



Математическая
статистика

1-лекция

Постановка статистической задачи

1. Выборка. Статистическая структура (Модель)
2. Эмпирическое распределение. Вариационный ряд.
3. Выборочные моменты и другие характеристики.
4. Графическое изображение статистических данных

1. Выборка. Статистическая структура (Модель)



МС-прикладная дисциплина, родственная с ТВ, опирающаяся на понятия и факты (утверждения) ТВ, решает свои специфические задачи своими методами.

МС делает выводы на основе числовых данных полученных в результате случайных экспериментов.

(Ω, A, P) -вероятностная тройка, где (Ω, A) -измеримое пространство

$$\xi\text{-случайная величина} \Rightarrow \xi: (\Omega, A) \rightarrow (X, B)$$

X -пространство (множество) всевозможных значений ξ .

$B = \sigma(X)$ -борелевская сигма-алгебра (сигма алгебра борелевских множеств).

1. Выборка. Статистическая структура (Модель)



$\{P^\xi(B) = P(\omega: \xi(\omega) \in B), B \in \mathcal{B}\}$ - вероятностное распределение, порожденное с.в. ξ .

$(\mathcal{X}, \mathcal{B}, P^\xi)$ - тройка порожденная с.в. ξ , вероятностное пространство.

Предполагаем, что P^ξ образует класс:

$\{P^\xi\} = \mathcal{P}$ -семейство вероятностных распределений с.в. ξ .

Оказывается, мы должны работать в следующей тройке: $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ -семейство вероятностных пространств (случайный эксперимент).

1. Выборка. Статистическая структура (Модель)



В некотором эксперименте наблюдаем с.в. ξ :

$\xi = x$ -реализация с.в. ξ .

x_1, x_2, \dots, x_n -с.в. означающие реализации с.в. ξ в n независимых экспериментах. $x_i \sim P^\xi$ (имеют тоже самое распределение P^ξ)

$X^{(n)} = (x_1, x_2, \dots, x_n)$ -случайный вектор со случайными координатами в n -мерном пространстве.

$(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{P}^{(n)})$ -случайный эксперимент порожденный выборкой $X^{(n)}$, где $\mathcal{X}^{(n)} = \mathcal{X} \times \mathcal{X} \times \dots \times \mathcal{X}$, $\mathcal{B}^{(n)} = \sigma(\mathcal{X}^{(n)})$,

$\mathcal{P}^{(n)} = \{P^{X^{(n)}}\}$, $\mathcal{P}^{(n)}$ -семейство, порожденное выборкой $X^{(n)}$.

1. Выборка. Статистическая структура (Модель)



$$\begin{aligned} P^{X^{(n)}}(B^{(n)}) &= P(\omega: x_1 \in B_1, \dots, x_n \in B_n) = \\ &= \prod_{k=1}^n P(\omega: x_k \in B_k) = \prod_{k=1}^n P^\xi(B_k) \end{aligned}$$

где $B^{(n)} = B_1 \times B_2 \times \dots \times B_n$, $B_i \in \mathcal{B} \Rightarrow B^{(n)} \in \mathcal{B}^{(n)}$.

Тройка $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{P})$ - называется статистической структурой (модель).

2. Эмпирическое распределение. Вариационный ряд.



$$\mathcal{P} = \{P^\xi\} \Rightarrow (\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{P})$$

P^ξ -неизвестное распределение;

$$X^{(n)} = (x_1, x_2, \dots, x_n) \quad P^\xi(B) = P(\xi \in B);$$

Соберем ср.ариф. индикаторов тех событий которые попали в $B \Rightarrow$ возможным приближением для $P^\xi(B)$ является след:

$$P_n^\xi(B) = \frac{1}{n} \sum_{k=1}^n I_{X_k}(B), \quad (1)$$

$$\text{где } I_{X_k}(B) = \begin{cases} 1, & \text{если } x_k \in B \\ 0, & \text{если } x_k \notin B \end{cases}$$

2. Эмпирическое распределение. Вариационный ряд.



