Теория вероятностей и математическая статистика

Лекция 14

Основные понятия выборочной теории

Основные определения

Математической статистикой называется раздел математики, изучающий законы распределения СВ по результатам экспериментов над ними.

Пусть над СВ X проведена серия независимых экспериментов, результаты которых называются ***экспериментальными*** (***статистическими***) ***данными***. Для их изучения выбраны результаты n экспериментов (***выборочный метод***). Эти n чисел образуют случайный вектор $\vec{X}_n = (X_1, X_2, \dots, X_n)$.

Случайной выборкой называется случайный вектор $\vec{X}_n$, координаты которого независимы в совокупности и каждая координата имеет такое же распределение, что и СВ X.

Множество всех возможных значений СВ X называется ***генеральной совокупностью***, а её функция распределения $F_X(x)$ – ***теоретической функцией распределения***.

Множество всех возможных значений случайного вектора $\vec{X}_n$ называется ***выборочным пространством***, каждое конкретное значение $\vec{x}_n = (x_1, \dots, x_n)$ – ***реализацией случайной выборки*** (***выборкой***), а число n – ***объёмом*** выборки.

Функция распределения случайной выборки:

$$F_{\vec{X}_n}(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n) = \prod_{i=1}^{n} P(X_i < x_i) = \prod_{i=1}^{n} F_X(x_i).$$

Вариационный ряд

Вариационным рядом выборки называется расположение элементов выборки в неубывающем порядке:

$$x_{(1)} \le x_{(2)} \le \ \dots \ \le x_{(n)}.$$

Вариационным рядом случайной выборки называется совокупность случайных величин $X_{(1)}, X_{(2)}, \dots, X_{(n)}$, где $X_{(i)}$ при каждой реализации принимает значение, равное i-му члену вариац. ряда выборки $x_{(i)}$.

Распределение крайних членов вариационного ряда:

$$P\big(X_{(1)} < x\big) = 1 - \big(1 - F(x)\big)^n, \qquad P\big(X_{(n)} < x\big) = \big(F(x)\big)^n.$$

В самом деле:

$$P\big(X_{(n)} < x\big) = P(X_1 < x, \dots, X_n < x) = \prod_{i=1}^{n} P(X_i < x) = \big(F(x)\big)^n,$$

$$P\big(X_{(1)} < x\big) = 1 - P\big(X_{(1)} \ge x\big) = 1 - P(X_1 \ge x, \dots, X_n \ge x) = 1 - \prod_{i=1}^{n} P(X_i \ge x) =$$

$$= 1 - \big(1 - F(x)\big)^n.$$

Статистическое распределение

Пусть в выборке $x_1, x_2, \ldots, x_n$ встречается m различных значений

$$z_1 < z_2 < \ \ldots \ < z_m,$$

каждое из которых встречается соответственно $n_1, n_2, \ldots, n_m$ раз. Тогда каждое значение z_i называется ***вариантой*** выборки, а число n_i – ***частотой*** варианты z_i.

Относительной частотой варианты z_i называется отношение её частоты к объёму выборки:

$$\omega_i = \frac{n_i}{n}.$$

Статистическим рядом (***распределением***) выборки называется перечень всех её вариант и их частот или относительных частот.

Статистический ряд графически изображается в виде ***полигона***.

Пример

Для выборки

7 5 2 5 7 7 5 7 7 7

составим её статистические распределения по частотам и относительным частотам:

Изобразим эти распределения в виде полигонов:

Выборочная функция распределения

Выборочной функцией распределения называется функция от числа x и случайной выборки $\vec{X}_n$

$$\hat{F}\left(x, \vec{X}_n\right) = \frac{n\left(x,\ \vec{X}_n\right)}{n},$$

где $n\left(x,\ \vec{X}_n\right)$ – число элементов случайной выборки, меньших x.

Для фиксированного числа x получаем случайную величину $\hat{F}\left(x, \vec{X}_n\right)$ со значениями 0, $\frac{1}{n}$, $\frac{2}{n}$, …, $\frac{n-1}{n}$, 1 и биномиальным распределением с вероятностью успеха $p = F(x)$.

Теорема: Для фиксированного x имеем $\hat{F}\left(x, \vec{X}_n\right) \xrightarrow{P} F(x)$ при $n \to +\infty$.

Доказательство: ЗБЧ в форме Бернулли $\square$

Эмпирическая функция распределения

Выборочной функцией распределения называется функция от числа x и случайной выборки $\vec{X}_n$

$$\hat{F}\left(x, \vec{X}_n\right) = \frac{n\left(x,\ \vec{X}_n\right)}{n},$$

где $n\left(x,\ \vec{X}_n\right)$ – число элементов случайной выборки, меньших x.

Для фиксированной реализации $\vec{x}_n$ случайной выборки получаем ***эмпирическую функцию распределения***

$$F_n(x) = \frac{n(x)}{n},$$

где $n(x)$ – число элементов выборки, меньших x.

Пример

Для выборки

7 5 2 5 7 7 5 7 7 7

найдём эмпирическую функцию распределения

Интервальное распределение

Отрезок $[x_{(1)}; x_{(n)}]$ разбивают на m промежутков $J_1, \dots, J_m$ с шагом Δ.

