Конспект по курсу «Основы статистики и A/B-тестирования»

Описательная статистика

Описательная статистика помогает:

- Познакомиться с данными и визуализировать их.
- Проанализировать имеющиеся наблюдения с разных сторон. Для этого рассчитывают среднее, медиану, квантиль и дисперсию.
- Исследовать связи между переменными, чтобы делать более точные выводы и прогнозы. Изучать взаимосвязи числовых данных помогают ковариация и корреляция. В некоторых случаях взаимодействие более наглядно, если одну из переменных мы переводим из числовой в категориальную. Для этого используют бинаризацию.

Среднее

Типы данных:

- **Категориальные** данные 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** данные — измеримые или счётные значения. Например, возраст, доход, рост.

Методы визуализации: столбчатая диаграмма и гистограмма.

Чтобы вычислить выборочное среднее, нужно сложить все значения и разделить полученную сумму на их количество:

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

где n — количество элементов в выборке.

Медиана

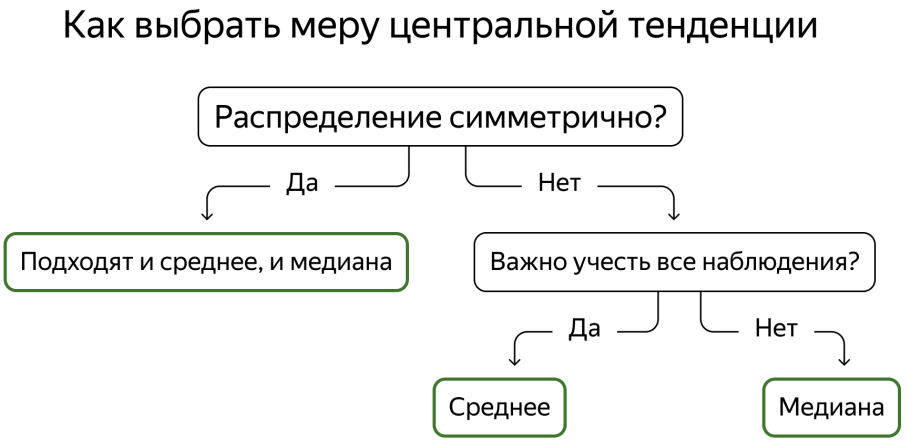
Медиана — это наблюдение, которое делит весь набор данных на две равные части: меньше него 50% наблюдений и больше него тоже 50% наблюдений.

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.

Распределение называется **симметричным**, если оно выглядит одинаково с обеих сторон от своей середины.



Квантиль

Число X является α-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 α-квантиля

1. Отсортируйте набор данных по возрастанию.
2. Найдите позицию квантиля по формуле: $n \cdot \alpha$, где n — количество элементов в наборе, α — доля, которая нас интересует.

3. Определите значение квантиля:
   а) Если позиция квантиля — целое число, α-квантиль равен значению, которое соответствует этой позиции в упорядоченном наборе данных.
   б) 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:

- $\mathrm{Q1}$ (первый квартиль) — это 0.25-квантиль,
- $\mathrm{Q2}$ (второй квартиль) — это 0.5-квантиль (медиана),
- $\mathrm{Q3}$ (третий квартиль) — это 0.75-квантиль.

Межквартильный размах $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}.$

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 $\mathrm{Q1}$ (0.25-квантиль) и третий квартиль $\mathrm{Q3}$ (0.75 -квантиль).
3. Рассчитайте межквартильный размах: $\mathrm{IQR} = \mathrm{Q3} - \mathrm{Q1}.$
4. Определите границы выбросов:
   - Нижняя граница: $\mathrm{Q1} - 1.5 \cdot \mathrm{IQR},$
   - Верхняя граница: $\mathrm{Q3} + 1.5 \cdot \mathrm{IQR}.$
5. Отсейте все значения, которые лежат за пределами этих границ, — они считаются выбросами.

Диаграмма «ящик с усами»:

Ящик:

- Левая граница ящика — первый квартиль ($\mathrm{Q1}$), то есть 25% данных лежат ниже этой точки.
- Медиана (второй квартиль, $\mathrm{Q2}$) представлена горизонтальной линией внутри ящика. Медиана делит данные на две равные части: половина данных лежит ниже этой линии, а половина — выше.

