Раздел: Теория вероятностей и математическая статистика

Тема: ***Статистические оценки параметров распределения***

Лектор Рожкова С.В.

2021 г.

§15. Точечные статистические оценки параметров распределения

Статистическое распределение выборки дает первоначальное представление о закономерностях, имеющих место в генеральной совокупности.

Предположение о характере распределения приводит к необходимости определения параметров этого распределения, его числовых характеристик.

Статистические данные не позволяют найти параметры распределения точно, они позволяют их только *оценить*.

Существует два вида оценок – точечные и интервальные.

Выборочная характеристика

$$\Theta^* = f(x_1, x_2, \dots, x_n)$$

используемая для нахождения приближённого значения неизвестной генеральной характеристики Θ, называется её ***точечной статистической оценкой***.

$$\Theta \approx \Theta^*$$

Чтобы статистическая оценка давала хорошее приближение, она должна удовлетворять следующим требованиям:

1. **Несмещённость**: $M(\Theta^*) = \Theta$
2. **Эффективность:** Θ^* имеет наименьшую дисперсию среди других оценок Θ (при заданном объеме выборки).

3. Состоятельность: при увеличении объёма выборки Θ^* стремится по вероятности к Θ, т.е.

$$\lim_{n\to\infty} P(\left|\Theta^* - \Theta\right| < \varepsilon) = 1$$

а) Выборочная средняя:

x_i	x_1	x_2	$\dots$
n_i	n_1	n_2	$\dots$

$$\boxed{\bar{x} = \frac{\sum n_i x_i}{n}}$$

– оценка ***математического ожидания*** генеральной совокупности (не смещенная и состоятельная)

x_i	0	2	3	7
n_i	6	6	2	6

Объём выборки: $n = 20$

$$\bar{x} = \frac{0 \cdot 6 + 2 \cdot 6 + 3 \cdot 2 + 7 \cdot 6}{20} = 3$$

б) Выборочная дисперсия:

x_i	x_1	x_2	$\dots$
n_i	n_1	n_2	$\dots$

$$D_в = \frac{\sum n_i (x_i - \overline{x})^2}{n}$$

$$D_в = \overline{x^2} - (\overline{x})^2$$

– оценка ***дисперсии*** (смещенная)

x_i	0	2	3	7
n_i	6	6	2	6
$(x_i - \overline{x})^2$	9	1	0	16
$(x_i)^2$	0	4	9	49

$n = 20$

$\overline{x} = 3$

$$D_в = \frac{9 \cdot 6 + 1 \cdot 6 + 0 \cdot 2 + 16 \cdot 6}{20} = 7.8$$

$$D_в = \frac{0 \cdot 6 + 4 \cdot 6 + 9 \cdot 2 + 49 \cdot 6}{20} - 3^2 = 7.8$$

в) Исправленная выборочная дисперсия:

$$s^2 = \frac{n}{n-1} D_в = \frac{n}{n-1} \frac{\sum n_i (x_i - \overline{x})^2}{n} = \frac{\sum n_i (x_i - \overline{x})^2}{n-1}$$

г) Выборочное среднее квадратическое отклонение:

$$\sigma_в = \sqrt{D_в}$$

д) Исправленное выборочное среднее квадратическое отклонение:

$$s = \sqrt{s^2}$$

е) Мода

Для ДСВ: мода – наиболее вероятное значение СВ.

Для дискретного статистического ряда мода – наиболее часто встречающаяся варианта

Обозначается M_0.

x_i	$\boxed{0}$	1	2
n_i	$\boxed{5}$	2	3

$$M_0 = 0$$

x_i	2	$\boxed{3}$	7	9	$\boxed{14}$
n_i	5	$\boxed{8}$	7	5	$\boxed{8}$

$$M_0 = 3, \quad M_0 = 14$$

У случайной величины может быть несколько мод.

Как оценить моду, если выборка задана интервальным статистическим рядом?

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	10	15	25	15	5

Для непрерывной случайной величины мода – это значение, при котором плотность распределения $f(x)$ достигает максимума.

Гистограмма относительных частот даёт представление о плотности распределения генеральной совокупности.

