

# Конспект по основам статистики

Константин Кобранов

Основано на курсе Яндекс.Практикума, курса А. Карпова и собственных знаний.

## Оглавление

Конспект по основам статистики .....	1
Глоссарий .....	3
1. Описательная статистика .....	4
1.1. Среднее .....	4
1.2. Медиана .....	4
1.3. Квантили и квартили .....	4
1.4. Межквартильный размах и выбросы .....	5
1.5. Дисперсия и стандартное отклонение .....	6
1.6. Доверительные интервалы .....	7
2. Взаимодействия между переменными .....	8
2.1. Ковариация .....	8
2.2. Корреляция .....	9
Коэффициент корреляции Пирсона .....	9
Коэффициент корреляции Спирмана .....	9
Пример .....	10
2.3. Сравнение категорий .....	10
3. Типы распределений .....	12
3.1. Генеральная совокупность, выборка, случайная величина .....	12
3.2. Равномерное распределение .....	13
3.3. Нормальное распределение .....	14
3.3. Правило Трех Сигм .....	15
3.4. Проверка распределения на нормальность .....	16
QQ-plot (квантиль-квантиль) .....	16
Shapiro–Wilk test .....	17
4. Статистические тесты .....	17
4.1. Центральная предельная теорема .....	18
4.1. Стандартная ошибка среднего .....	18

4.2. Статистические гипотезы.....	18
4.3. Общий принцип проверки гипотез .....	19
Способ 1: Через p-value .....	19
Способ 2: Через критические значения (z-критическое, t-критическое и т.д.) ....	20
Способ 3: Через доверительные интервалы .....	20
4.4. Ошибки I и II рода, уровень значимости, мощность .....	20
4.4.1. Определения .....	20
4.4.2. Пример .....	21
4.5. Z-тест (Критерий Фишера) .....	22
4.5.1. Общее.....	22
4.5.2. Пример (Двусторонний z-тест) .....	23
4.5.3. Пример (Односторонний z-тест. Левый).....	25
4.6. T-тест.....	26
4.7. Сравнение групп.....	27
Одновыборочные и двухвыборочные тесты .....	27
Двухвыборочный T-тест .....	27
Бакетный тест .....	28
Z-тест для пропорций .....	28
Пример. Проверка гипотезы с помощью z-теста для пропорций .....	28
Сравнение методов .....	29
5. Дисперсионный Анализ (ANOVA [Analysis of Variance]) .....	30
5.1. Однофакторный дисперсионный анализ .....	30
6. Критерий Хи-квадрат (Критерий Пирсона).....	32
Пример применения к данным A/B тестирования .....	32
7. A/B Тестирования .....	34
7.1. Что такое A/B тест .....	34
Терминология A/B-тестирования .....	34
Последовательность шагов при проведении A/B-теста: .....	35
Шаг 1. Выбираем метрики и формулируем гипотезы.....	35
Шаг 2. Выбираем способ рандомизации и определяем параметры выборки (СЕТЕВОЙ ЭФФЕКТ) .....	36
Шаг 3. Определяем необходимый размер выборки .....	37
Шаг 4. Запускаем эксперимент и собираем данные (ПРОБЛЕМА ПОДГЛЯДЫВАНИЯ) .....	38

Шаг 5. Проверяем валидность эксперимента (AA-TECT, Sample Ratio Mismatch) .	38
Шаг 6. Рассчитываем результаты и принимаем решение о раскатке фичи .....	38
7.2. Количественные метрики .....	38
Пример применения бакетного теста для APRU.....	39
7.3. Конверсии и метрики-отношения.....	41
7.4. MDE и мощность.....	43
7.5. Объем групп и продолжительность теста .....	43
7.6. Проверка валидности эксперимента.....	46
7.7. Расчет и интерпретация результатов.....	49

---

## Глоссарий

**Параметрические тесты:** делают предположения о параметрах распределения (обычно — нормальность) [t-тест, z-тест, ANOVA, корреляция Пирсона, линейная регрессия].

**Непараметрические тесты:** не предполагающие конкретного распределения данных [U-тест Манна–Уитни, тест Краскела–Уоллиса, тест Вилкоксона, корреляция Спирмена].

## 1. Описательная статистика

### 1.1. Среднее

#### Типы данных:

- **Категориальные данные (номинальные)** описывают качественные характеристики и могут быть разделены на различные группы или категории. Например, пол, цвет глаз или марка автомобиля.
- **Порядковые данные**, как и категориальные, описывают качественные характеристики, но их значения можно проранжировать. Например, размер одежды, уровень образования.
- **Числовые (количественные) данные** — измеримые или счётные значения. Например, возраст, доход, рост. Могут быть дискретные и непрерывные.

Методы визуализации: столбчатая диаграмма и гистограмма.

Формула выборочного среднего:

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

где  $n$  — количество элементов в выборке.

### 1.2. Медиана

**Медиана** — это наблюдение, которое делит весь набор данных на две равные части: меньше него 50% наблюдений и больше него тоже 50% наблюдений.

#### Алгоритм вычисления медианы:

1. Упорядочить элементы в списке по возрастанию.
2. Посчитать количество элементов в списке.
3. а) Если число элементов в списке нечётное, найти число, стоящее посередине. б) Если число элементов чётное, найти два числа, которые находятся посередине, сложить их и результат разделить пополам.

### 1.3. Квантили и квартили

Число  $X$  является  $\alpha$ -квантилем набора данных, если оно делит этот набор данных таким образом, что  $\alpha\%$  наблюдений меньше или равны  $X$ , и  $(100 - \alpha)\%$  наблюдений больше или равны  $X$ .

#### Алгоритм определения $\alpha$ -квантиля:

1. Отсортируйте набор данных по возрастанию.
2. Найдите позицию квантиля по формуле:  $n \cdot \alpha$ , где  $n$  — количество элементов в наборе,  $\alpha$  — доля, которая нас интересует.
3. Определите значение квантиля:
  - Если позиция квантиля — целое число,  $\alpha$ -квантиль равен значению, которое соответствует этой позиции в упорядоченном наборе данных.
  - Если позиция квантиля — дробное число, возьмите среднее значение между двумя ближайшими соседями.

**Перцентиль** — это то же самое, что и квантиль, но в процентах.

**Квартили** делят выборку на 4 равные части: - Q1 (первый квартиль) — это 0.25-квантиль, - Q2 (второй квартиль) — это 0.5-квантиль (медиана), - Q3 (третий квартиль) — это 0.75-квантиль.

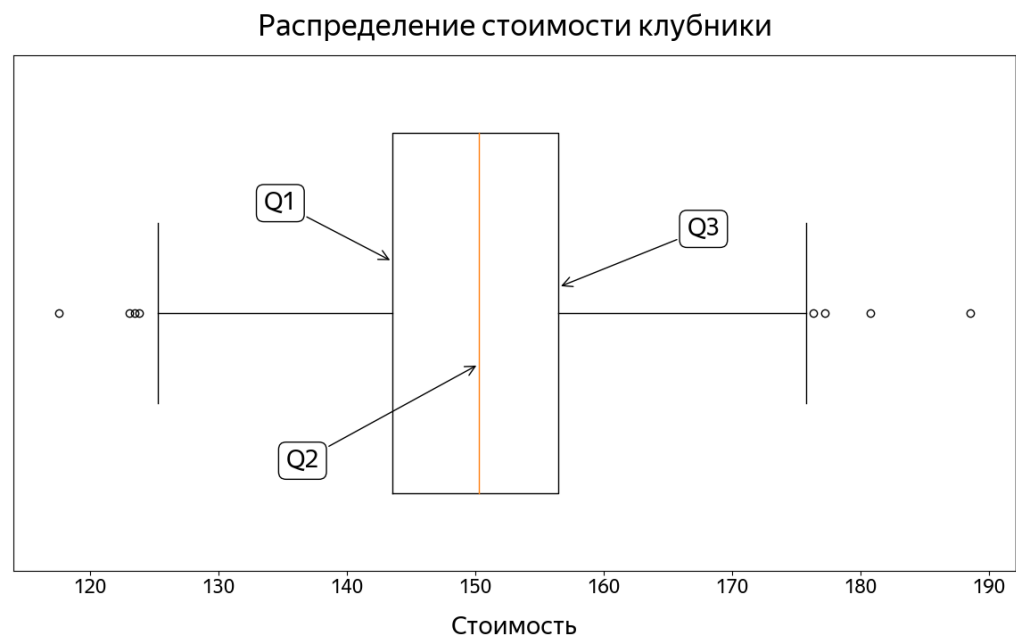
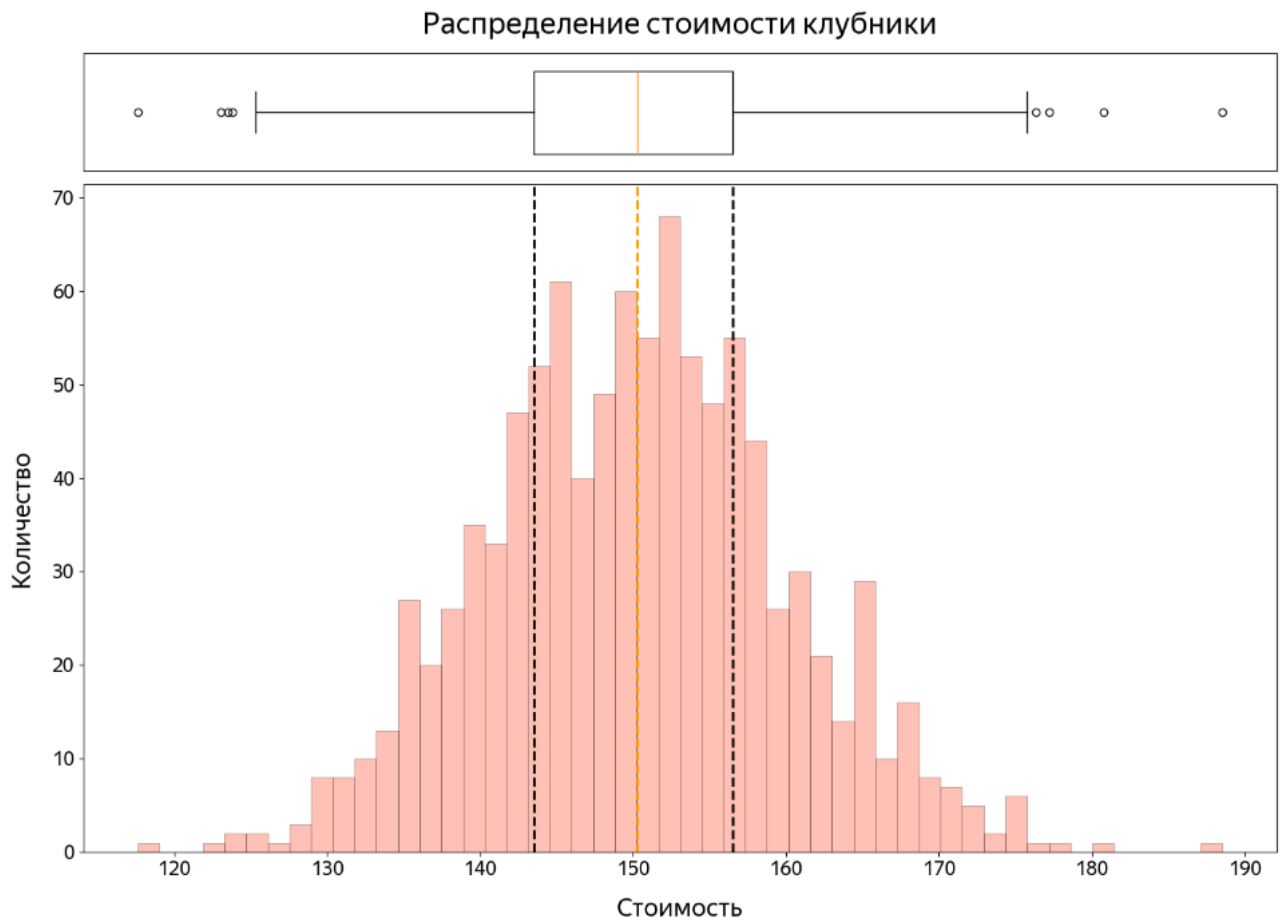
#### 1.4. Межквартильный размах и выбросы

$$IQR = Q3 - Q1$$

**Выброс** — это значение или набор значений в наборе данных, который сильно отличается от остальных.

##### Алгоритм отсеивания выбросов:

1. Отсортируем данные в возрастающем порядке.
2. Вычислим первый квартиль Q1 (0.25-квантиль) и третий квартиль Q3 (0.75-квантиль).
3. Рассчитайте межквартильный размах:  $IQR = Q3 - Q1$ .
4. Определите границы выбросов:
  - Нижняя граница:  $Q1 - 1.5 \cdot IQR$ ,
  - Верхняя граница:  $Q3 + 1.5 \cdot IQR$ .
5. Отсейте все значения, которые лежат за пределами этих границ, — они считаются выбросами.



