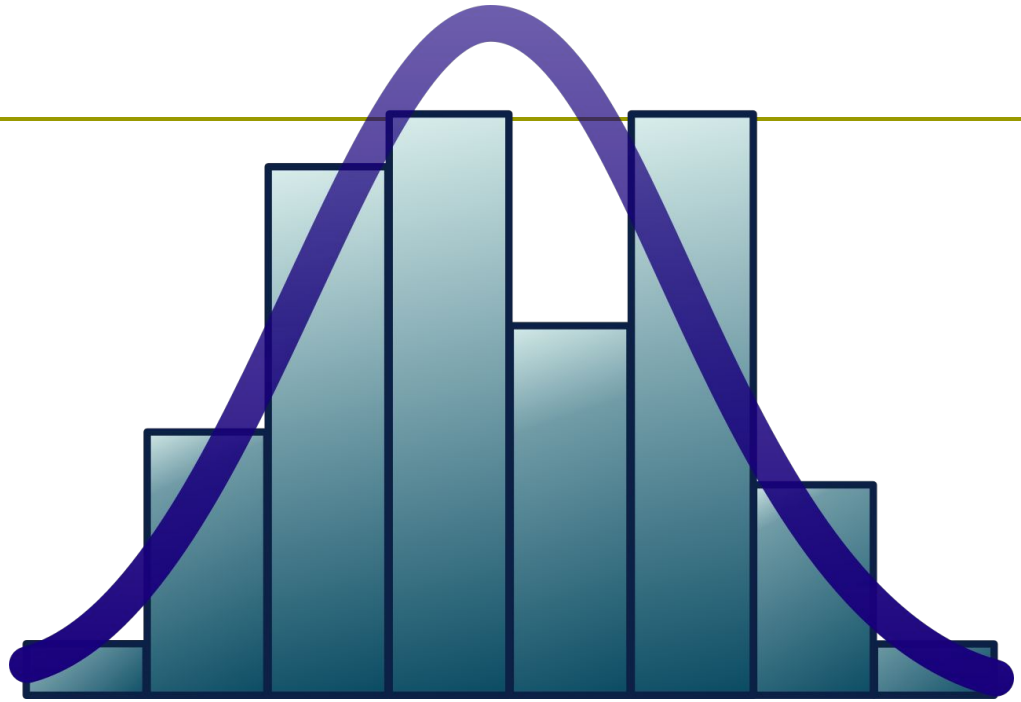
Описательная статистика и компьютерные технологии статистической обработки эмпирических данных



ПОНЯТИЕ О МАТЕМАТИКО-СТАТИСТИЧЕСКОМ АНАЛИЗЕ ДАННЫХ

СТАТИСТИКА

- Слово «статистика» имеет латинское происхождение (от status — состояние)
- XVII –XVIII в. – «государствоведение»

Первой опубликованной статистической информацией можно считать глиняные таблички Шумерского царства (III — II тысячелетия до н. э.).

«Существуют три вида обмана: ложь, наглая ложь и статистика»

Б. Дизраэли, премьер-министр Великобритании

Статистика как область деятельности

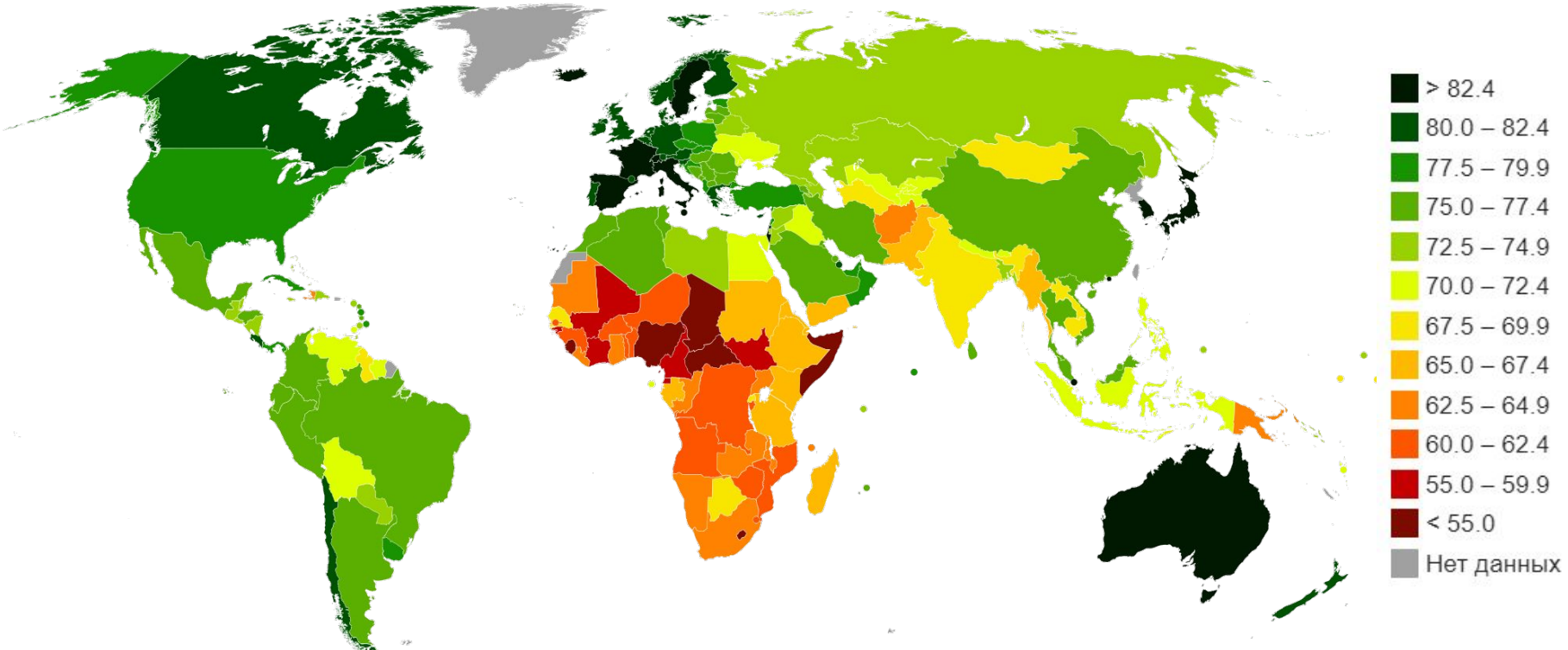
- статистика — отрасль практической деятельности, целью которой является ***сбор, обработка и анализ данных*** о разнообразных явлениях общественной жизни

- полученная в результате статистического исследования информация позволяет решать задачи выявления реально существующих ***закономерностей,*** свойственных описываемым процессам и явлениям

- Пример. Ожидаемая продолжительность жизни при рождении

Ожидаемая продолжительность жизни при рождении в РБ

	1995	2000	2005	2010	2015
Все население	69,7	70,8	72,3	73,5	76,3
мужчины	63,9	65,3	67	68,0	71,1
женщины	75,1	76,0	77,2	78,4	80,5



Совокупность и закономерность

- Предметом изучения в статистике являются ***совокупности***: группы населения, потребительские товары, районы страны и т.п.
- Статистика дает ***количественную*** характеристику исследуемой закономерности
- Пример. Продолжительность жизни для закономерности «женщины живут дольше мужчин»

Признаки совокупности

- Статистика изучает явления через **признаки**: возраст, образование, пол для человека; форма собственности, уставной капитал для предприятия

- Признаки различаются способами их **измерения** и некоторыми другими особенностями

Измерения и шкалы

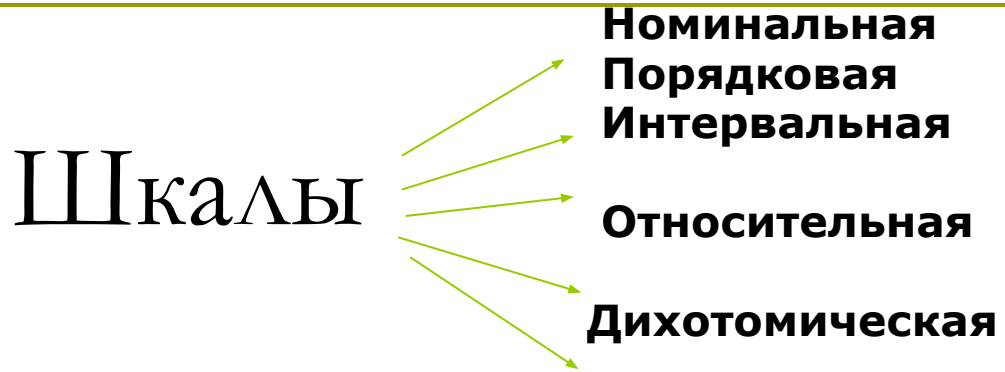
- **Измерение** означает присвоение чисел характеристикам изучаемых объектов, явлений согласно некоторому правилу

- **Шкала** (лат. *scala* – лестница) – упорядоченное множество действительных чисел (индексов), соответствующих последовательному ряду возможных значений измеряемой величины

Шкалы

- **Номинальная** Содержит только категории, данные не могут упорядочиваться
  Хобби студента
- **Порядковая** Содержит категории, которые могут упорядочиваться, разности не имеют смысла
  Место на соревнованиях
- **Интервальная** Разности между значениями могут быть вычислены, но отсутствует точка отсчета
  Температура тела
- **Относительная** Имеется точка отсчета, возможны отношения между значениями
  Рост студента
- **Дихотомическая** Содержит две категории
  Пол студента

Пример. Какой тип шкалы?



