**Математическая статистика**

**Введение в математическую статистику**

**Разница между описательной и математической статистикой**

**📊 Описательная статистика (Descriptive Statistics)**

**Цель:**  
Описать и суммировать данные, которые уже собраны (конкретную выборку).

**Что включает:**

* **Центральные тенденции**: среднее, медиана, мода
* **Разброс**: дисперсия, стандартное отклонение, диапазон
* **Форма распределения**: симметричность, скошенность (skewness), эксцесс (kurtosis)
* **Визуализация**: гистограммы, boxplot, диаграммы рассеяния и т. д.

**Пример:**  
В классе из 30 учеников средний рост = 165 см, стандартное отклонение = 8 см. Эти характеристики описывают именно этот класс.

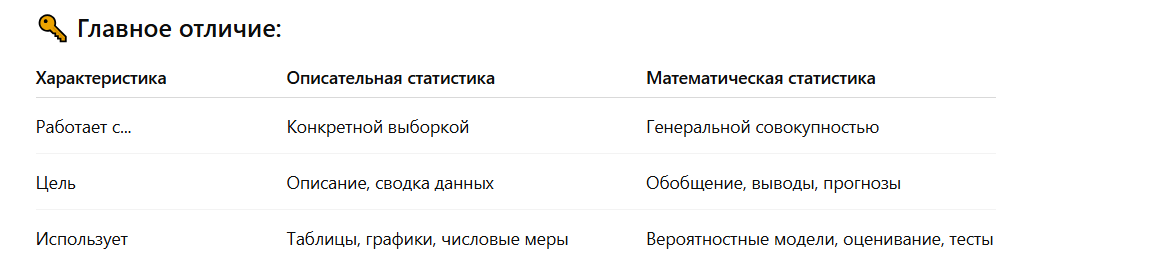
**📈 Математическая статистика (Inferential Statistics)**

**Цель:**  
Сделать выводы о **всей генеральной совокупности** на основе данных из выборки.

**Что включает:**

* **Оценивание** параметров (например, среднего значения по всей популяции)
* **Доверительные интервалы**
* **Проверка гипотез**
* **Регрессионный анализ**
* **Тесты значимости** (t-тест, χ²-тест и др.)

**Пример:**  
На основе роста 30 учеников из одной школы делается вывод о среднем росте всех школьников в городе — с определённой степенью доверия.

****

**Генеральная совокупность и выборка**

**🔹 Генеральная совокупность (Population)**

**Определение:**  
Это **всё множество объектов**, явлений или наблюдений, которые исследуются. Генеральная совокупность может быть конечной или бесконечной.

**Примеры:**

* Все жители Москвы — при исследовании уровня дохода.
* Все произведённые детали на заводе за год — при контроле качества.
* Все броски идеальной монеты — теоретическая бесконечная совокупность.

**Обозначения:**

* Размер: N (если конечная)
* Параметры: например, **среднее** μ, **дисперсия** σ^2, **доля** p и т. д.

**🔹 Выборка (Sample)**

**Определение:**  
Это **часть** генеральной совокупности, отобранная для изучения, измерений и анализа. Именно выборка используется для того, чтобы **делать выводы** о всей генеральной совокупности.

**Примеры:**

* 1000 случайных москвичей, опрошенных о доходе.
* 50 деталей, проверенных из партии в 10 000 штук.

**Обозначения:**

* Размер: n
* Статистики выборки: выборочное среднее xˉ, выборочная дисперсия s^2, выборочная доля p^.



**🔹 Зачем нужна выборка?**

Исследовать всю совокупность часто **невозможно** или **нецелесообразно** (дорого, долго, бесконечно). Поэтому:

* изучается выборка,
* рассчитываются выборочные характеристики,
* делаются **выводы о параметрах всей совокупности** — это и есть **индуктивная статистика**.

📌 **Смещение выборки** — это ситуация, когда **отобранные данные не репрезентативны**.  
В данном случае есть **риск смещения по добровольности** (self-selection bias):  
только те, кто **захотели** ответить, попали в выборку. Возможно, это более **недовольные** или наоборот, **сильно довольные** клиенты. Следовательно, результаты **не обязательно отражают мнение всей совокупности**.

**Ключевые ошибки в статистике**

## 🔻 1. **Ошибка выборки (Sampling Bias)**

### Что это:

Когда выборка **не отражает** генеральную совокупность — она **непредставительна**.

### Примеры:

* Опрос о политике проводят только в Telegram-канале — пропущены пожилые люди и офлайн-аудитория.
* Анкеты по email → отвечают только активные пользователи.

### Итог:

Выводы искажаются, так как часть совокупности не представлена или **представлена непропорционально**.

## 🔻 2. **Смещение по добровольности (Self-selection Bias)**

### Что это:

Когда **люди сами решают**, участвовать или нет. Обычно участвуют **эмоционально вовлечённые** — недовольные или фанаты.

### Примеры:

* Оценки на сайтах (5 и 1 звезда — нормальные оценки почти не ставят).
* Только мотивированные пациенты участвуют в опросах об эффективности терапии.

## 🔻 3. **Ошибка выжившего (Survivorship Bias)**

### Что это:

Мы видим **только тех, кто “выжил” или дошёл до результата**, и игнорируем тех, кто выбыл.

### Примеры:

* Успешные стартапы в интервью → никто не показывает тысячи провалов.
* Исследование долгожителей: мы не видим умерших раньше по другим причинам.

## 🔻 4. **Ошибка экстраполяции (Overgeneralization / External Validity Error)**

### Что это:

Выводы делают на **всю популяцию**, хотя выборка была **узкой**.

### Пример:

* Изучали только программистов, а вывод: “кофе улучшает продуктивность всех”.
* Исследование студентов Гарварда → «все молодёжь думает вот так».

## 🔻 5. **Смешение переменных (Confounding)**

### Что это:

Когда третья переменная влияет **и на причину, и на следствие**, и мы путаем её с причиной.

### Пример:

* Люди, пьющие кофе, продуктивнее?  
  → А может, у них просто **меньше проблем со сном** или **более активный характер**?

## 🔻 6. **Ошибка формулировки (Leading Questions)**

### Что это:

Формулировка вопроса **влияет на ответ** — вопрос “подталкивает” к нужному мнению.

### Пример:

* “Вы согласны, что нужно снизить вредное налогообложение?”  
  (внушает, что оно вредное)
* “Как сильно вас устраивает новая политика компании?”  
  (предполагается, что она устраивает)

## 🔻 7. **Ошибка обратной причинности (Reverse Causation)**

### Что это:

Перепутаны причина и следствие.

### Пример:

* “Люди с хорошим настроением часто улыбаются → значит, улыбка вызывает хорошее настроение”.
* “Больные принимают больше лекарств → лекарства вызывают болезнь”.

**Типы данных: количественные и категориальные**

## 🧮 1. **Количественные (числовые) данные**

Это данные, которые выражаются в **числах** и могут быть **измерены**.

### 🔹 a) ****Непрерывные (интервальные/дискретные)****

* **Примеры:** рост, вес, температура, доход, время.
* **Свойства:**
  + Можно **выполнять арифметические операции**: сложение, деление и т.д.
  + Имеют **порядок** и **расстояния между значениями** имеют смысл.

### 🔹 b) ****Дискретные****

* Это числовые значения, которые **могут принимать только отдельные значения** (обычно целые).
* **Примеры:** количество детей, число автомобилей, баллы по тесту.

📌 **Итого:**  
→ Количественные данные **можно измерить**.  
→ Они часто используются при **расчётах среднего, стандартного отклонения и т.д.**

## 🧩 2. **Категориальные (качественные) данные**

Это данные, которые описывают **категории или группы**. Значения не измеряются, а **отнесены к классу**.

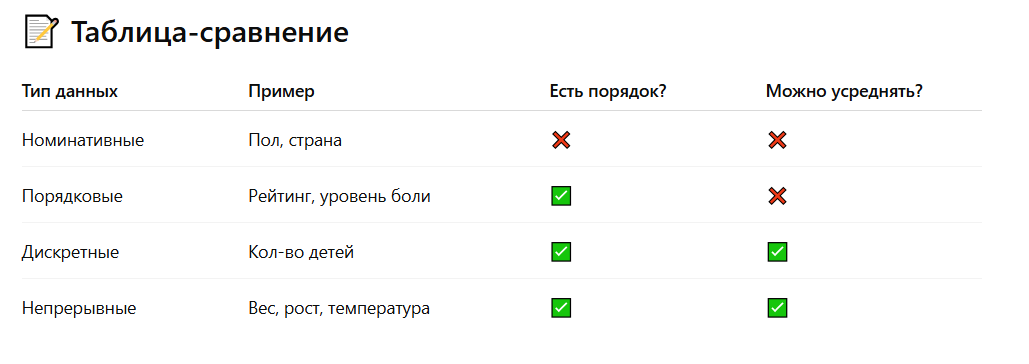
### 🔹 a) ****Номинативные (номинальные)****

* Просто **имена** категорий, **без порядка**.
* **Примеры:** пол (м/ж), страна, цвет глаз, марка машины.

### 🔹 b) ****Порядковые (ордиальные)****

* Категории с **естественным порядком**, но **без равных интервалов**.
* **Примеры:** уровень образования (школа < бакалавр < магистр), оценка удовлетворенности (низкий, средний, высокий).

📌 **Итого:**  
→ Категориальные данные нельзя **усреднять** или **вычитать**, но можно **сравнивать** или **сортировать**, если есть порядок.



**Оценивание и доверительные интервалы**

В математической статистике **точечная** и **интервальная оценка** — это два подхода к **оцениванию неизвестных параметров** генеральной совокупности (например, среднего или дисперсии), используя данные выборки.

## 🔹 Точечная оценка

### ❓ Что это такое:

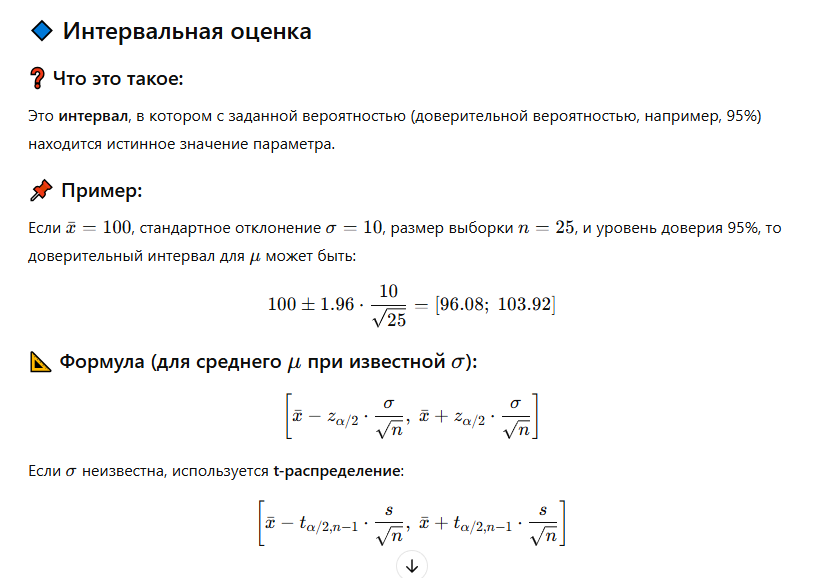
Это **одно число**, которое используется как "лучшая догадка" (оценка) неизвестного параметра.

