09.04.03 Прикладная информатика

Профиль «Машинное обучение и анализ данных»

Дисциплина «Математические основы анализа данных»

Лекция 11

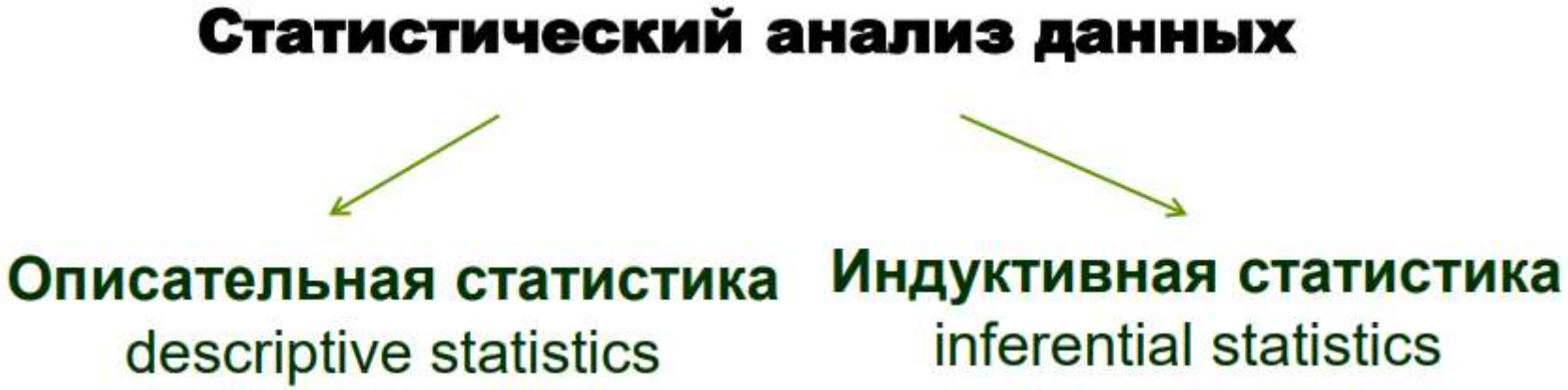
Введение в статистические методы анализа данных

План лекции

- Предмет математической статистики
- Классификация шкал измерения
- Статистические методы анализа данных
- Выборочные характеристики
  - Меры положения или центральной тенденции
  - Меры вариации или изменчивости
- Доверительные интервалы

О статистике

- ***Математическая статистика*** — это раздел математики, посвящённый методам сбора, анализа и обработки статистических данных для научных и практических целей.
- **Статистические данные** — данные, полученные в результате обследования **большого числа** объектов или явлений (то есть, математическая статистика имеет дело с массовыми явлениями).
- Предметом изучения в статистике являются изменяющиеся (варьирующие) признаки, которые иногда называют **статистическими признаками.**



- Наличие общего признака является основой для образования статистической совокупности.
- Таким образом, **статистическая совокупность** — это результат описания или измерения общих признаков объектов исследования.
- **Измерение** – это приписывание числовых форм объектам или событиям в соответствии с определёнными правилами.
- **Признаки и переменные** – это измеряемые явления. Значения признака определяются при помощи специальных шкал наблюдения.
- **Данные** – результаты некоторого количества измерений какой-либо одной или нескольких ПЕРЕМЕННЫХ (признаков).
- Например: размеры, вес, температура, прибыль, количество, …

Классификация шкал измерения

В 1946 году Стенли Стивенсом предложена классификация из 4 типов шкал измерения:

1) номинативная (номинальная) или шкала наименований;

2) порядковая (ординальная) шкала;

3) интервальная (шкала равных интервалов);

4) шкала равных отношений.