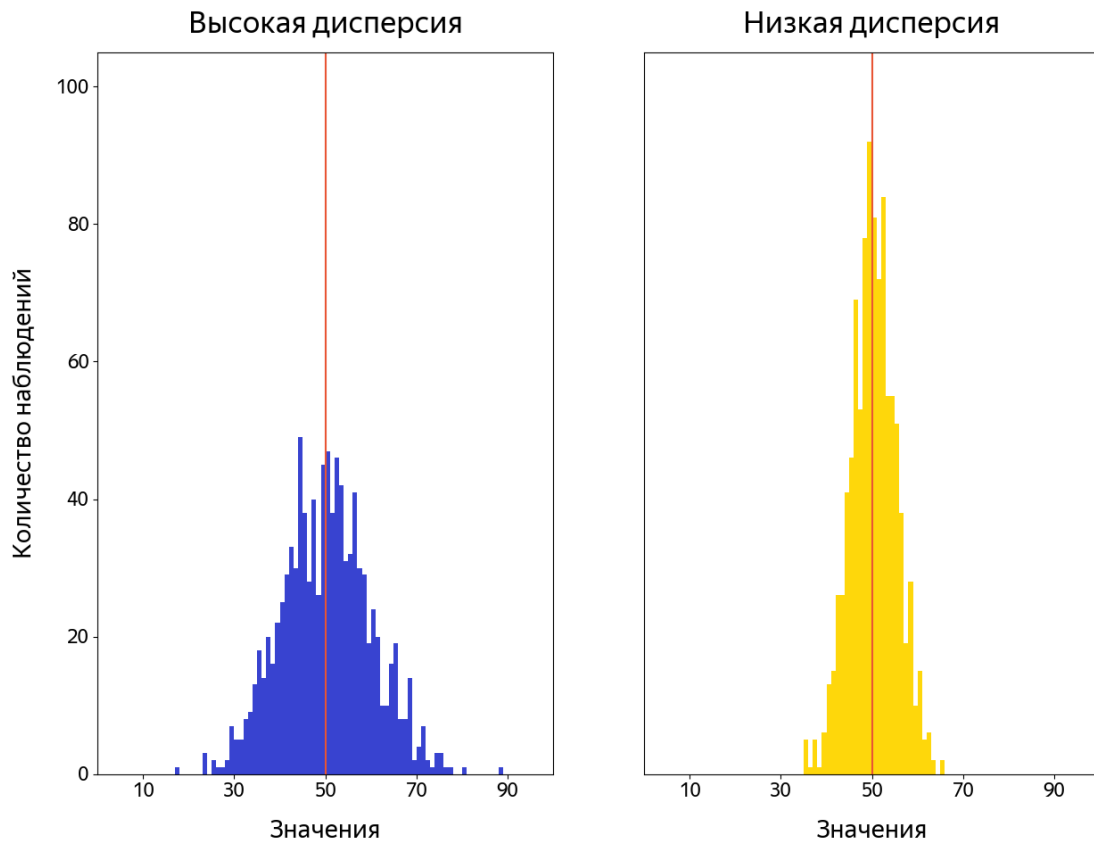
### 1.5. Дисперсия и стандартное отклонение

**Дисперсия** — это статистический показатель, который описывает разброс значений в наборе данных относительно среднего значения.

$$Var(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Стандартное отклонение** — это квадратный корень дисперсии, оно измеряется в тех же единицах, что и исходные данные.

$$s_X = \sqrt{Var(X)}$$



## 1.6. Доверительные интервалы

ДИ бывают разных видов и зависят от:

- Типа статистики (для среднего, для разницы средних, для доли)
- Известного стандартного отклонения (если известно, z-распределение, неизвестно – t-распределение)
- Распределения исследуемой величины

✓ Для среднего ( $\sigma$  известно):

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

✓ Для среднего ( $\sigma$  неизвестно):

$$\bar{x} \pm t_{\alpha/2, df} \cdot \frac{s}{\sqrt{n}}$$

✓ Для доли:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

✓ Для разницы средних (две независимые выборки):

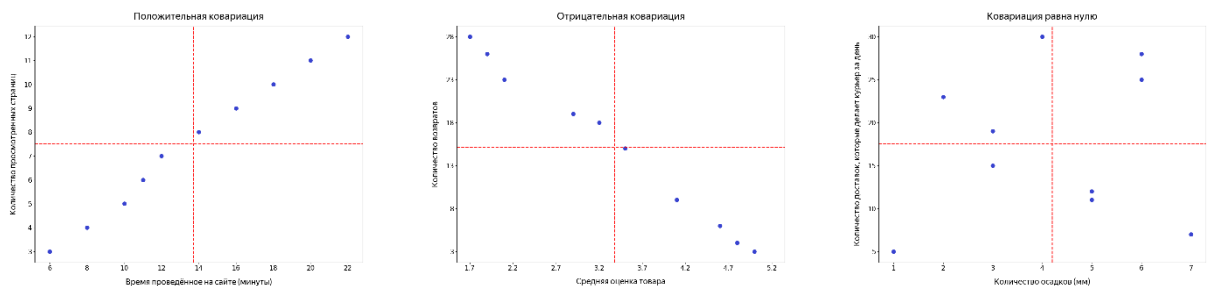
$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

## 2. Взаимодействия между переменными

### 2.1. Ковариация

**Диаграмма рассеивания** — это график, который позволяет визуализировать взаимоотношение двух числовых величин.

**Ковариация** — это мера совместной изменчивости двух величин. Она бывает положительной, отрицательной и равной нулю.



Формула ковариации:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



## 2.2. Корреляция

**Корреляция** — это мера силы и направления линейной взаимосвязи между двумя величинами. Она принимает значения от  $-1$  до  $1$ . Чем ближе абсолютное значение коэффициента корреляции к  $\pm 1$ , тем сильнее связь.

### Коэффициент корреляции Пирсона

Коэффициент корреляции Пирсона:

$$r_{X,Y} = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

**Важно:** При наличии выбросов использовать коэффициент корреляции Пирсона нельзя!

А что делать?

Надо либо удалять выбросы из данных, либо использовать другие методы, которые более устойчивы к выбросам. Например, ранговый коэффициент корреляции Спирмана.

### Применение и интерпретация

Коэффициент корреляции помогает определить, насколько сильно связаны величины. Но он не помогает определить, что было причиной, а что следствием.

Поэтому если исходные данные в задаче не удовлетворяют формальным требованиям (линейность и отсутствие выбросов), то часто аналитики выбирают не другой метод, а стараются решить эти проблемы в данных и дальше использовать именно коэффициент корреляции Пирсона.

### Коэффициент корреляции Спирмана

- Непараметрический аналог корреляции Пирсона
- Измеряет монотонную (не обязательно линейную) связь между переменными
- Основан на рангах, а не на самих значениях
- Подходит, когда данные не нормально распределены, есть выбросы или сильная нелинейность

$$\rho = \frac{\text{Cov}(R_x, R_y)}{s_{R_x} s_{R_y}} = \frac{\sum (R_{x_i} - \bar{R}_x) * (R_{y_i} - \bar{R}_y)}{\sqrt{\sum (R_{x_i} - \bar{R}_x)^2 * \sum (R_{y_i} - \bar{R}_y)^2}}$$

Где  $R_{x_i}$ ,  $R_{y_i}$  – ранги значений  $X$  и  $Y$ .

Данная формула применима даже при наличии одинаковых значений (тайов).

### Пример

**X Y Ранг X Ранг Y**

1	2	1.0	2.5
2	1	2.5	1.0
2	2	2.5	2.5
3	3	4.0	4.0

#### 1. Присваиваем ранги.

##### Ранги для X:

- 1 → ранг 1
- 2, 2 → ранг  $(2 + 3)/2 = 2.5$
- 3 → ранг 4

##### Ранги для Y:

- 1 → ранг 1
- 2, 2 → ранг  $(2 + 3)/2 = 2.5$
- 3 → ранг 4

#### 2. Считаем коэффициенты.

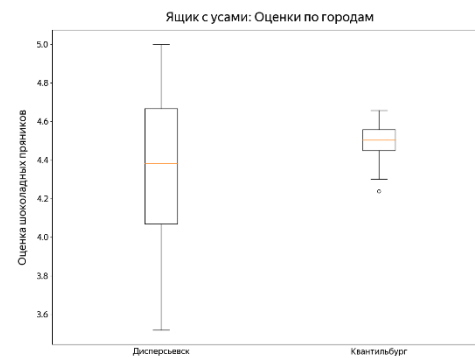
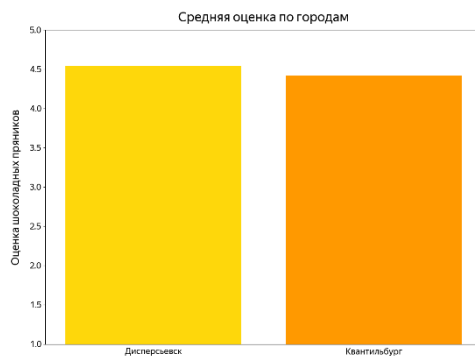
- Пирсон:  $r=0.5$
- Спирман:  $\rho=0.5$

В этом примере результат совпал, но при других данных с выбросами или сильной нелинейностью они могут различаться существенно.

### 2.3. Сравнение категорий

Для сравнения категориальных и числовых данных используют:

Метод	Преимущества	Недостатки
Столбчатая диаграмма со средними	Простота восприятия	Показывает только средние значения
Ящик с усами	Показывает распределение, медиану, выбросы	Требуется обучения для интерпретации



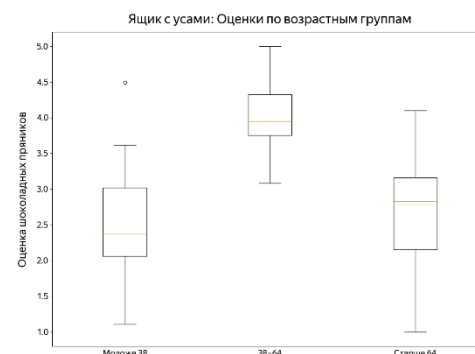
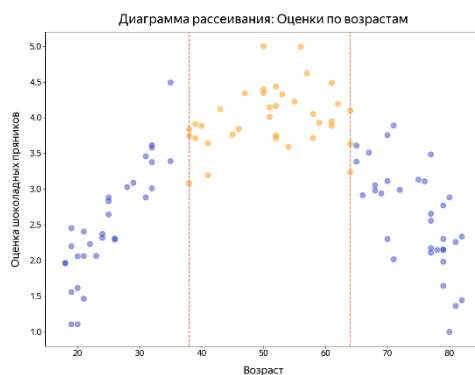
## 2.4. Бинаризация

**Бинаризация** — процесс преобразования числовой переменной в категориальную путём разделения её на интервалы.

Зачем? Например, когда зависимость нелинейная и трудно сделать какой-то вывод.

Пример (Бинаризация вручную)

Разобьем данные о возрасте на интервалы произвольно, чтобы построить box plots.



Методы бинаризации:

1. **Вручную:** на основе экспертных знаний
2. **Равные интервалы:**  $x_i = a + \frac{b-a}{k} i$  (Данные равномерно распределены и нет особого значения, насколько точно определены границы интервалов)
3. **Равные частотности:** одинаковое количество наблюдений в каждом интервале (Важно, чтобы интервалы содержали примерно одинаковое количество точек, например, для сбалансированного сравнения групп) [Обычно используют квантили]
4. **Кластерный анализ:** автоматическое определение интервалов

### 3. Типы распределений

#### 3.1. Генеральная совокупность, выборка, случайная величина

**Генеральная совокупность (Population)** — это полный набор всех элементов, которые исследуют в рамках задачи.

Правила, которым должна соответствовать ГС:

- Чёткое определение: четкие критерии принадлежности объектов к ней.
- Релевантность: соответствует целям исследования и включает все необходимые объекты.
- Достижимость: данные доступны и их можно было собрать.
- Воспроизводимость
- Временные рамки: если генеральная совокупность связана с конкретным временным интервалом, указывайте его в определении.