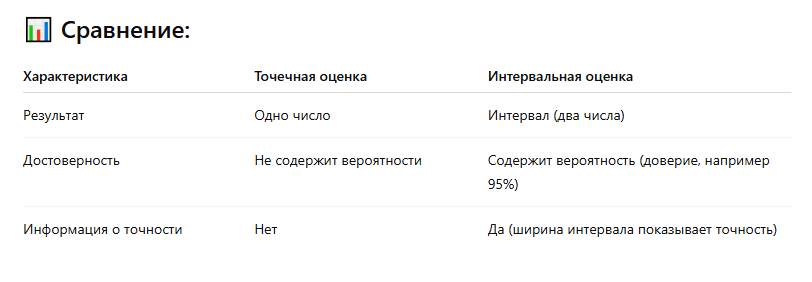
### 📌 Примеры:

* Среднее значение выборки (xˉ) — точечная оценка математического ожидания μ.
* Выборочная дисперсия s^2 — точечная оценка истинной дисперсии σ2.

### 🧠 Свойства хорошей оценки:

1. **Несмещённость**: математическое ожидание оценки равно истинному значению параметра.
2. 
3. **Состоятельность**: при увеличении объёма выборки оценка стремится к истинному значению.
4. **Эффективность**: имеет наименьшую дисперсию среди всех несмещённых оценок.
5. **С достаточной статистикой**: использует всю доступную информацию из выборки.

****

****

## 🔸 Представь ситуацию

Ты хочешь узнать **средний вес всех арбузов** на складе. Их там **тысячи**, взвешивать каждый — долго.  
Поэтому ты берёшь **случайные 10 арбузов**, взвешиваешь их, и находишь, что **средний вес = 5 кг**.

Но ты понимаешь:  
«Ну вдруг мне просто повезло? Может, я случайно выбрала лёгкие или тяжёлые?..»

## 🔸 Вот тут и появляется интервальная оценка

Она говорит:

«Мы **не уверены, что 5 кг — это точный ответ**, но **почти уверены**, что **средний вес всех арбузов лежит где-то между 4.5 и 5.5 кг**.»

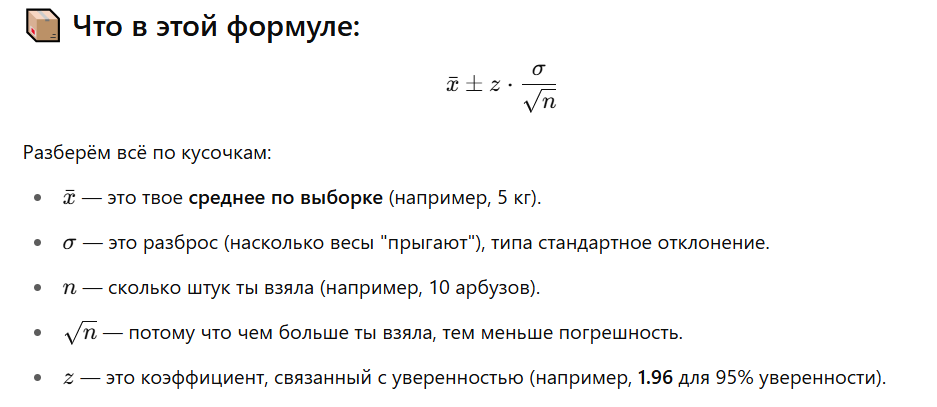
## 🔸 Как это устроено (на пальцах):

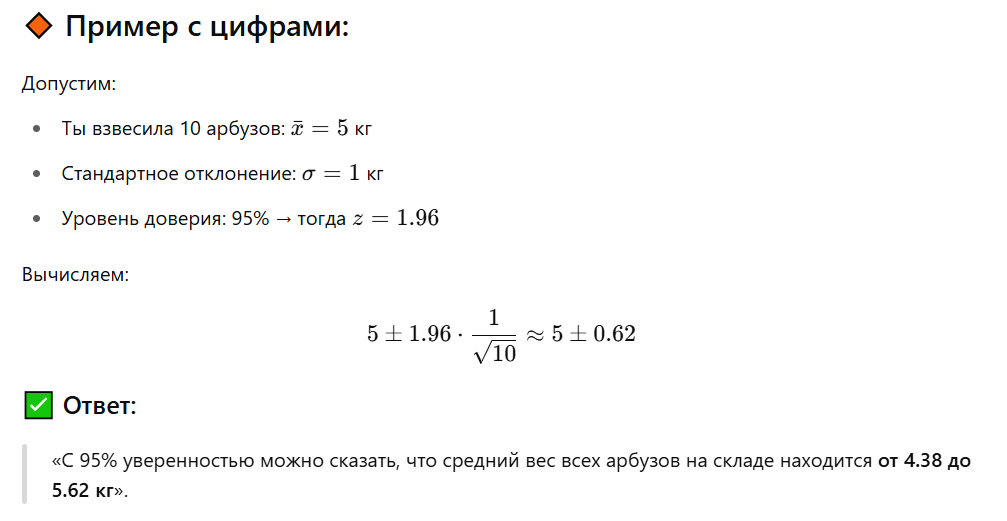
Интервал всегда строится так:

среднее значение выборки

±

запас на ошибку

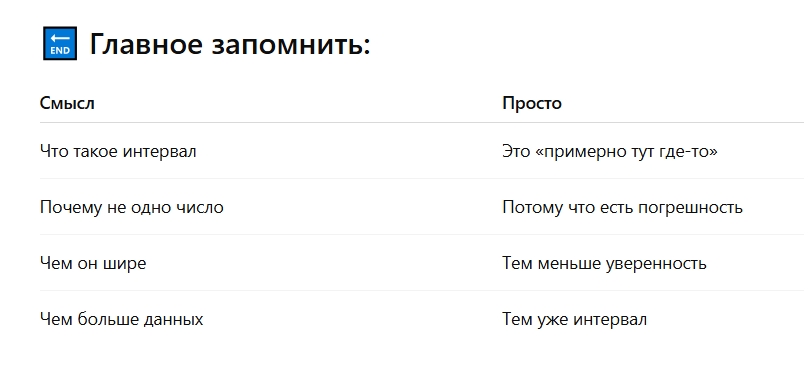
****

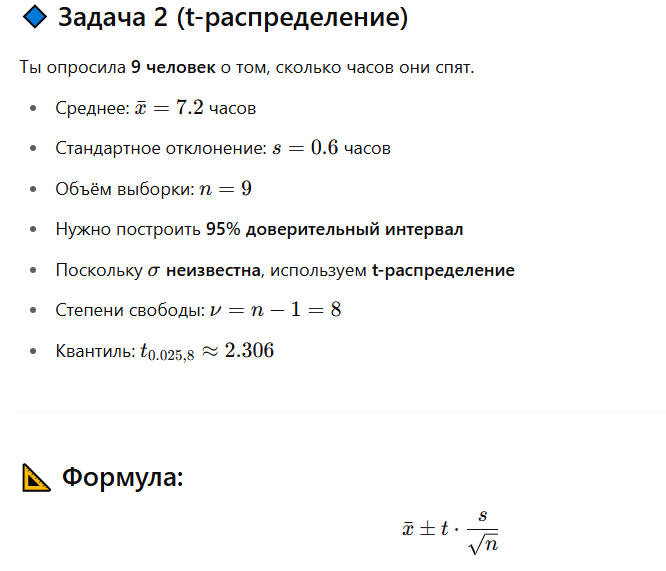
****

## 🔸 Что значит “95% уверенности”?

Это не значит, что 95% арбузов попадают в этот интервал.  
Это значит:

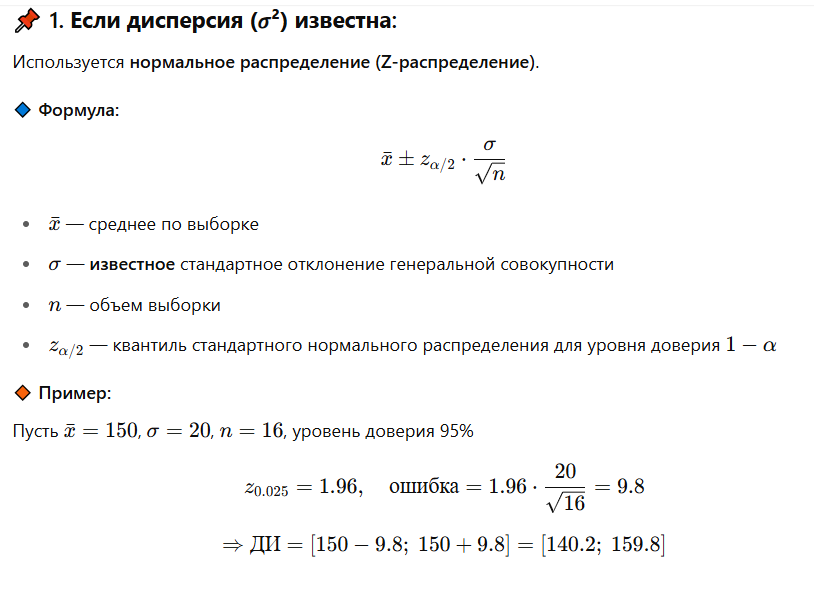
«Если бы я **много раз** брала по 10 арбузов и каждый раз строила такой интервал, то **в 95% случаев** в него попадало бы **истинное среднее**».

