

# Математическая статистика

## 1

Баллы:

- баллы: задание + теор тест + посещение
- очно:  $45 + 20 + 20$
- дист:  $45 + 40 + 0$
- экзамен: 25
- распределение: очно/дистанционно по формату – старосты

#todo

☐ сделать опрос дист/очно

Необходимо знать:

- законы больших чисел
- типы распределений

---

Задача Математической Статистики: вы знаете только частично о том, что вы изучаете. Возникает чёрный ящик. На основе экспериментальных данных нужно дать оценки на числовые характеристики распределения. Отвечает не на теоретические вопросы, а на практические с некоторой вероятностью.

Возможно построение модели, которая с некоторой надёжностью предсказывает распределение.

Работаем с экспериментальными данными.

**Генеральная совокупность** – все результаты данной серии экспериментов (или экспериментальных значений случайно величины).

Если эксперимент чистый, независимый, то эти данные должны в точности соответствовать случайной величине. Но нюанс – вы не можете посмотреть всевозможные результаты экспериментов.

**Выборочная совокупность** – имеющиеся у нас данные (выборка из генеральной совокупности, возможно неполная).

Выборочная совокупность может не отражать реальное поведение случайно величины. *байка про самолёты-бомбардировщики в вmw.* (ошибка выжившего)

**Репрезентативная выборка** – выборка, имеющая то же самое распределение, что и теоретическая.

В дальнейшем предполагаем, что все выборки репрезентативные.

**Выборка** объёма  $n$  – набор экспериментальных данных  $(x_1, \dots, x_n)$ .

**Выборка** объёма  $n$  – набор  $(X_1, \dots, X_n)$  независимых одинаково распределённых случайных величин.

Note: поэтому  $EX_i = EX_2, DX_i = DX_2$  и используем обозначения  $EX_1, DX_1$  и т.д.

### Выборочные характеристики

Выборку можно рассматривать как дискретную случайную величину

$X$	$x_1$	$\dots$	$x_n$
$p^*$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

**Среднее выборочное** – число

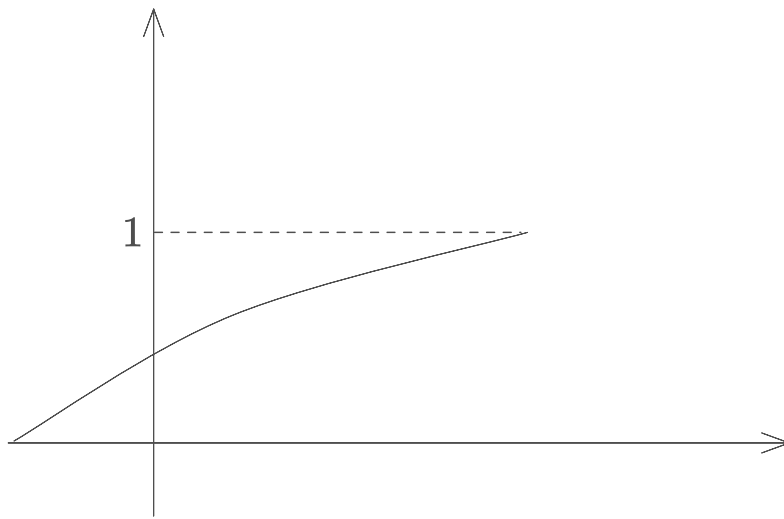
$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

**Выборочная дисперсия** – число

$$D_B = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Выборочная функция распределения** –

$$F^*(y) = \frac{\text{число данных } x_i \in (-\infty, y)}{n}$$



### Теорема Гливенко–Кантелли:

$\bar{X} = (x_1, \dots, x_n)$  объёма  $n$ ,  $F^*(y), F(y)$  – выборочная и теоретическая функции распределения

Тогда  $\sup_{y \in \mathbb{R}} |F^*(y) - F(y)| \rightarrow 0$  при  $n \rightarrow \infty$

### Начальная обработка статистических данных

Данные могут быть неоднородные.

Пусть однородные, тогда её можно **ранжировать** (упорядочить данные по возрастанию). В результате получаем **вариационный ряд**

- $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

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

Статистика – некоторая процедура, обрабатывающая статистические данные.