- Температура воздуха в лекционной аудитории?
- Возраст студента?
- Пол студента?
- Семейное положение?
- Религиозные предпочтения?
- Время на подготовку домашнего задания?
- Трудолюбие?
- Традиционная система педагогических оценок (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)?

ТИПЫ ДАННЫХ

Качественные

номинальная шкала

Ранговые

порядковая шкала

Количественные

шкала отношений

интервальная шкала

дискретные

непрерывные

Допустимые преобразования:
всегда возможен переход от более мощной шкалы к менее мощной, но не наоборот

Потеря информации и точности

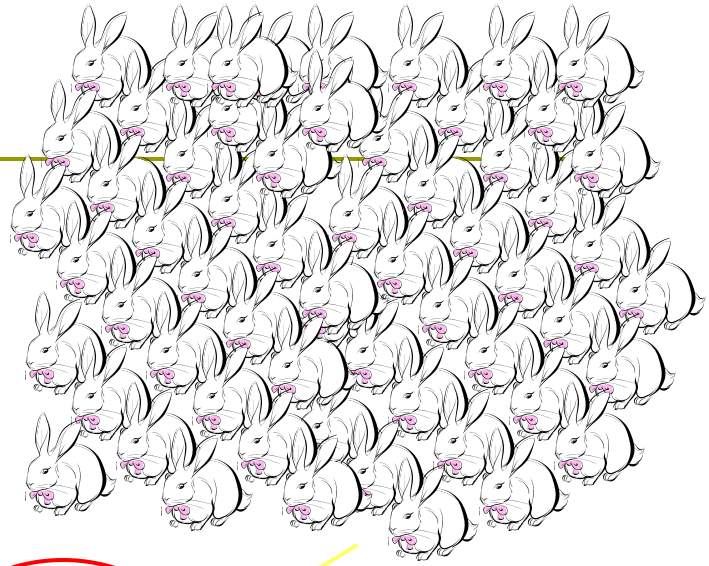
Популяция и выборка

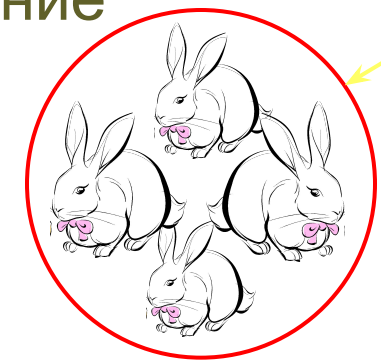
Популяция (population) - *совокупность всех субъектов, обладающих интересующим исследователя признаком (признаками) или свойством (свойствами).*

• **Выборка (sample)** *– это часть популяции, которая исследуется на практике, и результаты оценки которой исследователь распространяет на всю популяцию.*

наблюдение

популяция





выборка

Популяция ≠ выборка

Формирование выборки

Простая случайная выборка (simple random sample) – *это выборка, полученная путем случайного отбора членов генеральной совокупности методом жеребьевки при помощи генератора случайных чисел или таблиц случайных чисел.*

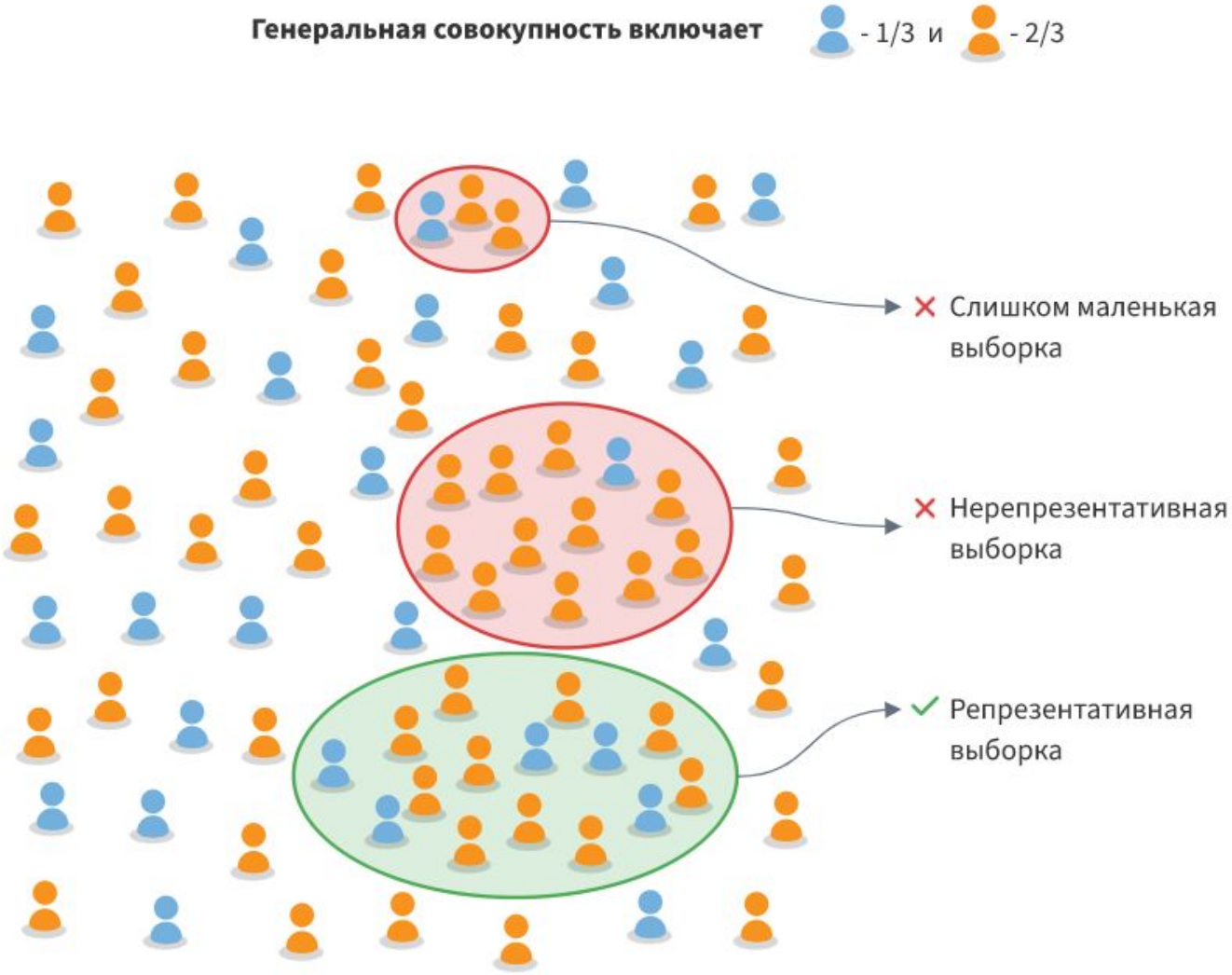
Типическая выборка (стратифицированная) – *предполагает разделение неоднородной генеральной совокупности на типологические группы по какому-либо признаку, после чего из каждой группы производится случайный отбор единиц*

Механическая выборка – *отбор единиц через равные промежутки (по алфавиту, через временные промежутки, по пространственному способу)*