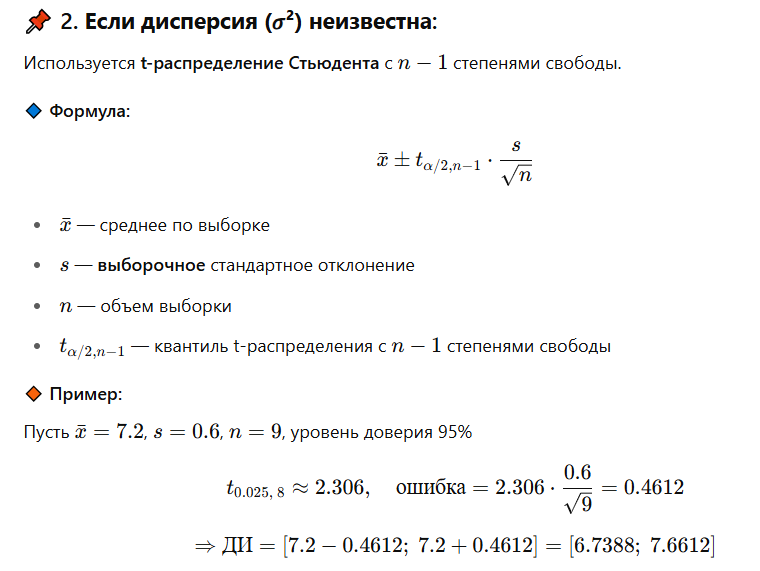
****

****

**Доверительные интервалы для среднего при известной и неизвестной дисперсии**

Доверительный интервал (ДИ) для среднего строится по-разному в зависимости от того, **известна ли дисперсия (или стандартное отклонение)** генеральной совокупности.

****

****

****

**Концепция уровня доверия (напр. 95%)**

**Уровень доверия (например, 95%)** — это **вероятность того, что интервал, построенный по определённой процедуре, накроет истинное значение параметра** (например, среднего 𝜇).

### 🔍 Простой пример:

Допустим, вы измеряете рост студентов и хотите оценить **средний рост** всех студентов университета. У вас есть выборка, и вы по ней строите **доверительный интервал**:  
**[170; 174] при уровне доверия 95%**.

**Что это значит?**  
Если бы вы **повторили эксперимент 100 раз** (взяли бы 100 разных выборок и построили 100 интервалов по такой же формуле), то примерно **в 95 из них интервал накрыл бы истинное значение 𝜇**.  
Остальные 5 интервалов — "неудачные", они промахнулись.

### ⚠️ Частая ошибка:

"С вероятностью 95% параметр лежит в этом интервале"

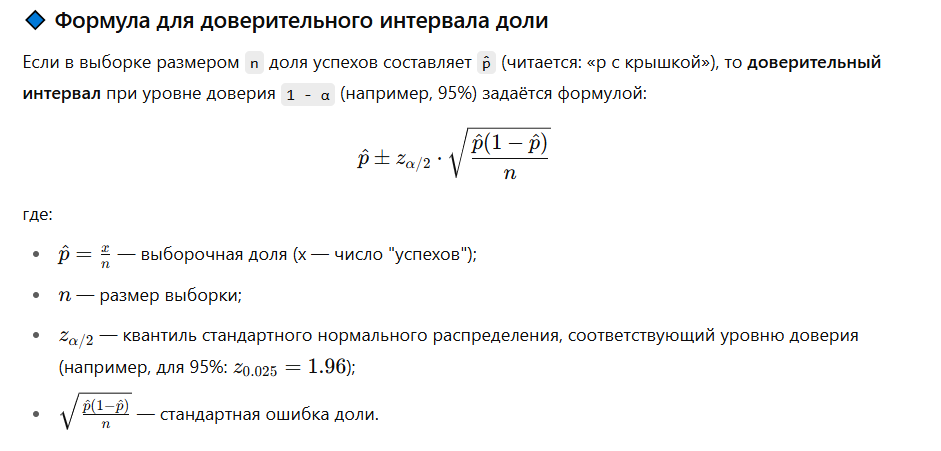
❌ Это **неправильно** с точки зрения строгой частотной статистики. Параметр (например, 𝜇) **фиксирован**, и он **либо в интервале, либо нет**.  
А **95% — это характеристика метода**, а не одного конкретного интервала.

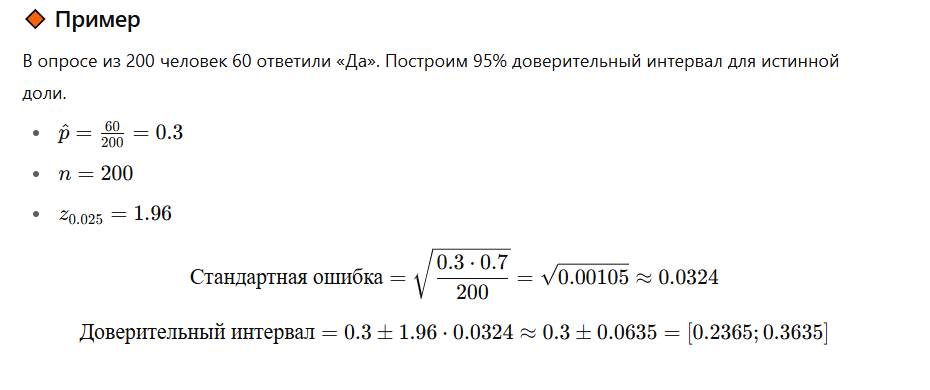
**📌 Как это работает:**

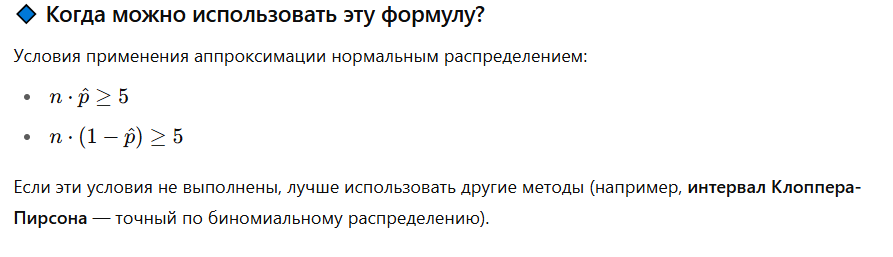
1. **Уровень доверия** задаёт **насколько широким будет интервал**:
   * 90% → более узкий
   * 95% → чуть шире
   * 99% → ещё шире  
     Шире — значит **больше уверенность, но меньше точность**.
2. Используются **квантильные значения** из Z- или t-распределения:
   * для 95%: z ≈ 1.96 (если σ известна)
   * или t-квантиль с нужными степенями свободы (если σ неизвестна)

**Доверительный интервал для доли**

Доверительный интервал для **доли** (например, доли голосов, дефектных изделий и т.п.) — это интервал, в котором, с заданной степенью уверенности (например, 95%), находится **истинная доля** в генеральной совокупности на основе данных из выборки.

****

****

****

**Проверка гипотез и t-тест**

**Нулевая и альтернативная гипотезы**

В математической статистике **нулевая гипотеза** и **альтернативная гипотеза** — это два взаимоисключающих предположения, которые используются в проверке статистических гипотез. Давай разберёмся по порядку.

**🔹 Нулевая гипотеза (обозначается H₀)**

Это **исходное предположение**, которое мы считаем **правильным по умолчанию**, пока не появятся достаточные доказательства обратного.

* Формулируется как «никаких изменений нет», «различий нет», «эффекта нет».
* Примеры:
  + H₀: Средний рост мужчин в стране = 175 см.
  + H₀: Новое лекарство **не** отличается по эффективности от старого.
  + H₀: Доля голосующих за кандидата = 50%.

📌 **Важно**: Нулевая гипотеза — это то, что **пытаются опровергнуть**.

**🔹 Альтернативная гипотеза (обозначается H₁ или Hₐ)**

Это **гипотеза-«противник»**, которая утверждает, что **есть различия**, **эффект присутствует**, **изменения произошли**.

* Формулируется как «есть эффект», «есть различия».
* Примеры (антиподы предыдущих):
  + H₁: Средний рост мужчин ≠ 175 см.
  + H₁: Новое лекарство **лучше** (или **хуже**) старого.
  + H₁: Доля голосующих ≠ 50%.

**🔸 Типы альтернативных гипотез:**

1. **Двусторонняя (двухсторонняя)**:  
   H₁: μ ≠ μ₀ (любое отличие от нулевой гипотезы)
2. **Левосторонняя**:  
   H₁: μ < μ₀ (интересует, уменьшилось ли значение)
3. **Правосторонняя**:  
   H₁: μ > μ₀ (интересует, увеличилось ли значение)

**🔹 Пример:**

Ты хочешь проверить, влияет ли кофе на продуктивность.

* H₀: Кофе **не влияет** на продуктивность.
* H₁: Кофе **влияет** на продуктивность.

Если данные покажут **значимое отличие**, мы отвергаем H₀ и принимаем H₁. Если нет — оставляем H₀.

**🧠 Почему это важно?**

* Отказ от нулевой гипотезы должен быть **обоснованным**: мы не «принимаем» альтернативную гипотезу просто так.
* Решение принимается с учётом уровня значимости (α), например, 0.05 (5% риск сделать ложный вывод).

**Пошаговая проверка статистических гипотез**

**🔹 1. Формулируем гипотезы**

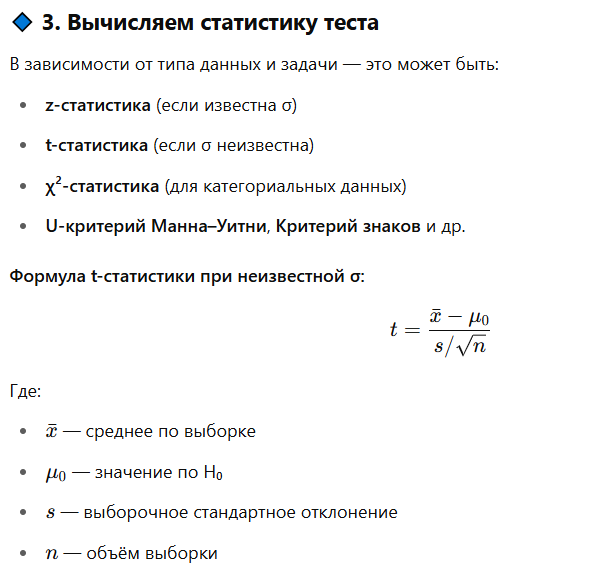
* **Нулевая гипотеза H₀**: предполагает **отсутствие эффекта** или **различий**.
* **Альтернативная гипотеза H₁**: предполагает, что **эффект есть** или **значимые различия существуют**.

**Пример:**  
Проверяем, отличается ли средний рост студентов от 170 см.

* H₀: μ = 170
* H₁: μ ≠ 170 (двусторонняя альтернатива)

**🔹 2. Выбираем уровень значимости α**

* Обычно берут **α = 0.05** (5%) — вероятность того, что мы **ошибочно отвергнем H₀**, если она на самом деле верна.
* Иногда используют 0.01 (1%) или 0.10 (10%) в зависимости от контекста.