Тип шкалы	Действия в шкале	Пример
Отношений	Определение отношений между свойствами (качествами) объектов	Кофе в два раза дороже, чем мороженое.
Интервальная	Определение интервала между свойствами (качествами) объектов	Кофе дороже, чем мороженое на 50 рублей.
Порядковая	Становление отношений между качествами объектов	Кофе дороже, чем мороженое.
Номинативная	Наделение объектов именами	Это кофе. Это мороженое.

Стэнли Смит Стивенс


ДАННЫЕ
Категориальные (качественные)
Количественные (числовые)
Номинальные
nominal
Категории взаимоисключающие (альтернативные) и неупорядоченные (их нельзя выстроить в последователь-ность)
Порядковые
ordinal
Категории взаимоисключающие (альтернативные) и упорядоченные (могут быть упорядочены; размер интервалов на шкале неодинаковый)
шкала отношений
ratio scale
интервальная шкала
interval scale
Дискретные
discrete
Целочисленные значения, типичные для счета
Непрерывные
continuous
Любые значения в определенном интервале
Потеря информации и точности

Статистические методы анализа данных

- **Статистические методы** (методы, основанные на использовании математической статистики) являются эффективным инструментом сбора и анализа информации. Применение этих методов не требует больших затрат и позволяет с заданной степенью точности и достоверностью судить о состоянии исследуемых явлений (объектов, процессов), прогнозировать и регулировать проблемы на всех этапах их жизненного цикла и на основе этого вырабатывать оптимальные решения.
- **Графические методы** основаны на применении графических средств анализа статистических данных.
- В эту группу могут быть включены такие методы, как ***контрольный листок, диаграмма Парето, схема Исикавы, гистограмма, диаграмма разброса, расслоение, контрольная карта, график временного ряда*** и др.
- Данные методы не требуют сложных вычислений, могут использоваться как самостоятельно, так и в комплексе с другими методами. Находят самое широкое применение в промышленности, особенно в работе групп качества.

Статистические методы анализа данных

- **Методы анализа статистических совокупностей** служат для исследования информации, когда изменение анализируемого параметра носит случайный характер.
- Основными методами, включаемыми в данную группу, являются: регрессивный, дисперсионный и факторный виды анализа, метод сравнения средних, метод сравнения дисперсий и др.
- Эти методы позволяют установить зависимость изучаемых явлений от случайных факторов как качественную (дисперсионный анализ), так и количественную (корреляционный анализ); исследовать связи между случайными и неслучайными величинами (регрессивный анализ); выявить роль отдельных факторов в изменении анализируемого параметра (факторный анализ) и т.д.

Статистические методы анализа данных

- **Экономико-математические методы** представляют собой сочетание экономических, математических и кибернетических методов.
- Центральным понятием методов этой группы является ***оптимизация***, т. е. процесс нахождения наилучшего варианта из множества возможных с учётом принятого критерия (критерия оптимальности).
- Строго говоря, экономико-математические методы не являются чисто статистическими, но они широко используют аппарат математической статистики, что даёт основание включить их в рассматриваемую классификацию статистических методов.
- Из обширной группы экономико-математических методов следует выделить: математическое программирование (линейное, нелинейное, динамическое); планирование эксперимента; имитационное моделирование: теория игр; теория массового обслуживания; теория расписаний; функционально-стоимостной анализ и др.

Этапы анализа данных и их статистические методы

№ п/п	Этапы анализа данных	Статистические методы исследования
1.	Описание данных	Описательная статистика, определение необходимого объема выборки.
2.	Изучение сходств и различий	Статистические критерии: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи–квадрат, Фишера и др.
3.	Исследование зависимостей	Корреляционный анализ, дисперсионный анализ, регрессионный анализ.
4.	Снижение размерности	Факторный анализ, метод главных компонент.
5.	Классификация и прогноз	Дискриминантный анализ, кластерный анализ, группировка.

Выборочный метод

Пусть для получения опытных данных необходимо провести обследование некоторых объектов.

Примеры:

- Проверить качество выпускаемого некоторым предприятием продукта.
- Определить среднюю заработную плату по региону.
- Оценить заболеваемость населения данной болезнью.