Интервальным статистическим рядом (***распределением***) выборки называется перечень всех промежутков, на которые разбит отрезок $[x_{(1)}; x_{(n)}]$, с указанием количества элементов выборки, попавших в каждый промежуток.

Эмпирической плотностью распределения выборки называется функция

$$p_n(x) = \begin{cases} \dfrac{n_i}{n \cdot \Delta}, & x \in J_i, \\ 0, & x \notin J_i. \end{cases}$$

График эмпирической плотности распределения называется ***гистограммой***.

По ЗБЧ в форме Бернулли $\dfrac{n_i(\vec{X}_n)}{n} \xrightarrow{P} P(X \in J_i) = \int_{J_i} p(x)\,dx = p(\tilde{x}_i) \cdot \Delta$, где $\tilde{x}_i \in J_i$.

При больших n имеем $\dfrac{n_i}{n} \approx p(\tilde{x}_i) \cdot \Delta$, откуда $\dfrac{n_i}{n \cdot \Delta} \approx p(\tilde{x}_i)$,
т. е. $p_n(x)$ – статистический аналог плотности распределения $p(x)$.

Пример

Для выборки с данным статистическим распределением

x	0	1	2	4	6	7	9	10	11
n	2	4	1	5	2	6	6	2	2

найдём интервальное распределение с шагом $\Delta = 2$ и с шагом $\Delta = 3$

построим графики эмпирической плотности распределения

Для определения оптимального числа промежутков используют ***эмпирическое правило Стёрджеса***: $m \approx \log_2 n + 1$.

Выборочные характеристики

Статистикой (***выборочной характеристикой***) называется любая функция от случайной выборки $\varphi(\vec{X}_n)$.

Выборочное среднее

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Выборочная дисперсия

$$\hat{\sigma}^2(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Выборочное среднеквадр. отклонение

$$\hat{\sigma}(\vec{X}_n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Выборочный начальный момент k-го порядка

$$\hat{\mu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Выборочный центральный момент k-го порядка

$$\hat{\nu}_k(\vec{X}_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$$

Легко проверяется, что

$$\hat{\mu}_1 = \bar{X}, \qquad \hat{\nu}_1 = 0, \qquad \hat{\nu}_2 = \hat{\sigma}^2 = \hat{\mu}_2 - (\hat{\mu}_1)^2.$$

Выборочные характеристики

Статистики для случайной выборки $\left(\vec{X}_n, \vec{Y}_n\right)$ объёма n из двумерной генеральной совокупности (X, Y).

Выборочный корреляционный момент:

$$\widehat{K}\left(\vec{X}_n, \vec{Y}_n\right) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Выборочный коэффициент корреляции:

$$\hat{\rho}\left(\vec{X}_n, \vec{Y}_n\right) = \frac{\widehat{K}\left(\vec{X}_n, \vec{Y}_n\right)}{\hat{\sigma}\left(\vec{X}_n\right) \cdot \hat{\sigma}\left(\vec{Y}_n\right)}.$$

Пример

Для выборки

$$7\ \ 5\ \ 2\ \ 5\ \ 7\ \ 7\ \ 5\ \ 7\ \ 7\ \ 7$$

найдём выборочные характеристики:

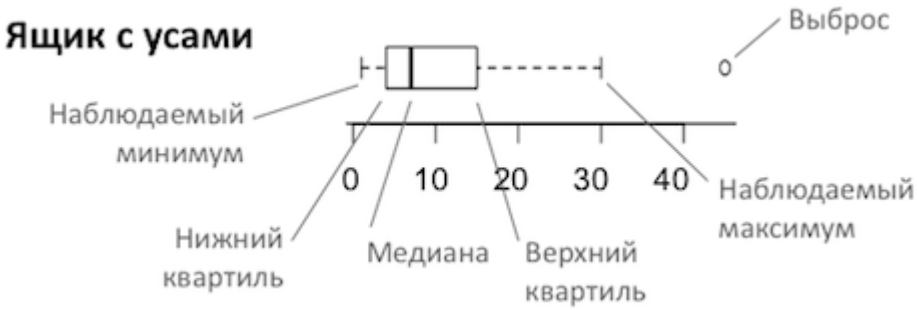
$$\bar{x} = \frac{7+5+2+5+7+7+5+7+7+7}{10} = 2 \cdot 0{,}1 + 5 \cdot 0{,}3 + 7 \cdot 0{,}6 = 5{,}9,$$

$$\hat{\sigma}^2 = (2-5{,}9)^2 \cdot 0{,}1 + (5-5{,}9)^2 \cdot 0{,}3 + (7-5{,}9)^2 \cdot 0{,}6 =$$

$$= 2^2 \cdot 0{,}1 + 5^2 \cdot 0{,}3 + 7^2 \cdot 0{,}6 - 5{,}9^2 = 2{,}49, \qquad \hat{\sigma} = \sqrt{2{,}49} \approx 1{,}578,$$

$$Q_{0{,}25} = 5, \qquad Q_{0{,}5} = m_e = 7, \qquad Q_{0{,}75} = 7.$$

Минимальную и максимальную варианты, а также квартили выборки часто изображают в виде ***«ящика с усами»***.



Пример вычислений в MS Excel