**🔹 4. Находим p-значение (p-value)**

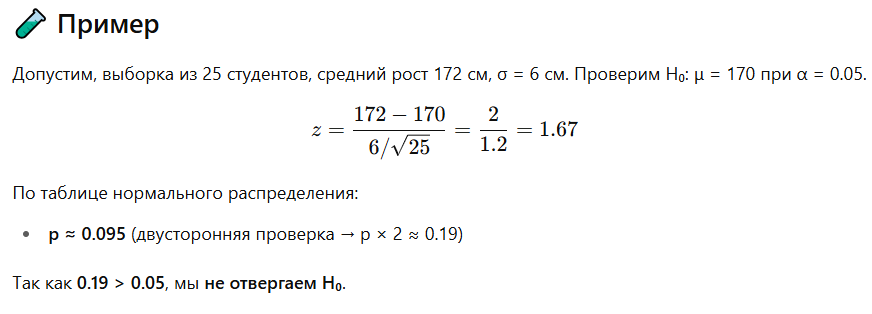
**p-value** — вероятность получить такое же или более экстремальное значение статистики, если **H₀ верна**.

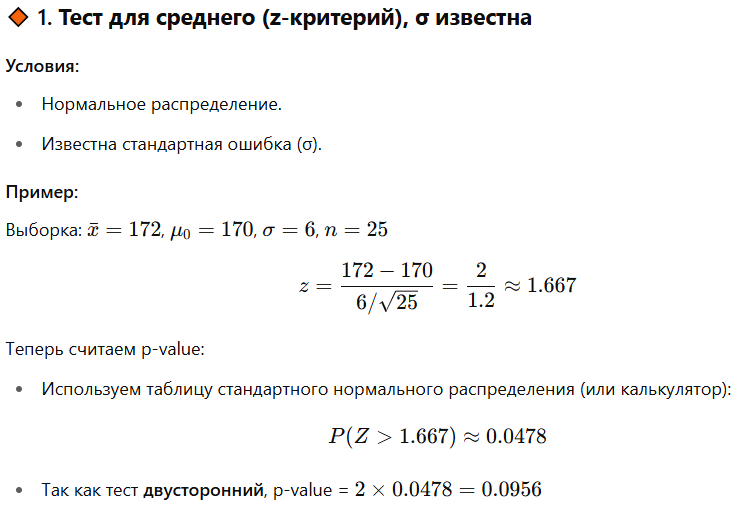
Проще говоря:  
**Насколько удивительны наши данные**, если предположить, что H₀ — правда?

### 🔹 Как интерпретировать?

* Если **p-value мало** (например, < 0.05), значит, такие данные **маловероятны** при H₀ → мы **отвергаем H₀**.
* Если **p-value велико**, данные **могут быть объяснены** H₀ → мы **не отвергаем H₀**.
* Если **p-value < α**, мы **отвергаем H₀** — данные значимы.
* Если **p-value ≥ α**, мы **не отвергаем H₀** — данных недостаточно.

📌 Важно: *«Не отвергнуть H₀» ≠ «Принять H₀».* Мы просто не доказали, что она ложна.

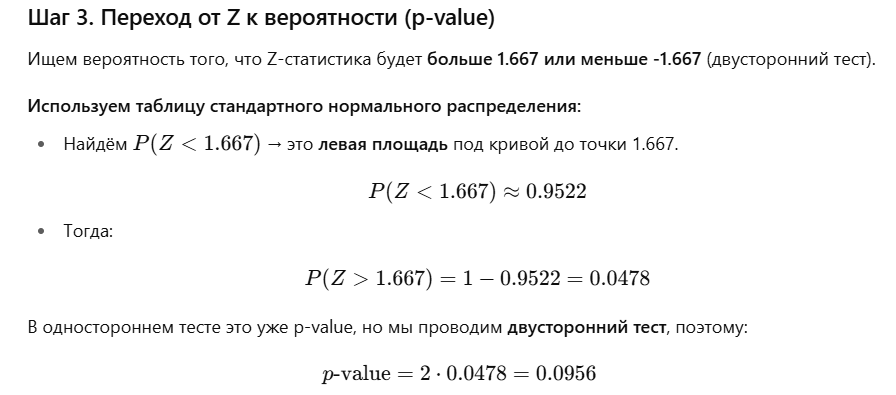




**Что значит Z = 1.667?**

Это означает, что наше выборочное среднее (172) **находится на 1.667 стандартных ошибок выше**, чем ожидаемое среднее по H₀ (170).

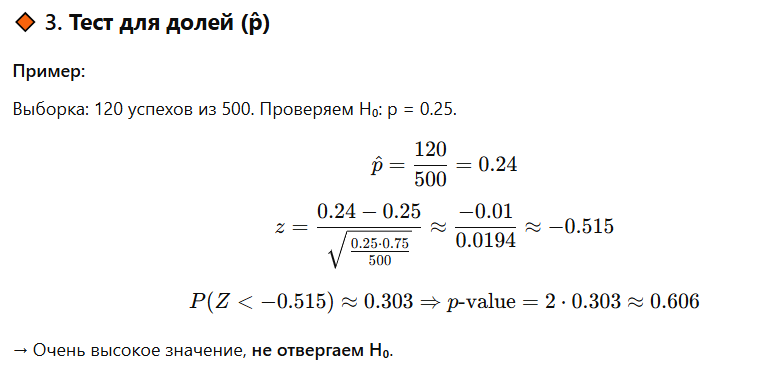
Теперь надо понять: **насколько "редко" получить такое отклонение при условии, что H₀ верна?**



**🔸 2. Тест для среднего (t-критерий), σ неизвестна**

Всё то же самое, но:

* Используем **t-распределение** с n−1n - 1n−1 степенями свободы.
* Считаем **t-статистику** и используем таблицу t-распределения или Python/Excel/калькулятор.



**🔸 Как посчитать p-value в Python?**

from scipy.stats import norm

z = 1.667

p\_value = 2 \* (1 - norm.cdf(z)) # двухсторонний тест

print(p\_value)

**🔸 Как в Excel?**

* =NORM.S.DIST(z, TRUE) — даёт значение CDF (левая сторона).
* =2 \* (1 - NORM.S.DIST(z, TRUE)) — для двухстороннего p-value.

❗ Важное предупреждение

p-value — не вероятность того, что H₀ верна.

Это только вероятность получить такие данные при условии, что H₀ верна.

Низкое p-value не доказывает H₁ — оно лишь говорит, что H₀ плохо согласуется с наблюдаемыми данными.

**Ошибки I и II рода**

**🧪 Контекст: проверка гипотез**

Когда мы проверяем гипотезу, есть два возможных решения:

1. **Отклонить нулевую гипотезу H₀**
2. **Не отклонять нулевую гипотезу H₀**

Но при этом **мы можем ошибиться**, ведь мы делаем выводы на основе выборки, а не всей совокупности.

**❗ Ошибка I рода (False Positive)**

📌 Это **ошибка, при которой мы отвергаем H₀, хотя на самом деле она верна**.

**Обозначается:** α (альфа) — **уровень значимости**, который ты задаёшь заранее (например, 0.05).

**Пример:**

* H₀: Пациент **здоров**
* H₁: Пациент **болен**
* Ошибка I рода: ты говоришь, что пациент болен (отверг H₀), хотя он на самом деле здоров.

Это ложная тревога. Мы увидели "эффект", которого нет.

**❗❗ Ошибка II рода (False Negative)**

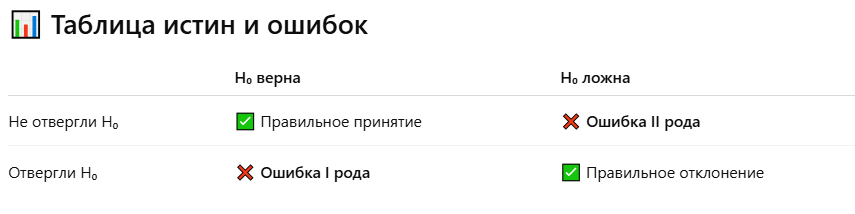
📌 Это **ошибка, при которой мы не отвергаем H₀, хотя на самом деле она ложна**.

**Обозначается:** β (бета)

**Пример:**

* H₀: Пациент **здоров**
* H₁: Пациент **болен**
* Ошибка II рода: ты говоришь, что пациент здоров (не отвергаешь H₀), хотя он болен.

Это пропущенный эффект. Мы **не заметили**, хотя он был.



**🎯 Что такое мощность теста?**

**Мощность теста** = 1−β1 - \beta1−β

То есть:

* Вероятность **обнаружить эффект**, если он есть.
* Чем выше мощность, тем **меньше риск ошибки II рода**.

## 📌 Как уменьшить ошибки?

* Уменьшить **α** → меньше риск ошибки I рода, но **больше шанс ошибки II рода**.
* Увеличить **размер выборки** → уменьшается **обе ошибки**.
* Выбрать правильный **тест** и **двусторонность**.

**Тестирование гипотез: пошаговый алгоритм**

### 🔹 ****Шаг 1. Сформулируй гипотезы****

* **Нулевая гипотеза (H₀)**: утверждение, что ничего не изменилось, различий нет, эффекта нет.
* **Альтернативная гипотеза (H₁)**: противоположная идея — есть эффект, есть отличие.