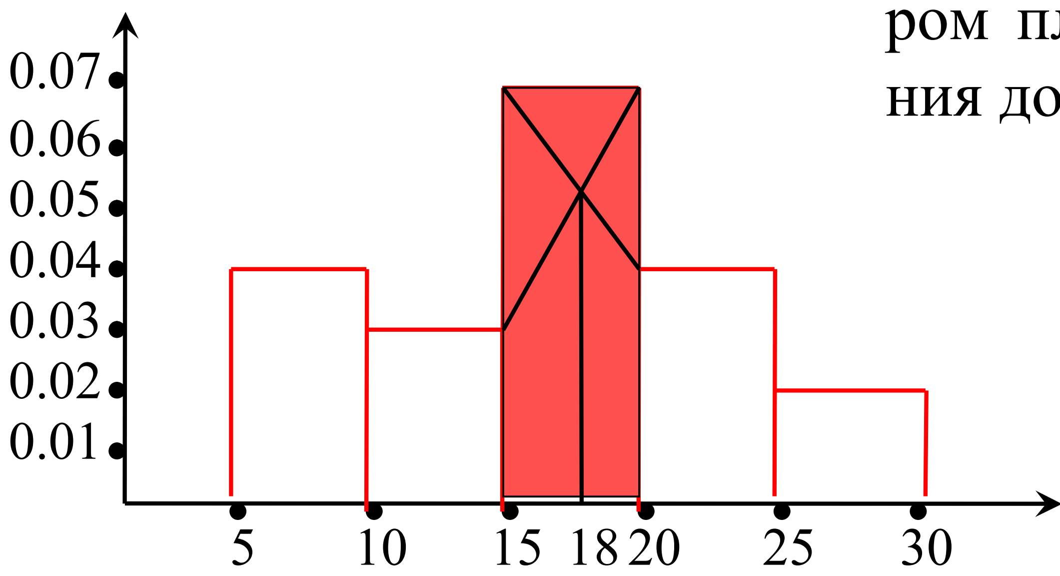
Построим гистограмму относительных частот.

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	20	15	35	20	10
w_i	0.2	0.15	0.35	0.2	0.1
w_i/h	0.04	0.03	0.07	0.04	0.02