- Обычно исследуют **не всю совокупность объектов**, а отбирают из неё некоторое количество объектов и исследуют только их.
- В этом и заключается **выборочный метод**.

- **Генеральной совокупностью** называют совокупность всех объектов, над которыми производят наблюдение.
- **Выборочной совокупностью (выборкой)** называют часть отобранных из генеральной совокупности объектов.
- **Объёмом совокупности** называют количество объектов в ней.

По выборке судят о генеральной совокупности.

- Выборка должна правильно представлять генеральную совокупность, то есть быть **репрезентативной**.
- Это обеспечивается способом отбора и увеличением объёма выборки.

Репрезентативность выборки

Способы отбора

1. Отбор, не требующий разделения генеральной совокупности на части:

    а) простой случайный бесповторный отбор,

    б) простой случайный повторный отбор.

2. Отбор, при котором генеральная совокупность разбивается на части:

    а) типический,

    б) механический,

    в) серийный.

3. Комбинированный отбор.

Выборки

- Выборка, в которой менее 30-ти элементов, называется **малой.**
- В противном случае, выборка называется **большой.**
- Для выборок малого объёма необходимо выбирать специально разработанные методы.

Выборочные данные делятся на:

а) качественные;

б) количественные.

- Качественные данные представляются (кодируются) определённым числом в соответствии с некоторым свойством.
- Для работы с качественными данными используются специально разработанные методы, которые называются **непараметрическими.**
- Методы, разработанные для количественных данных (**параметрические методы**), не могут использоваться для качественных данных.

Первичная обработка результатов наблюдений

После того, как данные получены:

1. Упорядочим по возрастанию значения переменной;
2. Разобьём их на группы по равным интервалам;
3. Получаем **вариационный ряд.**

- **Вариационный ряд** – ряд, в котором сопоставлены (по степени возрастания или убывания) варианты и соответствующие им частоты.
- **Вариантами** (x_i) считаются отдельные значения признака, которые он принимает в вариационном ряду.
- **Частота** (n_i) – число, показывающее, сколько раз повторяется варианта.

Относительной частотой (частотностью) варианты называют отношение частоты к объёму выборки:

$$w_i = n_i / n.$$

Пример 1:

0 1 2 2 1 2 0 0 0 0 – выборка

Объём выборки: $n = 10$

x_i	0	1	2	перечень значений
n_i	5	2	3	частота встречаемости
w_i	0.5	0.2	0.3	относительная частота

Замечания по описанию вариационного ряда

1. Сумма всех частот равна объёму выборки:

$$\sum_i n_i = n$$

2. Сумма всех относительных частот равна 1:

$$\sum_i w_i = 1$$

3. Относительная частота варианты даёт приближённое значение вероятности этой варианты:

$$p(X = x_i) \approx w_i$$

Непрерывные значения варианты

- Если наблюдаемые данные имеют непрерывный характер, то перечень вариант обычно очень велик и одинаковые значения вариант встречаются очень редко $\Rightarrow$ дискретный статистический ряд неудобен
- Составляют **интервальный статистический ряд:**

1. Разбивают весь интервал, в который попадают варианты, на частичные интервалы:

На сколько интервалов разбивать выборку?

Формула Стерджесса (Sturges): $k = 1 + [3.332 \cdot \lg n]$ или $k \leq 5 \cdot \lg n$, где n – объём выборки

2. В верхнюю строку записывают полученные интервалы;
3. В нижнюю строку записывают частоту попадания в соответствующий интервал.