Репрезентативная выборка (representative sample) -

корректно отражает генеральную совокупность

Репрезентативность выборки

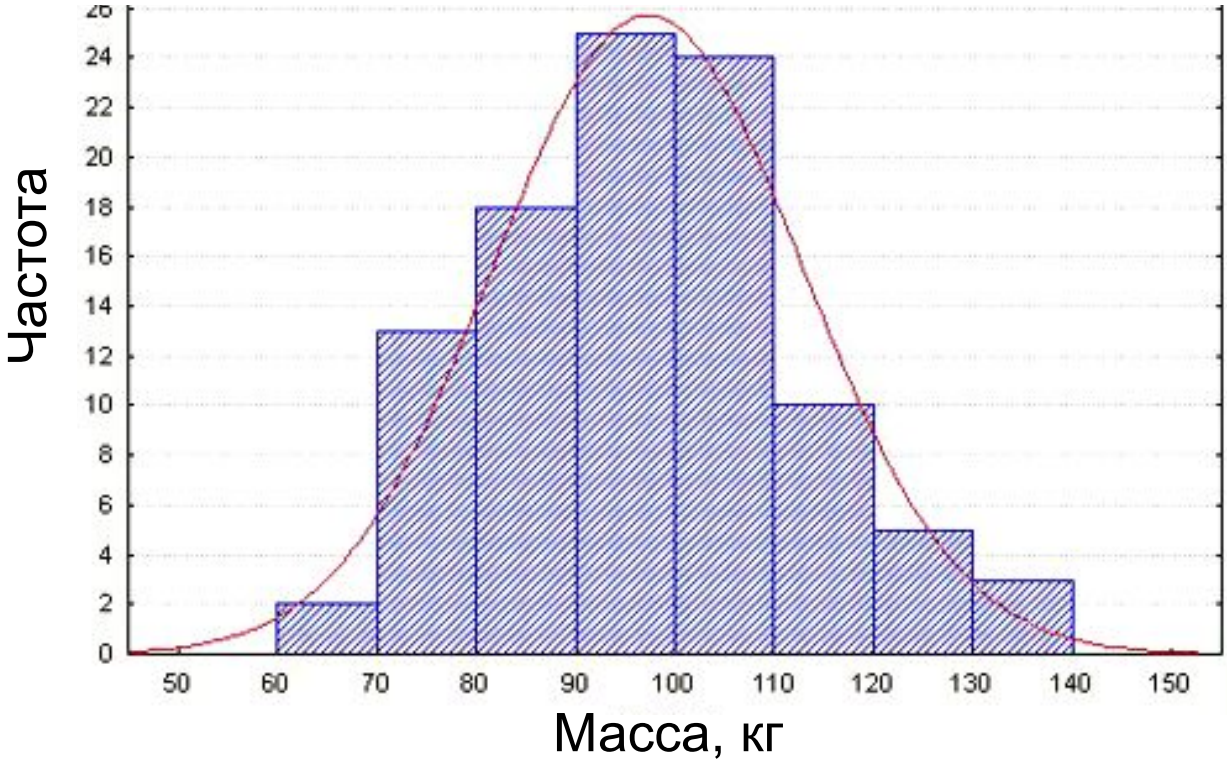


Репрезентативность выборки

Частотное распределение переменной

Частота – сколько раз встретилось данное значение переменной

Гистограмма – графическое представление частотного распределения, разбитого по интервалам, где высота столбика отражает **ЧАСТОТУ** появления варианты в выборке



Разделы исследовательского анализа данных

Исследовательский анализ данных - Exploratory Data Analysis (EDA) представляет собой применение статистических методов для представления, упорядочения данных и понимания их важнейших характеристик.

Основными разделами анализа являются:

1. **Показатели, характеризующие центральную тенденцию**. Вычисление и анализ среднего, моды, медианы.
2. **Показатели, характеризующие вариации вокруг центральной тенденции**. Нахождение дисперсии, стандартного отклонения.
3. **Меры положения**. Минимум, максимум, размах, нахождение квартилей.
4. **Выбросы**. Нахождение и анализ выбросов.
5. **Форма распределения**. Асимметрия и эксцесс.

Анализ данных: измерение центральной тенденции

Мера центральной тенденции – это числовой показатель, которых характеризует наиболее типичные значения переменной в выборке или популяции.

Измерение центральной тенденции состоит в выборе одного числа, которое **наилучшим образом** описывает все значения признака из набора данных.

- **Мода**
- **Медиана**
- **Среднее значение**

Мода

Мода – наиболее часто встречающееся значение в выборке, наборе данных. Обозначается ***Mo***.

Выборка: 5,4 1,2 0,42 1,2 0,48

Мода=1,2

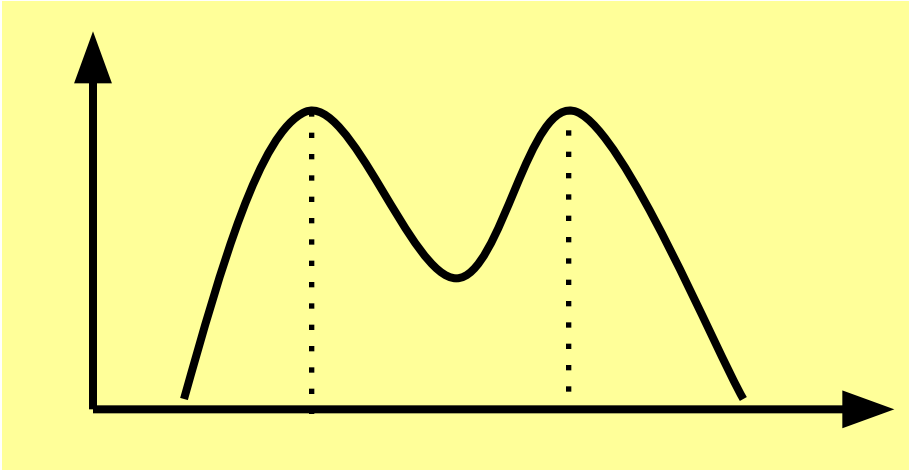
Для данных, расположенных в таблице частот, мода определяется как значение, имеющее наибольшую частоту.

Найдите моду:

1,2,2,3,3,3,3,4,4,4,4,4,5,5,5,5,5,5,5,6,6,6,6,7,7

Одна ли мода?

Если наибольшую частоту имеет два значения выборки, выборочное распределение называется **бимодальным**.



Если наибольшую частоту имеет более двух значений выборки, выборочное распределение называется **мультимодальным**.

Если ни одно из значений не повторяется, **мода отсутствует**.

Свойства моды

1. Наличие одного или двух крайних значений, сильно отличающихся от остальных, не влияет на значение моды.

2. Мода совпадает с точкой наибольшей плотности данных.

3. Мода может иметь несколько значений.

4. Мода может существовать для всех типов данных. Это единственная мера, которая работает в номинальной шкале!

Медиана

Медиана есть значение серединного элемента для набора данных. Для нахождения медианы требуется составить вариационный ряд, то есть расположить все значения признака в порядке возрастания или убывания. Медиана расположена в середине вариационного ряда.

Для набора из n значений, если n нечетно, средний элемент имеет номер:

$$\frac{n-1}{2}$$

Если n четно, медиана находится как среднее арифметическое двух соседних серединных элементов:

$$\frac{n}{2} \quad \frac{n}{2}+1$$

Пример вычисления медианы

Для набора данных из семи чисел:

6 1 3 7 1 7 3

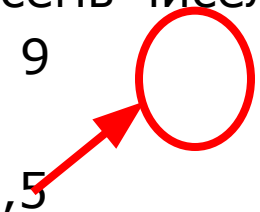
После упорядочения получим вариационный ряд:

1 1 3 3 6 7 7

Медиана есть средний элемент. Его номер четвертый.

Если набор данных включает восемь чисел:

1 1 3 3 6 7 7 9



Тогда медиана равна (3+6)/2=4,5

Свойства медианы

1. Сильно отличающиеся от остальных данных крайние значения не влияют на величину медианы.

2. Значение медианы является единственным для каждого набора данных.

3. Медиана может быть определена не из полного набора данных. Достаточно знать их расположение, общее число и несколько значений, расположенных в середине вариационного ряда.

4. Медиана может быть определена для числовых данных и данных, измеряемых порядковой шкалой. Для порядковой шкалы в случае четного количества элементов оба серединных значения объявляются медианой.