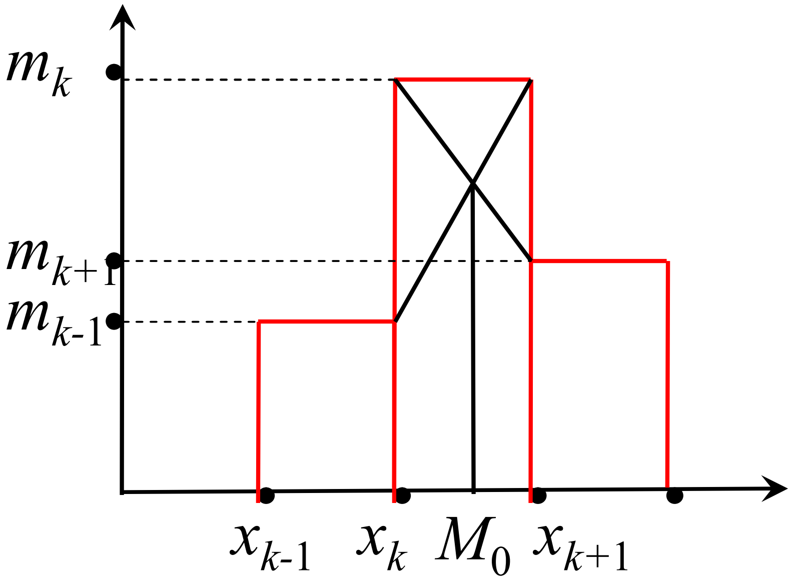
Объём выборки: $n = 100$

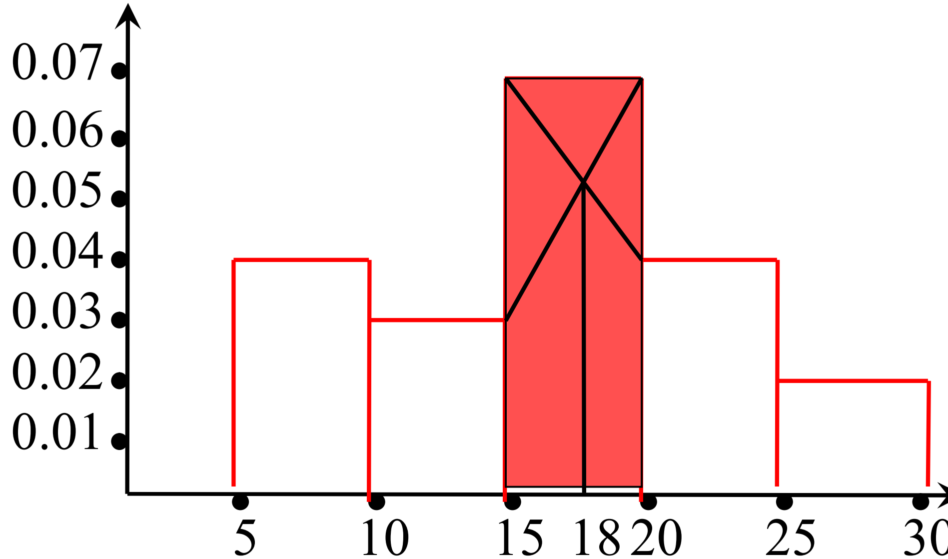
Длина интервала: $h = 5$

Мода – значение, при котором плотность распределения достигает максимума.



$M_0 = 18$





$$M_0 = x_k + \frac{m_k - m_{k-1}}{2m_k - (m_{k-1} + m_{k+1})}(x_{k+1} - x_k)$$

$$x_{k-1} = 10 \qquad x_k = 15 \qquad x_{k+1} = 20$$

$$m_{k-1} = 0.03 \qquad m_k = 0.07 \qquad m_{k+1} = 0.04$$

$$M_0 = 15 + \frac{0.07 - 0.03}{2 \cdot 0.07 - (0.03 + 0.04)}(20 - 15) \approx 17.857$$

Медиана

Медиана генеральной совокупности – такое число x, что

$$p(X < x) = p(X > x) = 0.5$$

Как оценить медиану генеральной совокупности?

– такое число M_e, что количество вариант, меньших M_e, равно количеству вариант, больших M_e

$$\underbrace{0,\ 0,\ 1,\ 2,\ 2,\ 2,}\ \boxed{4},\ \underbrace{5,\ 5,\ 5,\ 5,\ 6,\ 6} \qquad M_e = 4$$

$$\underbrace{0,\ 0,\ 1,\ 2,\ 2,\ 2,\ \boxed{3},}\ \underbrace{\boxed{4},\ 5,\ 5,\ 5,\ 5,\ 6,\ 6} \qquad M_e = ?$$

$$M_e = (3+4)/2 = 3.5$$

Если n – нечётное, то $M_e = x_{(n+1)/2}$ (средняя варианта).

Если n – чётное, то $M_e = (x_{n/2} + x_{(n/2)+1})/2$

x_i	(x_1, x_2)	(x_2, x_3)	$\ldots$
n_i	n_1	n_2	$\ldots$

n – объём выборки

h – длина интервала

Находим такое число l, что $\sum_{i=1}^{l} n_i \le n/2, \quad \sum_{i=1}^{l+1} n_i > n/2.$

Пусть $f = \sum_{i=1}^{l} n_i.$

$$\boxed{M_e = x_{l+1} + \frac{n/2 - f}{n_{l+1}} \cdot h}$$

x_i	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	20	15	35	20	10

$n/2 = 50$

$h = 5$

$35 < 50$

$70 > 50$

$l = 2 \qquad f = 35$

$x_{l+1} = x_3 = 15 \qquad n_{l+1} = n_3 = 35$

$$M_e = 15 + \frac{50 - 35}{35} \cdot 5 \approx 17.143$$

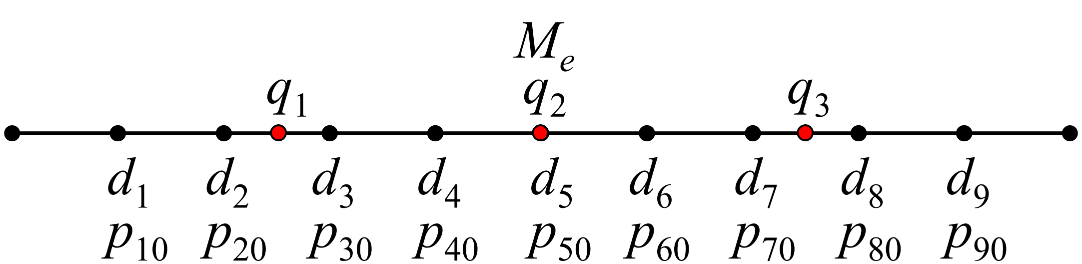
$$\underbrace{0,\ 0,\ 1,\ 2,\ 2,\ 2,}\ \boxed{4},\ \underbrace{4,\ 5,\ 5,\ 5,\ 5,\ 6} \qquad M_e = 4$$

Ряд наблюдений делится на 2 части, равные по количеству вариант.

