

# Содержание

<b>Лекция 1.</b>	<b>2</b>
Выборки . . . . .	2
Выборочные характеристики . . . . .	2
Начальная обработка статданных . . . . .	3
Геометрическая интерпретация данных . . . . .	4
<b>Лекция 2.</b>	<b>6</b>
Точечная оценка . . . . .	6
Свойство точечных оценок . . . . .	6
Точечные оценки моментов . . . . .	6
Метод моментов (Пирсона) . . . . .	8

## Лекция 1.

Теория вероятности изучает характеристику случайных величин, тогда как математическая статистика решает обратную задачу

Допустим, что у нас есть случайная величина, по ней мы можем найти математическое ожидание, моменты и оценить, какое распределение имеет случайная величина.

### Выборки

**Def. Выборка** - набор данных, полученных в ходе экспериментов. Тогда количество экспериментов  $n$  - объем Выборки

**Def. Генеральной совокупностью** называются все результаты проведенных экспериментов

**Def. Выборочной совокупностью** называются наблюдаемые данные экспериментов

Не все данные экспериментов мы можем наблюдать, например, выборы, тогда опросы голосовавших - выборочная совокупность, а результаты выборов - генеральная. Очевидно, что выборочная и генеральная совокупности могут иметь различные распределения.

**Def.** Выборка называется **репрезентативной**, если ее распределение близко к распределению генеральной совокупностью

Пример - **ошибка выжившего**. Во время Второй Мировой стал вопрос, в каких местах стоит бронировать корпус самолета. Самолеты возвращались с пулевыми отверстиями, и интуитивно казалось, что стоит бронировать те места, которые больше всего пострадали. Однако не были учтены те самолеты, которые не вернулись, а те, которые выжили, выжили благодаря тому, что были прострелены в нелетальных местах, поэтому было принято решение бронировать фюзеляж в менее пострадавших местах

В дальнейшем считаем, что все выборки репрезентативны

**Def. 1.** Выборкой объема  $n$  называется набор из  $n$  экспериментальных данных  $\vec{X} = (x_1, x_2, \dots, x_n)$  (апостериорное определение)

**Def. 2.** Выборкой объема  $n$  называется набор из  $n$  независимых одинаково распределенных случайных величин  $\vec{X} = (X_1, X_2, \dots, X_n)$  (априорное определение)

### Выборочные характеристики

Можно выборку рассматривать как дискретную случайную величину с одинаковыми вероятностями  $p_i = \frac{1}{n}$  и вычислить для нее математическое ожидание, дисперсию и функцию распределения

**Def.** Выборочным средним  $\bar{x}$  называется величина  $\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$

**Def.** Выборочной дисперсией  $D^*$  называется величина  $D^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2$  (или  $D^* = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{x}^2$ )

По закону больших чисел выборочное среднее будет сходиться к матожиданию

**Def.** Исправленной дисперсией называется величина  $S^2 = \frac{n}{n-1} D^* = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})^2$

**Def.** Выборочной функцией распределения  $F^*(x)$  называется функция  $F^*(x) = \frac{\text{число данных } x_i < x}{n}$

**Th.** Выборочная функция распределения поточечно сходится к теоретической функции распределения:

$$\forall y \in \mathbb{R} F^*(y) \xrightarrow{p} F(y)$$

$$F(y) = P(X < y)$$

$$F_y^* = \frac{1}{n} \sum_{i=1}^n I(X_i < y) \xrightarrow[\text{по ЗБЧ}]{p} EI(X_i < y) = P(X_i < y) = P(X_1 < y) = F_{X_1}(y)$$

Усилим теорему

**Th. Гливенко-Кантелли.**  $\sup_{x \in \mathbb{R}} |F^*(x) - F(x)| \xrightarrow{p} 0$

**Th. Колмогорова.**  $\sqrt{n} \sup_{x \in \mathbb{R}} |F^*(x) - F(x)| \rightrightarrows K$  - распределение Колмогорова с функцией распределения  $F_K(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 x^2}$ ,  $x \in [0; \infty)$