Среднее значение

Выборочное среднее будем называть среднее арифметическое выборки, то есть сумму всех значений выборки, деленную на ее объем.

Формула: $$\bar{x} = \frac{\sum x}{n}$$

где $\sum x$ = сумма всех значений выборки

n = объем выборки

Свойства среднего

1. Вычисляется только в числовых шкалах.

2. При ее вычислении необходимо использовать все данные.

3. Имеется для каждого набора данных только одно значение средней.

4. Средняя есть единственная мера центральной тенденции, для которой сумма отклонений каждого значения от нее равна нулю:

$$\sum(x - \bar{x}) = 0$$

Среднее для сгруппированных данных

Среднее для сгруппированных данных вычисляется по формуле:

$$\bar{x} = \frac{\sum f \cdot x}{\sum f}$$

где $\sum f \cdot x$ = сумма всех значений выборки

$\sum f$ = сумма частот, равна объему выборки

Если данные сгруппированы по интервалам, в качестве значения выбирается середина интервала.

Пример вычисления среднего для сгруппированных данных

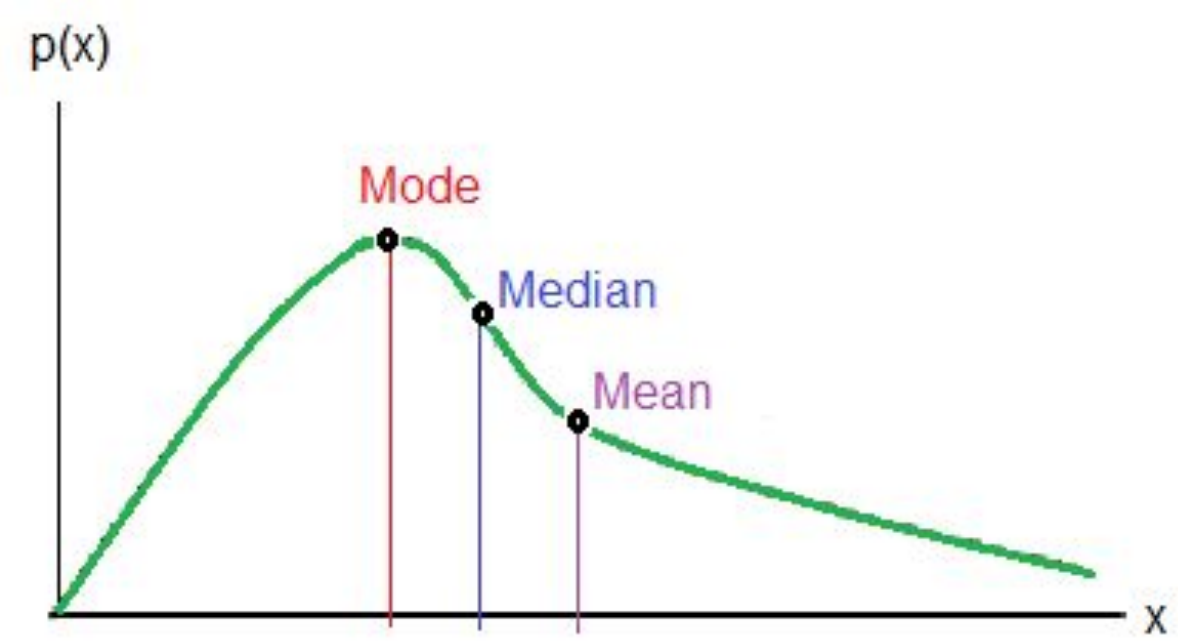
Имеются результаты экзамена. Найти среднее значение.

x	f	f·x
0	1	0
1	2	2
2	6	12
3	12	36
4	3	12
5	1	5
	25	67

$$\bar{x} = \frac{\sum f \cdot x}{\sum f} = \frac{67}{25} = 2{,}68$$

«Середина» распределения

Мода, медиана и среднее СОВПАДАЮТ для симметричного унимодального распределения



К появлению перекоса чувствительнее всего среднее значение

Три меры и тип шкалы

Три меры меры центральной тенденции накладывают ограничения на тип шкалы, в которой измеряется переменная.

Типическое значение	Номинальные данные	Порядковые данные	Интервальные данные
Мода			
Медиана			
Среднее			

Среднее для дихотомической шкалы

Среднее может также применяться и для переменной, измеренной в дихотомической шкале.

Если два значения признака кодируются 0 и 1, то среднее указывает долю (относительную частоту) единиц в выборке.

Пример.

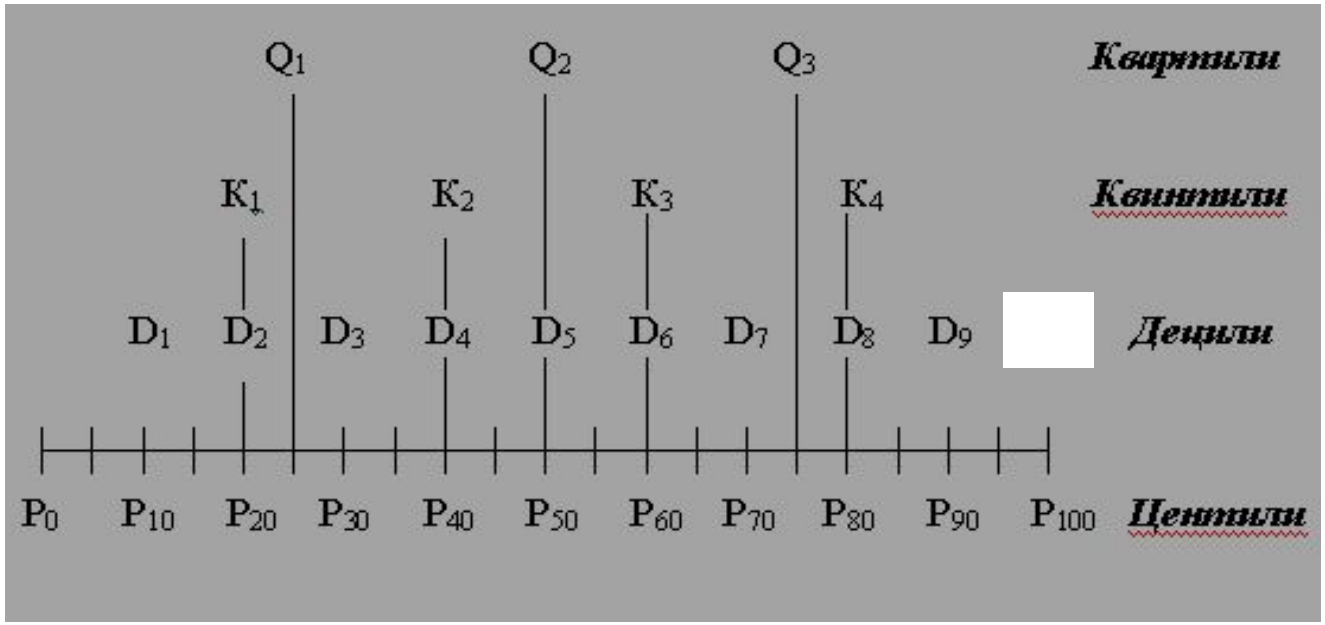
1, 0, 0, 0, 1, 1, 1, 1, 1, 0

Среднее равно 0,6. То есть 60% значений выборки принимают значение, равное единице.

Другие меры положения

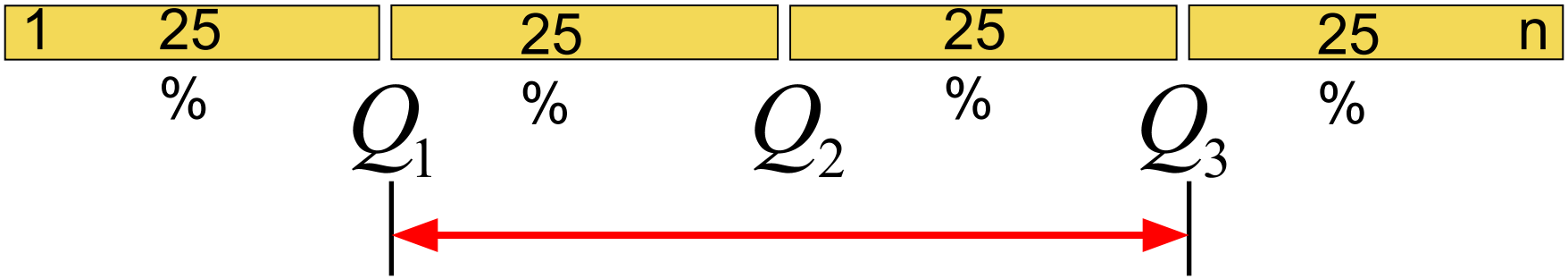
Квантиль – это точка на числовой оси, на которой откладываются результаты наблюдений. Эта точка делит всю совокупность наблюдений на части (группы) с определенными пропорциями между ними.