- Правая граница ящика — третий квартиль ($\mathrm{Q3}$), то есть 75% данных лежат ниже этой точки.

Усы:

- Левый ус ограничивается значением $\mathrm{Q1} - 1.5 \cdot \mathrm{IQR}$.
- Правый ус ограничивается значением $\mathrm{Q3} + 1.5 \cdot \mathrm{IQR}$.

Выбросы обозначают кружочками.

Дисперсия

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

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

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

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

Типы распределений

На практике обычно важно не только исследовать имеющиеся данные, но и понять, чего ждать от новых. В этом помогают статистические методы.

В этой теме собраны ключевые теоретические понятия и инструменты, которые используют при работе со статистическими методами.

Случайная величина

Генеральная совокупность — это полный набор всех элементов, которые исследуют в рамках задачи.

Выборка — это отдельный набор элементов, отобранных из генеральной совокупности некоторым случайным процессом.

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 $\hat{F}(x) = P(X \leqslant x)$.

Формула	Вероятность	Описание
$\hat{F}(x)$	$P(X \leqslant x)$	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.
$\hat{F}(x_2) - \hat{F}(x_1)$	$P(x_1 < X \leqslant x_2)$	Вероятность того, что случайная величина примет значение в определённом диапазоне.

Равномерное распределение

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

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

Функция распределения определяет вероятность того, что случайная величина примет значение меньше или равное заданному. Обозначается как $F(x) = P(X \leqslant x)$.

Математическое ожидание — это взвешенное среднее значение случайной величины, где веса представляют собой вероятности возможных значений этой случайной величины. Оно считается как $\mathrm{E}(X) = \sum_i x_i \cdot P(X = x_i)$.

Дисперсия случайной величины — это математическое ожидание квадрата отклонения случайной величины от её математического ожидания. Её формула выглядит так: $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]$.

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, определяющими минимальное и максимальное возможные значения. Короткое обозначение: $X \sim \mathrm{U}(a, b)$.

Говорят, что дискретная случайная величина имеет равномерное распределение с параметрами a, b, если для каждого целого значения в интервале $a \leqslant x \leqslant b$, она описывается функцией вероятности: $P(X = x) = \frac{1}{n}$.

Функция распределения $F(x)$ для всех целых значений из интервала $a \leqslant x \leqslant b$ в случае равномерного распределения будет иметь вид: $F(x) = \frac{x - a + 1}{n}$.

Математическое ожидание случайной величины, имеющей равномерное распределение: $\mathrm{E}(X) = \frac{a + b}{2}$.

Дисперсия случайной величины, имеющей равномерное распределение: $\mathrm{Var}(X) = \frac{n^2 - 1}{12}$.

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 $f_X(x)$.

Свойства функции плотности вероятности:

- Если взять всю площадь под графиком функции плотности вероятности, то получится 1.
- Значения функции плотности вероятности всегда больше или равны нулю.

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 μ и дисперсией σ^2.