## Начальная обработка статданных

1. Ранжирование данных - упорядочиваем выборки по возрастанию. В результате получаем вариационный ряд  $\vec{X} = (X_{(1)}, X_{(2)}, \dots, X_{(n)})$

$$X_{(1)} = \min X_i; \quad X_{(n)} = \max X_i$$

$X_{(i)}$  -  $i$ -ая порядковая статистика

2. Объединим повторяющиеся данные - получаем т.н. частотный вариационный ряд

$X_i$	$X_{(1)}$	$\dots$	$X_{(r)}$	$\sum$
$n_i$	$n_1$	$\dots$	$n_r$	$n$

Иногда часть данных отбрасывается сверху и снизу (по 5, по 10, по 5% и так далее), чтобы сделать выборку репрезентативной

$$\text{Тогда } \bar{x} = \frac{1}{n} \sum X_i n_i, \quad D^* = \frac{1}{n} \sum (X_i - \bar{x})^2 n_i$$

3. Чтобы уменьшить количество вычислений или сделать гистограмму, делают интервальный вариационный ряд: разбиваем данные на интервалы и считаем, сколько данных  $n_i$  попало в интервал.

Тогда  $n_i$  - частота интервала  $A_i$

Есть два основных способа разбиения на интервалы:

- (а) Интервалы одинаковой длины
- (б) Равнонаполненные интервалы (в каждом интервале примерно одинаковое количество данных)

Число интервалов  $K$  такое, что  $\frac{K(n)}{n} \rightarrow 0$  и  $K(n) \xrightarrow{n \rightarrow \infty} 0$

Обычно применяют формулу Стерджесса  $K \approx 1 + \log_2 n$  или  $K \approx \sqrt[3]{n}$

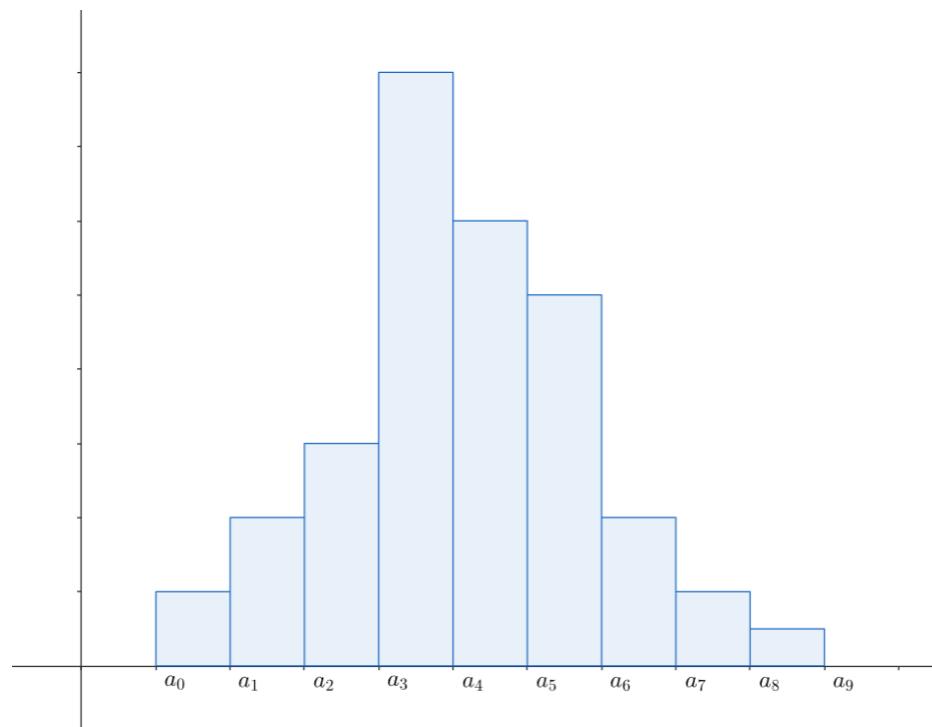
Пусть получили интервальный вариационный ряд

интервалы	$[a_0; a_1)$	$[a_1; a_2)$	$\dots$	$[a_{K-1}; a_K]$	$\sum$
частоты	$n_1$	$n_2$	$\dots$	$n_K$	$n$

## Геометрическая интерпретация данных

- Гистограмма

Строится ступенчатая фигура из прямоугольников, основание  $i$ -ого прямоугольника - интервал, высота прямоугольника -  $\frac{n_i}{nl_i}$ , где  $l_i$  - длина интервала