- Квартили
- Центили
- Квинтили
- Децили



Квартили (Quartile)

Под квартилями понимаются значения, которые делят вариационный ряд на четыре равные части:



Ниже первого квартиля расположено 25% всех данных. Между первым и вторым квартилем также расположено 25% данных. Второй квартиль совпадает с медианой.

Размах квартилей (InterQuartile Range) вычисляется по формуле:

$$IQR = Q_3 - Q_1$$

Пример определения медианы и квартилей

Определим медиану и квартили для признака X – суммы, баллов, набранной студентами при изучении дисциплины по группе из 15 человек.

X	30	41	42	**44**	45	47	50	**54**	55	55	58	**59**	60	62	65
№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Первая квартиль= 25–я процентиль

Медиана

Третья квартиль = 75-я процентиль

Мера изменчивости

Мера изменчивости – это числовой показатель, который характеризует вариацию (разброс) значений совокупности:

- размах,
- интерквартильный размах,
- дисперсия,
- стандартное отклонение,
- коэффициент вариации.

Пример: рассмотрим три вариационных ряда:

а) 999, 1000, 1001

б) 900, 1000, 1100

в) 1, 1000, 1999

В каком случае разброс значений больше?

Как выразить степень разброса одним числом?

Размах (Range)

Размах – разность между наибольшим значением набора данных и наименьшим.

$$R = x_{\max} - x_{\min}$$

Пример: Для набора данных 27, 8, 3, 12, 10, 26, 6, 19 размах равен R = 27 – 3 = 24.

Дисперсия (Variance)

Дисперсия выборки – среднее арифметическое квадратов отклонений значений выборки от их среднего.

Вычисляем по формуле:

$$D(x) = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Стандартное отклонение (standard deviation)вычисляется как корень из дисперсии:

$$s = \sqrt{D(x)}$$

Вторая формула для дисперсии

- Дисперсия вычисляется также по равносильной формуле:

$$D(x) = \frac{n \cdot \sum x^2 - \left(\sum x\right)^2}{n \cdot (n-1)}$$

- Считается, что эта формула более пригодна для практических вычислений при ручном счете и при использовании электронных таблиц.
- Не требуется вычислять среднее!!!

Коэффициент вариации

Коэффициент вариации вычисляется как отношение стандартного отклонения к среднему:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Коэффициент вариации считается

- слабым, если $CV \leq 10\%$,
- средним, если $10\% < CV \leq 33\%$,
- значительным, если $CV > 33\%$.

Пример для коэффициента вариации

Какие данные имеют большую вариацию:

имеющие стандартное отклонение 20 при среднем 200 или имеющие стандартное отклонение 3 при среднем 30?

$$CV = \sigma / \bar{x} = 3/30 = 0{,}1 \ \ (10\ \%)$$

$$CV = \sigma / \bar{x} = 20/200 = 0{,}1 \ \ (10\ \%)$$

Ответ. Коэффициенты вариации равны. Вариация одинакова.

Коробковая диаграмма (Boxplot)

Диаграмма, основывающаяся на вычислении и построении пяти характеристик. Удобна для анализа данных и используется очень часто.

Минимум Q_1 Медиана Q_3 Максимум

Коэффициент асимметрии

Асимметрия является мерой несимметричности распределения. Если этот коэффициент значительно отличается от 0, распределение является асимметричным

Коэффициент асимметрии находится по следующей формуле:

$$A = \frac{\sum (x - \bar{x})^3}{N s^3}$$

Изменяется в пределах от -3 до 3.

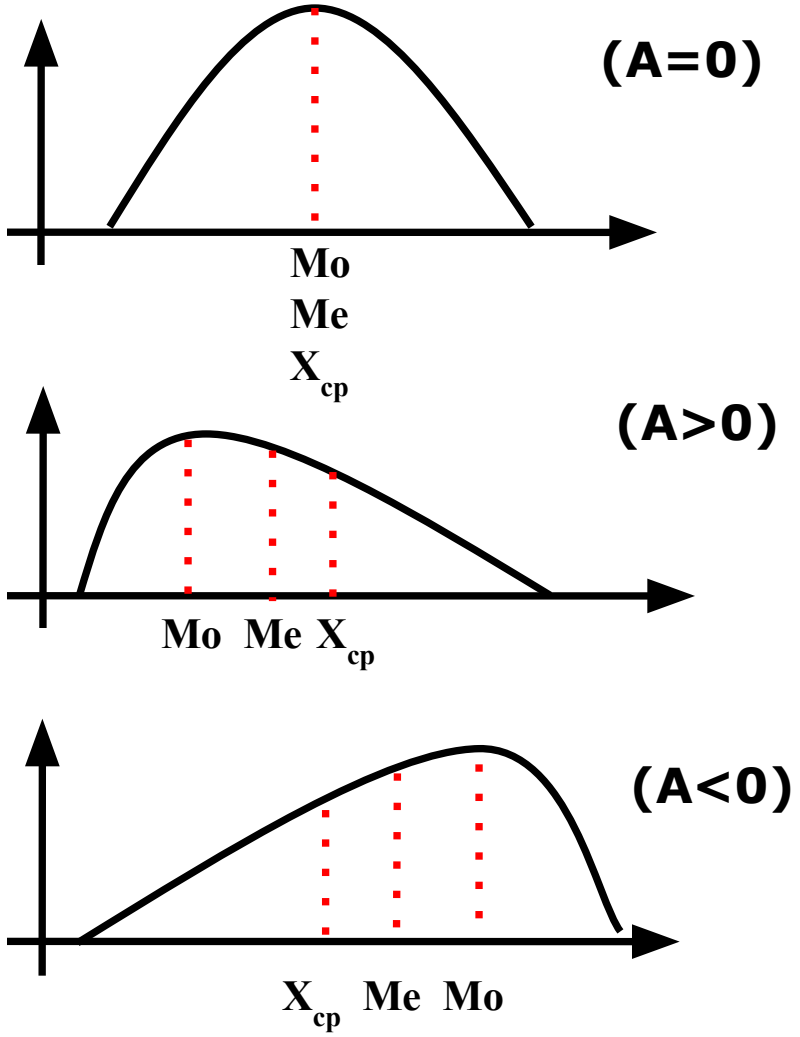
- |A |≤ 0,25 – слабая асимметрия,
- 0,25 < |A| ≤ 0,5 – умеренная асимметрия,
- |A| > 0,5 – крайне асимметричное распределение.

Асимметрия (Skewness)

- Если распределение симметрично, **асимметрия равна нулю**. В этом случае совпадают значения моды, медианы и среднего арифметического.

- Если одно или несколько значений существенно превышают остальные, имеется **положительная асимметрия**. Средняя больше моды и медианы.

