**Тема 1. Введение в математическую статистику**

## 1. От теории к практике: зачем нужна статистика?

До этого момента случайные величины и их распределения могли рассматриваться как теоретические объекты:

пусть $X \sim F$, и функция распределения $F(x)$ нам известна.

В рамках теории вероятностей задача формулируется так:

если известно распределение $F(x)$, найти вероятности событий, математическое ожидание, дисперсию и другие характеристики.

Однако в реальности нам практически никогда не дана «истинная» функция распределения.

Например, если $X$ — ежемесячная заработная плата гражданина страны, то распределение $F_X(x)$ нам неизвестно. Нет готовой формулы, которую можно подставить в интеграл и получить ответ.

Следовательно, возникает обратная задача:

мы наблюдаем конечное число значений случайной величины

$$X_1, X_2, \dots, X_n,$$

и должны по этим данным восстановить свойства распределения $F_X$.

Именно здесь начинается статистика.

Если теория вероятностей отвечает на вопрос:

«Что произойдёт, если распределение известно?», то статистика отвечает на противоположный вопрос:

«Как восстановить распределение, если оно неизвестно?».

Таким образом, предмет статистики — это переход от абстрактной вероятностной модели к анализу реальных данных.

## 2. Основные понятия: генеральная совокупность и выборка

### Наблюдение

Пусть $X$ — случайная величина, описывающая исследуемый объект (например, зарплату).

Каждое зафиксированное значение

$$x_1, x_2, \dots, x_n$$

называется наблюдением.

Важно различать:

- $X$ — случайная величина (модель),
- $x_i$ — конкретная реализация.

**Генеральная совокупность (Population)**

Генеральная совокупность — это множество всех возможных наблюдений исследуемого объекта.

Пусть её объём равен $N$.

Если бы нам были известны все значения

$$x_1, x_2, \dots, x_N,$$

мы могли бы вычислить истинные параметры:

математическое ожидание

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

дисперсию

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2.$$

Но в большинстве практических задач $N$ очень велико или бесконечно.

**Выборка (Sample)**

Выборка — это часть генеральной совокупности:

$$X_1, X_2, \dots, X_n, \quad n \ll N.$$

Статистика основана на предположении, что наблюдения являются реализациями независимых одинаково распределённых случайных величин:

$$X_1, \dots, X_n \sim F_X.$$

Почему мы работаем с выборками?

Потому что:

- сбор данных обо всей совокупности слишком дорог;
- иногда невозможен физически;
- часто нет доступа ко всем объектам.

Задача статистики — по данным $n$ наблюдений сделать выводы о всей совокупности объёма $N$.

### 3. Две ветви статистики

Статистический анализ условно делится на два больших направления.

#### 1. Описательная (дескриптивная) статистика

Её цель — структурировать и описать имеющиеся данные.

Основные инструменты:

выборочное среднее

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

выборочная дисперсия

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

медиана, квартили, гистограммы, диаграммы рассеяния.

Описательная статистика отвечает на вопрос:

«Что происходит в нашей выборке?».

#### 2. Статистический вывод (инференция[1])

Здесь мы делаем шаг дальше:

[1] Статистическая инференция — это процесс получения обоснованных выводов о параметрах распределения или свойствах генеральной совокупности на основе анализа выборочных данных.

на основе выборки оцениваем параметры генеральной совокупности.

Например, если предполагается, что

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

то по выборке строятся оценки:

$$\hat{\mu} = \overline{X}, \quad \widehat{\sigma^2} = S^2.$$

Также строятся доверительные интервалы и проверяются гипотезы, например:

$$H_0: \mu = \mu_0.$$

Статистический вывод отвечает на вопрос:

«Что можно сказать о всей совокупности, зная только выборку?».

**4. Типы данных**

Перед применением методов исследователь должен понять, с каким типом данных он работает.

**Количественные (числовые)**

Наблюдения принимают числовые значения.

Они бывают:

- дискретные (например, число ошибок);
- непрерывные (например, температура или доход).

**Качественные (категориальные)**

Наблюдения — это категории:

- пол,
- специальность,
- цвет глаз.

Для них применяются иные методы анализа (частоты, доли, таблицы сопряжённости).

## Раздел 2. Графическое представление данных

*«Говорят, что нельзя сделать ничего такого, что вы не можете представить себе заранее.»*
*— неизвестный автор*

Графики используются для быстрого визуального представления данных.