Пример 2: Пусть имеются следующие данные

27 3,5 21,1 0,8 12,3 18 11 3,4 1,2 5,2 22 17,2 18,1 11,1 0,7 7,9 19 3,2 4,9 25,4 6,1 21,6 22,3 3,4 18,4 3,4 23,2 13,1 6,5 2,4 18,4 14,1 2,1 24,8 17,4 15,1 4,8 19,8 10,4 16,1 3,7 29,4 3,1 28,7 16,4 22,2 1,7 12,4 17 15,3 3,3 14 16,8 10,1 2,4 20 14,1 19 19,8 5,4 2,5 4,1 24,4 0,4 24,7 1,3 13,7 0,1 28 24 17,1 15 3,1 19 0,4 23,1 6,7 4,6 14,8 20,7 16,2 9,4 21,3 13,4 16,1 15,7 11,3 5,1 1,9 2,8 17 2 20,8 3,4 16,7 9,3 15,2 8,7 10,7

Диапазон значений: числа от 0 до 30.

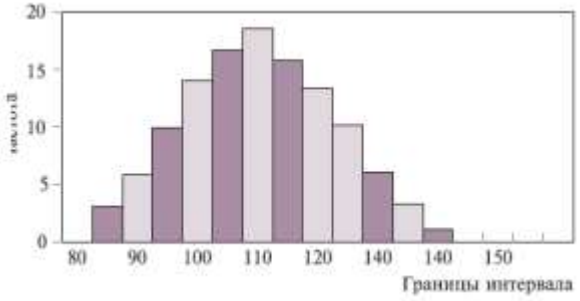
6 интервалов: [0, 5);[5, 10);[10, 15);[15, 20); [20, 25);[25, 30)

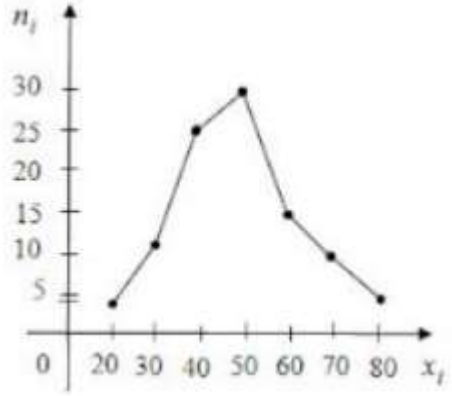
x_i	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
n_i	30	10	15	25	15	5
w_i	0.3	0.1	0.15	0.25	0.15	0.05

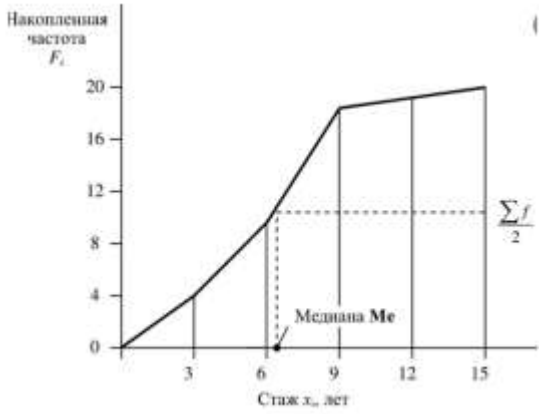
Визуализация данных

Наиболее употребительными графиками для изображения вариационных рядов, т. е. соотношений между значениями признака и соответствующими частотами или относительными частотами, являются:

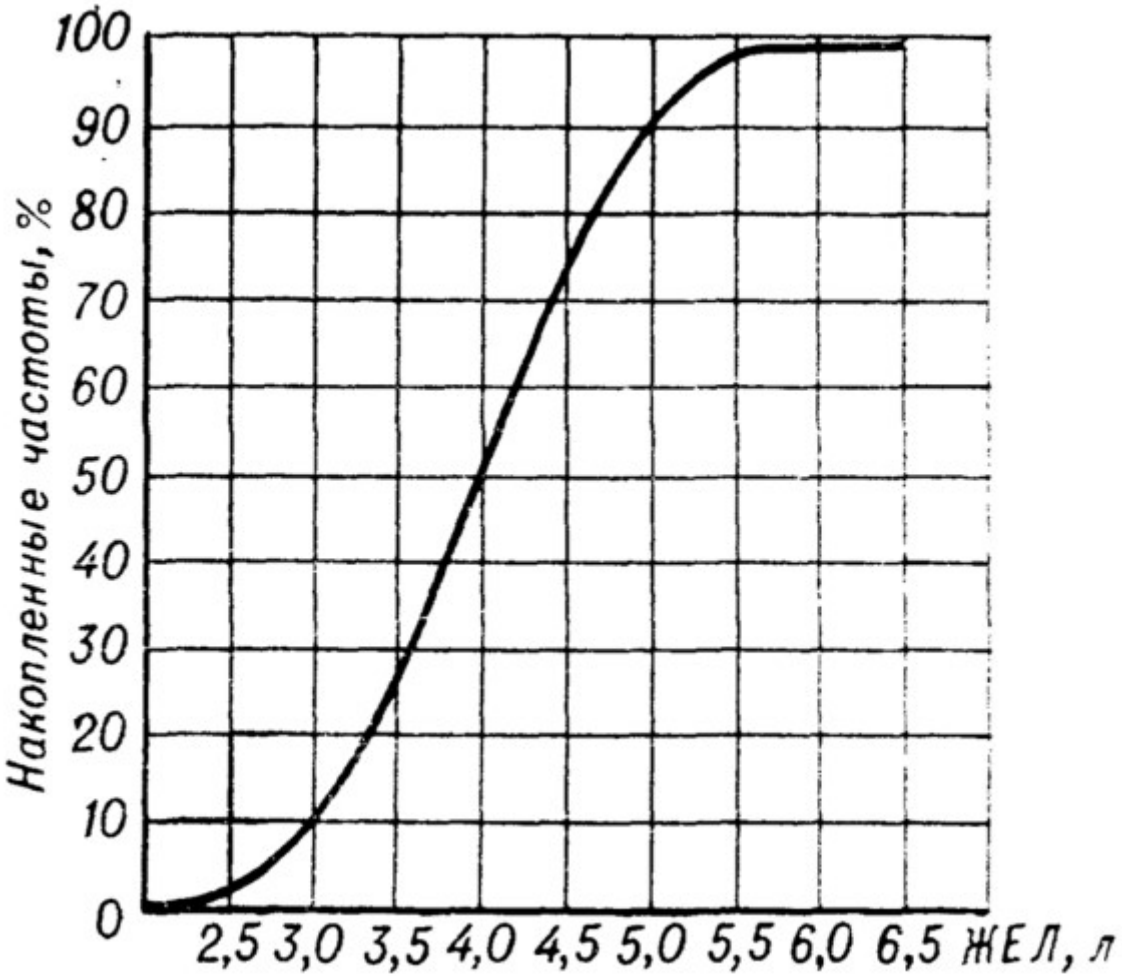
- гистограмма
- полигон
- кумулята







Кумулята служит для графического изображения кумулятивного (с накоплением) вариационного ряда.