**Выборка (Sample)** — это отдельный набор элементов, отобранных из генеральной совокупности некоторым случайным процессом.

Аспекты, которые нужно учитывать при создании выборки:

- Репрезентативность: содержать характеристики, которые отражают генеральную совокупность в целом.
- Размер
- Способ отбора выборки

**Способы создания репрезентативной выборки:**

- Simple Random Sample (простая случайная выборка)
- Stratified Sample (Стратифицированная выборка): ГС разбивается на обособленные и различные по своей природе группы. Из каждой группы случайным образом выбираются элементы.

Пример: ГС разбивается на М и Ж. Из групп М и Ж случайно извлекаются элементы для формирования выборки.

- Cluster Sample (Групповая выборка): ГС разбивается на группы, похожие между собой.

Пример: Город разбивается на географические районы.

**Случайная величина** — это переменная, значение которой определяется случайными факторами и которая может принимать разные значения с определёнными вероятностями.

**Вероятность события** — это отношение числа случаев, когда событие произошло, к общему числу испытаний или наблюдений.

**Функция вероятности** определяет вероятность того, что случайная величина примет определённое значение. Обозначается как  $P(X = x)$ .

**Эмпирическая функция распределения** определяет вероятность того, что случайная величина примет значение, меньшее или равное заданному. Считается как  $\hat{F}(x) = P(X \leq x)$ .

Формула	Вероятность	Описание
$\hat{F}(x)$	$P(X \leq x)$	Вероятность того, что случайная величина примет значение, меньшее или равное заданному.
$1 - \hat{F}(x)$	$P(X > x)$	Вероятность того, что случайная величина примет значение больше заданного.
$\hat{F}(x_2) - \hat{F}(x_1)$	$P(x_1 < X \leq x_2)$	Вероятность того, что случайная величина примет значение в определённом диапазоне.

Обозначение:  $\hat{F}$  – эмпирическая,  $F$  – теоретическая функция распределения.

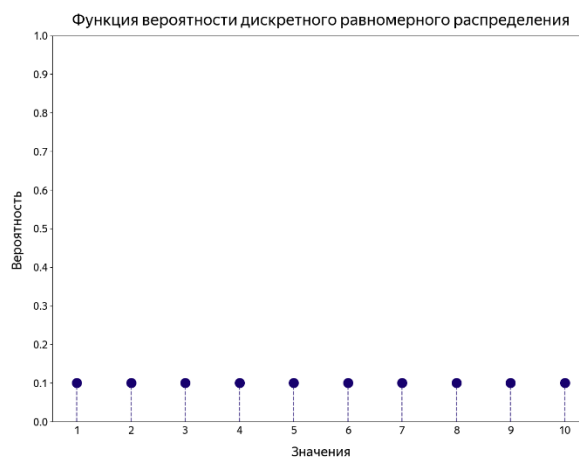
### 3.2. Равномерное распределение

**Теоретические распределения** — это математические модели, которые позволяют получить полное представление о данных.

**Равномерное дискретное распределение** — это тип вероятностного распределения, в котором каждое возможное значение случайной величины  $X$  имеет одинаковую вероятность и лежит в пределах от  $a$  до  $b$ , где  $a$  и  $b$  являются параметрами распределения. Короткое обозначение:  $X \sim U(a, b)$ .

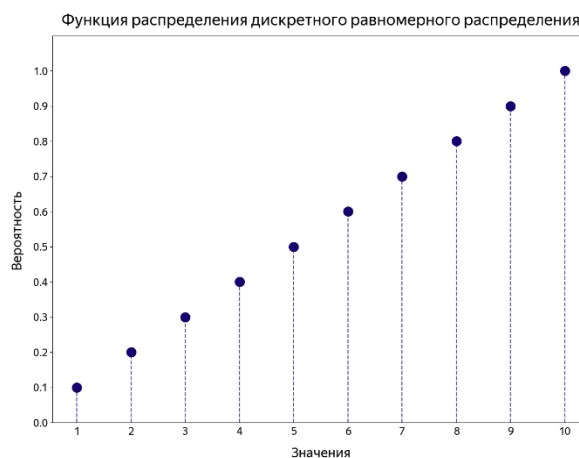
Функция вероятности:

$$P(X = x) = \frac{1}{n}$$



Функция распределения:

$$F(x) = \frac{x - a + 1}{n}$$



Математическое ожидание:

$$E(X) = \frac{a + b}{2}$$

Дисперсия:

$$Var(X) = \frac{n^2 - 1}{12}$$

### 3.3. Нормальное распределение

**Дискретная случайная величина** — это тип случайной величины, которая может принимать только определённые значения (обычно целые числа), например количество людей в очереди.

**Непрерывная случайная величина** — это тип случайной величины, которая может принимать любое значение внутри определённого интервала, например вес яйца.

**Функция плотности вероятности** — это функция, которая описывает вероятность того, что непрерывная случайная величина примет значение в определённом интервале. Она обозначается как  $f(x)$ .

**Свойства функции плотности вероятности:** -  $\int_{-\infty}^{\infty} f(x)dx = 1$  -  $f(x) \geq 0$  для всех  $x$

**Нормальное распределение** — это тип теоретического распределения, в котором значения в основном сосредоточены вокруг среднего. Это распределение имеет форму колокола и описывается двумя параметрами: средним значением  $\mu$  и дисперсией  $\sigma^2$ .

Функция плотности нормального распределения:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

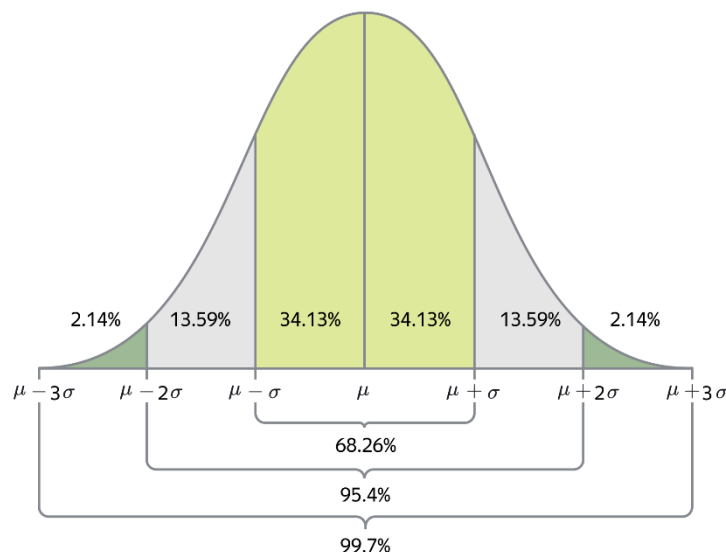
**Стандартное нормальное распределение** — частный случай нормального распределения, когда  $\mu = 0$ ,  $\sigma = 1$ .

$$z = \frac{x - \mu}{\sigma}$$

### 3.3. Правило Трёх Сигм

Для нормального распределения:

- В нормальном распределении примерно 68.3% всех значений находятся в пределах одного стандартного отклонения от среднего ( $\mu \pm \sigma$ ).
- Примерно 95.4% значений находятся в пределах двух стандартных отклонений ( $\mu \pm 2 \cdot \sigma$ ).
- И примерно 99,7% значений находятся в пределах трёх стандартных отклонений ( $\mu \pm 3 \cdot \sigma$ ).



Для стандартного нормального распределения:

95% наиболее вероятных значений такой случайной величины будут располагаться в интервале между  $-1.96$  и  $1.96$ .



### 3.4. Проверка распределения на нормальность

Существует три способа:

- Визуально по гистограмме.
- Визуально по QQ-plot.
- Тесты: Колмогорова-Смирнова, Шапиро-Вилки

#### QQ-plot (квантиль-квантиль)

Q-Q plot (Quantile-Quantile plot) — это график, на котором сравниваются квантили наблюдаемого распределения и квантили теоретического нормального распределения.

#### Вид Q-Q графика

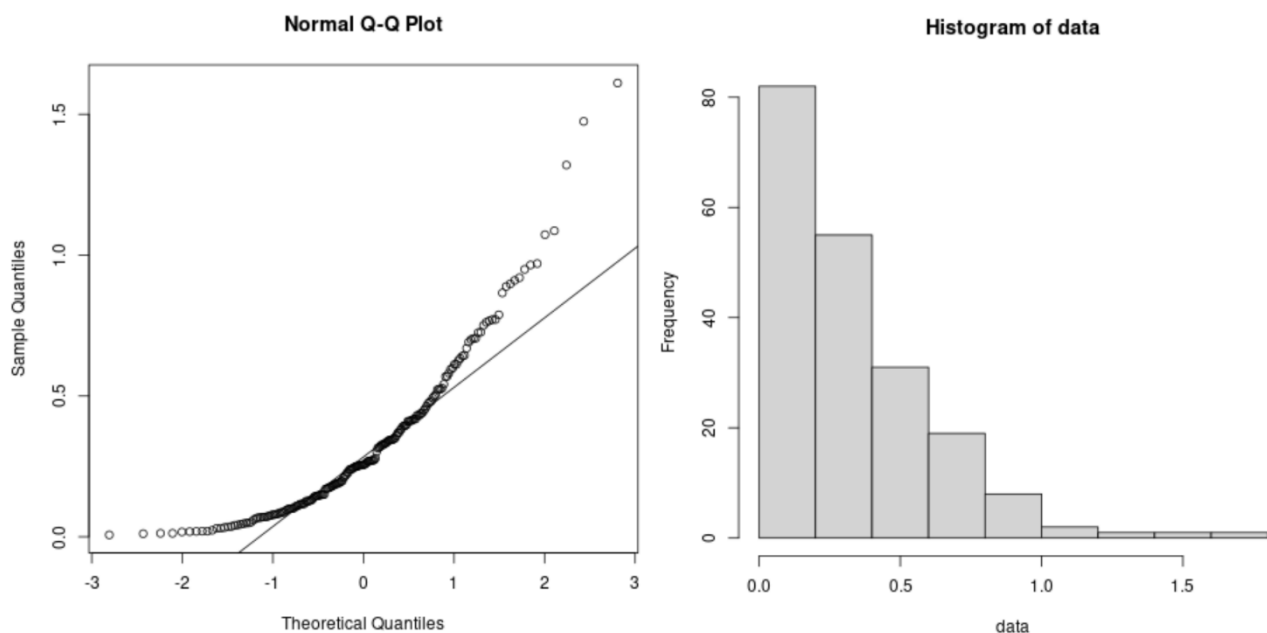
#### Интерпретация

Точки примерно лежат на прямой      Данные близки к нормальному распределению

Точки выгнуты вверх или вниз      Данные имеют **тяжёлые** или **лёгкие** хвосты

Точки S-образно отклоняются      Распределение **асимметрично**





### Shapiro–Wilk test

$H_0$ : данные имеют нормальное распределение

$H_1$ : данные не из нормального распределения

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$x_{(i)}$  – упорядоченные наблюдения (отсортированные по возрастанию).

$\bar{x}$  – среднее.

$a_i$  - константы, зависящие от ковариационной матрицы нормального распределения.

- **Преимущества:** Очень мощный для маленьких выборок ( $n < 50$ ), но работает и до  $n \approx 2000$ .
- **Когда использовать:** Малые и средние выборки.
- **p-value**  $< \alpha \rightarrow$  распределение не нормальное.

## 4. Статистические тесты

На практике мы не можем делать выводы по генеральной совокупности, а выбираем некоторую выборку.

## 4.1. Центральная предельная теорема

**Распределение выборочных средних** — это распределение, показывающее, какие значения принимает среднее значение случайной выборки из генеральной совокупности при многократном повторении эксперимента.

**Центральная предельная теорема:**

- Если мы многократно извлекаем выборки размера  $n$  и вычисляем среднее значение для каждой выборки ( $\bar{a}_j$ ) то распределение этих средних ( $A$ ) значений будет приближаться к нормальному распределению;
- Среднее значение ( $\bar{a}$ ) этого распределения будет практически совпадать со средним значением генеральной совокупности ( $\mu$ );
- А стандартное отклонение распределения выборочных средних ( $\sigma_a$ ) будет меньше стандартного отклонения генеральной совокупности ( $\sigma$ ), причем в  $\sqrt{n}$  раз!

$$\sigma_a = \frac{\sigma}{\sqrt{n}}$$

То есть:

$$\sqrt{n} * \frac{(\bar{a}_j - \mu)}{\sigma} \xrightarrow{n \rightarrow \infty} N(0, 1)$$

## 4.1. Стандартная ошибка среднего

Стандартное отклонение  $\sigma_a$  также называется стандартной ошибкой среднего (SE — Standard error). Когда мы оцениваем среднее генеральной совокупности с помощью среднего по выборке при заданных  $n$  и  $\sigma$ , стандартная ошибка помогает понять погрешность этой оценки.