Визуальное восприятие позволяет получить мгновенное впечатление о структуре данных и заметить такие характеристики, как:

- форма распределения,
- расположение центра,
- степень разброса,
- наличие выбросов,
- возможные закономерности.

Графическое представление даёт возможность сделать предварительные выводы уже при первом взгляде на данные.

### Основные методы графического представления данных

К основным способам визуализации относятся:

1. Точечная диаграмма (Dot plot)
2. Столбчатая диаграмма (Bar chart)
3. Диаграмма «стебель и листья» (Stem-and-leaf plot)
4. Гистограмма (Histogram) — с отображением абсолютной или относительной частоты
5. График накопленной частоты (Cumulative frequency plot)
6. Ящик с усами (Boxplot) — рассматривается в разделе описательной статистики

### Точечная диаграмма (Dot Plot)

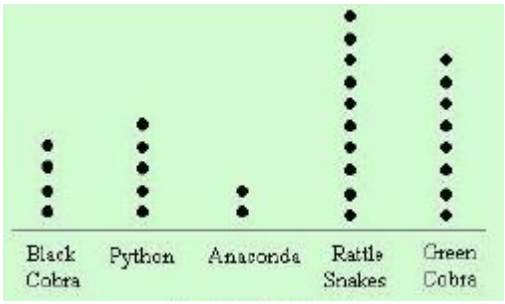
Точечная диаграмма является одним из самых простых способов представления данных. Каждое наблюдение отображается отдельной точкой на числовой оси.

Если некоторое значение повторяется несколько раз, точки располагаются друг над другом.

Точечная диаграмма особенно полезна при небольшом объёме выборки, поскольку позволяет:

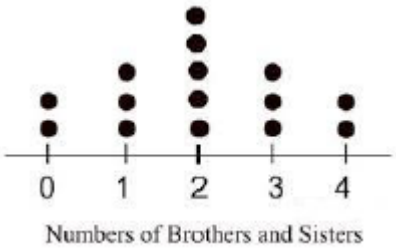
- увидеть форму распределения,
- заметить повторяющиеся значения,
- обнаружить выбросы,
- оценить приблизительное расположение центра.

Она даёт непосредственное представление о распределении данных без предварительного агрегирования.




Types of snakes in the Zoo.

Точечный график для количественных данных. Набор данных был получен следующим образом: 15 человек спросили, есть ли у них братья и сестры, и если да, то сколько. На приведенном ниже точечном графике показан результат.




Numbers of Brothers and Sisters

**Столбчатая диаграмма (Bar Chart)**

Столбчатая диаграмма используется для представления частот или относительных частот категориальных или дискретных данных.

В отличие от точечной диаграммы, где отображается каждое наблюдение, bar chart агрегирует данные по категориям.

**Формальная постановка**

Пусть имеется выборка:

$$x_1, x_2, \dots, x_n,,$$

где значения принимают дискретные или категориальные значения $a_1, a_2, \dots, a_k$.

Определим абсолютную частоту категории $a_j$:

$$f(a_j) = \sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{x_i = a_j}\}}.$$

Относительная частота:

$$p(a_j) = \frac{f(a_j)}{n}.$$

На столбчатой диаграмме по оси $x$ располагаются категории $a_j$, а по оси $y$ — либо $f(a_j)$, либо $p(a_j)$.

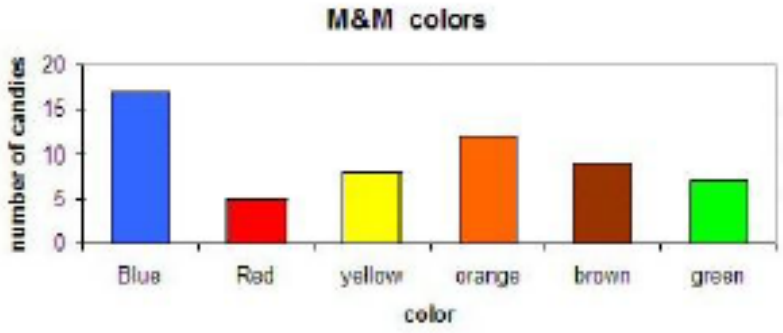
Высота каждого столбца равна соответствующей частоте.

**Важная особенность**

Столбцы в bar chart **не соприкасаются друг с другом**.