(1)-эмпирическое –дискретное распределение ξ (выборки):

- $P_n^\xi(\emptyset) = \frac{0}{n} = 0$
- $P_n^\xi(\overline{B}) = 1 - P_n^\xi(B)$
- $P_n^\xi(\mathcal{X}) = \frac{n}{n} = 1$, играет роль достовер.события
- $B_1 \subseteq B_2: P_n^\xi(B_1) \leq P_n^\xi(B_2)$

$$\sum I_{X_k}(B_1) \leq \sum I_{X_k}(B_2)$$

Если $B = (-\infty; x]$, $x \in R^1 \Rightarrow$

$$\Rightarrow P^\xi(B) = P(\xi \in (-\infty; x]) = P(\xi \leq x) = F(x), \quad x \in R^1$$

2. Эмпирическое распределение. Вариационный ряд.



Тогда $P_n^\xi(B) = \frac{1}{n} \sum_{k=1}^n I(x_k \leq x) = F_n(x)$, $x \in R^1$, где

$F_n(x)$ – эмпирическая функция распределения.

Насколько $F_n(x)$ близко к $F(x)$?

Свойства эмпирической функции распределения

$$1. MF_n(x) = \frac{1}{n} n M I(x_k \leq x) = 1 \cdot F(x) + 0(1 - F(x)) = F(x) \Rightarrow$$

$$M(F_n(x) - F(x)) = 0, x \in R^1 \quad (2)$$

2. Эмпирическое распределение. Вариационный ряд.



$$\begin{aligned} 2. \quad DF_n(x) &= \frac{1}{n^2} \cdot n \cdot DI(x_k \leq x) = \frac{1}{n} (M I(x_k \leq x) - F^2(x)) = \\ &= \frac{1}{n} F(x)(1 - F(x)) \leq \frac{1}{4n} \text{ Место для уравнения.} \end{aligned}$$

$$\max_{0 \leq a \leq 1} [a(1 - a)] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

По неравенству Чебышева, $\varepsilon > 0$:

$$P(|F_n(x) - F(x)| \geq \varepsilon) \leq \frac{DF_n(x)}{\varepsilon^2} = \frac{\sigma_F^2(x)}{\varepsilon^2} \leq \frac{1}{4n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0$$

$$\Rightarrow F_n(x) \xrightarrow{p} F(x) \text{ по теореме Бернулли} \quad (3)$$

3. Выборочные моменты и другие характеристики.



Выборочные моменты - моменты, вычисляемые с помощью эмпирического распределения.

Моменты k -го порядка:

$$v_{nk} = \int x^k dF_n(x) = \frac{1}{n} \sum_{i=1}^k x_i^k$$

$$v_{n1} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i - \text{сред. ариф.}$$

Центральный момент k -го порядка:

$$\mu_{nk} = \int (x - \bar{x})^k dF_n(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

$$\mu_{n2} = S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \text{дисперсия выборки}$$

$$S^2 = v_{n2} - v_{n1}^2 = \overline{x^2} - (\bar{x})^2 \geq 0$$

3. Выборочные моменты и другие характеристики.



Абсолютный момент k -го порядка:

$$\delta_{nk} = \int |x|^k dF_n(x) = \frac{1}{n} \sum_{i=1}^k |x_i|^k$$

Абсолютный центральный момент k -го порядка:

$$d_{nk} = \int |x - \bar{x}|^k dF_n(x) = \frac{1}{n} \sum_{i=1}^k |x_i - \bar{x}|^k$$

Медиана – середина вариационного ряда: $x_{Me} = Me$

$$Me = \begin{cases} x_{(i)}, & \text{если } m - \text{нечет.} \\ \frac{x_{(i)} + x_{(i+1)}}{2}, & \text{если } m - \text{чет.} \end{cases}$$

3. Выборочные моменты и другие характеристики.



Мода – варианта x_i с максимальной частотой n_i $x_{M_0} = M_0$

Квантиль уровня p ($0 < p < 1$)

$$\zeta_{(p)n} = \inf\{x: F_n(x) \geq p\}$$

$$\zeta_{\left(\frac{1}{2}\right)n} = \text{Me-медиана,}$$

$$\zeta_{\left(\frac{1}{4}\right)n} = \text{квартиль}$$

Вариация (изменчивость) - $V = \frac{S}{\bar{x}} \cdot 100\%$, где $S = \sqrt{S^2}$.