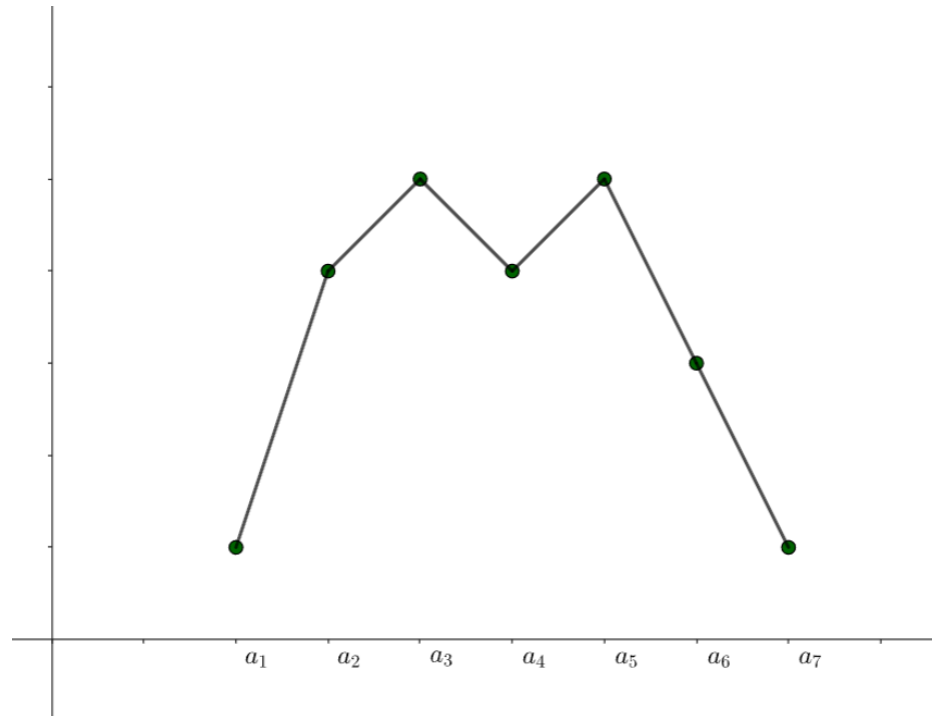


Визуально можно сделать гипотезу, как ведет себя распределение.

**Th.** Гистограмма поточечно сходится к теоретической плотности

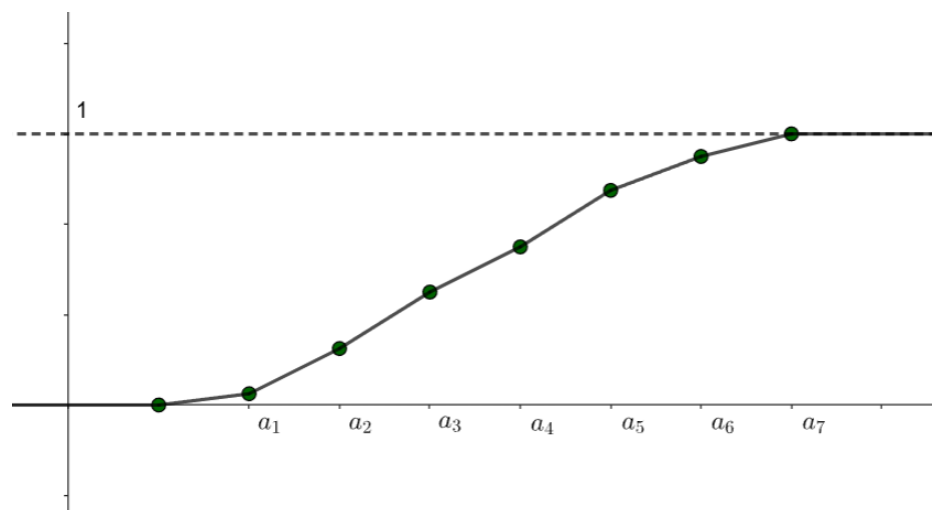
## • Полигон

На оси абсцисс отмечаем значения частотного вариационного ряда, по оси ординат - их частоты. Получившиеся точки соединяем отрезками



## • Выборочная функция распределения

На основе таблицы строится график функции распределения



Она может быть ступенчатой, ломаной или соединена по усмотрению

## Лекция 2.

### Точечная оценка

Пусть имеется выборка  $\vec{X} = (X_1, X_2, \dots, X_n)$  объемом  $n$

Пусть требуется найти приближенную оценку  $\theta^*$  неизвестного параметра  $\theta$