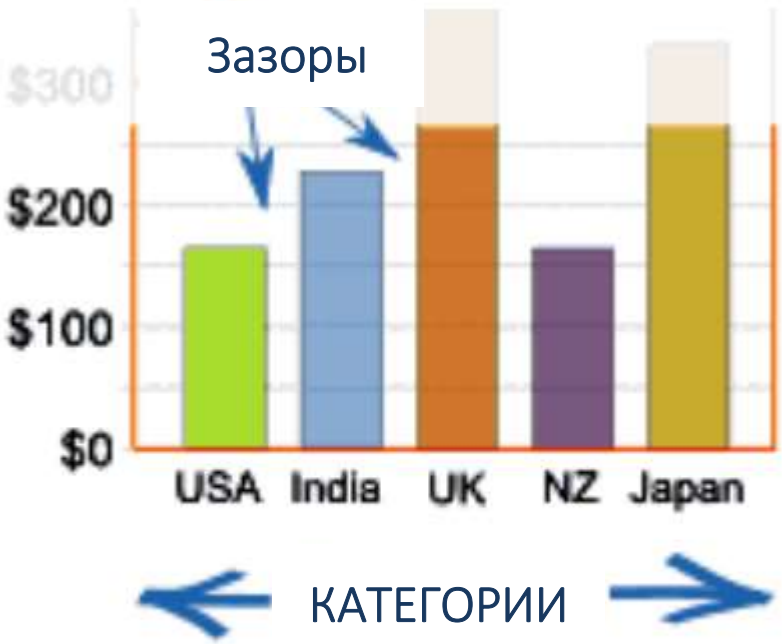


На оси абсцисс откладывают значения аргумента, а на оси ординат - накопленные частоты.

Строят точки, абсциссы которых равны вариантам (в случае дискретных рядов) или верхним границам интервалов (в случае интервальных рядов), а ординаты - соответствующим частотам (накопленным частотам).

Эти точки **соединяют отрезками прямой**.

Отличие графического представления распределения качественных и количественных признаков



ГИСТОГРАММА

Основные характеристики выборки

1. Меры положения или центральной тенденции
2. Меры вариации или изменчивости
3. Меры формы

- Цель – оценить параметры генеральной совокупности с помощью показателей из выборки

Меры положения или центральной тенденции распределения

Все они могут служить оценками выборочного среднего.

Среднее в выборке – наиболее эффективная и **несмещённая оценка**.

Основные характеристики выборки

1. **Среднее значение** – сумма всех значений переменной, делённая на количество значений в выборке

$$\overline{X} = \frac{\sum_i X_i}{n}$$

2. **Медиана** – значение, которое делит распределение пополам: половина значений больше медианы, половина – не больше.
3. **Мода** – наиболее часто встречающееся значение.

Пример 3: Для данных

27 3,5 21,1 0,8 12,3 18 11 **3,4** 1,2 5,2 22 17,2 18,1 11,1 0,7 7,9 19 3,2 4,9 25,4 6,1 21,6 22,3 **3,4** 18,4 **3,4** 23,2 13,1 6,5 2,4 18,4 14,1 2,1 24,8 17,4 15,1 4,8 19,8 10,4 16,1 3,7 29,4 3,1 28,7 16,4 22,2 1,7 12,4 17 15,3 3,3 14 16,8 10,1 2,4 20 14,1 19 19,8 5,4 2,5 4,1 24,4 0,4 24,7 1,3 13,7 0,1 28 24 17,1 15 3,1 19 0,4 23,1 6,7 4,6 14,8 20,7 16,2 9,4 21,3 13,4 16,1 15,7 11,3 5,1 1,9 2,8 17 2 20,8 **3,4** 16,7 9,3 15,2 8,7 10,7

Среднее значение: 12,48383838

Медиана 13,7

Мода 3,4

- Если распределение не симметричное, то медиана лучше характеризует центр распределения.
- Медиана содержит меньше информации, чем среднее, поскольку определяется только рангом измерений, а не их значениями.
- Медиана может применяться даже в случае, если измерения в выборке не точные.