Данные могут повторяться. Например считать оценки (2,3,4,5) среди большого количества людей.

Note: Если мы объединяем повторяющиеся результаты с учётом числа повторов, то получаем частотный вариационный ряд. (надо ли это делать или нет – смотря по ситуации)

Другой нюанс: всё таки могут быть искажения в выборке. Бывает, что отбрасывается часть первых и часть последних (уже в вариационном ряду).

Если данных много и они не повторяются разумно разбить выборку на интервалы и составить так называем **интервально-вариационный ряд**.

Интервалы:

- равной длины (гистограммы, выдвижение гипотезы о типе распределения)
- равнонаполненные (проверка гипотез о типе распределения)

Число интервалов обычно (не всегда) берётся по формуле

$$K \approx 1 + \log_2 n$$

(иногда  $\sqrt[3]{n}$  или что-то зависящее не от  $n$ , а от системы, например система оценивая итмо)

В результате получаем  $K$  интервалов  $[a_{i-1}, a_i)$ .

$\nu_i$  — число данных попавших в  $i$ -ый интервал.

$\frac{\nu_i}{n}$  — относительная частота. (оценка теоретической вероятности попадания случайно величины в данный интервал)

Чтобы получить из этого дискретную величине можно взять середины каждого интервала, которым будет соответствовать вероятность  $\frac{\nu_i}{n}$ .

$$\bar{x} = \frac{1}{n} \sum c_i \nu_i \quad c_i = \frac{a_{i-1} + a_i}{2}$$

$$D_B = \frac{1}{n} \sum (c_i - \bar{x})^2 \cdot \nu_i$$

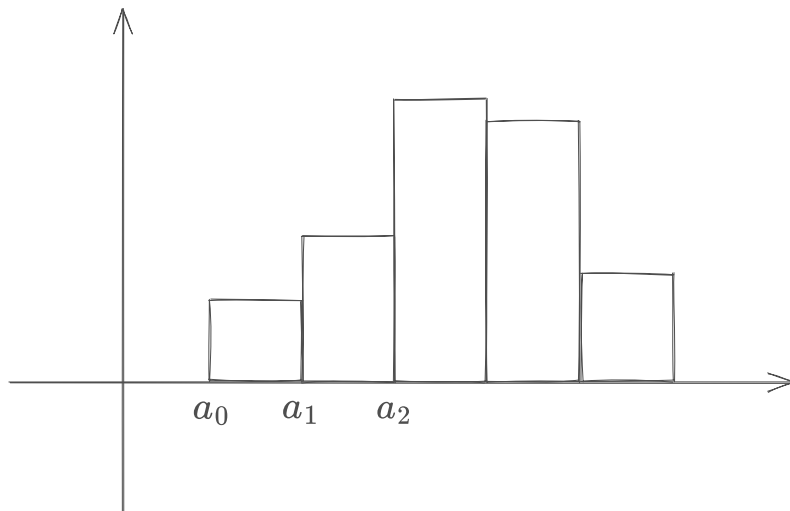
## Геометрическая интерпретация данных

Обычно удобнее рисовать гистограммы.

**Гистограмма** — набор прямоугольников для каждого интервала. Основание  $[a_{i-1}, a_i)$  длины  $l = a_i - a_{i-1}$ , а высоту берём пропорционально частоте, причём, чтобы суммарная площадь равнялась 1. высота:  $\frac{\nu_i}{nl}$

Гистограмма является приближением плотности распределения (если она непрерывная) и по её виду можно

выдвинуть гипотезу о типе распределения. Именно поэтому для этих целей лучше брать интервалы одинаковой длины.



Теорема:

Если число интервалов  $k(n) \rightarrow \infty$  и при этом  $\frac{k(n)}{n} \rightarrow 0$ , то гистограмма по вероятности поточечно сходится к теоретической плотности.

Полигон — кусочно-линейная функция соединяющая точки вида  $(x_i, \nu_i)$

## 1

### ✓ Математическая статистика с 2022-09-05

Сделать то, что мы делали на практике с ценами акций

- разбить на промежутки длины  $1 + \log_2(n)$
- посчитать частоты
- посчитать частоты/ $n$
- центры отрезков
- функция распределения
- гистограмму + полигон