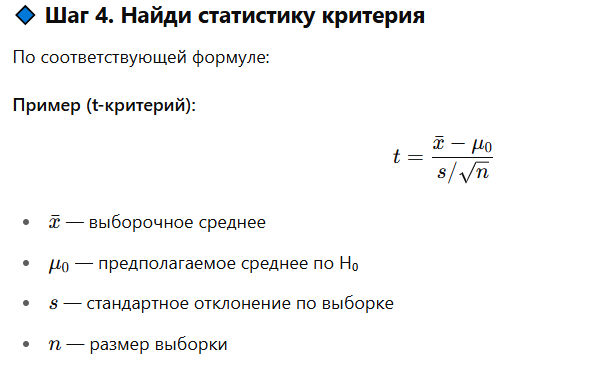
Пример:  
Исследуем, отличается ли средний рост студентов от 170 см:

* H₀: μ = 170
* H₁: μ ≠ 170 (двусторонняя альтернатива)

### ****Шаг 2. Выбери уровень значимости α****

* Это вероятность допустить **ошибку I рода** (ложно отклонить H₀).
* Типичные значения:
  + **0.05** — наиболее распространённый выбор
  + **0.01** — если нужно быть очень осторожным
  + **0.10** — если допускается чуть больше риска





**🔹 Шаг 5. Найди p-value или критическое значение**

**Два способа:**

**🔸 Через p-value:**

* Используем таблицу распределения или калькулятор, чтобы найти вероятность того, что статистика ≥ наблюдаемого значения.
* **Сравниваем p-value с α**:
  + Если **p < α** → отвергаем H₀
  + Если **p ≥ α** → не отвергаем H₀

**🔸 Через критическое значение:**

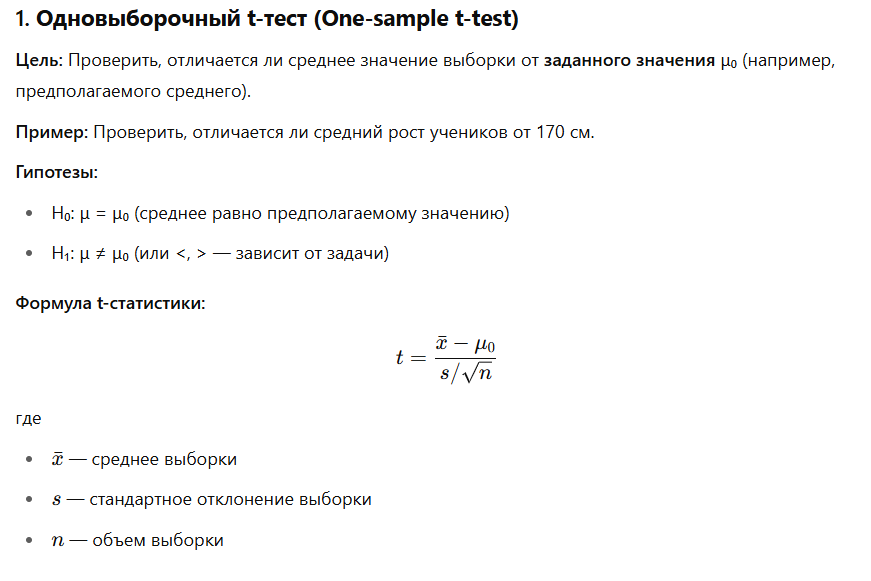
* Для заданного α находим **границы критической области**.
* Если статистика попала туда — отвергаем H₀.

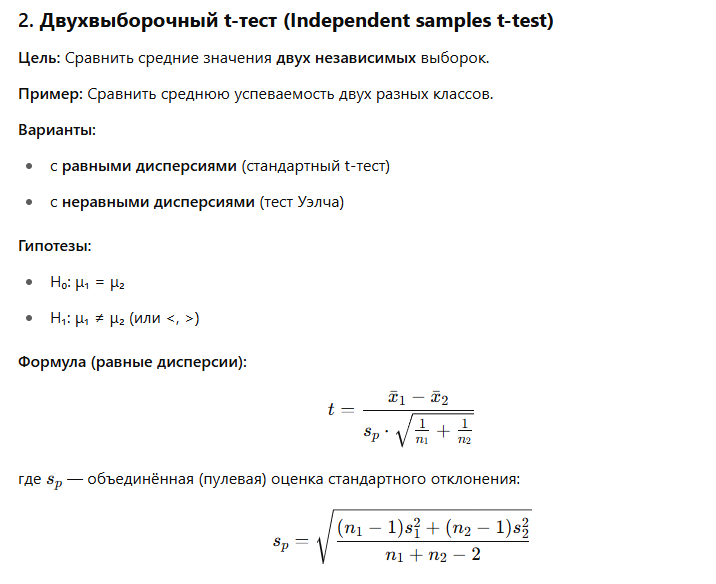
### 🔹 ****Шаг 6. Сделай вывод****

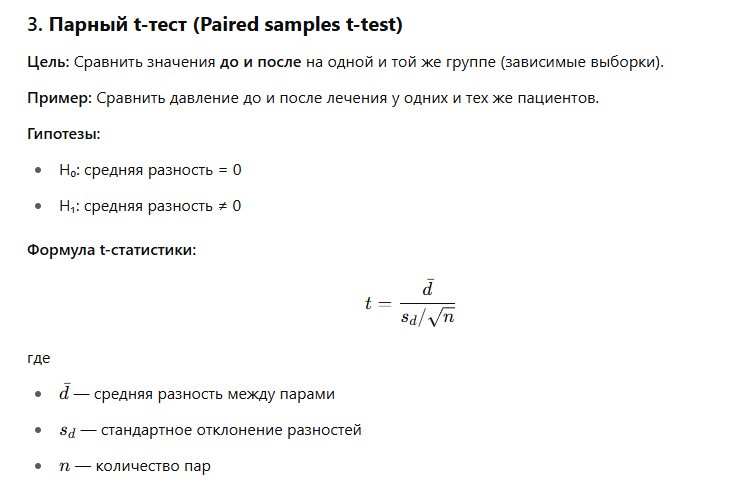
* Формально: "На уровне значимости 5% мы (не) отвергаем H₀."
* По сути: "**Обнаружены статистически значимые различия**" или "Различия **не значимы**".

**t-тест: одновыборочный, двухвыборочный, парный**

t-тест — это статистический тест, используемый для проверки гипотез о средних значениях. Он особенно полезен, когда объем выборки мал (обычно меньше 30) и неизвестна дисперсия генеральной совокупности. Существует **три основных типа t-тестов**:

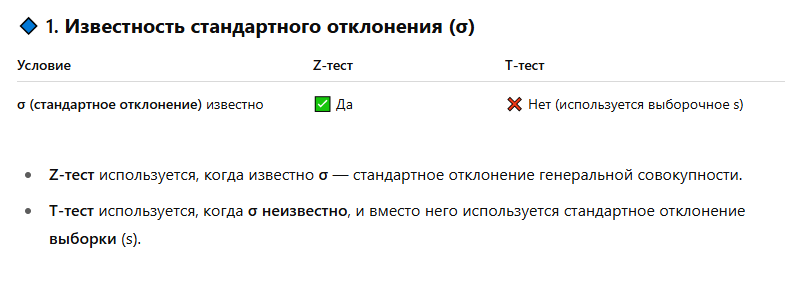






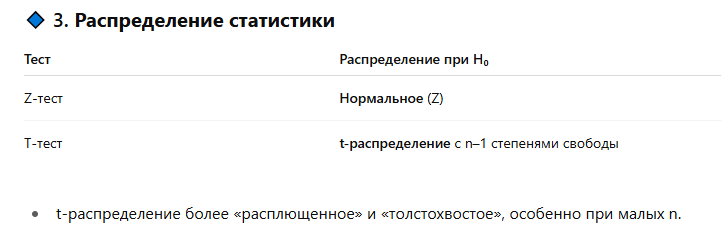
### Общие шаги проведения t-теста:

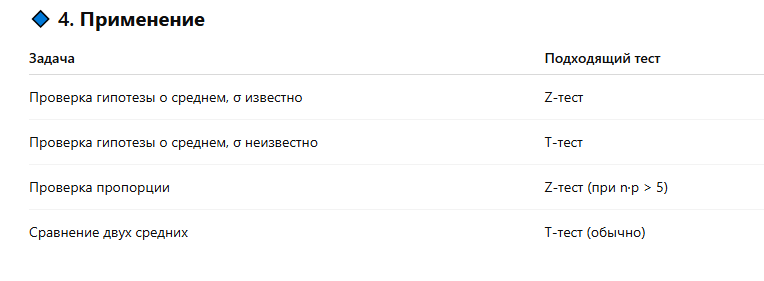
1. **Сформулировать гипотезы** (нулевую и альтернативную).
2. **Вычислить t-статистику.**
3. **Найти p-value** или критическое значение t по таблице (с нужными степенями свободы).
4. **Сравнить с уровнем значимости α** (например, 0.05).
5. **Сделать вывод**: отвергается ли нулевая гипотеза.



### 🔹 2. ****Размер выборки****

* **Z-тест** предпочтителен при **больших выборках** (обычно n > 30), особенно если распределение близко к нормальному.
* **T-тест** лучше работает при **малых выборках** (n ≤ 30), особенно если σ неизвестно.





### 🔹 5. ****Пример****

Допустим:

* Среднее в выборке: xˉ=170
* σ неизвестно, s = 5
* n = 20
* Тогда применим **t-тест**, т.к. σ нет и выборка мала.

**Непараметрические методы**

Непараметрические методы применяются в математической статистике **в ситуациях, когда стандартные (параметрические) методы неприменимы или ненадёжны**. Вот основные случаи, **когда стоит использовать непараметрические методы**:

**✅ 1. Неизвестно распределение данных**

Параметрические тесты (например, t-тест) требуют, чтобы данные были **нормально распределены**.  
Если распределение:

* неизвестно,
* сильно отличается от нормального (например, скошено, с выбросами),
* не удаётся нормализовать преобразованиями (логарифмами и т.д.),

→ используем непараметрические методы.

**Пример:**  
Если вы хотите сравнить медианы двух выборок, но одна из них с асимметрией, лучше использовать **тест Манна–Уитни**, а не двухвыборочный t-тест.

**✅ 2. Малый объём выборки**

На малых выборках нормальность трудно проверить и оценить.

→ Непараметрические методы более устойчивы к ошибкам при работе с малым количеством данных.

**✅ 3. Шкалы порядка и номинальные данные**

Параметрические методы работают с количественными (интервальными и шкалами отношений) данными.  
Если у вас:

* порядковая шкала (например, уровень боли: "нет", "средняя", "сильная"),
* номинативная шкала (например, цвета: "красный", "синий", "зелёный"),

→ используйте непараметрические методы (тесты согласия, ранговые критерии).

**✅ 4. Есть выбросы или шум**

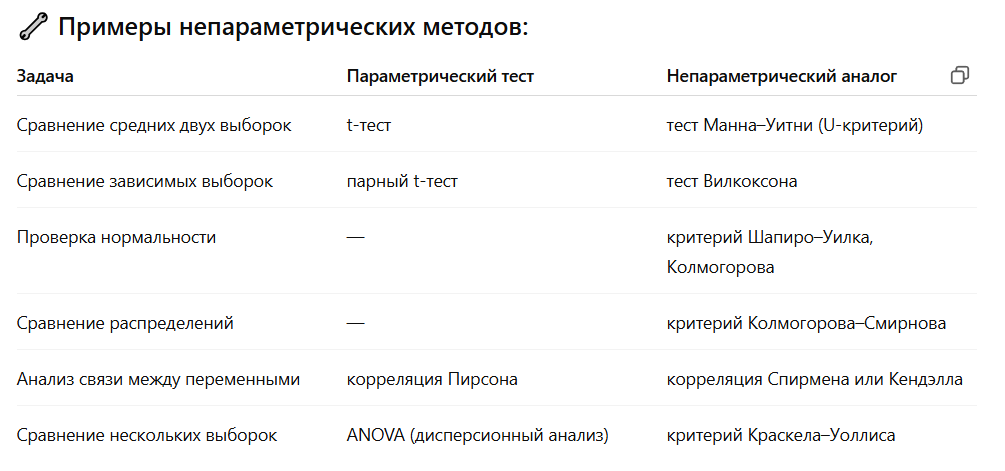
Непараметрические тесты не чувствительны к выбросам, т.к. часто работают с рангами, а не с самими значениями.