Чем ближе V к 0, тем хорошо.

Если вариация большая, то можно подозревать, что выборка неоднородная (специально подобранная)

4. Графическое изображение статистических данных.



Статистическое распределение изображается графически с помощью полигона и гистограммы.

Определение. *Полигоном частот* называют ломаную, отрезки которой соединяют точки с координатами (x_i, n_i) ; *полигоном отн.частот*

– с координатами (x_i, p_i^*) , где $p_i^* = \frac{n_i}{n}$, $i = \overline{1, m}$.

Полигон служит для изображения дискретного статистического ряда. Полигон отн.частот является аналогом многоугольника распределения дискретной случайной величины в теории вероятностей.

Определение. *Гистограммой частот* называют ступенчатую фигуру, состоящую из прямоугольников, основания которых расположены на оси Ox и длины их равны длинам частичных интервалов (h) , а высоты равны отношению:

$$\frac{n_i}{h} \text{ - для гистограммы частот; } \frac{n_i}{n \cdot h} \text{ - для гистограммы отн.частот.}$$

4. Графическое изображение статистических данных.



Пример 1. Дана выборка значений случайной величины X объема 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12

18, 17, 15, 13, 17, 14, 14, 13, 14, 16

Требуется: - построить дискретный вариационный ряд;
- найти размах варьирования R , моду M_0 , медиану M_e ;
- построить полигон отн.частот.

1) Ранжируем выборку : 12, 12, 13, 13, 13, 14, 14, 14, 14, 14,
15, 15, 16, 16, 17, 17, 17, 18, 18, 19.

2) Находим частоты вариантов и строим дискретный вариационный ряд (табл.1)

4. Графическое изображение статистических данных.

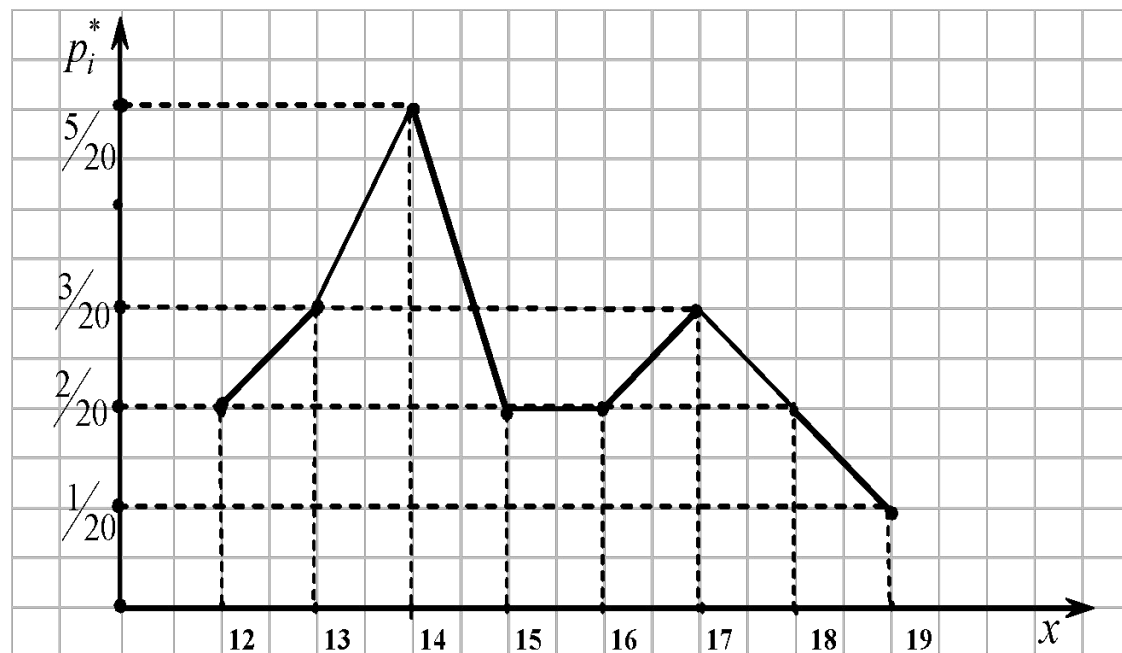


Значения вариантов	12	13	14	15	16	17	18	19
Частоты	2	3	5	2	2	3	2	1
Отн. частоты	$2/20$	$3/20$	$5/20$	$2/20$	$2/20$	$3/20$	$2/20$	$1/20$

3) По результатам таблицы находим:

$$R=19-12, M_0=14, M_e=(14+15)/2$$

4) Строим полигон отн.частот.