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

Статистические тесты

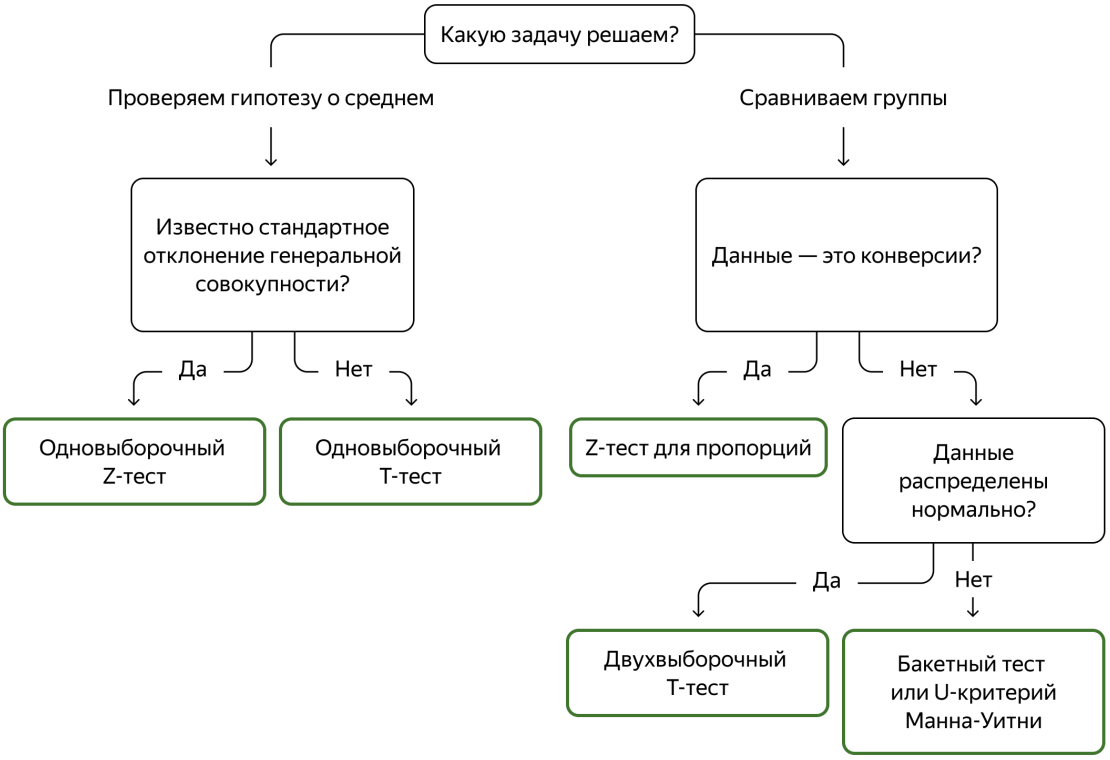
Основная задача статистики — по небольшой выборке сделать обоснованные выводы о генеральной совокупности. В этом помогают статистические тесты.

В основе многих идей статистического анализа лежит погрешность измерения. Оценить её позволяет, например, стандартная ошибка среднего.

Частый подход к анализу — тестирование гипотез, когда мы выдвигаем базовое предположение и проверяем его с помощью наших данных. Принять решение по гипотезам помогает p-value. Это универсальный инструмент, его используют в разных тестах.

У каждого теста есть свои условия применения, выбор конкретного зависит от задачи и данных.

Как выбрать подходящий тест к задаче



Стандартная ошибка среднего

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

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

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}.$$

Стандартная ошибка среднего (SE) указывает, насколько большая ожидается погрешность при измерении среднего значения по выборке. Чем больше дисперсия в самих данных и чем меньше выборка, тем больше погрешность.

Статистические гипотезы

Цель проверки **статистических гипотез** — на основе небольшой выборки сделать обоснованные предположения о генеральной совокупности.

Нулевая гипотеза, обозначаемая как H_0, это исходное предположение о данных.
Альтернативная гипотеза, обозначаемая как H_1, представляет собой предположение, которое противоположно нулевой гипотезе.

Критические значения формируют область допустимых значений для статистики, при условии, что нулевая гипотеза верна.

Ошибка первого рода — это ситуация, когда исследователь отвергает нулевую гипотезу, хотя она на самом деле верна.

Уровень значимости α — это вероятность совершения ошибки первого рода, то есть отклонить нулевую гипотезу, когда она верна.

Z-тест для среднего значения — это статистический тест, который используется, чтобы определить, отличается ли среднее значение выборки от предполагаемого среднего значения генеральной совокупности.

В процессе Z-теста вычисляется z-статистика, которая представляет собой отношение разности между средним значением выборки и средним значением генеральной совокупности к стандартной ошибке среднего.

$$z = \frac{\bar{x} - \mu}{SE},$$
$$SE = \frac{\sigma}{\sqrt{n}}.$$

Алгоритм проведения Z-теста:

1. Сформулировать H_0 и H_1.
2. Выбрать α, найти соответствующие критические значения.

3. Вычислить выборочное среднее.
4. Вычислить z-статистику.
5. Сравнить z-статистику с критическими значениями:
6. Принять решение:
    - Если z-статистика вываливается, отклонить нулевую гипотезу.
    - Если z-статистика не вываливается, принимаем решение, что отклонить нулевую гипотезу нельзя.