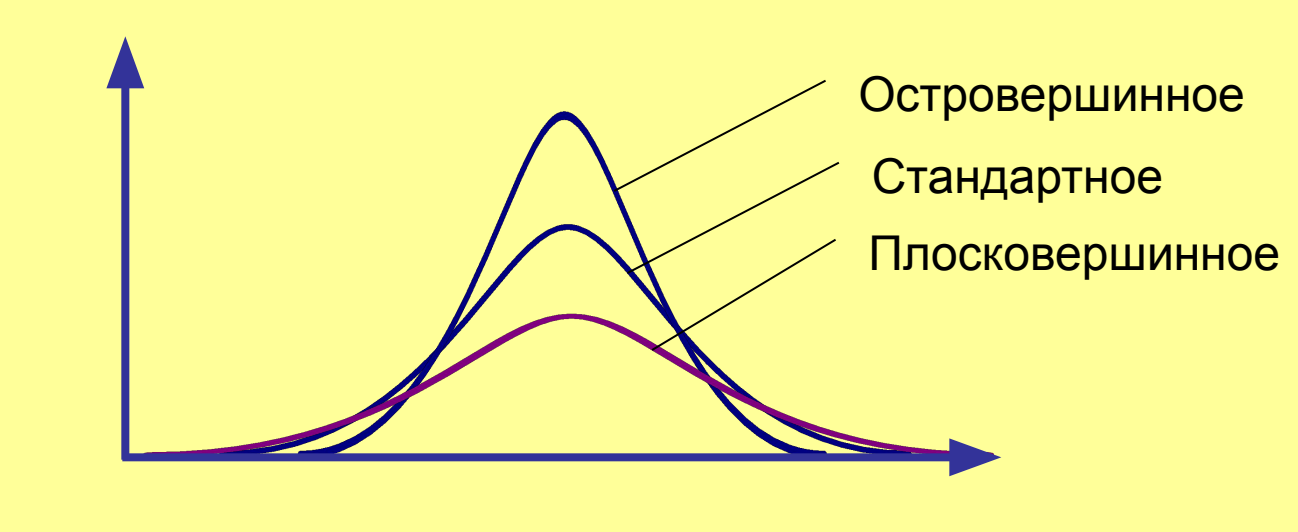
- Если одно или несколько значений существенно меньше остальных, имеется **отрицательная асимметрия**. Средняя меньше моды и медианы.



Эксцесс (Kurtosis)

$$E = \frac{\sum (x - \bar{x})^4}{N s^4} - 3$$

Эксцесс измеряет остроту пика распределения
Для нормального распределения E = 0.



