет биномиальное распределение с параметрами n и $p = P\{X_i \leq x\} = F(x)$, так что

$$P\left\{F_n^*(x) = \frac{k}{n}\right\} = C_n^k F(x)^k (1-F(x))^{n-k},$$

где $k = 0, 1, \dots, n$.

Докажем следующую теорему 2, показывающую, что $F_n^*(x)$ с увеличением n сближается в каждой точке x с теоретической ф.р. $F(x)$.

Теорема 2 Для любого $x \in (-\infty; +\infty)$ и любого $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\{|F_n^*(x) - F(x)| < \varepsilon\} = 1, \text{ т.е. } F_n^*(x) \xrightarrow{P} F(x).$$

Док-во. Рассмотрим систему независимых случайных величин

$$\eta_i = \begin{cases} 1, & \text{если } X_i \leq x \\ 0, & \text{если } X_i > x \end{cases}, \quad i = 1, 2, \dots, n.$$

Каждая из η_i имеет один и тот же закон распределения

$$\begin{array}{c|cc} \eta_i & 1 & 0 \\ \hline P & p & 1-p \end{array}, \quad \begin{aligned} M\eta_i &= p \\ D\eta_i &= p(1-p) \end{aligned}$$

где $p = P\{X_i \leq x\} = F(x)$. Так что выполнены все условия ЗБЧ в форме Чебышёва (см. теорему Чебышёва в курсе ТВ) и

$$\lim_{n \rightarrow \infty} P\left\{\left|\frac{\eta_1 + \dots + \eta_n}{n} - p\right| < \varepsilon\right\} = 1.$$

Но $p = F(x)$, $\frac{\eta_1 + \dots + \eta_n}{n} = F_n^*(x)$.

Теорема 2 доказана.

Замечание. Отметим (без доказательства), что имеет место и более сильный результат:

$$F_n^*(x) \xrightarrow{n.н.} F(x).$$

Эти свойства говорят о равномерной сходимости $F_n^*(x)$ к $F(x)$ при $n \rightarrow \infty$. Оказывается, справедливо и более сильное свойство, отражающее факт равномерной сходимости $F_n^*(x)$ к $F(x)$.

Теорема 3 (Тимверко) (без док-ва)

Введем последовательность СВ

$$Y_n = \sup_{-\infty < x < +\infty} |F_n^*(x) - F(x)|, \quad n=1, 2, \dots$$

Тогда

$$Y_n \xrightarrow[n \rightarrow \infty]{н.н.} 0.$$

Таким образом, теорема 2, 3 говорят о том, что $F_n^*(x)$ при больших значениях n является хорошим приближением для $F(x)$.

В заключение, получим еще один результат, позволяющий в вероятностном смысле оценить возможность приближения для $F(x)$ с помощью $F_n^*(x)$. Применяем ЦПТ (см. курс ТВ) к введенной при доказательстве теоремы 2 последовательности независимых СВ η_1, η_2, \dots . По ЦПТ

если
то $V_n = \eta_1 + \dots + \eta_n$, $\tilde{V}_n = \frac{V_n - M\{V_n\}}{\sqrt{D\{V_n\}}}$,

$$\lim_{n \rightarrow \infty} P\{a \leq \tilde{V}_n \leq b\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

В нашем случае

$$V_n = \eta_1 + \dots + \eta_n = n F_n^*(x),$$

$$M\{V_n\} = np = n F(x),$$

$$D\{V_n\} = n \cdot p(1-p) = n F(x)(1-F(x)),$$

тогда

$$\lim_{n \rightarrow \infty} P\left\{a \leq \frac{n F_n^*(x) - n F(x)}{\sqrt{n F(x)(1-F(x))}} \leq b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

В частности,

$$\lim_{n \rightarrow \infty} P\left\{ \frac{\sqrt{n} |F_n^*(x) - F(x)|}{\sqrt{F(x)(1-F(x))}} \leq \varepsilon \right\} = 2\Phi(\varepsilon),$$

где $\Phi(\varepsilon) = \int_0^\varepsilon \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$ — функция Лапласа.