Стандартная ошибка среднего (SE) – это стандартное отклонение выборочных средних, которое зависит от количества наблюдений в выборке и стандартного отклонения генеральной совокупности.:

$$SE = \sigma_a = \frac{\sigma}{\sqrt{n}}$$

## 4.2. Статистические гипотезы

В реальных исследованиях мы не знаем среднее значение генеральной совокупности, у нас есть только выборочное среднее, и мы хотим понять, можно ли ему верить. Нам может помочь один из методов статистики — тестирование гипотез. В нём мы будем учитывать погрешность измерений по отдельной выборке.

Гипотеза — это предположение, которое мы делаем о параметрах нашего распределения.

Гипотезы (тесты) бывают:

- односторонние (проверяем на равенство)
- двусторонние ( $>$ ,  $\geq$ ,  $<$ ,  $\leq$ ): левые ( $H_1: \mu < \text{some\_value}$ ), правые ( $H_1: \mu > \text{some\_value}$ ).

Тесты могут быть:

- Одновыборочные: сравнение среднего с некоторым числом в одной выборке.
- Двухвыборочные: применяется для сравнения средних значений двух независимых выборок.

Обозначения:

- $H_0$ : нулевая гипотеза (нет эффекта)
- $H_1$ : альтернативная гипотеза (есть эффект)
- $\alpha$ : уровень значимости (обычно 0.05) [вероятность ошибки I рода]
- $p$ -value: вероятность получить такие или более крайние результаты при верной  $H_0$

И так, у нас есть:

- истинное среднее значение генеральной совокупности, которое нам неизвестно;
- наше предположение о среднем значении;
- выборочное среднее.

По выборочному среднему мы хотим подтвердить предположение или опровергнуть его.

### 4.3. Общий принцип проверки гипотез

1. Формулируем гипотезы.
2. Выбираем уровень значимости.
3. Считаем статистику теста  $z$ ,  $t$ ,  $\chi^2$  и т.д.

Оцениваем результат, используя  $P$ -value или критические значения теста, или доверительные интервалы.

#### Способ 1: Через $p$ -value

**$P$ -value** — вероятность получить определённое или ещё более экстремальное значение статистического критерия при условии, что нулевая гипотеза верна.

То есть  $p$ -value – это площадь под кривой, ограниченная критическими значениями (см. рисунок ниже).



- Если  $p\text{-value} < \alpha \rightarrow$  **Отклоняем  $H_0$**  (значимый результат)
- Если  $p\text{-value} \geq \alpha \rightarrow$  **Не отклоняем  $H_0$**  (недостаточно доказательств против  $H_0$ )

*Способ 2: Через критические значения (z-критическое, t-критическое и т.д.)*

- Если статистика выходит за пределы критических значений, отклонить нулевую гипотезу.
- Если статистика не выходит за пределы критических значений, принимаем решение, что отклонить нулевую гипотезу нельзя.

*Способ 3: Через доверительные интервалы*

Идея: проверить входит ли тестируемое значение в доверительный интервал или нет.

- Если входит – не можем отклонить нулевую гипотезу.
- Если не входит – отклоняем нулевую гипотезу.

Например, если  $H_0: \mu_1 = 30$ , то проверяем входит ли 30. Если  $H_0: \mu_1 - \mu_2 = 0$ , то проверяем, входит ли 0.

## 4.4. Ошибки I и II рода, уровень значимости, мощность

### 4.4.1. Определения

**Ошибка первого рода** – отвергаем нулевую гипотезу, хотя она на самом деле верна (Вероятность совершить такую ошибку соответствует **уровню значимости  $\alpha$** )

*Пример:*

- метод контроля качества ошибочно обнаруживает отклонение от стандарта у таблетки, которая на самом деле соответствует стандарту.

**Ошибка второго рода** – не отвергаем нулевую гипотезу когда верна альтернативная (Вероятность совершить такую ошибку обозначим через  $\beta$ )

*Пример:*

- метод контроля качества не обнаруживает отклонение от стандарта у таблетки, которая на самом деле не соответствует стандарту.

**Уровень значимости  $\alpha$**  — это вероятность совершения ошибки первого рода, то есть отклонить нулевую гипотезу, когда она верна. Интерпретация: Если уровень значимости равен 0.05, это означает, что мы готовы принять риск ошибочного отклонения верной нулевой гипотезы в 5% случаев.

**P-value** — вероятность получить определённое или ещё более экстремальное значение статистического критерия при условии, что нулевая гипотеза верна.

**Мощность статистического теста,  $1-\beta$** — это вероятность правильно отклонить нулевую гипотезу, когда она действительно неверна. Иными словами, это вероятность обнаружить эффект, если он действительно существует. Мощность теста зависит от размера выборки и величины эффекта.

*Важно отметить, что мощность теста следует рассматривать ещё на этапе планирования исследования, так как она помогает определить необходимый размер выборки.*

#### 4.4.2. Пример

Пример:

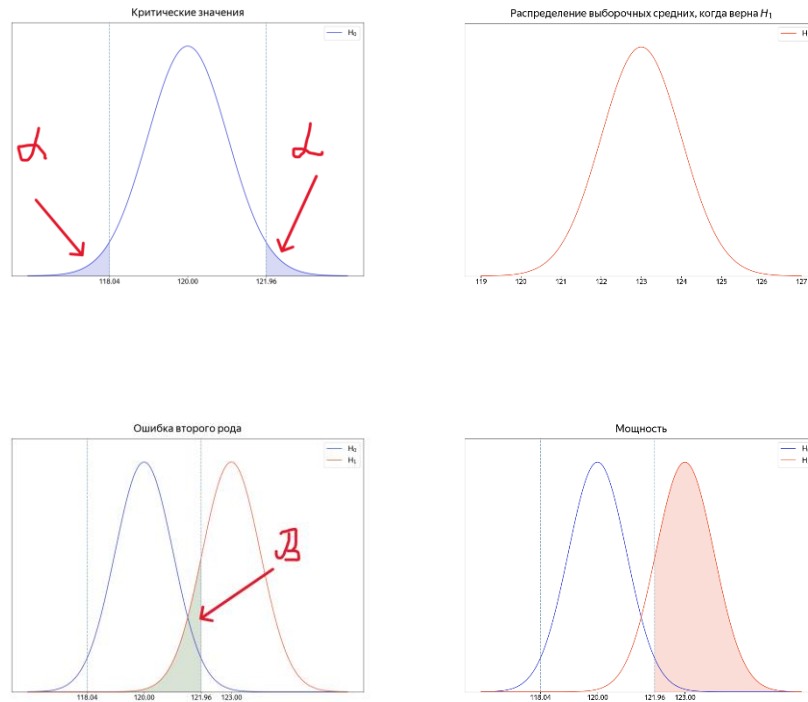
- $H_0: \mu = 120, H_1: \mu \neq 120, \alpha = 0.05, SE = 1$

Если мы выбираем стандартный уровень значимости,  $\alpha=0.05$ , то критические значения для z-статистики будут  $-1.96$  и  $1.96$ .

Найдём критические значения для средних:

- Левая граница:  $\mu - 1.96 \cdot SE = 120 - 1.96 \cdot 1 = 118.04$ ,
- Правая граница:  $\mu + 1.96 \cdot SE = 120 + 1.96 \cdot 1 = 121.96$ .

Многokrатно повторяем эксперимент, строим выборочные средние (1 график) со средним 120. Далее представим, что нулевая гипотеза неверна, а верна альтернативная. Пусть мы знаем, что настоящее среднее 123. Многократно повторяем эксперимент, строим выборочные средние, но уже со средним 123 (2 график). Далее совмещаем (3 график). Закрашенная область отражает ситуации, когда на самом деле верна альтернативная гипотеза. Далее просто строим  $1-\beta$  (4 график).



## 4.5. Z-тест (Критерий Фишера)

### 4.5.1. Общее

Формула z-статистики:

$$z = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\bar{x}$  – выборочное среднее;  $\sigma$  – стандартное отклонение ГС,  $n$  – размер выборки.

z- статистика подчиняется стандартному нормальному распределению.

#### Условия применимости теста:

- Размер выборки должен быть больше или равен 30.
- Известна дисперсия генеральной совокупности.
- Наблюдения независимы и распределение близко к нормальному.

#### Порядок действий для сценария использования p-value в рамках Z-теста:

- формулируем  $H_0$ ,  $H_1$
- выбираем  $\alpha$ , находим соответствующие критические значения;

- считаем выборочное среднее;
- считаем z-статистику;
- считаем p-value;
- сравниваем p-value с уровнем значимости:
  - Если p-value меньше уровня значимости, отклоняем нулевую гипотезу.
  - Если p-value больше уровня значимости, то оснований отклонять нулевую гипотезу нет.

#### 4.5.2. Пример (Двусторонний z-тест)

$H_0: \mu = 168, H_1: \mu \neq 168, \alpha = 0.05, \sigma = 3.9, n = 36, c = 0.95 \rightarrow \alpha = 0.05$

##### 1. Метод через z-value

Найдем z-статистику:

$z = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = 2.31$ . Значение положительно. Значит найдем критическое значение  $z_{critical}$ , которое ограничивает распределение справа (оно положительно).

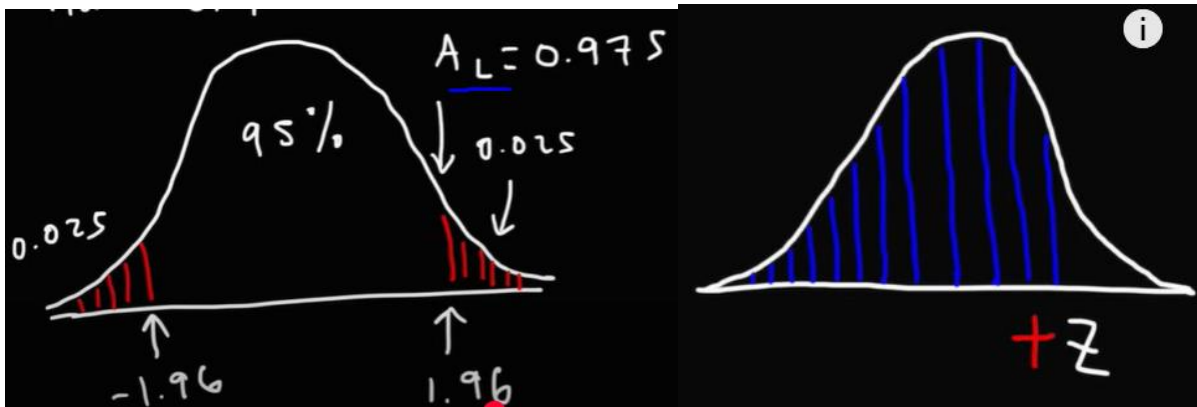
Поскольку мы знаем площадь, которую ограничивает критическое  $z$ , то по таблице можем найти само значение  $z_{critical}$  и сравнить с полученной  $z$ -статистикой.

Наше  $\alpha = 0.05$  делится пополам в закрашенной красной области. Чтобы воспользоваться таблицей, мы берем площадь  $A_t = 0.95 + \frac{\alpha}{2} = 0.975$ . Смотрим таблицу для положительных z-критических. Вероятности 0.975 соответствует  $z_{critical} = 1.96$ . Нормальное распределение симметричное  $\rightarrow$  слева  $z_{critical} = -1.96$ .