Это подчёркивает, что данные являются дискретными или категориальными.

*На приведенной ниже гистограмме показано распределение цветов конфет M&M в некоторых выбранных упаковках.*




**Stem and Leaf Plot (ствол с ветвями/ стебель и листья)**

Маша готовится к спортивным соревнованиям в Више. Чтобы оценить свои шансы, Маша собрала данные о результатах 20 университетских спортсменов (в секундах).

У нее есть следующий набор данных: 10.17, 10.23, 10.25, 10.28, 10.31, 10.32, 10.34, 10.35, 10.41, 10.44, 10.45, 10.46, 10.49, 10.52, 10.55, 10.64, 10.68, 10.69, 10.71, 11.

Ниже приведен график стебля и листьев, представляющий эти данные. График получил свое название из–за того, как он выглядит – как дерево - листья растут на основании “растущего” стебля.

**Time for 100-meters Sprint**

| *stem* | *leaves* |
|---|---|
| 10.0 | |
| 10.1 | 7 |
| 10.2 | 3 5 8 |
| 10.3 | 1 2 4 5 |
| 10.4 | 1 4 5 6 9 |
| 10.5 | 2 5 |
| 10.6 | 4 8 9 |
| 10.7 | 1 |
| 10.8 | |
| 10.9 | |
| 11.0 | 0 |

key: 10.7|1 = 10.71 seconds

единица измерения стебля: 1,00

единица измерения листа: 0,01

На графике представлены так называемые значения стебля (слева) и листьев (справа). Они разделены вертикальной линией. Каждый лист представляет собой одно наблюдение, которое можно интерпретировать как сумму значений стебля и листа.

Например, обозначение 10.2|3 означает 10.23 = 10.2+0.03.

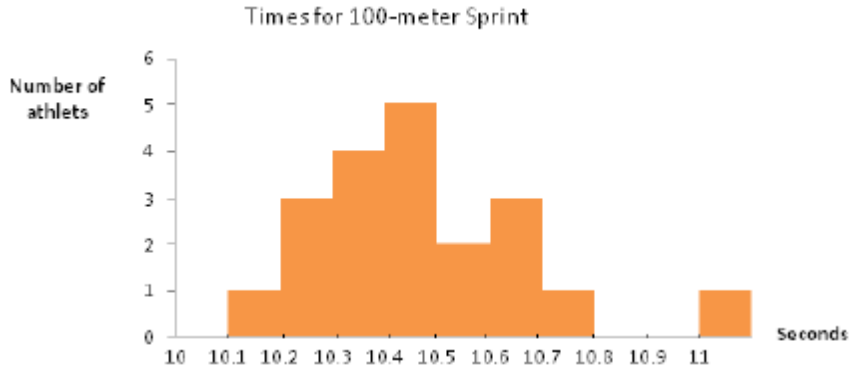
Основное преимущество графиков стебля и листа перед многими другими типами графиков заключается в том, что они отображают набор данных без потери информации, поскольку вы можете восстановить значение каждого конкретного наблюдения. Это невозможно, например, с гистограммой.

**Histogram (гистограмма)**

Гистограмма — это график, отображающий распределение наблюдаемых значений. Ниже вы можете увидеть гистограмму для набора данных ”спринтеры".

Высота каждого столбца наблюдений (частота) значений в соответствующем интервале.

На гистограмме сначала следует разделить данные на группы, или интервалы. Вот выдержка из 0.1. Например, интервал [10.1,10.2) содержит один элемент - 10.17. Это отражено в его высоте, которая равна 1. [10.2, 10.3) состоит из 3 элементов - 10.23, 10.25 и 10.28, и так далее для всех остальных интервалов. [10.3, 10.4) - 4 элемента, [10.4, 10.5) – 5 элементов, [10.5, 10.6) – 2, [10.6, 10.7) – 1, [10.7, 10.8) – 1, [11, 11.01) – 1.