**✅ 5. Проверка медиан или распределений**

Если интересует не среднее, а **медиана**, или **форма распределения** (например, однородность), непараметрические методы подойдут лучше.

**❌ Когда не стоит использовать непараметрические методы:**

* Когда данные чётко соответствуют требованиям параметрических тестов.
* Когда важны параметры (например, оценки среднего и дисперсии).
* Когда нужно максимальная мощность теста (непараметрические тесты обычно менее мощные при выполнении условий параметрических методов).



**📌 1. Оценка эффективности обезболивающего по шкале боли**

**Задача:** Пациенты оценивают боль до и после приёма лекарства по шкале от 0 до 10.  
**Проблема:** Шкала — **порядковая**, разница между "2" и "3" не обязательно равна разнице между "6" и "7".

✅ **Метод:** Тест Вилкоксона для парных выборок.  
❌ **Почему не t-тест:** t-тест требует количественных данных с равномерными интервалами и нормальностью распределения разностей.

**📌 2. Сравнение доходов в двух регионах**

**Задача:** Сравнить уровень дохода людей в двух разных регионах.  
**Проблема:** Доходы часто имеют **асимметричное распределение** с большими выбросами (очень богатые и очень бедные).

✅ **Метод:** Тест Манна–Уитни (U-критерий).  
❌ **Почему не t-тест:** t-тест чувствителен к выбросам и нормальности, которой здесь нет.

**📌 3. Сравнение реакции на три вида рекламы**

**Задача:** Людей просят ранжировать три рекламных ролика от наиболее до наименее понравившегося.  
**Проблема:** Данные — **ранги**, нет числовых значений.

✅ **Метод:** Критерий Фридмана (для повторных измерений).  
❌ **Почему не ANOVA:** ANOVA требует количественных данных и гомогенности дисперсий.

**📌 4. Проверка, нормально ли распределён вес новорождённых**

**Задача:** Оценить, подчиняется ли вес новорождённых нормальному распределению.  
**Проблема:** Распределение может быть скошенным из-за недоношенных детей.

✅ **Метод:** Тест Шапиро–Уилка или Колмогорова–Смирнова.  
❌ **Почему не использовать среднее и стандартное отклонение:** Они ничего не говорят о **форме** распределения.

**📌 5. Проверка связи между стрессом и успеваемостью**

**Задача:** Измеряется уровень стресса (в баллах) и средний балл студентов.  
**Проблема:** Стресс может быть оценён субъективно, шкала — **непропорциональна**, а данные — с выбросами.

✅ **Метод:** Корреляция Спирмена (ранговая).  
❌ **Почему не Пирсон:** Пирсон требует линейности и нормальности распределений.

**📌 6. Опрос: любимый цвет**

**Задача:** Определить, равномерно ли распределены предпочтения по цвету (красный, зелёный, синий, жёлтый).  
**Проблема:** Данные — **номинативные** (категории).

✅ **Метод:** Тест χ² (хи-квадрат) согласия.  
❌ **Почему не параметры:** Тут нет ни среднего, ни дисперсии — только частоты.

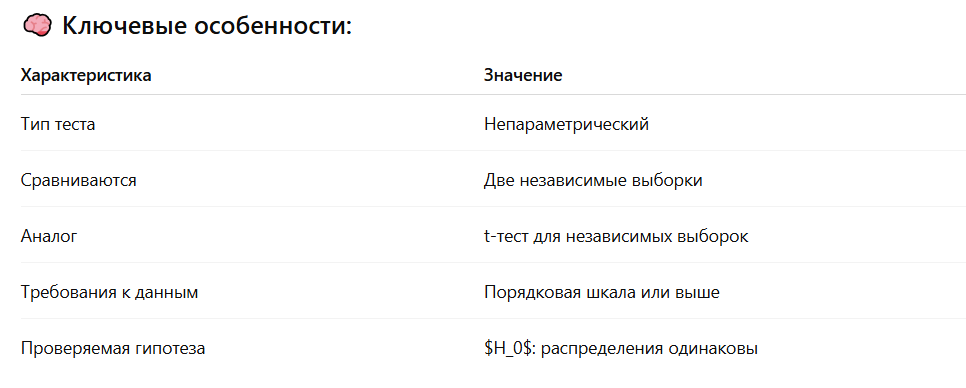
**Тест Манна–Уитни (U-тест)**

Тест **Манна–Уитни**, или **U-тест Манна–Уитни**, — это **непараметрический** статистический тест, используемый для сравнения **двух независимых выборок**. Его применяют, когда:

* выборки **нельзя считать нормально распределёнными**,
* или когда переменная **измерена хотя бы на порядковой шкале**,
* либо выборки имеют **маленький объём** (например, < 30 наблюдений).

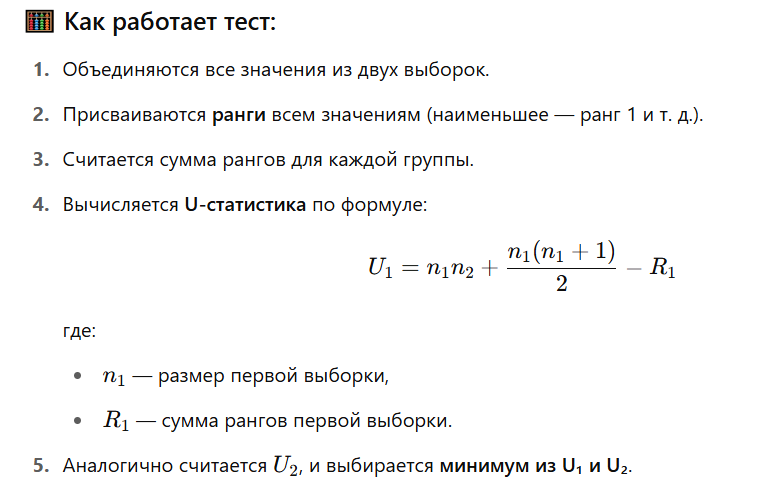
**📌 Основная цель:**

Проверить, различаются ли распределения (или медианы) двух независимых групп.



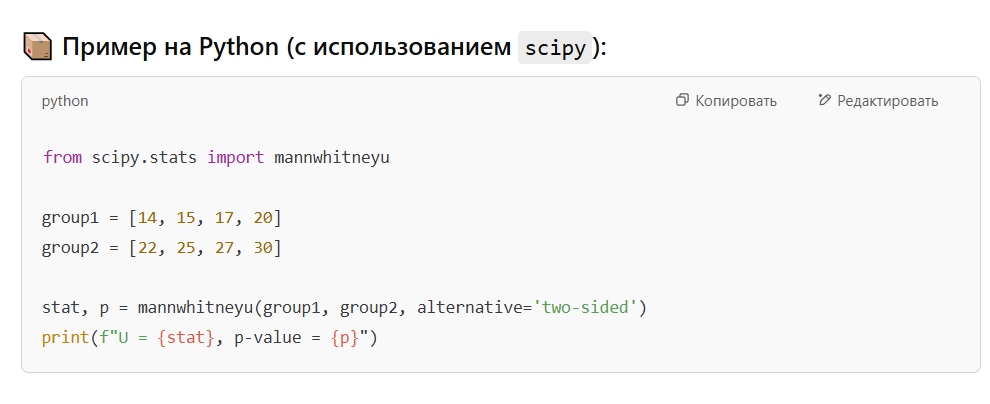
**📋 Формулировка гипотез:**

* **Нулевая гипотеза (H₀):** Распределения двух выборок одинаковы.
* **Альтернативная гипотеза (H₁):** Одно распределение систематически больше/меньше другого (в зависимости от направления теста).



**📊 Интерпретация:**

* Если p-значение меньше уровня значимости (например, 0.05), отвергаем H\_0​.
* Это значит, что есть статистически значимые различия между группами.



### ⚠️ Когда ****не стоит**** использовать U-тест:

* Если выборки зависимы → используйте **тест Вилкоксона**.
* Если данные дискретные с множеством повторяющихся значений — возможна потеря мощности теста.
* При больших выборках и нормальном распределении предпочтительнее **t-тест**.

**Тест Вилкоксона для зависимых выборок**

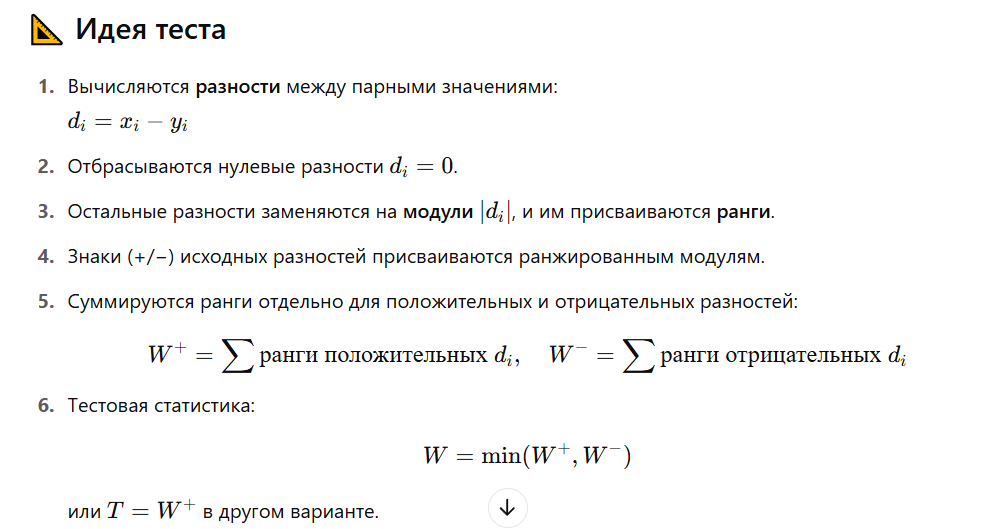
Тест Вилкоксона для зависимых выборок (точнее, **тест знаков рангов Вилкоксона**, англ. *Wilcoxon signed-rank test*) — это **непараметрический аналог парного t-теста**, который применяется для сравнения **двух связанных (зависимых) выборок**, например:

* до и после лечения;
* два метода измерения, применённые к одним и тем же объектам;
* оценки одного и того же субъекта в двух условиях.