Интервал, при котором не отвергаем  $H_0$ :  $[-1.96, 1.96]$

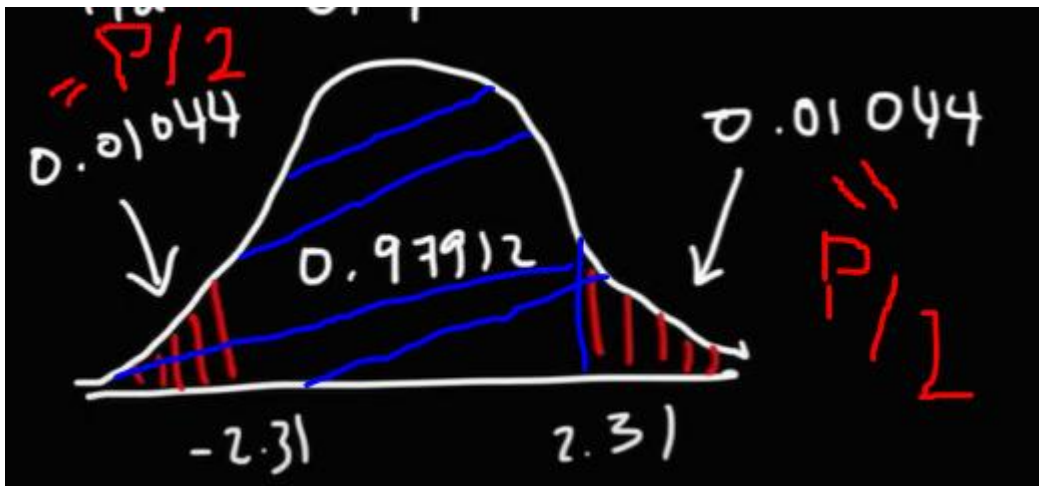
$z$  не принадлежит  $[-1.96, 1.96]$

$\rightarrow$  отвергаем  $H_0$



## 2. Метод через p-value

Визуально: нанесем теперь на график наши рассчитанные значения z-статистики. Они ограничиваются площадь, которая соответствует *p-value* (соответственно *p-value/2* слева и справа). Эту площадь и нужно сравнить с уровнем значимости.



Как и выше, воспользуемся положительной z таблицей, то есть рассмотрим **2.31**. Ему соответствует вероятность **0.98856** (как и выше, это будет полностью синяя область, только теперь до 2.31).

Т.к. мы хотим найти красную область. Следовательно:

$$(1 - 0.98956)/2 = 0.01044$$

Т.к. красных областей две, то  $p\text{-value} = 0.01044 * 2 = 0.02088$

$$p\text{-value} = 0.02088 < \alpha = 0.05 \rightarrow \text{reject } H_0$$

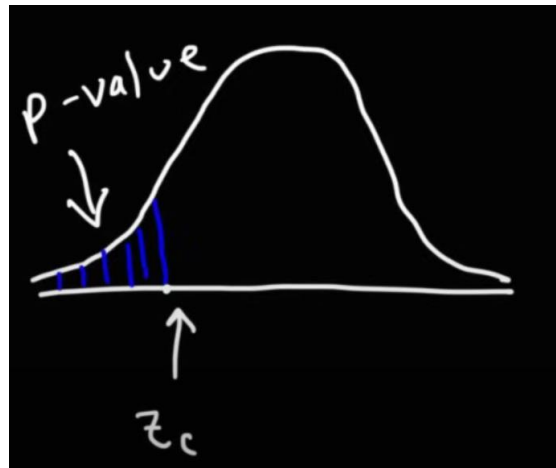


#### 4.5.3. Пример (Односторонний z-тест. Левый)

Основное отличие – не нужно делить p-value ни на что, т. к. тест односторонний.  
Область соответствующая p-value – слева.

$$\begin{array}{l} H_0: \mu \geq 5 \\ H_a: \mu < 5 \end{array} \quad n = 40 \quad \bar{x} = 4.8 \quad s = 0.50 \quad \mu_0 = 5$$

$\alpha = 0.02$   $C = 0.98$



$$z = \frac{\bar{x} - \mu_0}{SE} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -2.53 \rightarrow \text{using } z\text{-table (negative)} \rightarrow p\text{-value} = 0.0057$$

$$p\text{-value} < \alpha \rightarrow \text{reject } H_0$$

## 4.6. Т-тест

Формула t-статистики:

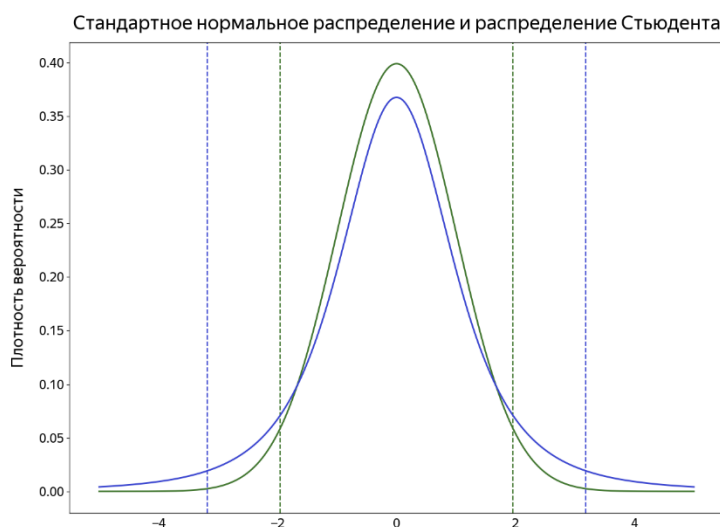
$$t = \frac{\bar{x} - \mu_0}{ESE} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$s$  – стандартное отклонение выборки.

Степени свободы:

$$df = n - 1$$

**Оценка стандартной ошибки** (Estimated Standard Error, ESE) — это стандартное отклонение выборочных средних, вычисленное на основе стандартного отклонения выборки.



Для t-статистики интервал получается немного шире, нам нужно раздвинуть границы, чтобы в него попало такое же количество возможных значений.

### Условия применимости:

- Необходимо, чтобы выборочные средние имели нормальное распределение и наблюдения были независимы.
- В случае применения двухвыборочного критерия для независимых выборок также необходимо соблюдение условия равенства дисперсий.
- Также не вполне корректно применять t-критерий Стьюдента при наличии в данных значительного числа выбросов.

*При несоблюдении этих условий при сравнении выборочных средних должны использоваться аналогичные методы непараметрической статистики, среди*

которых наиболее известными являются U-критерий Манна — Уитни (в качестве двухвыборочного критерия для независимых выборок), а также критерий знаков и критерий Уилкоксона (используются в случаях зависимых выборок).

## 4.7. Сравнение групп

### Одновыборочные и двухвыборочные тесты

**Одновыборочные тесты** - тесты, которые позволяют исследовать одну выборку.

**Двухвыборочные тесты** - помогают сравнить две выборки.

В двухвыборочных тестах хотим сравнить различаются ли средние двух генеральных совокупностей.

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

### Двухвыборочный Т-тест

Применяется когда: - Наблюдения независимы - Данные распределены нормально - Размеры выборок равны (для данной формулы ниже)

Формула t-статистики:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{ESE}$$

где:

$$ESE = \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}$$

Формула для степеней свободы:

$$df = \frac{(n-1)(s_1^2 + s_2^2)^2}{s_1^4 + s_2^4}$$

Обозначения: -  $\bar{x}_1, \bar{x}_2$  - выборочные средние -  $s_1, s_2$  - стандартные отклонения выборок -  $n$  - размер каждой выборки -  $\mu_1, \mu_2$  - математические ожидания.

**Напоминание:** Если рассматриваем двухсторонний тест, то  $\mu_1 - \mu_2 = 0$  в формуле выше.

**Для выборок разного размера:**

$$ESE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$df = \frac{(ESE_1^2 + ESE_2^2)^2}{\frac{ESE_1^4}{n_1 - 1} + \frac{ESE_2^4}{n_2 - 1}}$$

## Бакетный тест

Применяется когда распределение наблюдений значительно отличается от нормального:

1. Данные в каждой выборке разбивают на 100 бакетов (корзин)
2. Для каждого бакета вычисляют среднее значение. Получается для каждой выборки распределение выборочных средних (средних по бакетам). Размер такой выборки из бакетов, в данном случае, очевидно равен 100.
3. К полученным распределениям средних значений применяют **двухвыборочный Т-тест**

## Z-тест для пропорций

Данные должны быть бинарными (0, 1). Например, доля избирателей, голосующих за или против кандидата. Или еще лучше, подчиняться распределению Бернулли. Для таких данных дисперсия:  $Var = p \cdot (1 - p)$ .

P – доля (“вероятность”).

Используется для сравнения долей в двух выборках:

$$z = \frac{(p_1 - p_2)}{ESE}$$

где:

$$ESE = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

Обозначения: -  $p_1, p_2$  - расчётные доли в выборках -  $n_1, n_2$  - размеры выборок

*Пример. Проверка гипотезы с помощью z-теста для пропорций*

Рассматриваем результаты А/В теста:

Можно ли утверждать, что если реальный эффект и есть, то он точно меньше желаемого? Другими словами, проверим, есть ли у нас основания полагать, что эффект отличен от 0.04. Это значение выше желаемого.

$$H_0: CR_{test} = CR_{control} + 0.04$$

$$H_1: CR_{test} \neq CR_{control} + 0.04$$

- конверсия в заказ в тестовой группе: 71.15%,
- конверсия в заказ в контрольной группе: 68.9%,
- дисперсия конверсии в заказ в тестовой группе: 0.205,
- дисперсия конверсии в заказ в контрольной группе: 0.214,
- размер тестовой группы: 3126 пользователей,
- размер контрольной группы: 3099 пользователей.

✓ P-value > 0.05. Не отвергаем  $H_0$ .

Мы проверяем статистически значимое отличие от 0.04, а не от 0, так что z-статистику следует рассчитывать так:

$$z = \frac{(\bar{p}_{test} - \bar{p}_{control}) - (p_{test} - p_{control})}{\sqrt{\frac{Var_{test}}{n_{test}} + \frac{Var_{control}}{n_{control}}}} = \frac{(0.7115 - 0.689) - 0.04}{\sqrt{\frac{0.205}{3126} + \frac{0.214}{3099}}} \approx -1.51.$$

Соответствующее значение p-value  $\approx 0.131$ . Мы не отвергаем нулевую гипотезу.

## Сравнение методов

Тест	Назначение	Условия применения	Формула
Двухвыборочный t-тест	Сравнение средних	Нормальность, независимость	$t = \frac{\bar{x}_1 - \bar{x}_2}{ESE}$
Z-тест для пропорций	Сравнение долей	$np > 5, n(1 - p) > 5$	$z = \frac{p_1 - p_2}{ESE}$
Бакетный тест	Непараметрическое сравнение	Нарушение нормальности	Применение t-теста к бакетам

## 5. Дисперсионный Анализ (ANOVA [Analysis of Variance])

Статистический метод для проверки, **отличаются ли средние значения между тремя или более группами**.

Используется для оценки влияния независимых переменных – факторов (категориальных переменных) на зависимые переменные (количественных).

### 5.1. Однофакторный дисперсионный анализ

Это ANOVA, в которой есть один независимый фактор (категориальная переменная) с несколькими уровнями (группами), и одна зависимая количественная переменная.

**Гипотезы:**

- $H_0$ : Средние значения **во всех группах равны**  $\mu_1 = \mu_2 = \mu_3$
- $H_1$ : хотя бы одно среднее отличается.

**Пример:**

Ты хочешь проверить, отличается ли средняя производительность работников в трёх разных отделах компании: А, В и С.

- **Фактор:** отдел (А, В, С)
- **Зависимая переменная:** производительность (кол-во задач/час)

**Основная идея:**

Разбиваем всю изменчивость данных на:

1. **Межгрупповую дисперсию** (между группами) — зависит от различий между средними
2. **Внутригрупповую дисперсию** (внутри групп) — разброс внутри каждой группы

Затем считаем:

$$F = \frac{\text{Внутригрупповая дисперсия}}{\text{Внутригрупповую дисперсию}}$$

Сравниваем полученное F-значение с критическим по F-распределению.

Если  $F > F_{\text{крит}} \rightarrow \text{reject } H_0 \rightarrow$  хотя бы два средних различаются.

**Формулы:**

- Общая сумма квадратов:

$$SS_{total} = \sum (x_{ij} - \bar{x}_{\text{общ}})^2$$

$\overline{x_{общ}}$  – общее среднее по всем группам (складываем все значения наблюдений из каждой группы и делим на суммарное количество наблюдений)

- Межгрупповая сумма квадратов:

$$SS_{between} = \sum n_j (\bar{x}_j - \overline{x_{общ}})^2$$

$n_j$  – количество элементов в группе,  $\bar{x}_j$  – среднее группы.

- Внутригрупповая сумма квадратов:

$$SS_{within} = \sum \sum (x_{ij} - \bar{x}_j)^2$$

- Степени свободы:

- $df_{between} = k - 1$

- $df_{within} = N - k$

$N$  – суммарное число элементов по всем группам,  $k$  – число групп.

