

Основы математической статистики.  
Количественные характеристики  
популяции. Графическое представление  
данных

Теория вероятностей  
и математическая статистика / Урок 3



# Математическая статистика

На предыдущих уроках мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На предыдущих уроках мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На практике, как правило, мы имеем некоторую выборку. Мы ничего не знаем про случайную величину, из которой взята эта выборка. А хотелось бы.

На предыдущих уроках мы говорили о теоретических характеристиках случайных событий и величин. Зачем нам это нужно?

Любая выборка представляет собой значения некоторой случайной величины.

На практике, как правило, мы имеем некоторую выборку. Мы ничего не знаем про случайную величину, из которой взята эта выборка. А хотелось бы.

В дальнейшем мы научимся по выборкам:

- строить оценки для параметров случайных величин,
- проверять гипотезы о значениях этих параметров, а также об общих характеристиках и свойствах случайных величин,
- строить доверительные интервалы для параметров случайных величин.

Для точечного оценивания параметров случайной величины используются различные статистики. **Статистика** — это любая функция от выборки.

Пусть дана выборка  $X = (x_1, x_2, \dots, x_n)$  из значений случайной величины. Одной из наиболее естественных статистик таких выборок является **среднее арифметическое** (или **выборочное среднее**). Оно обозначается как  $\bar{X}$ :

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

Выборочное среднее является **оценкой** для математического ожидания. Это означает, что, как правило, чем больше элементов в выборке, тем ближе выборочное среднее этой выборки к математическому ожиданию соответствующей случайной величины.

Выборочная дисперсия, как следует из названия, оценивает дисперсию случайной величины:

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

Такая формула для оценки дисперсии называется несмещённой за счёт этой единицы в знаменателе. Существует также смещённая оценка дисперсии:

$$\sigma_{X, biased}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$$

Однако, она в практических задачах не используется.

Почему смещённые оценки — это плохо, а также почему вычитают именно 1 — читайте в дополнительных материалах к уроку.



Вообще, дисперсия является не очень наглядной мерой разброса, поскольку имеет другой масштаб. Поэтому часто наряду с дисперсией используют **среднее квадратическое отклонение**, равное корню из дисперсии.

Оценивается среднее квадратическое отклонение аналогично дисперсии:

$$\sigma_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2}$$

## Медиана, квантили

**Медиана** — такое значение  $t$ , что половина элементов из выборки меньше, либо равна  $t$ , и, соответственно, половина больше, либо равна  $t$ .

Медиана представляет собой *середину* выборки: если отсортировать элементы выборки по возрастанию, то медиана приходится на середину.

Медиана может приходиться как на промежуток между элементами выборки, так и на конкретный элемент.

Медиана является частным случаем более общего понятия — *квантиля*.

Пусть  $\alpha \in (0, 1)$ . Квантиль порядка  $\alpha$  — такое число  $t_\alpha$ , что « $\alpha$  процентов» всех элементов выборки меньше  $t_\alpha$  и, соответственно, « $(1 - \alpha)$  процентов» элементов — больше  $t_\alpha$ .

Как и в случае с медианой, квантиль может как приходиться на один из элементов выборки, так и лежать где-то между ними.

Из определения следует, что медиана является квантилем порядка 0.5. Кроме того, часто используют:

- **первый квартиль** — квантиль порядка 0.25 (т.е. значение, которое не превышают 25% значений из выборки),
- **второй квартиль** — то же, что и медиана,
- **третий квартиль** — квантиль порядка 0.75.

Из определения следует, что медиана является квантилем порядка 0.5. Кроме того, часто используют:

- **первый квартиль** — квантиль порядка 0.25 (т.е. значение, которое не превышают 25% значений из выборки),
- **второй квартиль** — то же, что и медиана,
- **третий квартиль** — квантиль порядка 0.75.

Также могут встречаться:

- **децили** — то же, что и квартили, но делим мы не на 4 части, а на 10. Например, медиана будет пятым децилем,
- **перцентили** — это просто другой способ задать квантиль. Здесь мы используем не долю  $\alpha \in (0, 1)$ , а процент. Например, третий квартиль будет 75-перцентилем.

Интерквартильный размах — это отрезок между первым и третьим квартилями. Это отрезок, в который попадают 50% значений выборки.

