**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.

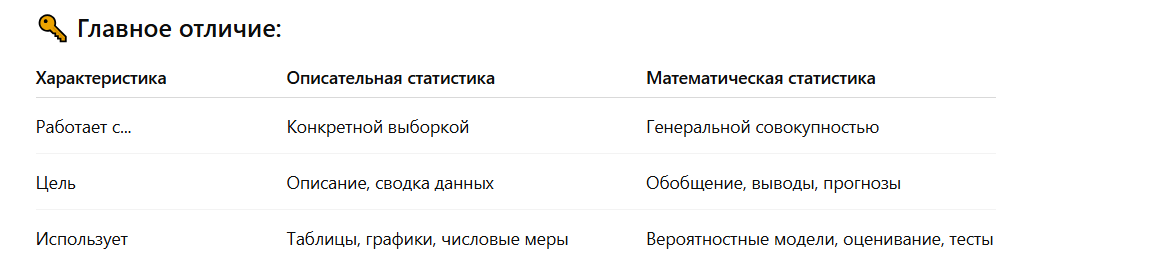
**📈 Математическая статистика (Inferential Statistics)**

**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**.

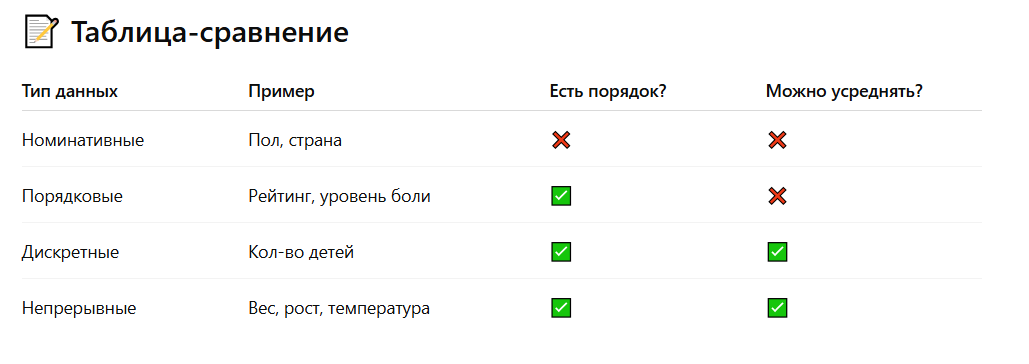
### 🔹 a) ****Номинативные (номинальные)****

* 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:

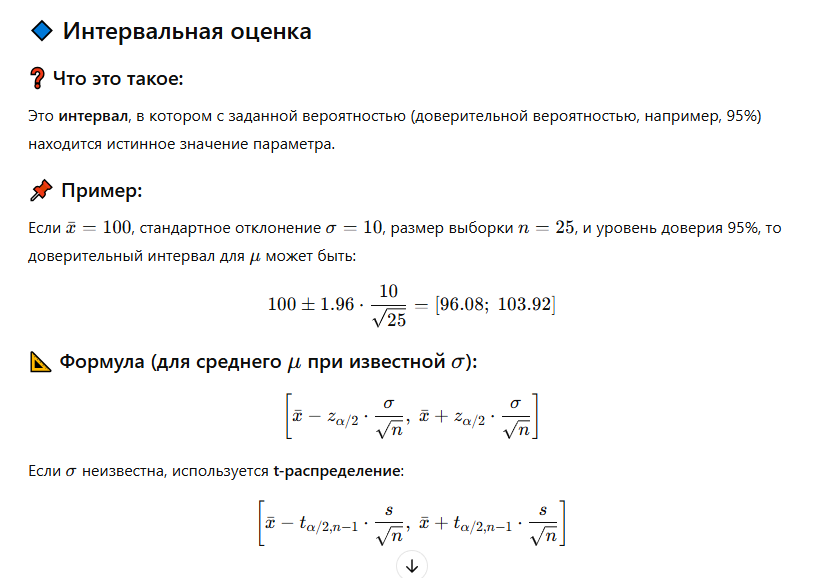
Это **одно число**, которое используется как "лучшая догадка" (оценка) неизвестного параметра.