- Дисперсии:

$$MS_{between} = \frac{SS_{between}}{df_{between}}, \quad MS_{within} = \frac{SS_{within}}{df_{within}}$$

- Статистика F:

$$F = \frac{MS_{between}}{MS_{within}}$$

## 6. Критерий Хи-квадрат (Критерий Пирсона)

Это статистический тест, с помощью которого проверяют, насколько наблюдаемые частоты отличаются от ожидаемых.

Есть 2 основных применения:

1.  **$\chi^2$  тест на согласие (goodness-of-fit)**  
→ Проверяет, соответствует ли распределение теоретическому (например, равномерному).
2.  **$\chi^2$  тест на независимость (chi-square test of independence)**  
→ Проверяет, есть ли связь между двумя категориальными переменными.  
📌 Именно он используется в **A/B тестах**.

**Статистика:**

$$\chi^2 = \frac{\sum_{i=1}^r \sum_{j=1}^c (O_{ij} - E_{ij})^2}{E_{ij}}$$

где  $i$  – номер строки (от 1 до  $r$ ),  $j$  – номер столбца (от 1 до  $c$ ),  $O_{ij}$  – фактическое количество наблюдений в ячейке  $ij$ ,  $E_{ij}$  – ожидаемое число наблюдений в ячейке  $ij$ .

Число степеней свободы:

$$df = (r - 1) * (c - 1)$$

Где  $r$  – число строк,  $c$  – число столбцов.

- Если  $\chi^2 >$  критического значения или  $p\text{-value} < \alpha \rightarrow$  **отвергаем  $H_0$**
- Иначе  $\rightarrow$  **принимаем  $H_0$**

**Условия применения:**

- данные **категориальные** (например, покупка: да/нет),
- есть **две или более групп**,
- *число наблюдений* (см. пример ниже) в ячейках достаточно большие (рекомендуется  $\geq 5$ ). Иначе применяют критерий **Фишера**.

### Пример применения к данным A/B тестирования

Ты тестируешь две версии email-рассылки:

- Группа **A** получает старый дизайн
- Группа **B** — новый дизайн

Хочешь понять, влияет ли дизайн на **клик по ссылке**.



### Группа Кликнули Не кликнули Итого

A	20	80	100
B	40	60	100
<b>Итого</b>	60	140	200

#### Шаг 1: Гипотезы

- **Нулевая гипотеза ( $H_0$ ):** дизайн **не влияет** на клики → группы A и B независимы
- **Альтернативная гипотеза ( $H_1$ ):** дизайн **влияет** на клики → группы зависимы

#### Шаг 2: Считаем ожидаемые частоты

$$E_{ij} = \frac{(\text{Сумма по строке}_i) * (\text{Сумма по столбцу}_j)}{\text{Общий итог}}$$

Например, ожидаемое число кликов в группе A:

$$E_{A,\text{Клик}} = \frac{100 * 60}{200} = 30$$

Полная таблица ожидаемых значений:

### Группа Кликнули (E) Не кликнули (E)

A	30	70
B	30	70

#### Шаг 3: Считаем статистику

$$\chi^2 = \frac{\sum (O - E)^2}{E}$$

Рассчитаем для каждой ячейки:

- А, клик:  $\frac{(20-30)^2}{30} = \frac{100}{30} \approx 3.33$
- А, не клик:  $\frac{(80-70)^2}{70} = \frac{100}{70} \approx 1.43$
- В, клик:  $\frac{(40-30)^2}{30} = \frac{100}{30} \approx 3.33$
- В, не клик:  $\frac{(60-70)^2}{70} = \frac{100}{70} \approx 1.43$

Суммируем:

$$\chi^2 \approx 3.33 + 1.43 + 3.33 + 1.43 = 9.52$$


Шаг 4: Число степеней свободы

$$df = (2-1)(2-1) = 1$$

При  $\alpha = 0.05$ , критическое значение из таблицы  $\chi^2$ :  $\chi_2^{crit} = 3.841$

Шаг 5: Выводы

- Наша  $\chi^2 = 9.52 > 3.841$
- Или p-value < 0.05 (можно посмотреть через калькулятор)

 **Отвергаем  $H_0$**  → есть статистически значимая разница между группами А и В

## 7. А/В Тестирования (доб про дельта-метод, бутстрэп или линеаризация)

### 7.1. Что такое А/В тест

**А/В-тест** — это инструмент, который позволяет делать надёжные выводы о влиянии изменения на продукт, за счёт использования статистических методов и параллельного сбора данных для сравниваемых групп.

*Терминология А/В-тестирования*

- **Контроль (группа контроля, контрольная группа)** — это группа пользователей, для которых не вводят изменений
- **Тест (группа теста, тестовая группа)** — это группа пользователей, для которых вводят тестируемое изменение.
- **Тритмент** (англ. treatment — лечение) — это изменение, которое вводят для тестовой группы. В нашем случае это вид экрана чекаута.

- **Фича** (англ. feature — особенность) — похожий по значению на «третмент», но более общий термин, который используется в IT-компаниях для обозначения нового функционала в продукте.
- **Раскатить фичу** — ввести изменение для всех пользователей или для части.
- **Метрика** — показатель, изменение которого анализируют по результатам эксперимента. Например, прибыль с пользователя или количество заказов.
- **Зелёный тест** — эксперимент, по результатам которого зафиксировали статистически значимый прирост метрики.
- **Красный тест** — эксперимент, по результатам которого зафиксировали статистически значимое падение метрики.
- **Серый тест** — эксперимент, по результатам которого не зафиксировали статистически значимого изменения метрики.
- Субъект, которого в эксперименте относят в одну из групп: тестовой или контрольной, называется **единицей рандомизации**. Чаще всего единица рандомизации — это пользователь.
- Субъект, для которого считают изменение метрики, называют **единицей анализа**. В формулах оно стоит в знаменателе. Например для ARPU – это пользователь.

#### *Последовательность шагов при проведении A/B-теста:*

1. Выбираем метрики и формулируем гипотезы.
2. Выбираем способ рандомизации и определяем параметры выборки.
3. Определяем необходимый размер выборки.
4. Запускаем эксперимент и собираем данные.
5. Проверяем валидность эксперимента.
6. Рассчитываем результаты и принимаем решение о раскатке фичи.

#### *Шаг 1. Выбираем метрики и формулируем гипотезы*

В классической ситуации бизнес будет вводить изменения в продукт, если они приносят какую-то выгоду. Поэтому мы могли бы использовать одностороннюю гипотезу. Однако это не всегда так. Иногда бизнесу важно ввести изменения для дальнейшего развития фичи. В этом случае подойдёт вариант, когда фича, по крайней мере, не ухудшает метрики.

Таким образом, для A/B-тестирования мы будем использовать двустороннюю гипотезу.

## *Шаг 2. Выбираем способ рандомизации и определяем параметры выборки (СЕТЕВОЙ ЭФФЕКТ)*

**Разбиение на группы** в рамках A/B-теста должно происходить параллельно и случайным образом.

На этом этапе важно ответить на четыре основных вопроса:

- Кого распределяем по группам,
- Как распределяем по группам,
- Кто попадает в тест,
- Каково соотношение групп.

### **Кого распределяем по группам?**

Для начала определимся с единицей разбиения или, как её чаще называют, *единицей рандомизации*. Обычно рандомизация происходит на основе **пользователей**. Это значит, что для каждого пользователя мы случайным образом определяем, в какую из групп он попадёт.

Но иногда простое разбиение по пользователям может привести к смещению итоговых результатов.

**Сетевым эффектом** называют ситуацию, когда поведение одних пользователей в рамках A/B-теста может влиять на поведение других пользователей.

### **Минимизировать сетевой эффект можно разными способами:**

- Предварительно выделить непересекающиеся кластеры пользователей с помощью алгоритмов машинного обучения и распределять всех пользователей из одного кластера в одну и ту же группу. Этот способ обычно используют для проведения тестов в социальных сетях.
- Проводить рандомизацию по географическому признаку. Это так называемые гео-A/B-тесты. Их дизайн отличается от классических экспериментов.
- Использовать switchback-тестирование. В этом варианте пользователей разбивают на тест и контроль не только по географическому признаку, но ещё и по временному.

### **Как распределяем по группам?**

Разбиение по группам происходит случайным образом, чтобы группы были максимально однородными по составу.

Например, иначе, в одну из групп попадут только пользователи, траты которых за месяц выше медианных трат всех пользователей, а в другую группу попадут только пользователи, траты которых ниже медианных.

**Валидность разбиения проверяют с помощью АА-тестов.**

### Кто попадает в тест?

Зависит от того, что именно мы тестируем. Например, этого могут быть все пользователи, заходящие в приложение; или только те, которые дошли до экрана чекаута; или группы распределены заранее.

### Каково соотношение групп?

Чаще всего при проведении классического А/В-теста, в рамках которого сравниваются два варианта, группы делают одинаковыми по количеству. Иногда бывают другие ситуации.

### *Шаг 3. Определяем необходимый размер выборки*

Размер выборки влияет на способность статистического критерия зафиксировать статистически значимые изменения там, где они действительно есть. Чем дольше мы держим эксперимент, тем большую выборку сможем набрать и тем более чувствительным к изменениям метрики становится критерий.

The diagram shows the formula for sample size  $n$  with detailed annotations for each part:

$$n = \frac{(\overbrace{Var_{control} + Var_{test}}^{\text{дисперсии контрольной и тестовой групп}}) \cdot (\overbrace{z_{\alpha/2} + z_{\beta}}^{\text{квантиль стандартного нормального распределения, соответствующий половине вероятности ошибки первого рода}})^2}{\underbrace{MDE^2}_{\substack{\text{минимальный размер эффекта, который мы хотим} \\ \text{зафиксировать в рамках А/В-теста}}}} \cdot \underbrace{z_{\beta}}_{\text{квантиль стандартного нормального распределения, соответствующий вероятности ошибки второго рода}}$$

**Так как размер выборки рассчитывается до того, как сама выборка будет набрана, эти дисперсии оцениваются путём усреднения дисперсии за несколько прошедших периодов.**

Мы предполагаем, что на исторических данных в тестовой и контрольной группах равны  $Var_{hist}$ :

$$n \approx \frac{2 * Var_{hist} * \left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2}{MDE^2}$$

#### *Шаг 4. Запускаем эксперимент и собираем данные (ПРОБЛЕМА ПОДГЛЯДЫВАНИЯ)*

Останавливать A/B-тест есть смысл только в критических ситуациях, которые явно ухудшают опыт пользователя или показывают, что тест проводится некорректно. Если же специалист останавливает эксперимент, как только видит положительное статистически значимое изменение, то вероятность ошибки первого рода будет выше, чем та, которую используют при применении статистического теста. Эта проблема известна как **проблема подглядывания** (peeking problem).

#### *Шаг 5. Проверяем валидность эксперимента (AA-ТЕСТ, Sample Ratio Mismatch)*

Цель — убедиться, что группы однородны по своей структуре. Сделать это можно с помощью A/A-теста.

**A/A-тест** — это тот же самый A/B-тест, но который проводится на данных, собранных на периоде, предшествующем периоду A/B-теста. Мы предполагаем, что тестируемая фича — единственное, что влияет на разницу в метриках между группой контроля и группой теста. Поэтому мы ожидаем, что на предпериоде, когда фича ещё не была раскатана ни на кого, значение метрики в обеих группах будет одинаково.

Также важно провести проверку на **Sample Ratio Mismatch**, то есть убедиться, что фактическая пропорция пользователей между контрольной и тестовой группами соответствует той, которую мы задали при планировании эксперимента.

#### *Шаг 6. Рассчитываем результаты и принимаем решение о раскатке фичи*

Рассчитывать p-value и определять статистическую значимость.

## 7.2. Количественные метрики

- Мы не можем просто использовать прибыль как метрику, так как ее распределение почти всегда скошено. Кто-то покупает много, кто-то ничего.
- Все метрики делят на количественные, конверсионные и метрики-отношения.
- При A/B-тестировании обычно сравнивают средние значения, значит рассматривают усредненные метрики.
- Для анализа изменений количественных метрик можно использовать **T-тест или бакетный тест**.
- **Бакетный тест используется при анализе денежных метрик (потому что распределение скошено, и нужно привести его к нормальному)**. Например, при анализе выручки на пользователя ARPU.

Примерами количественных метрик могут быть:

- денежные метрики — выручка, валовая прибыль, чистая прибыль;

- количественные метрики — количество созданных заказов, количество выкупленных заказов, количество заказанных вещей;
- технические метрики — время загрузки страницы, количество неудачных загрузок страницы.

Выручку можно считать в среднем на пользователя, платящего пользователя или заказ.