Разделим ряд наблюдений на 4 равные части.
Получим три числа $q_1,\ q_2,\ q_3,$ которые оценивают, соответственно, первый, второй и третий ***квартили***.

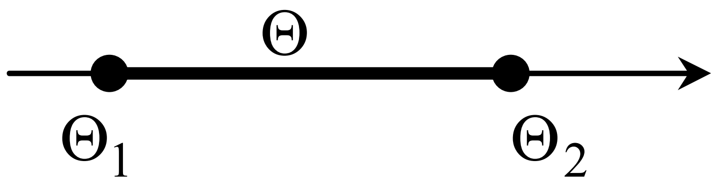
На 10 равных частей: $d_1, d_2, \ldots, d_9$ – ***децили***.

На 100 равных частей: $p_1, p_2, \ldots, p_{99}$ – ***процентили***.



§16. Интервальные статистические оценки параметров распределения

$\Theta \approx \Theta^*$ – точечная оценка



Интервальной называют оценку, которая определяется двумя числами – концами интервала:

$$\Theta \in (\Theta_1, \Theta_2)$$

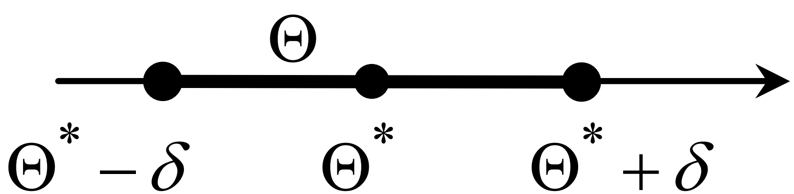
$$\Theta_1 = f_1(x_1, x_2, \ldots, x_n) \qquad \Theta_2 = f_2(x_1, x_2, \ldots, x_n)$$

– формулы для нахождения границ интервала по выборочным данным

Интервал $(\Theta_1;\ \Theta_2)$, который содержит в себе неизвестный параметр Θ с заданной вероятностью γ , называют ***доверительным интервалом***:

$$p(\Theta_1 < \Theta < \Theta_2) = \gamma$$

При этом вероятность γ называют ***доверительной вероятностью*** или ***надёжностью*** оценки.



$$(\Theta^* - \delta, \Theta^* + \delta)$$

$$p(\Theta^* - \delta < \Theta < \Theta^* + \delta) = p(-\delta < \Theta - \Theta^* < \delta) =$$

$$= p(|\Theta - \Theta^*| < \delta) = \gamma$$

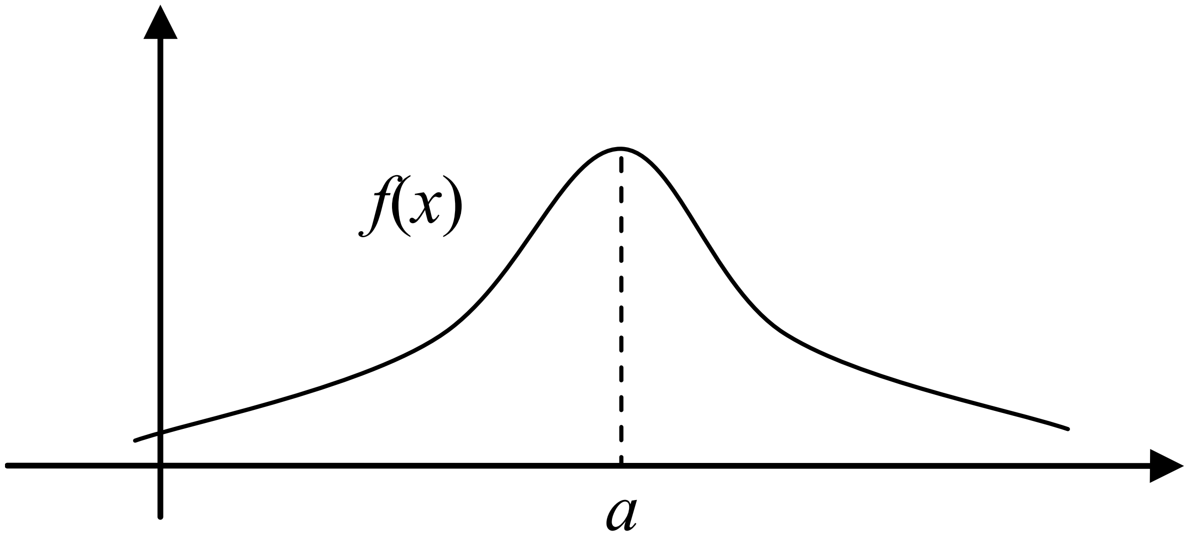
Число δ называют ***точностью*** оценки.