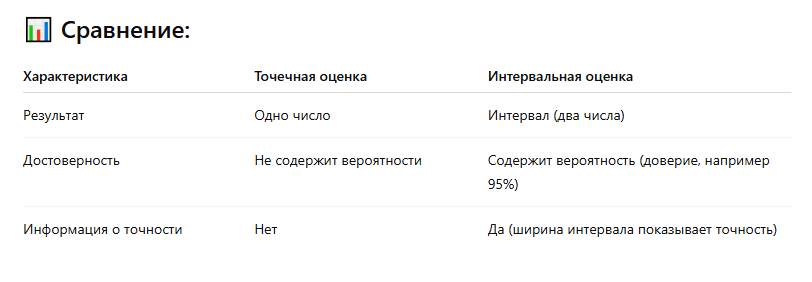
### 📌 Примеры:

* 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**.»

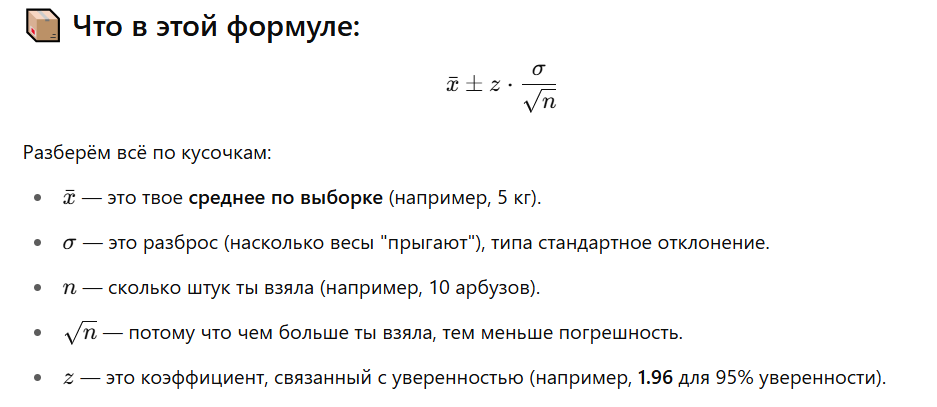
## 🔸 Как это устроено (на пальцах):

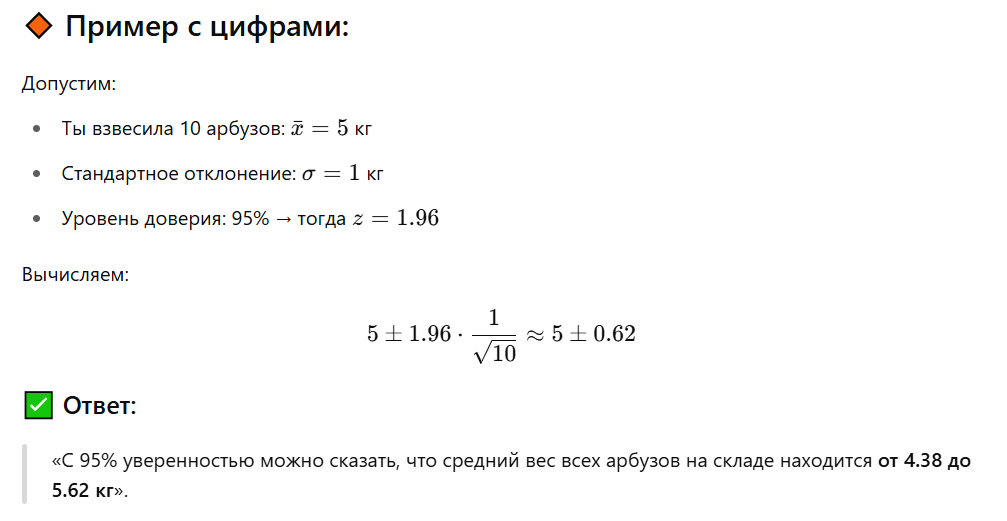
Интервал всегда строится так:

среднее значение выборки

±

запас на ошибку

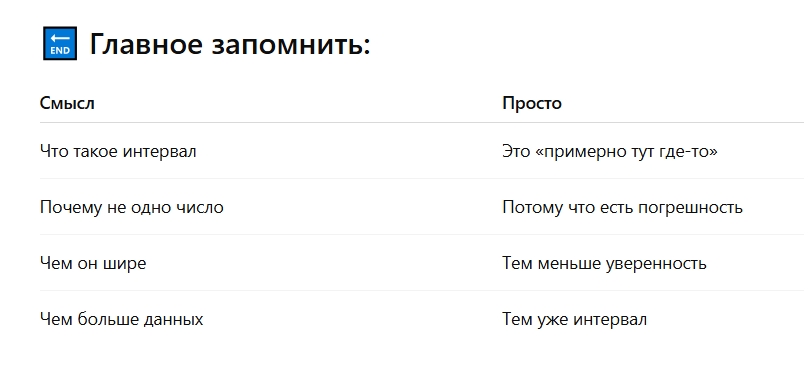
****

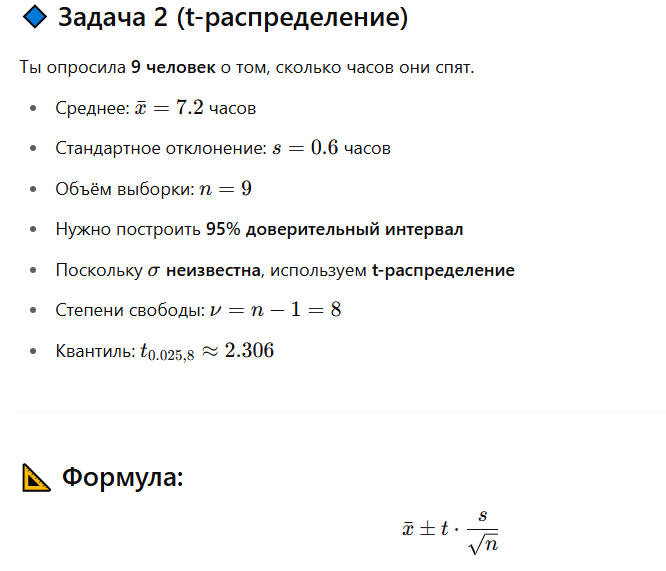
****

## 🔸 Что значит “95% уверенности”?

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

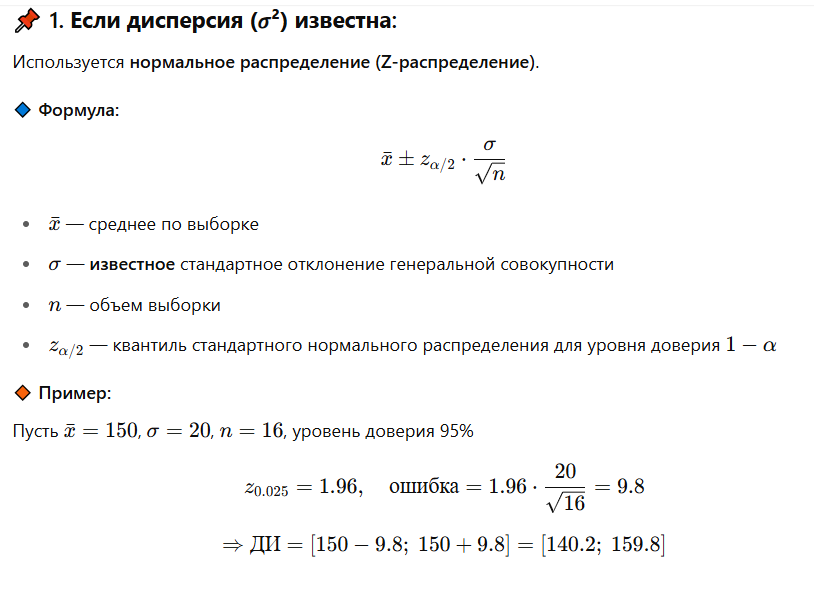
«Если бы я **много раз** брала по 10 арбузов и каждый раз строила такой интервал, то **в 95% случаев** в него попадало бы **истинное среднее**».