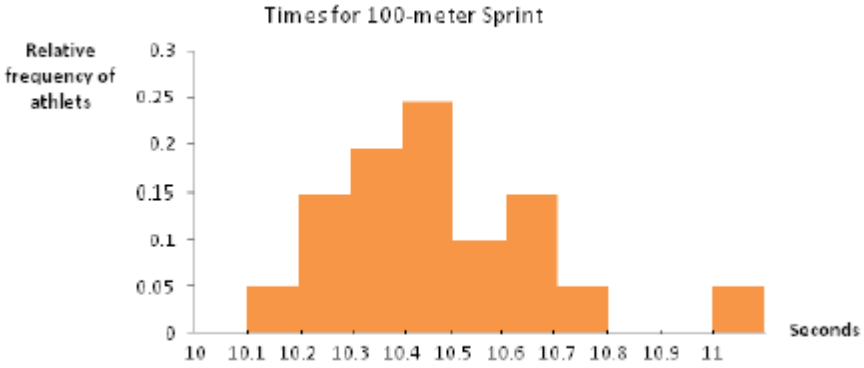



Обратите внимание, что эти группы обычно не задаются заранее. Итак, как это сделать?

Мы смотрим на весь диапазон значений (от min до max) в нашем наборе данных и решаем, на сколько интервалов равной длины мы будем его делить.

Самое главное, чтобы интервалы всегда были ОДИНАКОВОЙ ДЛИНЫ.

Сколько интервалов будет, зависит от вас, строгих правил на этот счет нет. Имейте в виду, что слишком малое количество интервалов и, следовательно, столбцов делает график менее информативным, а слишком большое – трудным для восприятия. Обычно оптимальным является значение от 5 до 10. Наши наблюдения начинаются с 10.17 до 11.00. секунд. Давайте для точности запишем на графике диапазон от 10.1 до 11.1 и разделим его на 10 интервалов продолжительностью 0,10 секунды.




Обратите внимание, что форма гистограммы не изменилась. Это всегда так, поскольку для перехода к относительным частотам мы просто разделили высоту столбцов (частот) на количество наблюдений n. Гистограмма позволяет увидеть закономерности распределения данных, такие как симметрия.

Обратите внимание, что: 1) гистограмма может быть построена только для количественных данных, 2) Столбцы на ней располагаются близко друг к другу, если в них нет промежутков (подробнее об этом ниже).

Если у вас есть качественный набор данных, аналогичный график следует называть столбчатой диаграммой, о чем говорилось ранее. В этом случае вы не можете использовать термин "гистограмма". Он отличается, за исключением типа используемых данных, тем, что между столбцами всегда есть пробелы. Эти пробелы подчеркивают идею о том, что категории представляют собой существенно разные реализации некоторого качества, которые нельзя однозначно поместить на ось непрерывных значений.