## 🔍 Когда используется

Тест Вилкоксона применяется, если:

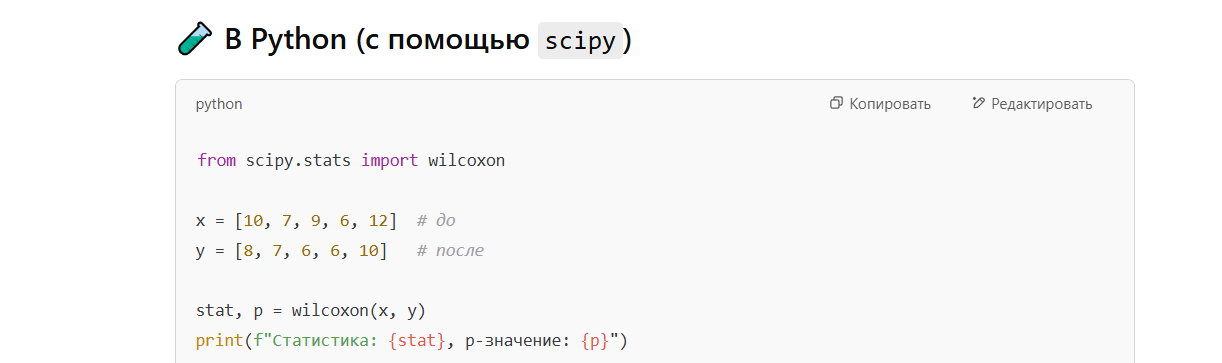
1. Данные парные (например, x1, x2 измерены на одних и тех же объектах);
2. Распределение разностей **не обязательно нормальное** (в отличие от парного t-теста);
3. Шкала данных — **порядковая или выше** (интервальная, но без требования к нормальности);
4. Размер выборки недостаточен для применения параметрических методов.



**📊 Гипотезы**

* **Нулевая гипотеза H\_0​**: медиана разностей равна нулю (нет эффекта).
* **Альтернативная гипотеза H\_1​**:
  + односторонняя: медиана > 0 или < 0;
  + двусторонняя: медиана ≠ 0.





**📌 Примечания**

* Для больших выборок используется **приближение нормальным распределением**.
* Если данных очень мало (например, n < 6), используют точные таблицы критических значений.

**Критерий знаков, критерий Краскела–Уоллиса**

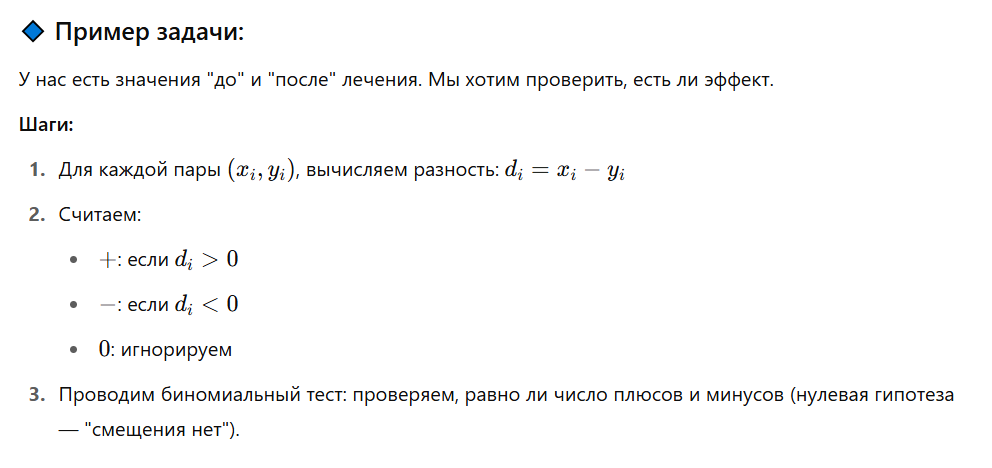
Разберём оба критерия — **критерий знаков** и **критерий Краскела–Уоллиса**, — которые относятся к **непараметрическим методам статистики**, то есть не требуют нормальности распределения данных.

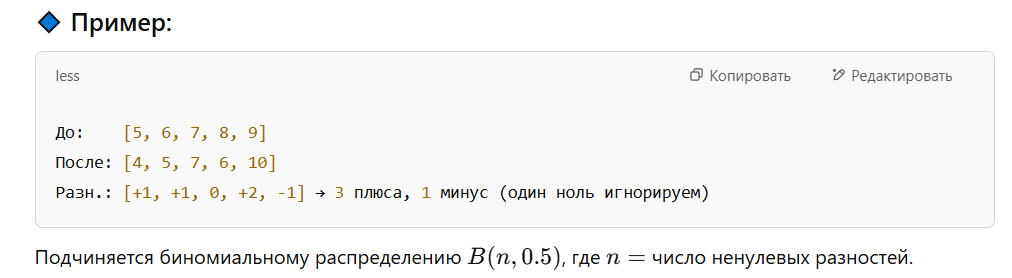
**📌 1. Критерий знаков (Sign test)**

**🔹 Применение:**

Для **одновыборочного случая** или для **парных (зависимых) выборок**, когда нас интересует, есть ли **смещение** в положительную или отрицательную сторону, **без учёта величины этого смещения**.

* Аналог парного t-теста.
* Используется, когда данные можно только сравнивать по знаку разности.

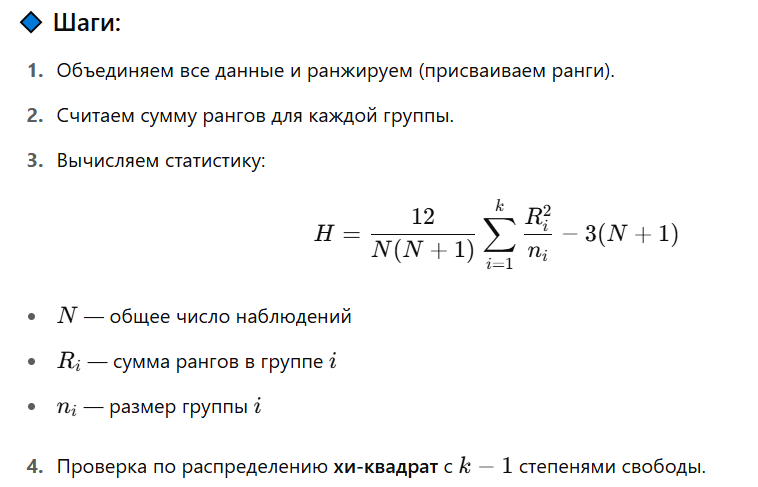


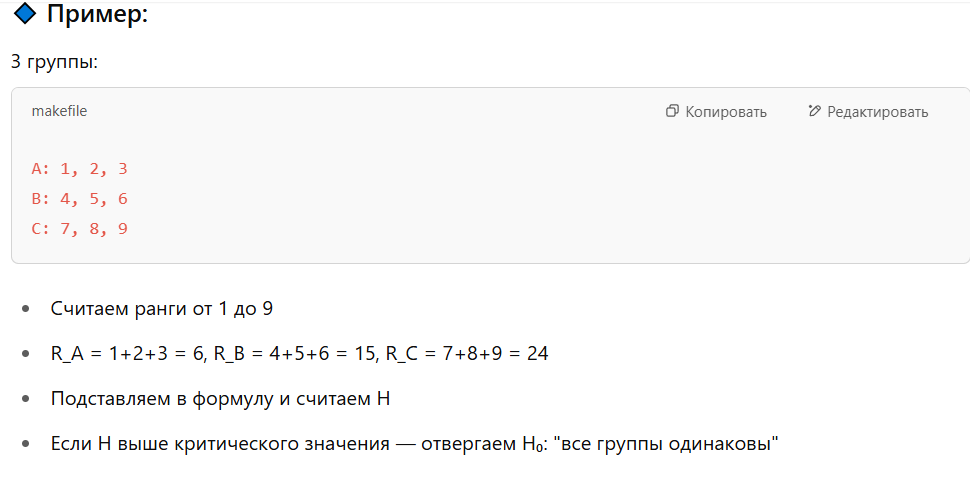


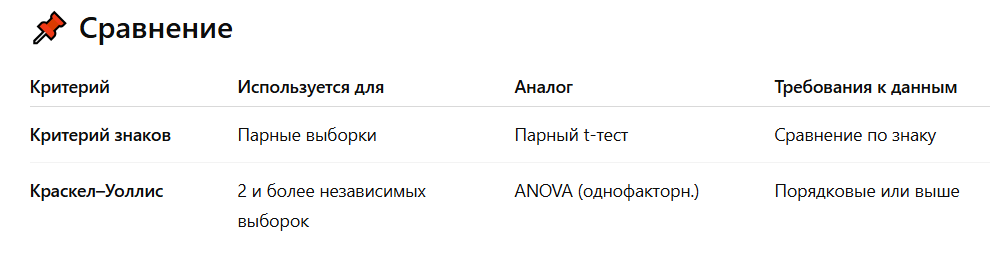
**📌 2. Критерий Краскела–Уоллиса (Kruskal–Wallis test)**

**🔹 Применение:**

Для **нескольких независимых выборок** (k-групп), если нужно проверить, одинаковы ли их распределения.  
Аналог однофакторного ANOVA (**дисперсионного анализа**), но **непараметрический**.