а) Нормальное распределение

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}$$

$$F(x) = \frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_0^x e^{\frac{-(x-a)^2}{2\sigma^2}} dx$$

a, σ – параметры распределения



$M(X) = a$ и $D(X) = \sigma^2$

Пусть генеральная совокупность имеет нормальное распределение

1) σ – известно. Оценить a.

Доверительным интервалом является интервал:

$$\left(\bar{x} - \frac{t_\gamma \cdot \sigma}{\sqrt{n}}, \bar{x} + \frac{t_\gamma \cdot \sigma}{\sqrt{n}} \right)$$

где t_γ определяется из условия $\Phi\left(t_\gamma\right) = \frac{\gamma}{2}$

Распределение Стьюдента

ОПРЕДЕЛЕНИЕ: Пусть $\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_n$ – независимые нормально распределенные СВ, для которых

$$M[\mathbf{X}_0] = M[\mathbf{X}_1] = \ldots = M[\mathbf{X}_n] = 0$$

$$D[\mathbf{X}_0] = D[\mathbf{X}_1] = \ldots = D[\mathbf{X}_n] = 1.$$

Тогда СВ

$$\mathbf{T} = \frac{\mathbf{X}_0}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(\mathbf{X}_k)^2}}$$

имеет распределение, называемое ***распределением Стьюдента с n степенями свободы***.

Плотность вероятностей СВ **T**, имеющей распределение Стьюдента:

$$S(t,n) = B_n \cdot \left[2 + \frac{t^2}{n}\right]^{-\frac{n+1}{2}}, \qquad B_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \cdot \Gamma\left(\frac{n}{2}\right)}$$

Числовые характеристика распределения Стьюдента:

1) M[**T**] = 0;

2) $D[\mathbf{T}] = \dfrac{n}{n-2}$ (n > 2)

$\Rightarrow$ Распределение Стьюдента определяется параметром n .

2) σ – неизвестно. Оценить a.

По данным выборки построим СВ **T**: $\mathbf{T} = \dfrac{(\overline{\mathbf{X}} - a)\sqrt{n}}{S}$

где СВ $\overline{X}$ – выборочная средняя, СВ S – «исправленное» среднее квадратическое отклонение

Тогда СВ **T** имеет распределение Стьюдента с k = n – 1 степенями свободы .

Учитывая четность функции S(t,k) получим:

$$P\left(\left| \frac{(\overline{\mathbf{X}} - a)\sqrt{n}}{S} \right| < t_\gamma \right) = 2\int_0^{t_\gamma} S(t,k)dt = \gamma$$

$$\Rightarrow \quad P\left(\overline{\mathbf{X}} - \frac{t_\gamma S}{\sqrt{n}} < a < \overline{\mathbf{X}} + \frac{t_\gamma S}{\sqrt{n}} \right) = 2\int_0^{t_\gamma} S(t,k)dt = \gamma$$

Доверительным интервалом является интервал:

$$\left(\bar{x} - \frac{t_\gamma \cdot s}{\sqrt{n}}, \bar{x} + \frac{t_\gamma \cdot s}{\sqrt{n}} \right)$$

где t_γ определяется из условия $\int_0^{t_\gamma} S(t,k)dt = \frac{\gamma}{2}$

s – исправленное выборочное среднее квадратическое отклонение