4. Графическое изображение статистических данных.



Пример 2. Результаты измерений отклонений от нормы диаметров 50 подшипников дали численные значения, приведенные в табл.

-0,158	1,701	0,634	0,720	0,490
1,531	-0,433	1,409	1,740	-0,266
-0,058	0,248	-0,095	-1,488	-0,361
0,415	-1,382	0,129	-0,361	-0,087
-0,329	0,086	0,130	-0,244	-0,882
0,318	-1,087	0,899	1,028	-1,304
0,349	-0,293	0,105	-0,056	0,757
-0,059	-0,539	-0,078	0,229	0,194
0,123	0,318	0,367	-0,992	0,529

Для данной выборки: - построить интервальный вариационный ряд;
- построить гистограмму и полигон отн.частот.

1. Строим интервальный ряд.

По данным таблицы 4 определяем: $x_{\min} = -1.76$; $x_{\max} = 1.74$

Для определения длины интервала h используем формулу Стерджеса:

$$h = (x_{\max} - x_{\min}) / (1 + 3.322 \lg n).$$

Число интервалов $m = 1 + 3.322 \lg 50$ (имеются также и формулы: $m = 5 \lg n$ или $m = \text{Ent}(\sqrt{n})$).

4. Графическое изображение статистических данных.



Примем $h=0,6$, $m=7$. За начало первого интервала примем величину

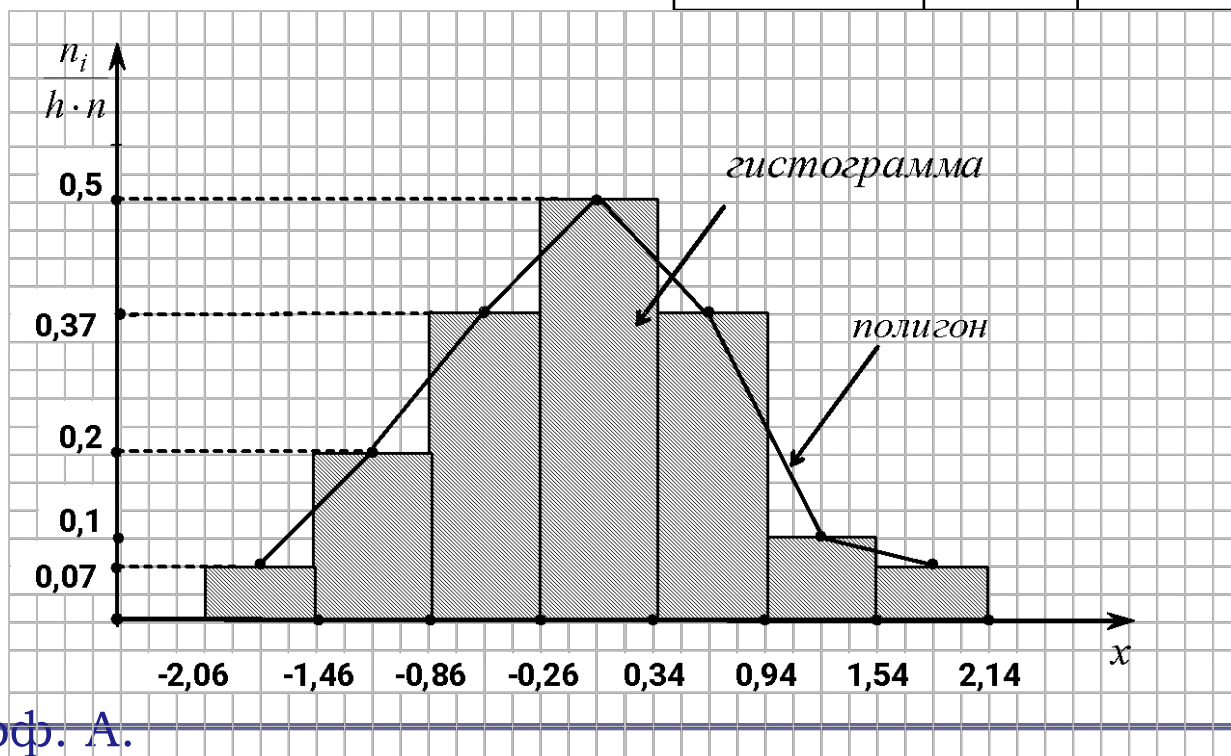
$$x_{\text{нач}} = x_{\text{min}} - h/2 = -2.06.$$