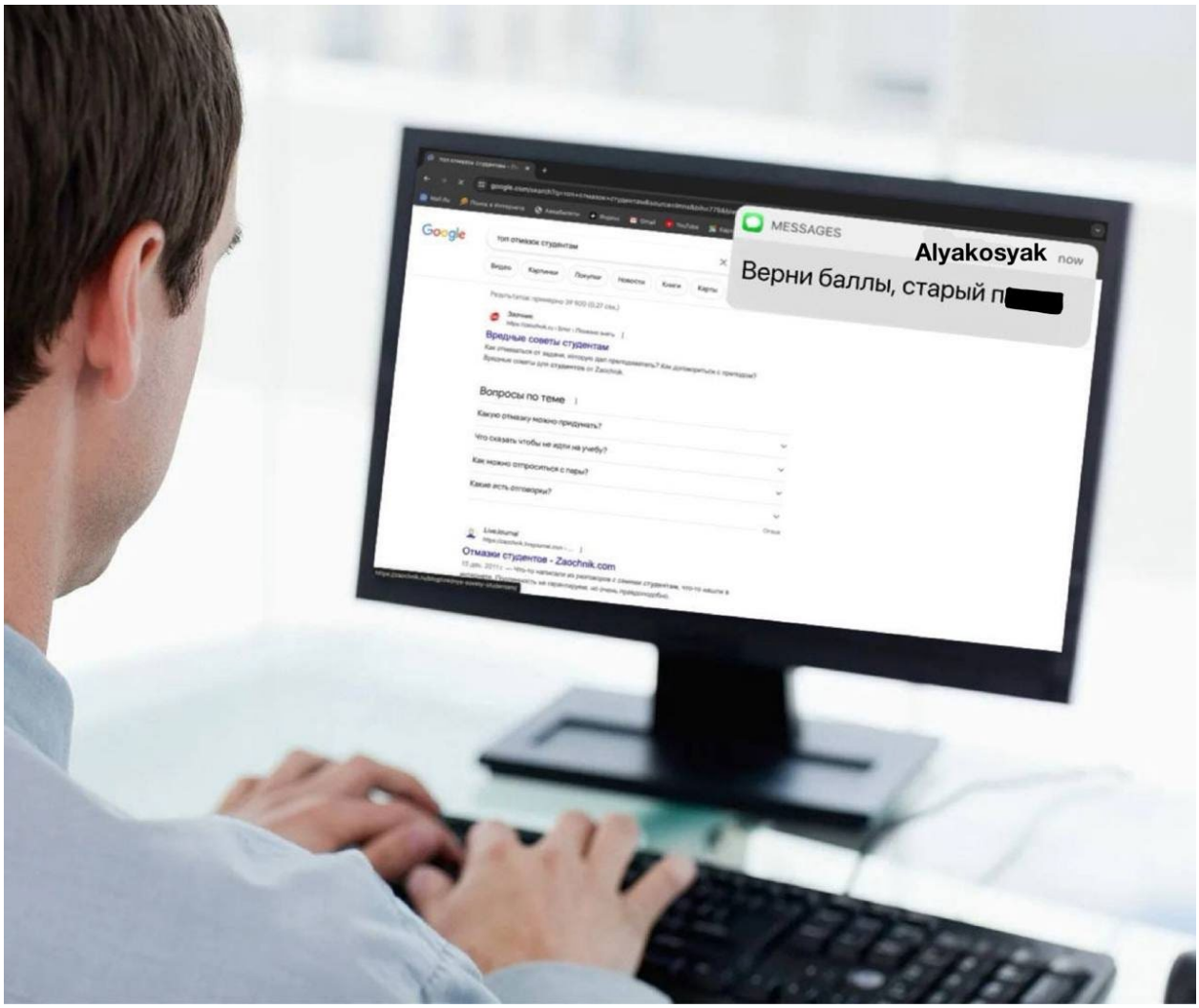
Представьте себе столбчатую диаграмму для университетских специальностей студентов. Значения на По оси X - названия отделов (экономика, медицина, СМИ и т.д.), по оси Y - частота встречаемости. Необходимо соблюдать дистанцию между категориями, чтобы показать, что между СМИ и Медицинские специальности в университете. Напротив, столбцы гистограммы расположены вплотную друг к другу. Промежуток между ними следует интерпретировать как пробел. Посмотрите на гистограмму выше: между столбцами нет пробелов, за исключением одного места, потому что между 10.8 и 11 нет наблюдений.

**Section 2. Descriptive Statistics**

*«Хорошо, все выстройтесь в алфавитном порядке по росту».*
*— Кейси Стенгел*

Просто глядя на набор данных, что можно сказать о его характеристиках? Например, посмотри на набор данных о количестве домашних заданий, сданных студентами.

**ПРИМЕР “КОЛИЧЕСТВО ДОМАШНИХ ЗАДАНИЙ”**




Маша решила проанализировать, насколько хорошо её одногруппники справляются с домашкой. Для этого она собрала данные о **общем количестве домашних заданий по статистике**, которые сдали 25 студентов её группы за 1-й семестр:

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4, 30, 20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32.

Просто глядя на эти числа, можешь ли ты сказать, насколько хорошо студенты в целом выполняют задания? Например: какое **среднее** число выполненных заданий, **сконцентрированы** ли результаты около какого-то значения или сильно **разбросаны**,

есть ли **разрыв** между низкими и высокими значениями. Сам список чисел не даёт очевидного вывода. Чтобы делать выводы, наблюдения нужно сначала упорядочить, а затем обработать. Чтобы получать знание из данных, нужны инструменты. **Описательная статистика** — это такие инструменты.

**Описательная статистика** — это формула, которая позволяет получить **одно число** (например, средний балл) из набора чисел (например, выборки баллов студентов). С помощью описательной статистики проще анализировать наборы данных и сравнивать их между собой.

Мы можем разделить описательную статистику на три основные категории:

1. **Положение (расположение) наблюдений на оси**: отвечает на вопросы о
   — центре набора данных (mean, median, mode),

   — крайних значениях (minimum, maximum),

   — промежуточных значениях (quartiles, percentiles).

2. **Вариативность / разброс**: отвечает на вопросы о

   — типичном разбросе (standard deviation, variance),

   — экстремальном разбросе (range),

   — разбросе “середины” данных (interquartile range).