- Среднее на вошедшего в приложение пользователя (Average Revenue per User)

$$ARPU = \frac{\text{Выручка}}{\text{Количество пользователей}}$$

- Среднее на совершившего заказ пользователя (Average Revenue per Paying User)

$$ARPPU = \frac{\text{Выручка}}{\text{Количество платящих пользователей}}$$

- Среднее на совершённый заказ (Average Order Value) [**Это метрика отношения**]

$$AOV = \frac{\text{Выручка}}{\text{Количество заказов}}$$

- Связь между ARPU и ARPPU описывается формулой:

$$ARPU = ARPPU * \text{Paying share}$$

где Paying share — это доля пользователей, совершивших покупку.

Для расчёта размера выборки необходимо оценить дисперсию в тестовой и контрольной группах. **Так как размер выборки рассчитывается до того, как сама выборка будет набрана, эти дисперсии оцениваются путём усреднения дисперсии за несколько прошедших периодов.**

Мы предполагаем, что на исторических данных в тестовой и контрольной группах равны  $Var_{hist}$ :

$$n \approx \frac{2 * Var_{hist} * \left( \frac{z_{\alpha}}{2} + z_{\beta} \right)^2}{MDE^2}$$

#### Пример применения бакетного теста для ARPU

Представим, что в контрольной группе у нас 3000 пользователей со значениями выручки. Делим на тестовую и контрольную. Для каждой группы выделяем бакеты (пусть будет 100 бакетов). Для каждого бакета считаем ARPU на бакет. Пусть получились такие значения:

Выделив 100 бакетов для каждой группы, мы получили следующие данные:

- Средний ARPU по 100 бакетам в тестовой группе: 3584 руб.
- Средний ARPU по 100 бакетам в контрольной группе: 3496 руб.
- Стандартное отклонение по 100 бакетам в тестовой группе: 302 руб.
- Стандартное отклонение по 100 бакетам в контрольной группе: 297 руб.

Сформулируем нулевую и альтернативную гипотезы и зафиксируем уровень значимости:

- $H_0: ARPU_{test} = ARPU_{control}$ ,
- $H_1: ARPU_{test} \neq ARPU_{control}$ ,
- Уровень значимости,  $\alpha = 0.05$ .

Для начала заполним таблицу с основными величинами, которые понадобятся для расчёта t-статистики:

$\bar{x}_{test}$	3584
$\bar{x}_{control}$	3496
$s_{test}$	302
$s_{control}$	297
$N$	100

Так как мы используем бакетный тест, количество наблюдений в обеих группах равно количеству бакетов. Тогда t-статистику рассчитывают так:

$$t = \frac{\bar{x}_{test} - \bar{x}_{control}}{\sqrt{\frac{s_{test}^2}{N} + \frac{s_{control}^2}{N}}} = \frac{3584 - 3496}{\sqrt{\frac{302^2}{100} + \frac{297^2}{100}}} \approx 2.08.$$

Теперь рассчитаем степени свободы. Так как количество бакетов в обеих группах одинаковое, можно использовать упрощённую формулу:

$$df = \frac{(N - 1) \cdot (s_{test}^2 + s_{control}^2)^2}{s_{test}^4 + s_{control}^4} = \frac{(100 - 1) \cdot (302^2 + 297^2)^2}{302^4 + 297^4} \approx 198.$$

Значение соответствующего p-value  $\approx 0.04$ . Это число меньше, чем уровень значимости 0.05, значит, мы отвергаем  $H_0$  и считаем, что фича статистически значимо повысила ARPU. Ура!



### 7.3. Конверсии и метрики-отношения

#### Конверсионные метрики

- Дисперсия описывается распределением Бернули:  $Var_{Berboulli} = \bar{p} * (1 - \bar{p})$ . [ $\bar{p}$  – рассчитанная конверсия]
- Для проверки гипотез используют Z-тест для пропорций.

**Конверсией** называют процент пользователей, совершивших целевое действие. Конверсию можно рассчитать по формуле:

$$CR_{X \text{ to } Y} = \frac{K}{N} * 100\%$$

где  $K$  — количество пользователей, которые совершили целевое действие  $Y$  (дошли до шага  $Y$ ),

$N$  — количество пользователей, которые совершили действие  $X$  (дошли до шага  $X$ ).

**Связь ARPU и конверсии:**

$$ARPU = ARPPU * \text{Paying share} = ARPPU * CR$$

Где  $CR$  – конверсия в заказ пользователей, которые вошли в приложение.

**Пример применения Z-теста для пропорций для конверсии:**

В рамках эксперимента получили такие данные:

- Конверсия в заказ в тестовой группе: 4.5%.
- Конверсия в заказ в контрольной группе: 4%.

Сформулируем нулевую и альтернативную гипотезы и зафиксируем уровень значимости:

- $H_0: CR_{test} = CR_{control}$ ,
- $H_1: CR_{test} \neq CR_{control}$ ,
- Уровень значимости,  $\alpha = 0.05$ .

Для начала рассчитаем дисперсии.

Для тестовой группы:

$$Var_{test} = \bar{p}_{test} \cdot (1 - \bar{p}_{test}) = 0.045 \cdot (1 - 0.045) \approx 0.043.$$

Для контрольной группы:

$$Var_{control} = \bar{p}_{control} \cdot (1 - \bar{p}_{control}) = 0.04 \cdot (1 - 0.04) = 0.0384.$$

Заполним таблицу с основными величинами, которые понадобятся для расчёта z-статистики:

$\bar{p}_{test}$	0.045
$\bar{p}_{control}$	0.04
$Var_{test}$	0.043
$Var_{control}$	0.0384
$n_{test}$	10 000
$n_{control}$	10 000

Рассчитаем z-статистику:

$$z = \frac{\bar{p}_{test} - \bar{p}_{control}}{\sqrt{\frac{Var_{test}}{n_{test}} + \frac{Var_{control}}{n_{control}}}} = \frac{0.045 - 0.04}{\sqrt{\frac{0.043}{10\,000} + \frac{0.0384}{10\,000}}} \approx 1.75.$$

Так как z-статистика распределена нормально, считать степени свободы не нужно. Значение соответствующего p-value  $\approx 0.08$ . Так как это число больше уровня значимости в 0.05, мы не отвергаем  $H_0$  и считаем, что влияние фиши на конверсию в заказ не является статистически значимым.

### Метрики-отношения (ratio-метрики)

- Единица рандомизации отличается от единицы анализа.
- Если целевая метрика является метрикой-отношением, то анализировать результаты А/В-экспериментов помогают **дельта-метод**, **бутстрэп** или **линеаризация**.
- Примеры метрик: AOV, количество добавлений товара за сессию.

### Почему нельзя использовать обычные тесты?

Если же мы считаем метрики-отношения, наблюдения нельзя считать независимыми. Например, когда мы считаем AOV, каждое наблюдение в выборке — это сумма заказа. При этом один и тот же пользователь может совершать несколько заказов в рамках эксперимента. А значит, наблюдения будут зависимы

## 7.4. MDE и мощность, ошибки

**MDE** (англ. Minimum Detectable Effect — минимальный детектируемый эффект) — это та разница математических ожиданий, которую мы сможем зафиксировать с заданным уровнем значимости  $\alpha$  и мощности  $1-\beta$ , если наберём две выборки размера  $n$ , и дисперсии этих выборок будут равны  $Var_{control}$  и  $Var_{test}$ .

Простая, но важная для понимания штука:

MDE – эффект (обычно прирост), который мы хотим зафиксировать для метрики. То есть это разница наших мат. ожиданий при тестировании гипотез. Когда тестируем есть ли вообще какой-то эффект (отличный от нуля), она равно 0. Когда проверяем, что эффект точно не меньше заданного, она соответственно равна какому-то числу (читай последнюю главу).

### Какое MDE использовать в формуле?

- Необходимо оценить затраты на внедрение фичи и взять такой прирост метрики, который обеспечит их покрытие. Это значение является лишь желаемым значением MDE, то есть нашим предположением относительно того, каким может быть реальный эффект.
- MDE по размерности **абсолютная** величина.
- MDE может быть больше или меньше реального (который мы получили уже постфактум). Это норм!
- В случае если реальный эффект окажется меньше желаемого, мы всё равно сможем его зафиксировать, но с меньшей вероятностью, чем предполагали изначально.

### Как оценить желаемый эффект?

Пусть общие затраты на внедрение новой фичи: 1300000 руб. Пусть усреднённое за несколько месяцев значение ARPU составляет 3000 руб.; в месяц в приложение заходит 100000 человек. Тогда, месячная выручка равна  $3000 * 100000 = 3E8$  руб.

Тогда процент затрат от месячной выручки:  $13E5 / 3E8 = 0.43$

Тогда  $MDE = 0.43 * ARPU = 0.43 * 3000 = 12.9$  руб

### Связь ошибок, мощности и MDE

**Вероятность ошибки первого рода,  $\alpha$**  — вероятность зафиксировать эффект там, где его на самом деле нет.

**Вероятность ошибки второго рода,  $\beta$**  — это вероятность не зафиксировать эффект там, где он на самом деле есть.

**Мощность,  $1-\beta$**  — это вероятность зафиксировать эффект там, где он на самом деле есть. Мощность также часто называют чувствительностью теста.

Мощность зависит от размера выборки, дисперсии в данных, уровня значимости и MDE:

$$z_{1-\beta} = \sqrt{\frac{n}{Var_{test} + Var_{control}}} * MDE + \frac{z_{\alpha}}{2}$$

Откуда взялось  $1 - \beta$ ? Нормальное распределение симметрично, можем в формуле заменить  $\beta$  на  $1 - \beta$ .

Рассмотрим графически:

Пусть выбраны гипотезы

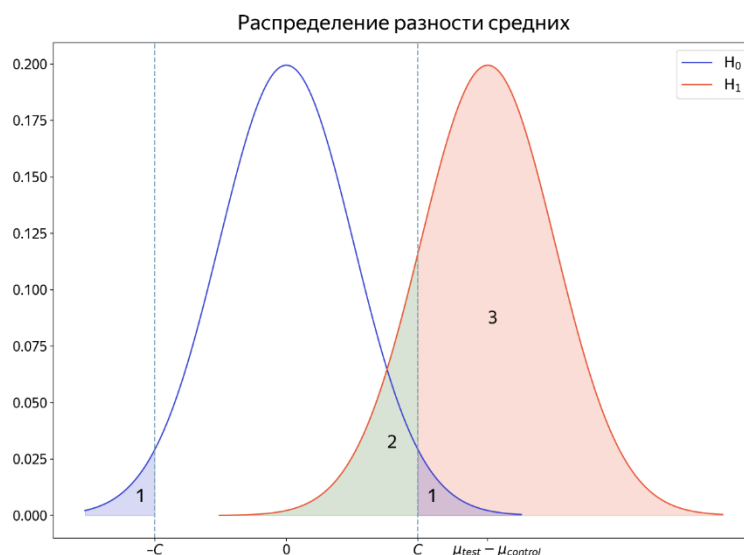
$H_0: \mu_{test} = \mu_{control}$ ;  $H_1: \mu_{test} \neq \mu_{control}$ ,  $\mu$  – выбранная метрика. если мы будем много раз брать пары выборок из двух генеральных совокупностей, считать разницу средних этих выборок и наносить её на график. Мы получим нормальное распределение с такими параметрами:

$$\mu = \mu_{test} - \mu_{control} \quad Var = \frac{Var_{control}}{n_{control}} + \frac{Var_{test}}{n_{test}}$$

Математическое ожидание этого распределения равно разности математических ожиданий генеральных совокупностей. В случае верной  $H_0$  эта разность будет равна 0, в случае верной  $H_1$  — какому-то другому числу.

Левый график соответствует нулевой гипотезе, правый — альтернативной. Значения  $C$  и  $-C$ , обозначенные вертикальными линиями, соответствуют пограничным значениям разности средних. Начиная с этих значений мы считаем отклонения от 0 статистически значимыми. Если наша разность окажется правее  $C$  или левее  $-C$ , то мы отвергнем  $H_0$ .

Значение  $C$  — это квантиль левого распределения, который соответствует величине  $1 - \frac{\alpha}{2}$ . В то же время значение  $C$  — это  $\beta$ -квантиль правого распределения. Если их приравнять, то можно вывести формулу для расчета размера выборки.



1- область (вероятность) ошибки 1 рода. 2–2 рода. 3 – мощность.

## 7.5. Объем групп и продолжительность теста. Эффект накопления метрик

Обычно длительность эксперимента выбирают кратной периоду, в рамках которого может наблюдаться сезонность в поведении метрики.