Находим ее при помощи некоторой функции обработки данных  $\theta^* = \theta^*(X_1, \dots, X_n)$

**Def.** Такая функция называется статистикой

**Def.** А оценка  $\theta^*$  называется точечной оценкой

### Свойство точечных оценок

#### 1. Состоятельность

**Def.** Статистика  $\theta^* = \theta^*(X_1, \dots, X_n)$  неизвестного параметра называется состоятельной, если  $\theta^* \xrightarrow{p} \theta$  при  $n \rightarrow \infty$

#### 2. Несмещенность

**Def.** Оценка  $\theta^*$  параметра  $\theta$  называется несмещенной, если математическое ожидание  $E\theta^* = \theta$

*Nota.* Оценка  $\theta^*$  называется асимптотически несмещенной, если  $E\theta^* \xrightarrow{p} \theta$  при  $n \rightarrow \infty$

#### 3. Эффективность

**Def.** Оценка  $\theta_1^*$  не хуже  $\theta_2^*$ , если  $E(\theta_1^* - \theta)^2 \leq E(\theta_2^* - \theta)^2$ . Или, если  $\theta_1^*$  и  $\theta_2^*$  несмещенные, то  $D\theta_1^* \leq D\theta_2^*$

**Def.** Оценка  $\theta^*$  называется эффективной, если она не хуже всех остальных оценок

*Nota.* Не существует эффективной оценки в классе всех возможных оценок

**Th.** В классе несмещенных оценок существует эффективная оценка

#### 4. Асимптотическая нормальность

**Def.** Оценка  $\theta^*$  параметра  $\theta$  называется асимптотически нормальной, если  $\sqrt{n}(\theta^* - \theta) \Rightarrow N(0, \sigma^2(\theta))$  при  $n \rightarrow \infty$

### Точечные оценки моментов

**Def.** Выборочным средним  $\bar{x}$  называется величина  $\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$

**Def.** Выборочной дисперсией  $D^*$  называется величина  $D^* = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2$

**Def.** Исправленной дисперсией  $S^2$  называется величина  $S^2 = \frac{n}{n-1} D^* = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})^2$

**Def.** Выборочным средним квадратическим отклонением называется величина  $\sigma^* = \sqrt{D^*}$

**Def.** Исправленным средним квадратическим отклонением называется величина  $S = \sqrt{S^2}$

**Def.** Выборочным  $k$ -ым моментом называется величина  $\bar{x}^k = \frac{1}{n} \sum_{i=1}^n X_i^k$

**Def.** Модой  $Mo^*$  называется варианта  $x_k$  с наибольшей частотой  $n_k = \max_i (n_1, n_2, \dots, n_m)$

**Def.** Выборочной медианой  $Me^*$  называется варианта  $x_i$  в середине вариационного ряда

$$\begin{cases} Me^* = X_{(k)}, & \text{если } n = 2k - 1 \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & \text{если } n = 2k \end{cases}$$

**Th.**  $\bar{x}$  - состоятельная несмещенная оценка теоретического матожидания  $EX = a$

1)  $E\bar{x} = a$

2)  $\bar{x} \xrightarrow{p} a$  при  $n \rightarrow \infty$

1)  $E\bar{x} = E\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \sum_{i=1}^n EX_i = \frac{1}{n} n EX_1 = EX_1 = a$

2)  $\bar{x} = \frac{\bar{x}_1 + \dots + \bar{x}_n}{n} \xrightarrow{p} a$  согласно Закону Больших Чисел

*Nota.* Если второй момент конечен, то  $\bar{x}$  - асимптотически нормальная оценка. По ЦПТ  $\frac{S_n - nEX_1}{\sqrt{n} \sqrt{DX_1}} = \sqrt{n} \frac{\bar{x} - EX_1}{\sqrt{DX_1}} \Rightarrow N(0, 1)$  или  $\sqrt{n}(\bar{x} - EX_1) \Rightarrow N(0; DX_1)$

**Th.** Выборочный  $k$ -ый момент является состоятельной несмещенной оценкой теоретического  $k$ -ого момента

1)  $E\bar{x}^k = EX^k$

2)  $\bar{x}^k \xrightarrow{p} X^k$

Это следует из предыдущей теоремы, если взять  $X^k$  вместо  $X$

**Th.** Выборочной дисперсией  $D^*$  и  $S^2$  являются состоятельными оценками теоретической дисперсией, при этом  $D^*$  - смещенная оценка, а  $S^2$  - несмещенная оценка