3. **Форма распределения**: описывает форму распределения (например, симметричность).

**Location of Observations along Dataset**

**Center**

Интуитивно первый шаг в описании любого набора данных — найти его **центральное** или **типичное** значение.

**Mean (среднее арифметическое)**

**Среднее (mean)** — это арифметическое среднее всех наблюдаемых значений. То есть сумма всех наблюдений, делённая на их количество.

Формула:

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

Маша хочет найти среднее число сданных домашних заданий в своей группе:

$$\bar{X} = \frac{23 + 45 + \cdots + 32}{25}.$$

Проверьте, правильно ли она посчитала среднее: $\bar{X} = 35.08$.

✅ Если среднее вычислено по выборочным данным, его называют **выборочным средним** (sample mean).

В отличие от него, **среднее генеральной совокупности** $\mu_x = E(X)$ основано на данных всей совокупности и требует знания истинного распределения X. Подробнее это различие будет дальше (в следующих главах).

**Median (медиана)**

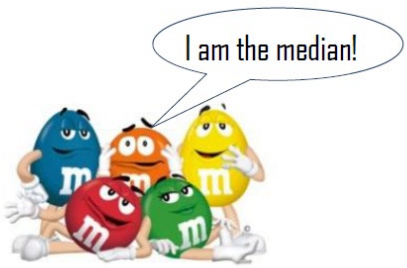
**Медиана (median)** — это число, которое отделяет нижнюю половину выборки от верхней. Если упорядочить все наблюдения и мысленно разделить их на две равные части, медиана будет границей между ними: **50% наблюдений выше медианы и 50% ниже**.

**Как найти медиану?**

1. Сначала упорядочь наблюдения по возрастанию (от минимального к максимальному) и пронумеруй их от 1 до n. Тогда 1-е наблюдение — это минимум, а n-е — максимум.
2. Если число наблюдений **нечётное (odd)**, медиана стоит ровно в середине списка:

$$\text{Med} = x_{\frac{n+1}{2}}.$$

Например, при $n = 5$ медиана — это 3-е наблюдение, потому что $\frac{5+1}{2} = 3$.



3. Если число наблюдений **чётное (even)**, в середине окажутся два наблюдения — с номерами $\frac{n}{2}$ и $\frac{n}{2}+1$. Тогда медиана равна их среднему:

$$\text{Med} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}.$$

Например, при n = 6 медиана — это среднее 3-го и 4-го наблюдения.

**Пример (домашние задания Маши)**

Маша упорядочила данные по возрастанию:

4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, 38, 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56.

Здесь n=25 (нечётное), значит

$$\text{Med} = x_{\frac{25+1}{2}} = x_{13}.$$

13-е наблюдение равно 38, поэтому

$$\text{Median} = 38.$$

**Mode (мода)**

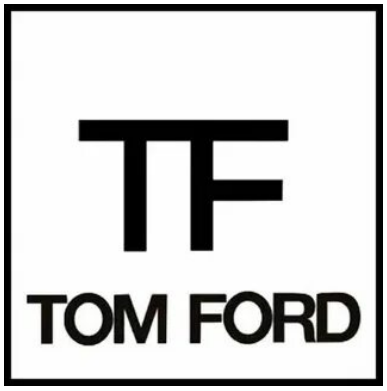
**Мода (mode)** — это наблюдение с **максимальной частотой**. То есть самое «популярное» значение в наборе данных.

Пример из лекции: если бы все носили **красные ботинки**, то они были бы «в моде», потому что встречаются чаще всего — значит, это и есть мода по максимальной частоте.

В нашем примере с домашними заданиями **мода равна 44**, потому что это единственное число, которое встретилось **три раза**.

Важно: мода не всегда является хорошей характеристикой «центра».

- У распределения может быть **больше одной моды** (если несколько значений имеют одинаковую или почти одинаковую частоту).
- Бывает, что самые частые значения — это **большие** значения на «хвосте» распределения, и тогда мода плохо отражает центр.
- Моду можно находить и для **количественных**, и для **качественных** данных (например, в задаче про зоопарк мода — это самый частый вид змей).




**Mean or Median: What is better? (Среднее или медиана: что лучше?)**

Рассмотрим данные о **месячном доходе** 29 сотрудников, работающих над стартап-проектом в Москве (в тысячах рублей): 13, 15, 21, 21, 22, 25, 25, 25, 27, 28, 28, 30, 30, 33, 34, 35, 35, 35, 39, 40, 40, 40, 41, 45, 45, 50, 55, 63, 1235.