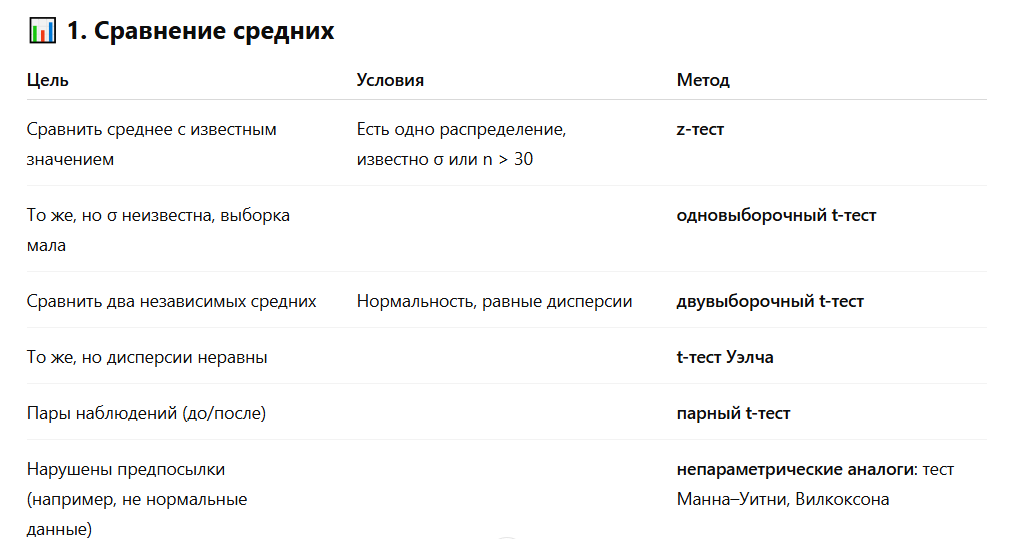


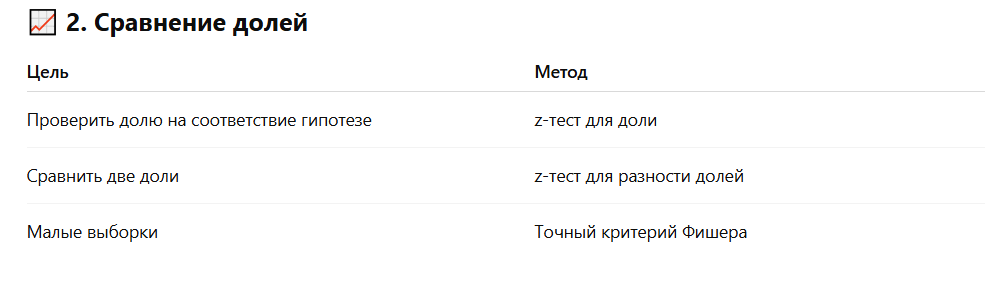




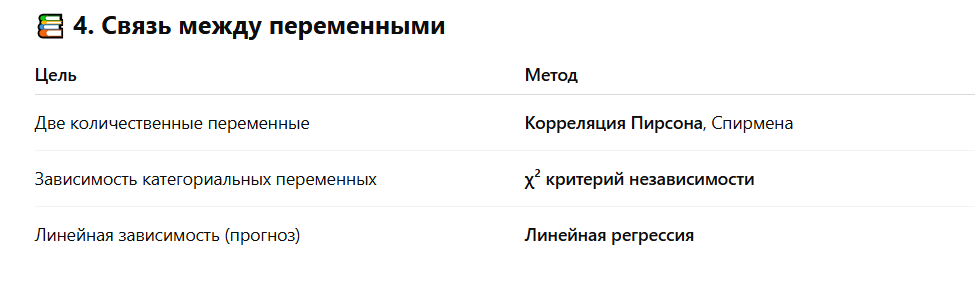
**Выбор метода в зависимости от задачи**

Выбор статистического метода зависит от **типа данных**, **цели анализа** и **гипотез**, которые ты проверяешь. Вот сводная таблица, помогающая выбрать подходящий метод в зависимости от задачи:

****

****

****

****

****

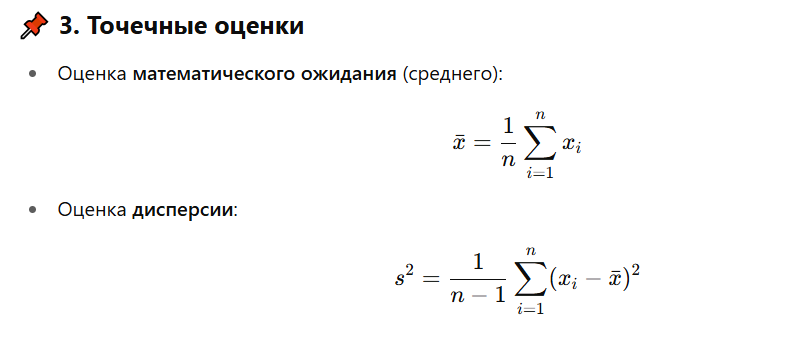
**Повторение ключевых понятий и формул**

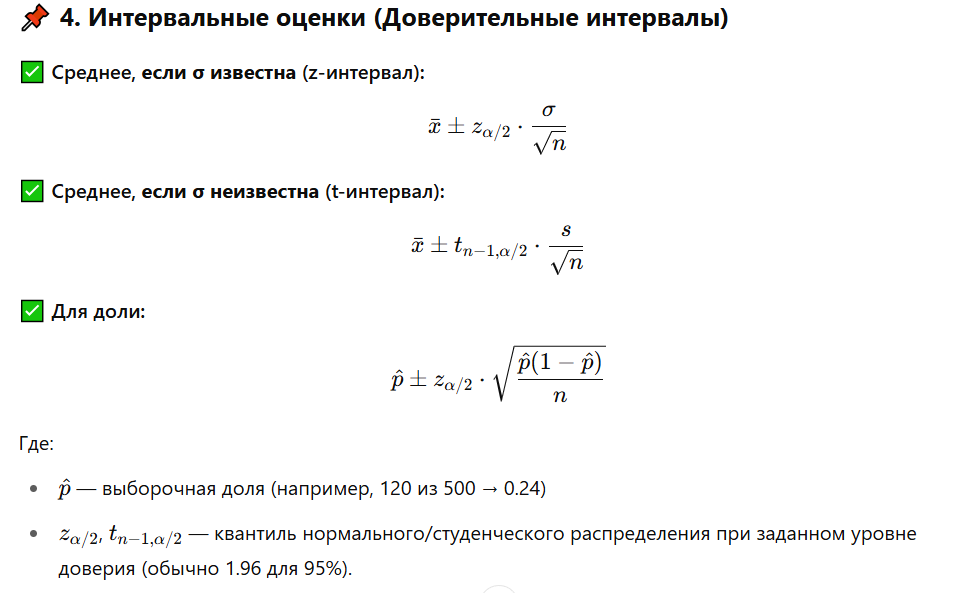
**📌 1. Типы данных**

* **Количественные:** дискретные (например, число детей), непрерывные (рост, вес).
* **Качественные (категориальные):** номинальные (цвет глаз), порядковые (оценка: плохо-средне-хорошо).

**📌 2. Выборка и генеральная совокупность**

* **Генеральная совокупность** — всё множество, которое изучается.
* **Выборка** — часть генеральной совокупности.
* **n** — объём выборки.

****

****

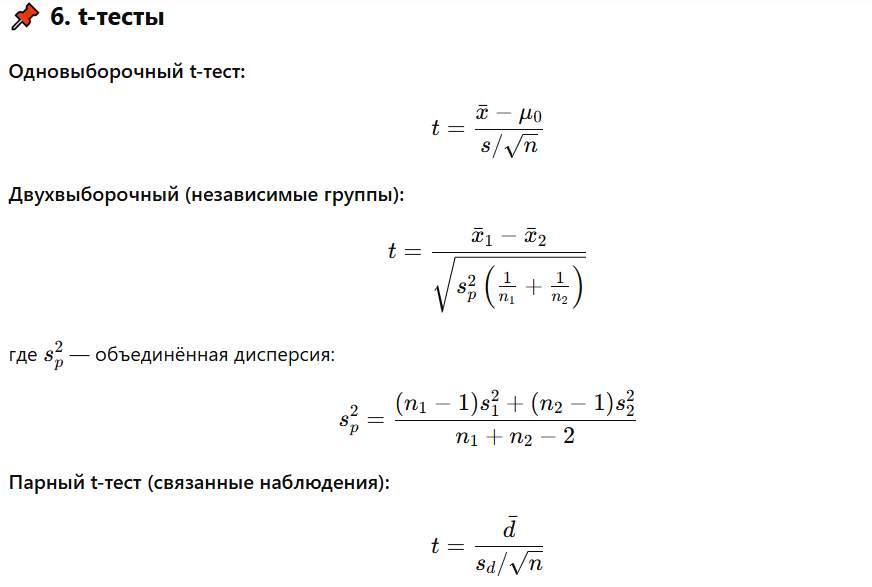
**📌 5. Проверка гипотез**

* **Нулевая гипотеза (H₀):** нет эффекта, разницы нет.
* **Альтернативная (H₁):** эффект есть, разница есть.

**Ошибки:**

* **I рода (α):** отвергли H₀, хотя она верна.
* **II рода (β):** не отвергли H₀, хотя она ложна.

**p-value:** вероятность получить результат не слабее наблюдаемого при условии, что H₀ верна.

****

****

**📌 8. Визуализация и описательная статистика**

* **Гистограмма** — распределение значений.
* **Boxplot (ящик с усами)** — медиана, квартили, выбросы.
* **Меры положения:** среднее, медиана, мода.
* **Меры разброса:** дисперсия, стандартное отклонение, размах.

**Интерпретация результатов в реальном контексте**

**📌 Общий алгоритм интерпретации статистических результатов**

1. **Понимание контекста задачи**  
   Что исследуется? Какая цель анализа? Кто участники? Какие последствия могут быть у разных выводов?
2. **Сопоставление результатов с гипотезой**
   * Что говорит p-value?
   * Каков результат теста? (отклоняем или не отклоняем H₀)
   * Что показывает доверительный интервал?
3. **Формулировка вывода человеческим языком**  
   Без сложных формул. Например:

«С вероятностью 95% доля довольных клиентов находится в пределах от 70% до 80%.»

1. **Оценка значимости результата в прикладном смысле**  
   Даже если результат статистически значим, **имеет ли он практическое значение**?
2. **Учёт ограничений анализа**  
   Была ли малая выборка? Нарушены ли предположения о нормальности? Есть ли смещения?

**🔍 Примеры интерпретации**

**✅ T-тест**

**Контекст:** сравнение среднего роста мужчин и женщин.  
**Результат:** p-value = 0.03  
**Интерпретация:**

«Разница в среднем росте между мужчинами и женщинами статистически значима на уровне 5%. Это значит, что с вероятностью 97% такая разница не могла возникнуть случайно.»

**✅ Доверительный интервал для доли**

**Контекст:** опрос 500 клиентов, 120 из них довольны.  
**95% ДИ:** [0.20; 0.28]  
**Интерпретация:**

«Мы можем быть уверены на 95%, что истинная доля довольных клиентов в популяции лежит в пределах от 20% до 28%.»

**✅ Непараметрический тест Манна–Уитни**

**Контекст:** сравнение уровня боли по шкале от 0 до 10 между группой, получившей обезболивающее, и группой плацебо.  
**Результат:** p-value = 0.001  
**Интерпретация:**

«Уровень боли в группе, получившей обезболивающее, **значительно ниже**, чем в группе плацебо. Вероятность случайного получения такого результата — менее 0.1%. Это подтверждает эффективность препарата.»

**⚠️ Типичные ошибки при интерпретации**

* Считать, что p-value = 0.05 означает 5% вероятность ошибочного вывода (это **не так**).
* Путать статистическую значимость с практической (например, разница в росте в 1 мм может быть значимой, но неважной).
* Игнорировать доверительные интервалы и ориентироваться только на p-value.