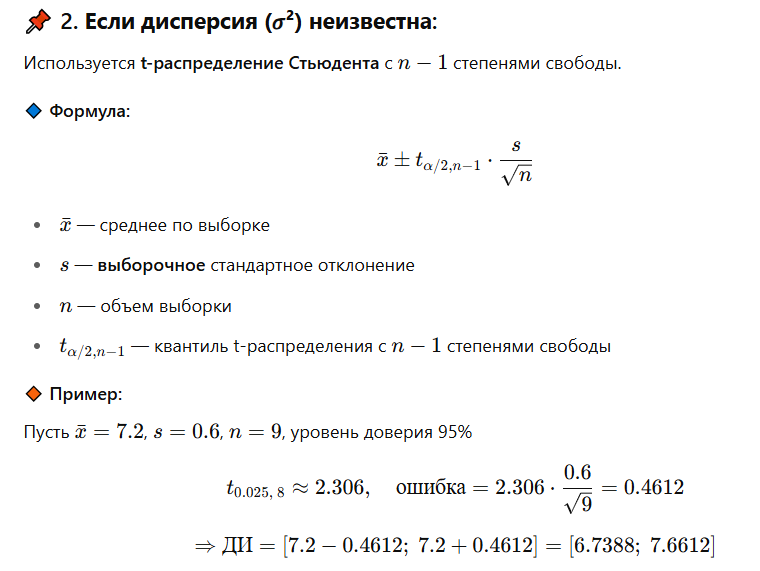
****

****

**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"

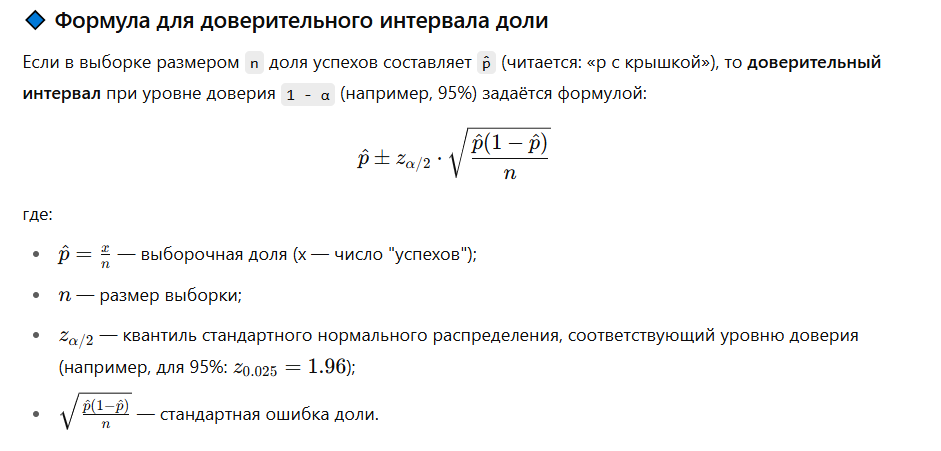
❌ Это **неправильно** с точки зрения строгой частотной статистики. Параметр (например, 𝜇) **фиксирован**, и он **либо в интервале, либо нет**.  
А **95% — это характеристика метода**, а не одного конкретного интервала.

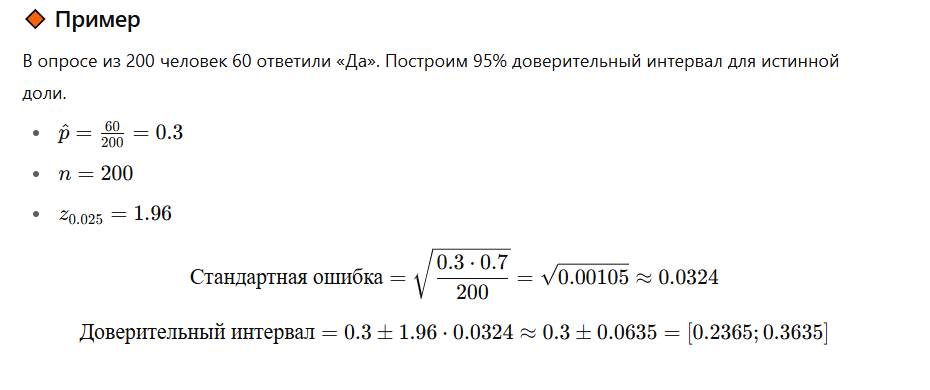
**📌 Как это работает:**

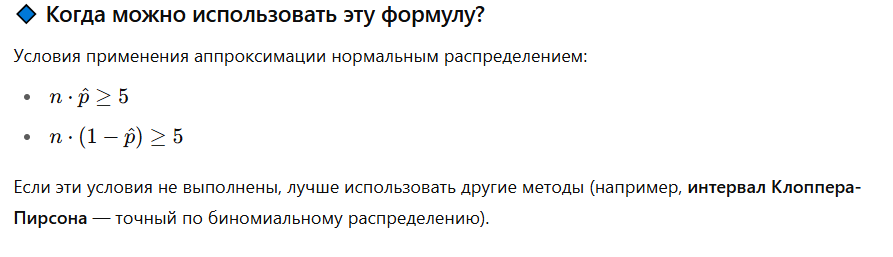
1. **Уровень доверия** задаёт **насколько широким будет интервал**:
   * 90% → более узкий
   * 95% → чуть шире
   * 99% → ещё шире  
     Шире — значит **больше уверенность, но меньше точность**.
2. Используются **квантильные значения** из Z- или t-распределения:
   * для 95%: z ≈ 1.96 (если σ известна)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

****

****

****