Последнее значение — это средний месячный доход владельца предприятия: он зарабатывает **больше миллиона рублей** в месяц.

Посчитаем средний доход и медиану:

$$\bar{X} = 75, \quad \text{Median} = 34.$$

Они сильно различаются. Что лучше описывает **типичный** доход в этом наборе данных?

В данном случае **медиана лучше**. Значение 34 000 рублей выглядит адекватной мерой центра распределения. А среднее 75 000 рублей — странная «типичность», потому что **в данных вообще нет человека с доходом около 75 000**.

Проблема среднего как меры центра возникает, когда в данных есть **малое число экстремальных наблюдений** (очень больших или очень маленьких). Среднее «взвешивает» все наблюдения одинаково, поэтому даже один миллионер может сильно сместить $\bar{X}$ вверх. Аналогично, если в выборке есть несколько карликов, средний рост будет смещён вниз. То есть **среднее чувствительно к значениям**, а медиана просто берёт середину упорядоченного списка и не «улетает» из-за одного экстремума.

Такое расхождение характерно для **асимметричных распределений** с так называемыми «тяжёлыми хвостами». В этом случае медиана обычно лучше, как мера центрального значения. Для **симметричных** распределений среднее и медиана дают примерно одинаковые числа.


Но даже при асимметрии среднее — не бессмыслица: оно уже не отражает «типичное» наблюдение, зато может быть полезно как характеристика **общего среднего уровня**. Например, сравнивая два стартапа, можно хотеть сравнить **доход на работника** (эффективность «прибыль/труд») — тогда логичнее смотреть на среднее. Аналогично экономисты используют **ВВП на душу населения**: абсолютный ВВП завышает «богатство» больших стран просто из-за размера, поэтому делят на численность населения, чтобы сравнение было осмысленным.

Итого: **медиана лучше**, когда ищем «центральное/типичное» значение X; **среднее лучше**, когда хотим меру «общего среднего уровня».

**Extreme values (экстремальные значения): Minimum и Maximum**

**Minimum (минимум)** и **Maximum (максимум)** — это **экстремальные значения** набора данных:

- минимум — самое маленькое наблюдение,
- максимум — самое большое наблюдение.

Если мы предварительно **упорядочили** наблюдения по возрастанию и пронумеровали их от 1 до n, то:

- $x_1$ — это минимум,
- $x_n$ — это максимум.

В примере с домашними заданиями Маши (упорядоченный список мы уже выписывали):

$$X_{min} = 4, \qquad X_{max} = 56.$$

**Quantiles (квантили)**

**Квантиль (quantile)** — это значение, которое “разделяет” набор данных: это такое число $x_a$, что некоторая фиксированная доля наблюдений **не превышает** его. Формальное определение:

$$P(X \leq x_a) = p,$$

где $x_a$ — квантиль, а $p$ — фиксированная вероятность (она же называется **уровнем** квантиля).

Если распределение **непрерывное**, то $x_a$ единственен и определяется уравнением:

$$F(x_a) = p,$$

и это должно напомнить тебе функцию распределения F(x), потому что это она и есть.

Если распределение **эмпирическое** (то есть у нас просто выборка из практики), то есть способы вычислять квантили по данным — их дальше и разбирают.

**Types of quantiles (виды квантилей)**

В зависимости от того, на сколько частей квантиль делит набор, выделяют:

- **quartiles** (квартили),
- **percentiles** (процентили),
- **deciles** (децили).

**Quartiles (квартили)**

Квартили — это дополнительные опорные точки положения данных на числовой оси. Всего выделяют три квартиля: **нижний**, **верхний** и **средний**.

**Lower quartile $LQ/Q_1$ (нижний квартиль)**

Нижний квартиль LQ — значение, которое отделяет **нижние 25%** наблюдений от **верхних 75%**.

Его также называют **первым квартилем** $Q_1$, потому что он отделяет первую четверть данных от остальных.

Интуитивно: если отсортировать наблюдения и разделить их на 4 равные части, $Q_1$ — граница между 1-й и 2-й четвертью; можно также воспринимать $Q_1$ как **медиану нижней половины** наблюдений.