|E| <0,2 – практически эксцесс отсутствует,

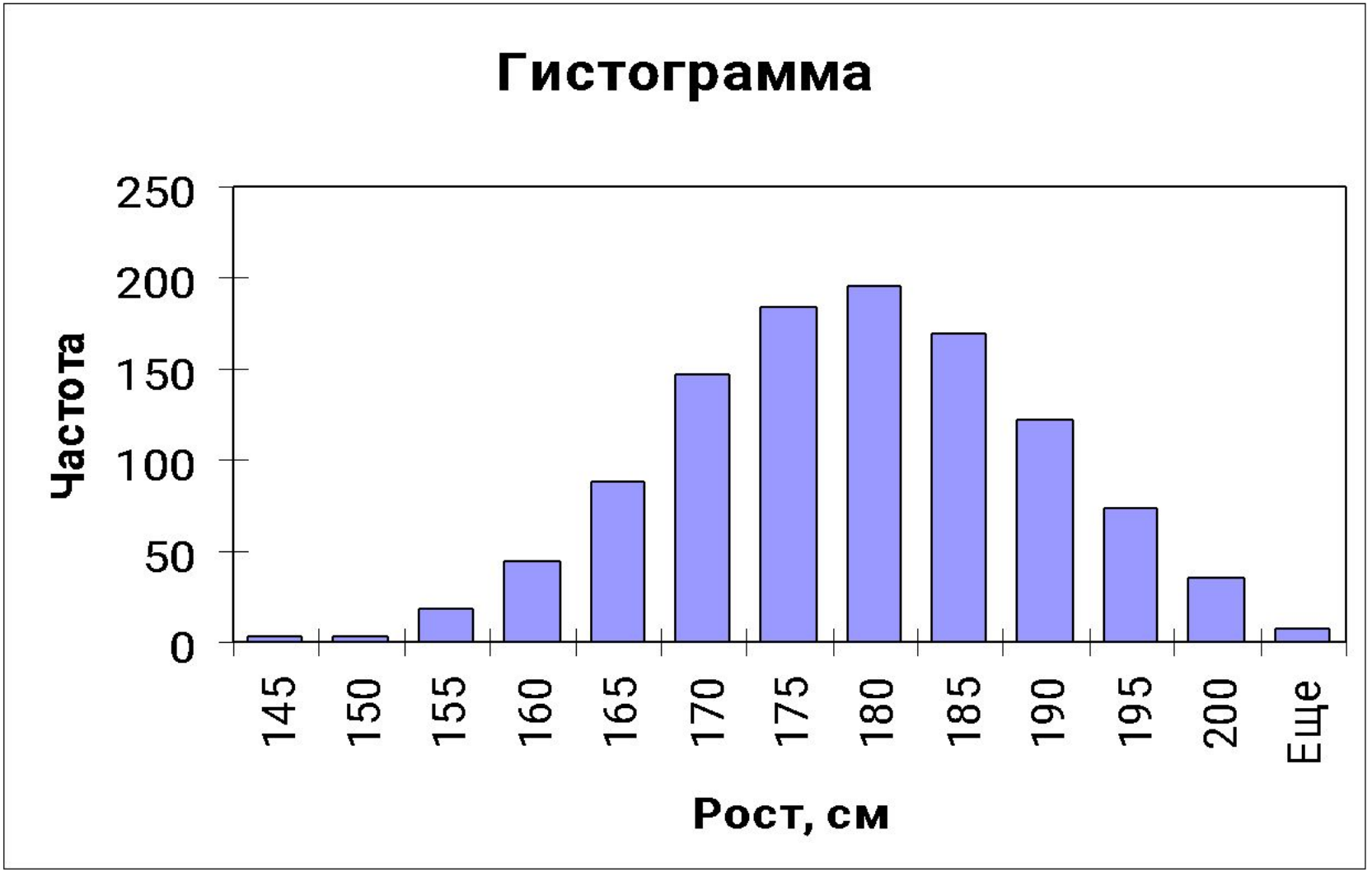
|E| = 0,2–0,3 – слабый эксцесс,

|E| = 0,3–0,6 – умеренный эксцесс

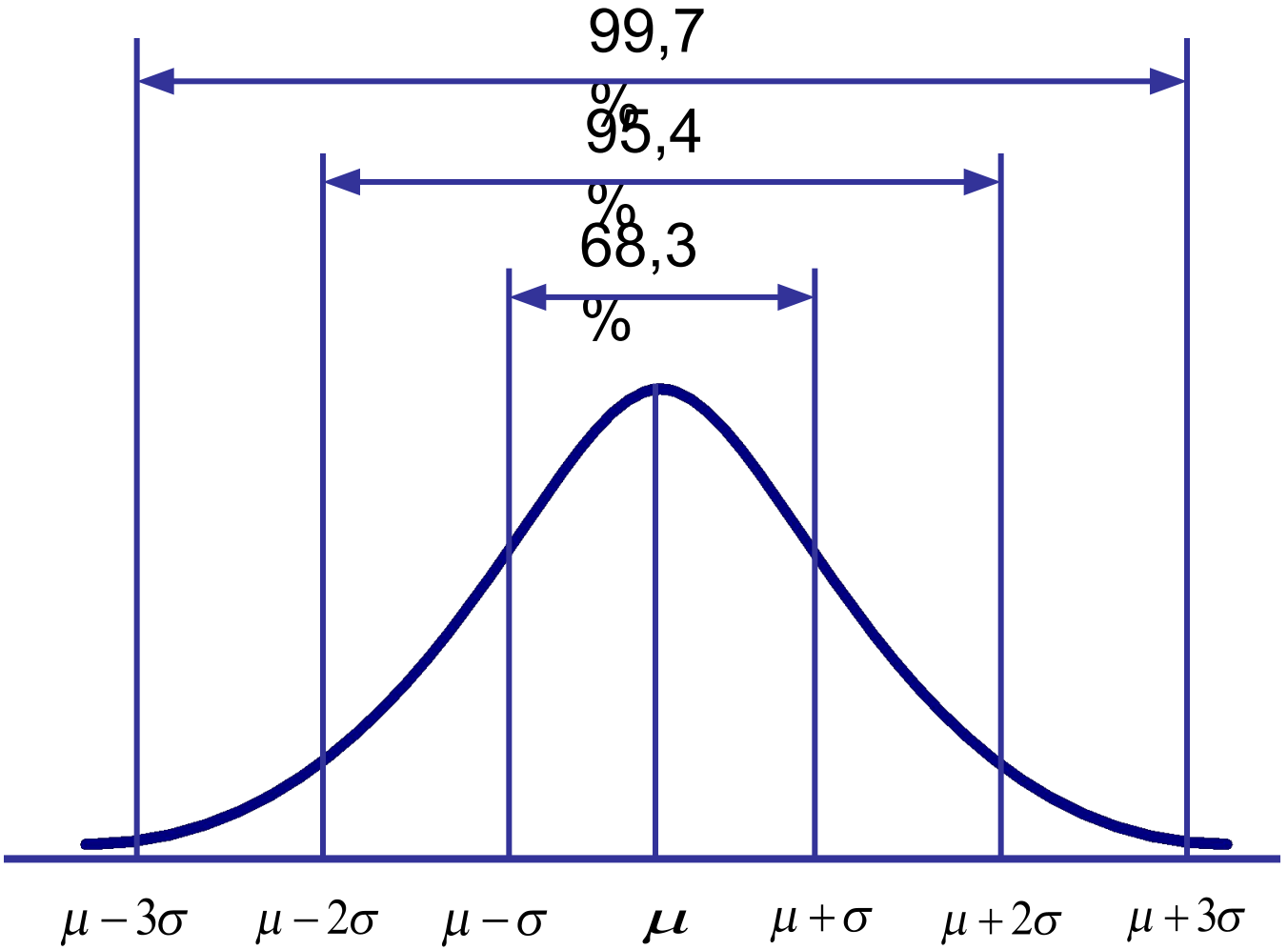
|E| = 0,6–1,0 – сильный эксцесс,

|E| > 1 – очень сильный эксцесс

Нормальное распределение



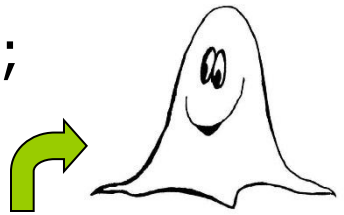
Нормальное распределение



Как определить, является ли распределение признака нормальным?

- ✔ Построить гистограмму, оценить визуально:
  - нормальное распределение симметрично относительно среднего значения;
  - асимметрия и эксцесс равны нулю;
  - среднее значение, мода и медиана совпадают.

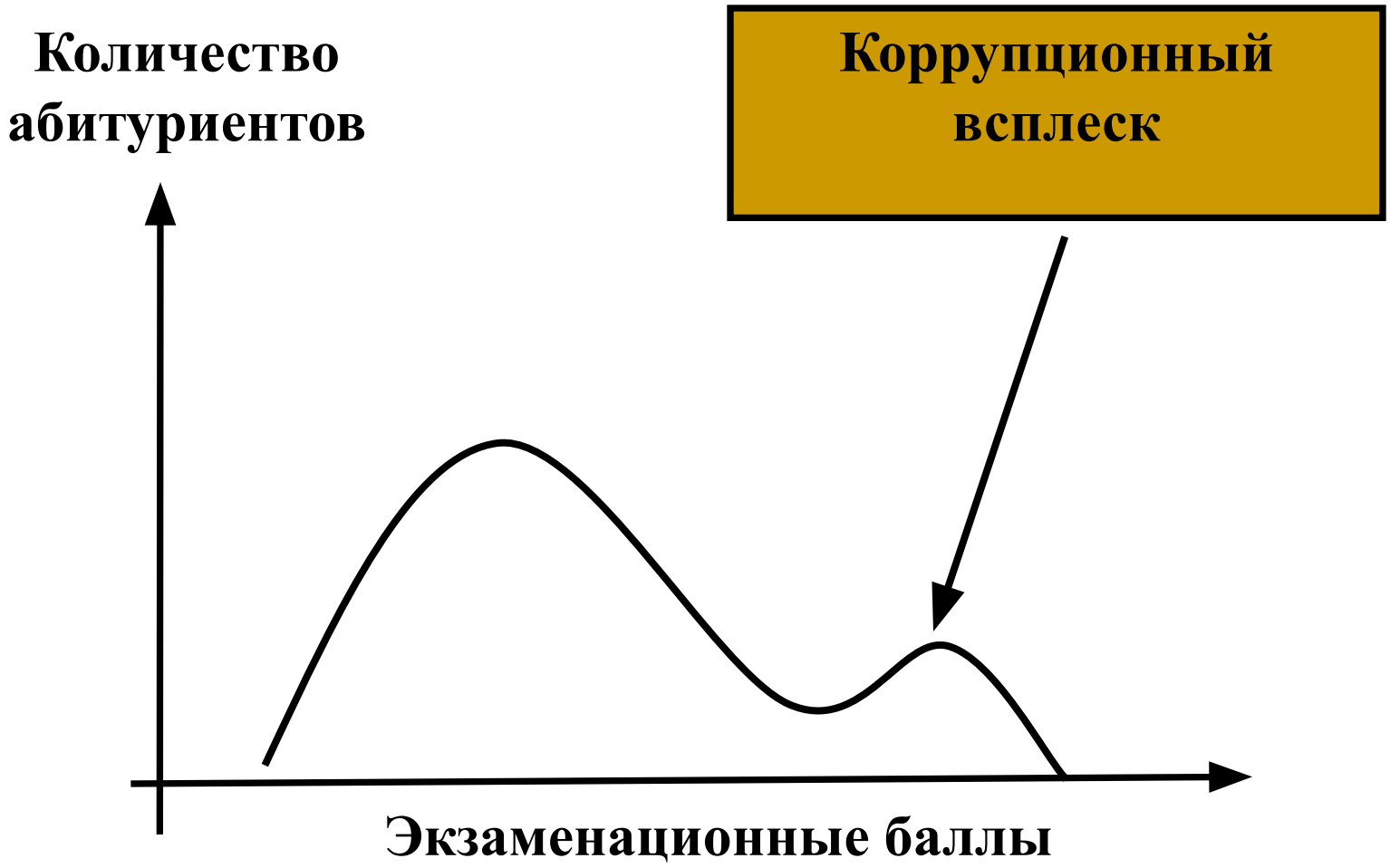
- ✔ Найти среднее значение $\overline{X}$ и стандартное отклонение σ, для нормального закона распределения приблизительно:
  - 68% значений находятся в интервале $\overline{X} \pm s$;
  - 95% – в интервале $\overline{X} \pm 2s$;
  - 99% – в интервале $\overline{X} \pm 3s$.



Форма, которую надо запомнить!

- ✔ Воспользоваться проверкой статистических гипотез о виде распределения.

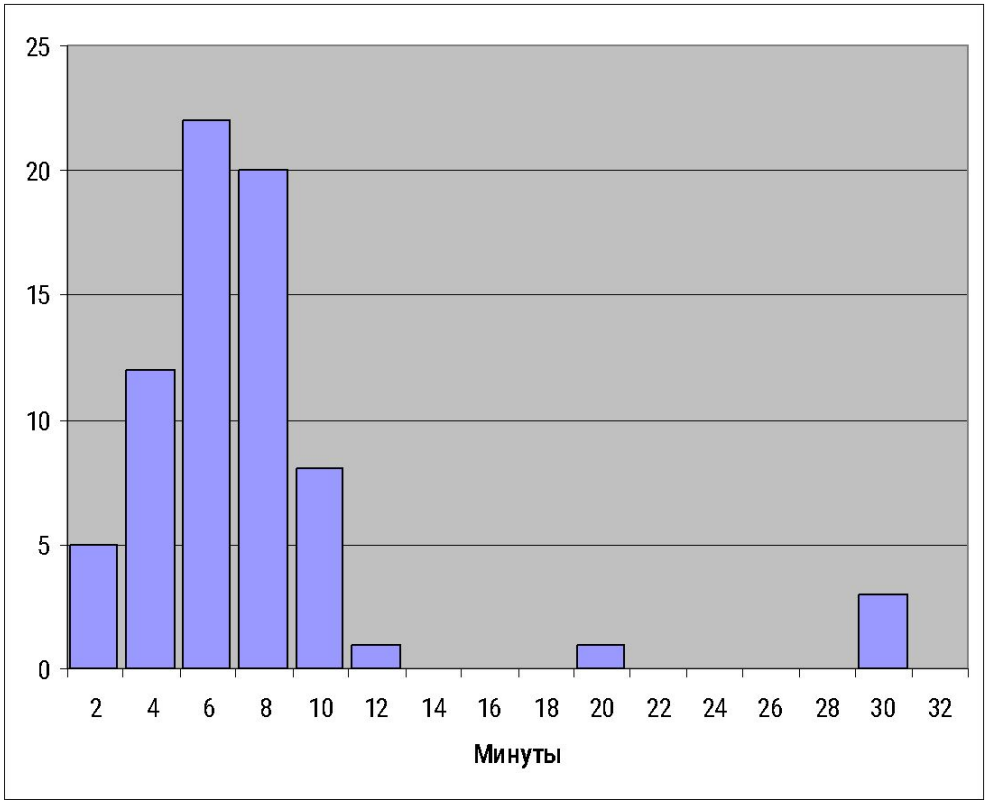
Меры формы


Количество абитуриентов
Коррупционный всплеск
Экзаменационные баллы

Выбросы

- Сильно отклоняющиеся значения называются **выбросами**.
- Являются ли эти наблюдения проявлением нормального разброса значений, случайностью или ошибкой ввода?