Конец последнего интервала должен удовлетворять условию: $x_{\text{кон}} - h \leq x_{\text{max}} < x_{\text{кон}}$.

Действительно, $2.14 - 0.6 \leq 1.17 < 2.14$;

Строим интервальный ряд

Частоты	2	6	11	15	11	3	2
Отн.частоты	2/50	6/50	11/50	15/50	11/50	3/50	2/50



Вершинами полигона являются середины верхних оснований прямоугольников гистограммы.

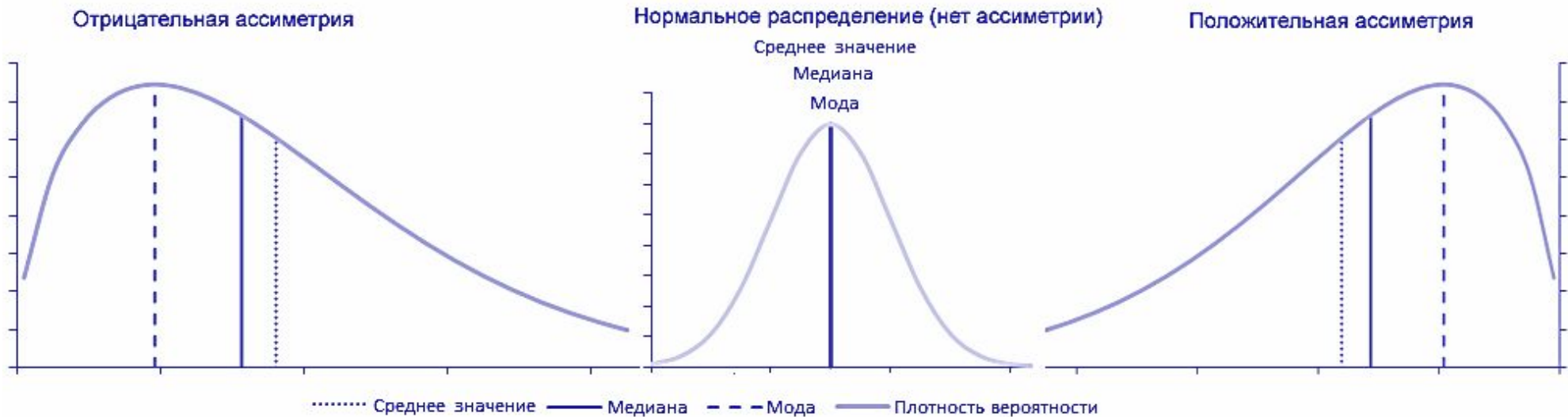
4. Графическое изображение статистических данных.



$$\text{Асимметрия} - A_s = \frac{\mu_{3n}}{S^3} \quad \mu_{3n} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

S^3 —желательно (рекомендовано) вычислять так : $S^3 = S^2 \cdot S$.

Идеальная симметрическая плотность- нормальная.



4. Графическое изображение статистических данных.



$$\text{Эксцесс} - E = \frac{\mu_{4n}}{S^4} - 3 \quad \mu_{4n} = \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^4, \quad S^4 = S^2 \cdot S^2$$

$E < 0$ – плосковершинный график

$E = 0$ – нормальный (стандартный) график

$E > 0$ – узковершинный график