Интерквартильный размах используется для измерения разброса значений выборки вокруг среднего. Иногда его использование оказывается более предпочтительным, чем использование среднего квадратического отклонения, поскольку не учитывает выбросы в данных.

Понятие квантиля также можно определить для случайной величины. Суть определения такая же, что и в случае выборки, но выглядит немного страшнее.

Квантилем порядка  $\alpha$  случайной величины  $X$  называется такое значение  $t_\alpha$ , что

$$P(X \leq t_\alpha) = \alpha, \quad P(X \geq t_\alpha) = 1 - \alpha$$

Идея та же: в доле  $\alpha$  всех случаев значение случайной величины  $X$  окажется меньше  $t_\alpha$  и в доле  $(1 - \alpha)$  случаев — больше  $t_\alpha$ .



Теоретические квантили постоянно используются в статистических методах (о таких мы поговорим на уроке 5). Они позволяют решить следующую задачу: имеется случайная величина  $X$  и значение вероятности  $\alpha \in (0, 1)$ . Требуется найти пороговое значение  $t$ , такое, что  $P(X \leq t) = \alpha$ .

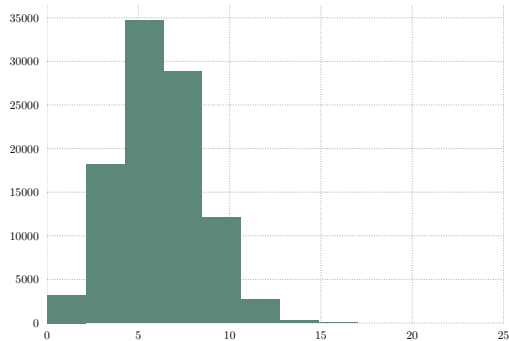
Это и есть квантиль порядка  $\alpha$ .

## Графическое представление данных

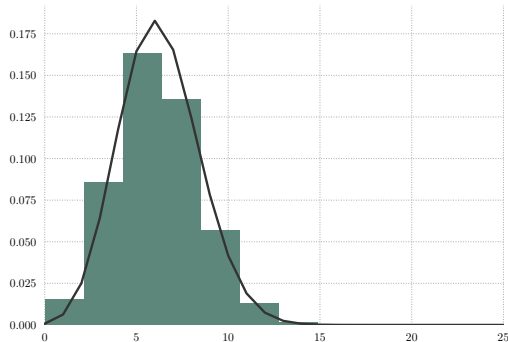
Для визуализации распределения значений выборки часто используется гистограмма. Как строится гистограмма?

- ① По оси  $x$  откладываются все возможные значения из выборки.
- ② Вся ось разбивается на какое-то заданное число одинаковых отрезков.
- ③ Для каждого отрезка вычисляется число значений выборки, которые лежат в этом отрезке, и это число откладывается по оси  $y$ .

Например, справа изображена гистограмма для выборки размера 1000 из биномиального распределения с параметрами  $n = 25$  и  $p = 0.25$ . Разбиение оси  $x$  здесь производится на 8 частей.



Гистограмма по форме напоминает график распределения вероятностей случайной величины. Нужно лишь нормировать откладываемые по оси  $y$  значения, чтобы сумма площадей колонок стала равной 1.



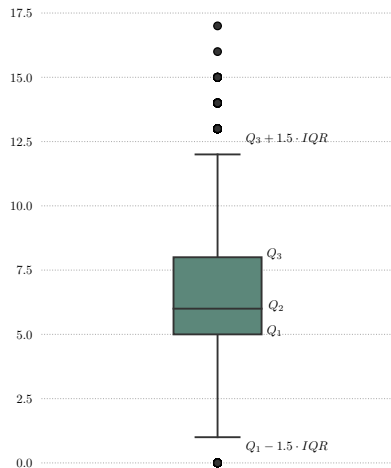
Другой способ визуализировать одномерные данные — **boxplot** или **ящик с усами**. В самом ящике отмечены квартили  $Q_1$ ,  $Q_2$  (медиана),  $Q_3$ . «Усы» здесь — границы отрезка


$$[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR],$$

где  $IQR$  — интерквартильное расстояние.

Всё, что выходит за границы этого отрезка, считается выбросами (отмечены кружками). Обычно это порядка 0.7% выборки.

Например, справа изображён boxplot той же выборки из биномиального распределения с параметрами  $n = 25$ ,  $p = 0.25$ .





Спасибо за внимание