### Алгоритм расчёта длительности:

**Шаг 0.** Выбрать уровень значимости и мощность.

**Шаг 1.** Составить список из потенциальных длительностей эксперимента: 11 неделя, 22 недели и так далее.

**Шаг 2.** Для каждой такой длительности рассчитать на данных за прошлые периоды

- усреднённую дисперсию метрики,
- усреднённое среднее значение метрики,
- усреднённое количество пользователей, посетивших сайт или приложение

**Шаг 3.** Для каждой длительности рассчитать значение MDE, которое можно обнаружить с заданной мощностью, при выбранном уровне значимости и оценённой на данных за прошедшие периоды дисперсии. Это можно сделать по формуле:

$$MDE = -\left(z_{\alpha} + z_{\beta}\right) * \sqrt{4 * \frac{Var_{hist}}{n_{hist}}}$$

$Var_{hist}$  – дисперсия, оценённая на данных за прошлые периоды,


$n_{hist}$  – количество пользователей, которые в среднем посещают сайт за период, равный выбранной длительности.

**Шаг 4.** Выбрать ту длительность, **для которой рассчитанный MDE наиболее близок, но не превышает минимальный желаемый эффект**. В таком случае мы будем уверены в том, что мощность теста для желаемого MDE будет не меньше, чем та, которую мы использовали на предыдущем шаге.

### Почему выбираем MDE, который не превышает?

Ты хочешь выявить +2% рост конверсии.

А твой текущий размер выборки позволяет заметить только эффекты  $\geq 5\%$  (MDE = 5%).

 Это значит, что тест может показать "нет эффекта", даже если рост на +2% есть — просто потому, что у тебя не хватает данных, чтобы это увидеть.

### Эффект Накопления Метрик

С ростом размера выборки должен падать MDE (смотри формулу). Но не всегда так. с увеличением длительности растёт не только количество пользователей, но и дисперсия метрики, поэтому **MDE может падать**.

**Эффект накопления метрик** — явление, при котором значение метрики на пользователя растёт со временем, что приводит к увеличению её математического ожидания и дисперсии.

#### Как побороть?

- изменить способ формирования выборок, то есть стратифицировать их.
- или трансформировать метрику (методы CUPED, CUPAC).
- **САМЫЙ ПРОСТОЙ СПОСОБ:** рассматривать относительный MDE, то есть деленное на усредненное значение метрики (усредненное в соответствии количества рассматриваемых периодов). Keep in mind: в формуле все равно надо использовать абсолютное значение, а относительное – только для сравнения резю.

#### Что значит периоды, усреднение и. т. д.

Мы рассчитываем на исторических данных. Например, мы проверяем, хватит **ли 2 недель** для проведения теста.

Для этого нужно выделить как минимум 2 (для усреднения) двухнедельных периода в прошлом, которые предшествуют дате начала теста.

Пусть, что сейчас вторник, 23 марта.

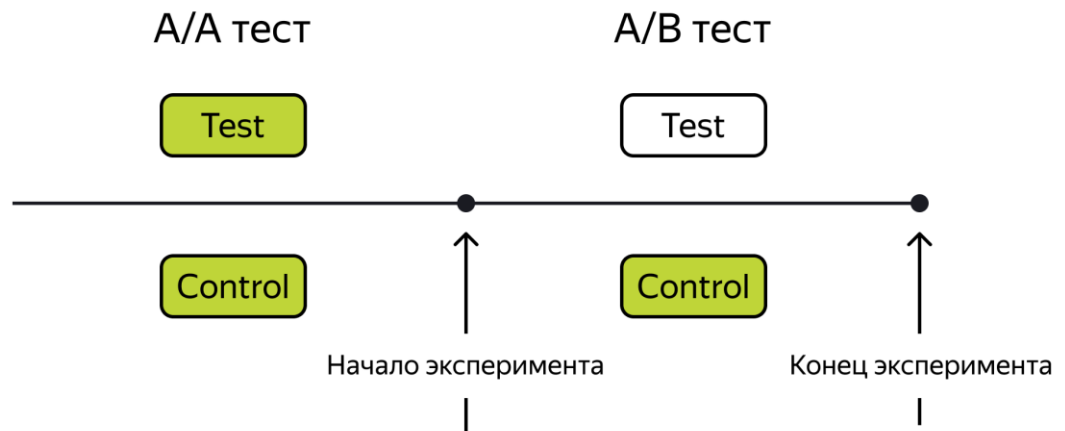
Тогда первый период: с 8 марта по 21 марта (2 недели). Второй период: с 22 февраля по 7 марта (2 недели). Для каждого периода считаем метрику, дисперсию, количество пользователей. Затем усредняем все.

## 7.6. Проверка валидности эксперимента. AA-тест, SRM (Sample Ratio Mismatch)

- Разбиение пользователей на группы происходит обычно с помощью хеш-функций. У этих функций иногда есть доппараметр – соль.
- Хеш-функции не всегда хорошо разбивают, группы оказываются неоднородными, то есть факторы, которые влияли на метрики еще до теста.
- Для проверки валидности используют AA-тест, SRM (Sample Ratio Mismatch).

### AA-тест

**A/A-тест** — это A/B-тест, применённый в ситуации, когда мы не ожидаем получить статистически значимых различий между группами, то есть до начала теста.



### Гипотезы:

$$H_0: metric_{test} = metric_{control}$$

$$H_1: metric_{test} \neq metric_{control}$$

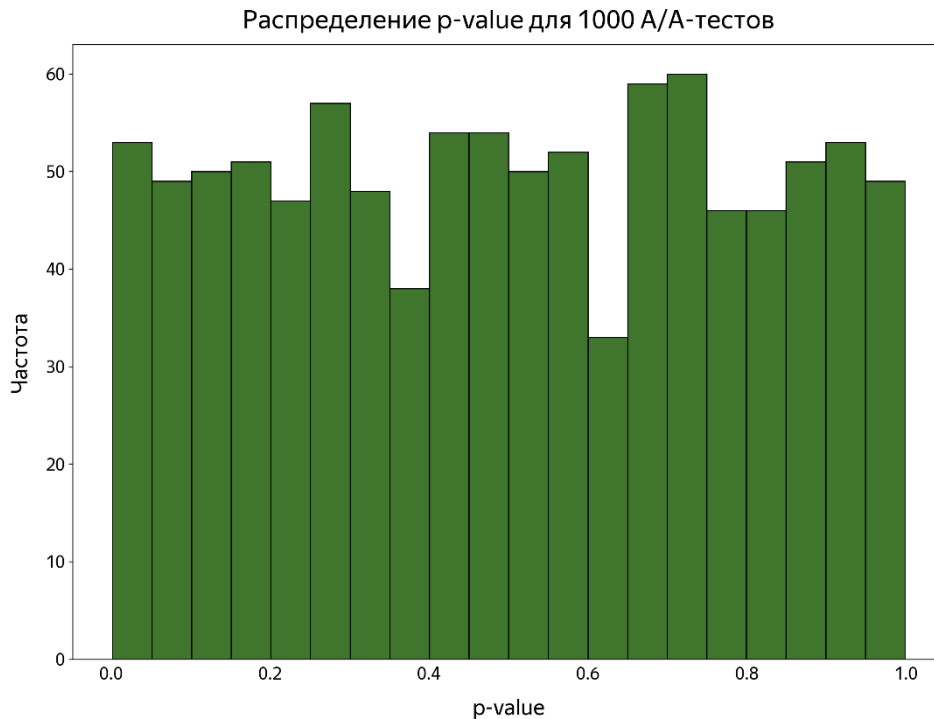
В отличие от АБ-теста, при АА-тесте, благоприятная гипотеза – это  $H_0$ , то есть не было эффекта до начала АБ теста.

### Как ещё можно использовать А/А-тест

Ещё А/А-тесты помогают обнаружить систематические ошибки. Для этого проводят много А/А тестов.

### Пример:

Возьмём пользователей за определённый период, например за месяц. Выбранных пользователей разделим на тестовую и контрольную группы с помощью хэш-функции с солью А. Сравним группы с помощью статистического критерия и запишем значение  $p$ -value. Повторим процедуру 10001000 раз, разбивая пользователей на группы с помощью разных солей. Нарисуем.



Если систематических ошибок нет, распределение будет близко к равномерному. Значения  $p\text{-value} \leq \alpha$  будут встречаться только в  $\alpha\%$  случаев. Если при применении статистического критерия мы использовали уровень значимости 0.05, мы получим статистически значимые отклонения примерно в 5% случаев.

Если же мы получаем статистически значимые отклонения существенно чаще, чем в 5% случаев, это значит, что в процессе проведения экспериментов есть проблема. Наиболее частых проблемы две:

- или группы, получаемые в результате разбиения хэш-функцией, неоднородны;
- или используемый статистический критерий не подходит для ваших нужд.

### *SRM (Sample Ratio Mismatch)*

Пусть на этапе планирования эксперимента мы запланировали один размер тестовой и контрольной выборки, а получился другой.

**SRM** (англ. Sample Ratio Mismatch — несоответствие пропорции выборки) — это ситуация, при которой полученная пропорция размеров тестовой и контрольной групп отличается от ожидаемой.

Чтобы определить, есть ли SRM, можно использовать критерий **хи-квадрат**.



Допустим, что мы ожидали увидеть распределение пользователей в пропорции 50/50. Тогда гипотезы можно сформулировать так:

- $H_0$ : Наблюдаемое отношение количества пользователей в тестовой группе к количеству пользователей в контрольной группе не отличается от ожидаемого отношения. 50/50.
- $H_1$ : Наблюдаемое отношение количества пользователей в тестовой группе к количеству пользователей в контрольной группе отличается от ожидаемого отношения 50/50.

**!!!Критерий достаточно чувствителен при выборках большого размера.** Это значит, что он будет с большой вероятностью определять даже небольшие отклонения как статистически значимые. Значит выбирают более малые уровни значимости, типа 0.01, 0.001 итд.

## 7.7. Расчет и интерпретация результатов. Проблема подглядывания

Результат теста может быть таким:

- зафиксировали статистически значимый прирост метрики — тест зелёный;
- зафиксировали статистически значимое падение метрики — тест красный;
- не зафиксировали статистически значимого изменения метрики — тест серый.

### *Интерпретация результатов серого A/B-теста*

В ситуации, когда **A/B-тест оказался серым**, мы говорим о том, что не нашли доказательства того, что фича оказала отличный от 0 эффект. Однако это не означает, что эффект в точности равен 0 или точно меньше, чем желаемый эффект, который мы использовали в качестве MDE.

### *Проблема подглядывания*

**Проблема подглядывания** – мы остановили тест раньше рассчитанной длительности.

A/B-тест не следует останавливать, как только была зафиксирована статистическая значимость, так как это ведёт к росту вероятности ошибки первого рода. Подводить итоги эксперимента следует только после того, как его длительность станет равна предрассчитанной.

### *Интерпретация результатов зеленого A/B-теста*

В ситуации, когда **A/B-тест оказался зелёным**, нам необходимо дополнительно проверить, что наблюдаемая разность средних позволяет сделать вывод о том, что истинный эффект не меньше желаемого. Для этого можно рассчитать t- или z-статистику (в зависимости от типа метрики) для желаемого эффекта по формулам:

$$z = \frac{(\bar{p}_{test} - \bar{p}_{control}) - (p_{test} - p_{control})}{\sqrt{\frac{Var_{test}}{n_{test}} + \frac{Var_{control}}{n_{control}}}},$$

$$t = \frac{(\bar{x}_{test} - \bar{x}_{control}) - (\mu_{test} - \mu_{control})}{\sqrt{\frac{s_{test}^2}{n_{test}} + \frac{s_{control}^2}{n_{control}}}}.$$

### Пример (зеленый тест):

Пусть желаемый эффект MDE = 0.033ю

Сначала проверяем, что эффект отличен от 0:

- $H_0: CR_{test} = CR_{control},$
- $H_1: CR_{test} \neq CR_{control},$

Пусть полученное p-value < alpha -> reject H0 -> есть эффект! Он называется статистический.

Теперь необходимо найти практический эффект, то есть проверить, что MDE не меньше желаемого 0.033. Тогда рассматриваем эти гипотезы и тестируем по формулам выше:

- $H_0: CR_{test} = CR_{control} + 0.033 \iff CR_{test} - CR_{control} = 0.033,$
- $H_1: CR_{test} \neq CR_{control} + 0.033 \iff CR_{test} - CR_{control} \neq 0.033.$