Заметим, что  $D^* = \overline{X^2} - \overline{X}^2$

$$ED^* = E(\overline{X^2} - \overline{X}^2) = E\overline{X^2} - E(\overline{X}^2) = EX^2 - E(\overline{X}^2)$$

$$\text{Так как } D\overline{X} = E(\overline{X^2}) - (E\overline{X})^2, \text{ то } EX^2 - E(\overline{X}^2) = EX^2 - ((E\overline{X})^2 + D\overline{X}) = (EX^2 - EX) - D\overline{X} =$$

$$DX - D\overline{X} = DX - D\left(\frac{X_1 + \dots + X_n}{n}\right) = DX - \frac{1}{n^2} \sum_{i=1}^n DX_i = DX - \frac{1}{n^2} n DX_1 = DX - \frac{1}{n} DX = \frac{n-1}{n} DX,$$

то есть  $D^*$  - смещенная вниз оценка

$$ES^2 = E\left(\frac{n}{n-1} D^*\right) = \frac{n}{n-1} \frac{n-1}{n} DX = DX \implies S^2 - \text{несмещенная вниз оценка}$$

$$2. D^* = \overline{X^2} - \overline{X}^2 \xrightarrow{p} EX^2 - (EX)^2 = DX - \text{состоятельная оценка}$$

$$S^2 = \frac{n}{n-1} D^* \xrightarrow{p} DX$$

*Nota.* Отсюда видим, что выборочная дисперсия - асимптотически несмещенная оценка. Поэтому при большом (обычно не меньше 100) объеме выборке можно считать обычную выборочную дисперсию

## Метод моментов (Пирсона)

Постановка задачи: пусть имеется выборка объема  $n$  неизвестного распределения, но известного типа, которое задается  $k$  параметрами:  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ . Требуется дать оценки данным неизвестным параметрам

Идея метода состоит в том, что сначала находим оценки  $k$  моментов, а затем с помощью теоретических формул из теории вероятности даем оценки этих параметров

Пусть  $\vec{X}$  - выборка из абсолютно непрерывного распределения  $F_\theta$  с плотностью известного типа, которая задается  $k$  параметрами  $f_\theta(x, \theta_1, \dots, \theta_k)$

Тогда теоретические моменты находим по формуле  $m_i = \int_{-\infty}^{\infty} x^i f_\theta(x, \theta_1, \dots, \theta_k) dx = h_i(\theta_1, \dots, \theta_k)$

Получаем систему из  $k$  уравнений с  $k$  неизвестными. В эти уравнения подставляем найденные оценки моментов и, решая получившуюся систему уравнений, находим нужные оценки параметров

$$\begin{cases} \bar{x} = h_1(\theta_1^*, \dots, \theta_k^*) \\ \overline{x^2} = h_2(\theta_1^*, \dots, \theta_k^*) \\ \dots \\ \overline{x^k} = h_k(\theta_1^*, \dots, \theta_k^*) \end{cases}$$

*Nota.* Оценки по методу моментов как правило состоятельные, но часто смещенные

*Ex.* Пусть  $X \in U(a, b)$ . Обработав статданные, нашли оценки первого и второго моментов:

$$\bar{x} = 2.25; \overline{x^2} = 6.75$$



Найти оценки параметров  $a^*, b^*$

Плотность равномерного распределения  $f_{(a,b)}(x) = \begin{cases} 0, & x < a \\ \frac{1}{b-a} & a \leq x \leq b, \\ 0, & x > b \end{cases}$

$$EX = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}$$

$$EX = \int_a^b x^2 \frac{1}{b-a} dx = \frac{a^2 + ab + b^2}{3}$$

Получаем:

$$\begin{cases} \bar{x} = \frac{a^*+b^*}{2} \\ \overline{x^2} = \frac{a^{*2}+a^*b^*+b^{*2}}{3} \end{cases} \iff \begin{cases} \frac{a^*+b^*}{2} = 4.5 \\ a^{*2} + a^*b^* + b^{*2} = 20.25 \end{cases} \iff \begin{cases} \frac{a^*+b^*}{2} = 4.5 \\ a^*b^* = 0 \end{cases} \iff \begin{cases} a^* = 0 \\ b^* = 4.5 \end{cases}$$