P-value

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

Ошибка первого рода, α — это ситуация, когда исследователь отвергает нулевую гипотезу, хотя она на самом деле верна.

Ошибка второго рода, β — это ситуация, когда исследователь не отвергает нулевую гипотезу, хотя верна альтернативная.

Мощность статистического теста, $1 - \beta$ — это вероятность правильно отклонить нулевую гипотезу, когда она действительно неверна.

T-тест и распределение Стьюдента

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

$$\mathrm{ESE} = \frac{s_X}{\sqrt{n}}.$$

В тесте для среднего, когда неизвестно стандартное отклонение генеральной совокупности, используется **t-статистика**. Это мера, которая помогает определить статистическую значимость различия между средним значением выборки и предполагаемым средним значением генеральной совокупности:

$$t = \frac{\bar{x} - \mu}{ESE}.$$

Распределение Стьюдента, или t-распределение, это вероятностное распределение, которое описывает поведение t-статистики. Оно имеет «тяжёлые» хвосты, это позволяет учесть большую неопределённость в оценках при малом объёме данных.

Степени свободы — это концепция в статистике, которая относится к количеству независимых значений в наборе данных, которые могут свободно варьироваться при расчёте некоторой статистической меры.

В контексте t-распределения Стьюдента степени свободы определяют форму распределения и рассчитываются по формуле:

$$df = n - 1.$$

Алгоритм проведения Т-теста:

1. Формулируем H_0, H_1
2. Выбираем α
3. Считаем выборочное среднее
4. Считаем выборочное стандартное отклонение
5. Считаем t-статистику
6. Считаем количество степеней свободы по формуле:
   $df = n - 1,$
   где n — количество наблюдений в выборке
7. Считаем p-value
8. Сравниваем p-value с уровнем значимости:
   - Если p-value меньше уровня значимости, принимаем решение, что надо отклонить нулевую гипотезу.
   - Если p-value больше уровня значимости, то оснований отклонять нулевую гипотезу нет.

Сравнение групп

Одновыборочные тесты — тесты, которые позволяют исследовать одну выборку. **Двухвыборочные тесты** помогают сравнить две выборки.

Двухвыборочный Т-тест используют, когда наблюдения независимы и распределены нормально.

Для двухвыборочного Т-теста в случае, если размеры выборок равны:

- t-статистику рассчитывают по формуле:
  $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{ESE},$$

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

- степени свободы рассчитывают по формуле:

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

  $\bar{x}_1$ и $\bar{x}_2$ — выборочные средние для первой и второй выборки соответственно,

  $\mu_1 - \mu_2$ — разность математических ожиданий двух выборок,

  s_1 и s_2 — значения стандартных отклонений, рассчитанных на основе выборок,

  где n — размер каждой выборки.

Бакетный тест используют, когда распределение наблюдений значительно отличается от нормального. Обычно данные в каждой выборке разбивают на 100 бакетов, далее вычисляют среднее значение по каждому бакету. К полученным данным применяют двухвыборочный T-тест.

Z-тест для пропорций помогает сравнить доли определённых наблюдений в двух выборках. Для него z-статистику рассчитывают так:

$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{ESE}$,

где $ESE = \sqrt{\frac{\bar{p}_1 \cdot (1 - \bar{p}_1)}{n_1} + \frac{\bar{p}_2 \cdot (1 - \bar{p}_2)}{n_2}}$,

$\bar{p}_1$ — расчётная доля в первой выборке,

$\bar{p}_2$ — расчётная доля во второй выборке,

n_1 — количество наблюдений в первой выборке,

n_2 — количество наблюдений во второй выборке.

A/B-тесты

A/B-тесты — популярный и полезный инструмент аналитика. Если его использовать корректно, то можно принять надёжные и обоснованные решения с опорой на данные.

Каждый A/B-тест включает:

- Формулировка гипотез и выбор метрик

    - метрики бывают количественные
    - и конверсионные
- Расчёт минимального желаемого эффекта (MDE) и мощности.
- Определение необходимого объёма группы и продолжительности теста.
- Проверка валидности эксперимента.
- Расчёт и интерпретация результатов.

Что такое A/B-тест

A/B-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.