**Upper quartile $UQ/Q_3$ (верхний квартиль)**

Верхний квартиль UQ отделяет **нижние 75%** наблюдений от **верхних 25%**. Его также называют **третьим квартилем** $Q_3$. Это граница между 3-й и 4-й четвертью отсортированных наблюдений; можно воспринимать как **медиану верхней половины**.

**Second quartile $Q_2$ (второй квартиль)**

Второй квартиль разделяет 2-ю и 3-ю четверть (то есть нижнюю и верхнюю половины). Это **ровно медиана**. Поэтому медиану иногда называют $Q_2$: медиана — это “средний” квартиль.

**How to calculate quartiles? (как считать квартили) — Method 1**

Маша хочет найти нижний и верхний квартили для данных по домашним заданиям.

**Метод 1:** квартили — это медианы нижней и верхней половины.

1. Она сортирует числа и находит медиану. Мы уже получили:

$$\text{Med} = x_{13} = 38.$$

2. Медиана делит набор на две части (саму медиану **исключаем** из обеих частей). Нижняя половина:

   4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36.

   Верхняя половина:

   38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56.

3. В каждой половине по 12 наблюдений (чётное число), значит медиана — среднее между 6-м и 7-м значениями. Тогда:

$$LQ = \frac{x_6 + x_7}{2} = \frac{23 + 30}{2} = 26.5.\ UQ = \frac{44 + 45}{2} = 44.5.$$

**Method 2**

Есть ещё один способ находить LQ и UQ, который опирается на более общее понятие — **процентиль (percentile)**.

**Percentile (процентиль)**

**Процентиль** — это число, которое делит набор данных на две части так, что заданная доля ppp наблюдений находится **ниже** этого числа, а доля 1−p1-p1−p — **выше**.

Например:

- для нижнего квартиля p=0.25,
- для медианы p=0.5,
- для верхнего квартиля p=0.75.

То есть:

- LQ = 25-й процентиль,
- $Median$ = 50-й процентиль,
- UQ = 75-й процентиль.

**30-й процентиль** — это число, выше которого лежит 70% наблюдений и ниже которого лежит 30% наблюдений (то есть оно “выше 30% нижних наблюдений и ниже 70% верхних”).

**Общий алгоритм нахождения p-процентиля**

1. Упорядочить наблюдения по возрастанию. Для домашек Маши: 4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, 38, 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56.
2. Посчитать число

$$p \cdot (n + 1),$$

где n — объём выборки.

Например, для нижнего квартиля: n=25, p=0.25, поэтому

$$p \cdot (n + 1) = 0.25 \cdot (25 + 1) = 6.5.$$

3. Представить это число как

$$p \cdot (n + 1) = k + a,$$

где k — целая часть, a — дробная ($0 \leq a < 1$).

В примере: 6.5=6+0.5, значит k=6, a=0.5.

Это означает: искомый процентиль лежит **между** $x_6$ и $x_7$.

4. Найти процентиль по формуле линейной интерполяции:

$$x_p = x_k + a \cdot (x_{k+1} - x_k).$$

**Применение к данным Маши**

**Нижний квартиль $LQ = x_{0.25}$**

Здесь $x_6 = 23$ $x_7 = 30, a = 0.5$:

$$LQ = x_{0.25} = x_6 + 0.5 \cdot (x_7 - x_6) = 23 + 0.5 \cdot (30 - 23) = 26.5.$$

**Медиана $\mathrm{Med} = x_{0.5}$**

$$\mathrm{Med} = x_{0.5} = x_{13} + 0 \cdot (x_{14} - x_{13}) = x_{13} = 38.$$

**Верхний квартиль $UQ = x_{0.75}$**

$$UQ = x_{0.75} = x_{19} + 0.5 \cdot (x_{20} - x_{19}) = 44 + 0.5 \cdot (45 - 44) = 44.5.$$

**Замечания из лекции**

- Разные методы иногда дают **слегка разные** значения квартилей. Это случается, когда квартиль попадает **между** двумя наблюдениями $x_k$ и $x_{k+1}$ (то есть когда $a \neq 0$). Это нормально: строго по определению любой номер между ними «подходит», просто по соглашению принято выдавать одно число.
- Есть ещё термин **quantile / q-quantile**: это то же самое, что процентиль, но q задаётся в долях от 0 до 1. То есть q-квантиль = 100-процентиль. Например, 0.25-квантиль = 25-й процентиль.