Квартили и процентили

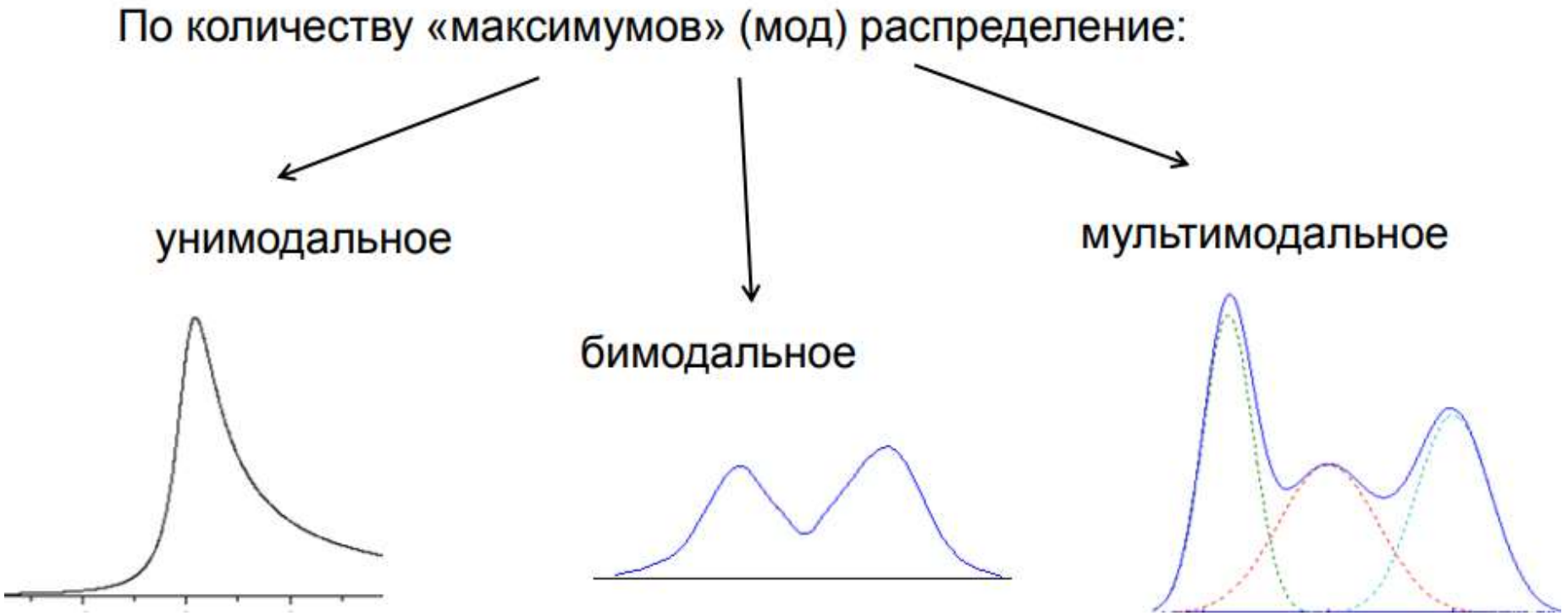
Медиана делит выборку на **две** части, но распределение можно поделить на:

- четыре (значения, стоящие на границах - квартили);
- восемь (... октили);
- сто (... процентили);
- N (... квантили).
- Квартили (quartiles) делят распределение на четыре части так, что в каждой из них оказывается поровну значений (2-я квартиль = медиана).

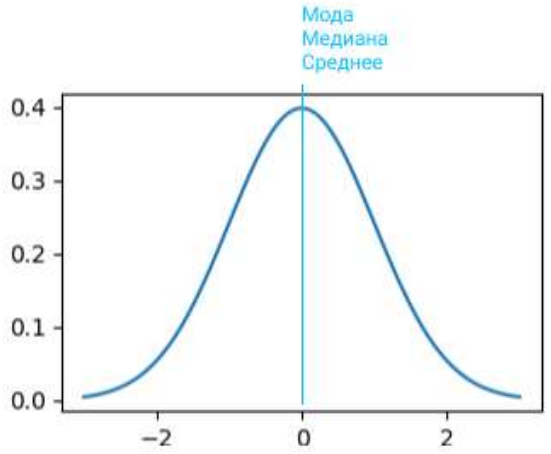
- Межквартильный размах – разница между третьей и первой квартилями.

Мода существует не только для количественных, но и для ранговых, и для качественных переменных.

В исследованиях интересует прежде всего количество мод в распределении, а не мода как таковая.



- Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения.



- К появлению перекоса чувствительнее всего среднее значение