Последовательность шагов при проведении A/B-теста:

Разбиение на группы в рамках A/B-теста должно происходить параллельно и случайным образом. Это позволяет устранить влияние отличных от тестируемого изменения эффектов на метрику.

При планировании эксперимента важно учесть потенциальное наличие **сетевого эффекта**. Если группы влияют друг на друга, то вместо простой рандомизации по пользователям необходимо использовать более сложные способы.

Прежде чем проводить тест, нужно **рассчитать необходимый объём выборки**. Только после того, как он наберётся, можно анализировать результаты.

Важно не только проанализировать непосредственно результаты теста, но и дополнительно провести проверки валидности эксперимента. Такими проверками являются **A/A-тест** и проверка на **SRM**.

Количественные метрики

Все метрики делят на количественные, конверсионные и метрики-отношения.

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

Связь между ARPU и ARPPU описывается формулой:

$$ARPU = ARPPU \cdot Paying\ share,$$

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.

Для анализа изменений количественных метрик можно использовать T-тест или бакетный тест.

Конверсии и метрики-отношения

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

$$CR_{X\ to\ Y} = \frac{K}{N} \cdot 100\%,$$

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

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

Дисперсия конверсионных метрик рассчитывается по формуле:

$$Var_{Bernoulli} = \bar{p} \cdot (1 - \bar{p}),$$

где $\bar{p}$ — рассчитанная конверсия.

Чтобы рассчитать необходимый размер выборки для конверсионных метрик, на этапе планирования эксперимента необходимо оценить дисперсии тестовой и контрольной групп. Для этого усредняют дисперсии конверсионной метрики за прошедшие периоды.

Метрики-отношения отличаются от количественных и конверсионных тем, что для них единица рандомизации, выбранная в рамках эксперимента, не совпадает с единицей анализа. Для таких метрик нельзя в чистом виде применять T-тест и Z-тест для пропорций.

//конец кат-контейнера//

MDE и мощность

Вероятность ошибки первого рода, α — вероятность зафиксировать эффект там, где его на самом деле нет.

Вероятность ошибки второго рода, β — 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:

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

Чтобы рассчитать MDE, 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.

Так как реальное значение MDE может отличаться от желаемого, то и реальная мощность также может отличаться от той, которую мы использовали при расчёте размера выборки.

В случае если реальный эффект окажется меньше желаемого, мы всё равно сможем его зафиксировать, но с меньшей вероятностью, чем предполагали изначально.

Объём групп и продолжительность теста

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: 1 неделя, 2 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:

$$MDE = -(z_{\alpha/2} + z_{\beta}) \cdot \sqrt{\frac{4 \cdot Var_{hist}}{n_{hist}}},$$

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.

Чтобы преодолеть эффект накопления метрик, нужно смотреть не на абсолютное значение MDE, а на относительное. Его можно рассчитать, разделив MDE на усреднённое значение метрики.

Проверка валидности эксперимента

Прежде чем анализировать результаты A/B-теста, необходимо убедиться, что кроме тестируемой фичи нет факторов, которые могли бы повлиять на целевую

метрику.

В этом помогают:

- A/A-тест на предпериоде, представляющий собой применение статистического критерия для сравнения значений метрики в тестовой и контрольной группах на периоде, предшествующем периоду эксперимента.
- Проверка на SRM, представляющая собой проверку того, что наблюдаемое соотношение количества пользователей в тестовой и контрольной группах не отличается от ожидаемого.

Расчёт и интерпретация результатов

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.

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 для желаемого эффекта по формулам:

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

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

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.

Таблица принятия решений по результатам A/B-теста

Зафиксировано ли статистически значимое отклонение от 0?	Зафиксировано ли статистически значимое отклонение от MDE?	Что делать
Да, отрицательное	Да, отрицательное	Оставить всё как есть. Фичу не раскатывать.
Нет	Да, отрицательное	Оставить всё как есть. Фичу не раскатывать.
Нет	Нет	Если есть время на повторный A/B-тест, то провести его с большей выборкой. Если времени на повторный тест нет, отказаться от внедрения фичи.
Да, положительное	Нет	Желательно перезапустить A/B-тест с большей выборкой. Возможно, рискнуть и раскатить фичу.
Да, положительное	Да, положительное	Раскатить фичу.