Пример. Время опроса одного студента

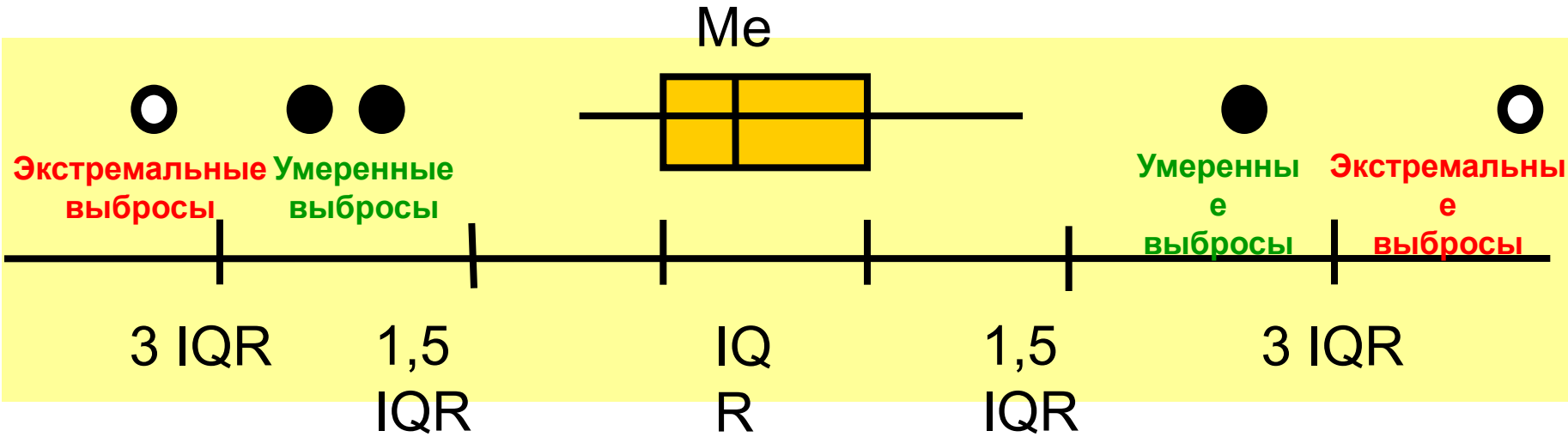


Выбросы (Критерий Тьюки)

Расширенная коробковая диаграмма строится с анализом выбросов. Для этого необходимо знать разброс квартилей IQR.

Умеренные выбросы изображаются темными точками и удалены ниже или выше медианы на 1,5 IQR, но не более 3 IQR.

Экстремальные выбросы изображаются светлыми точками и удалены ниже или выше медианы более чем на 3 IQR.



Статистические методы

- **Параметрические.** Примеряются для анализа нормально распределенных количественных признаков.

- **Непараметрические.** Применяются для анализа количественных признаков независимо от вида распределения и для анализа качественных признаков.

Описательная статистика

Параметрические методы:

- среднее значение;
- дисперсия;
- среднее квадратическое отклонение.

Непараметрические методы:

- медиана;
- интерпроцентильный размах (10-й и 90-й процентили);
- интерквартильный размах (значения 25-го и 75-го процентилей).

Восстановление пропущенных данных

- **Игнорирование пропусков.**

– для малых выборок с малым (<5%) числом пропусков

- **Заполнение средним значением.**

– для больших выборок с малым числом пропусков

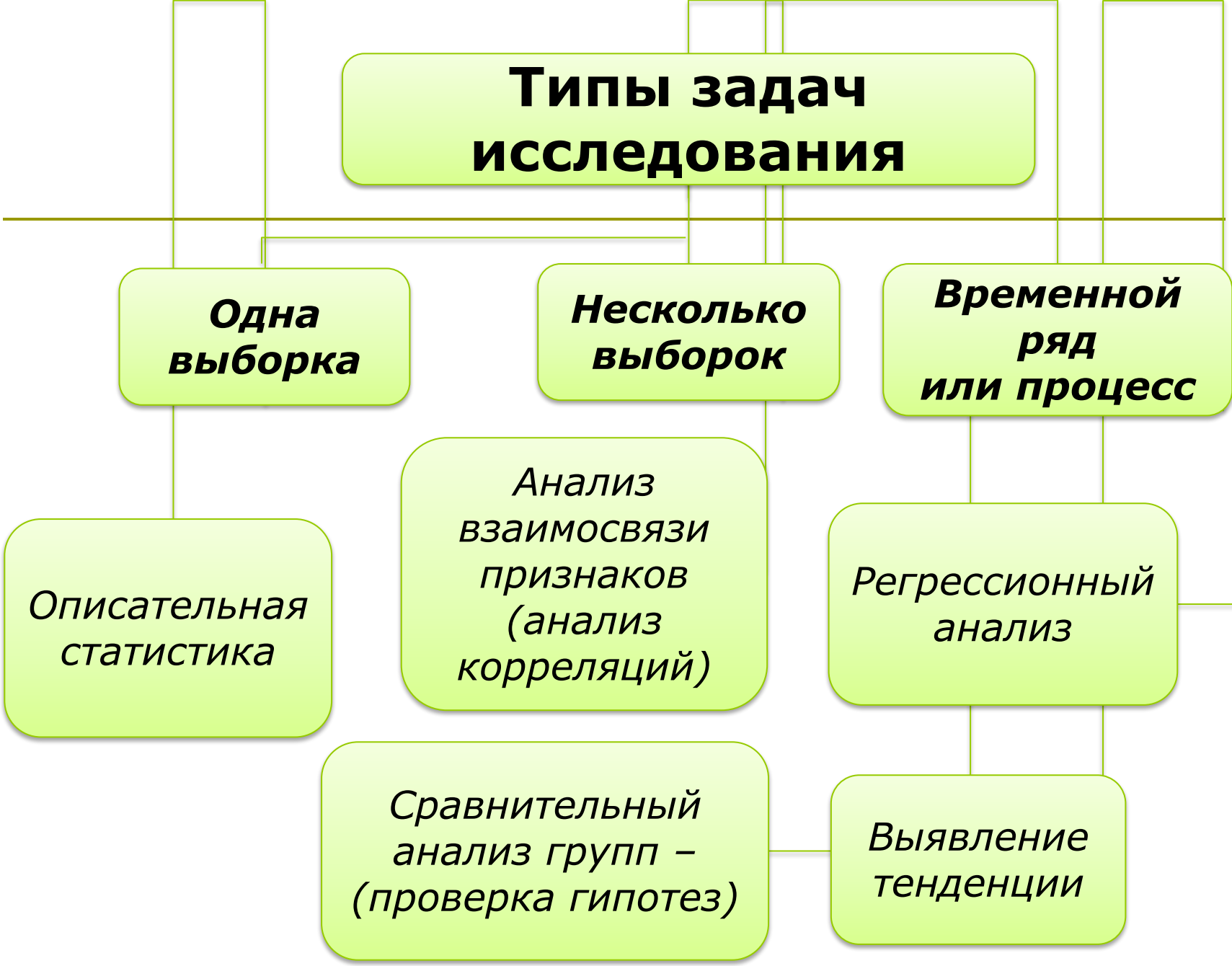
- **Заполнение регрессионными значениями.**

– для пар зависимых признаков

- **Заполнение случайными значениями**

– для больших выборок с малым числом пропусков

Типы задач исследования



Общая схема статистического анализа:

- Заполнение таблицы данными
- Обработка выбросов
- Обработка пропущенных данных
- Описательная статистика
- Определение вида распределения данных
- Корреляционный и регрессионный анализ
- Сравнительная статистика
- Углубленный анализ данных