Организация массового A/B-тестирования

Расчёт необходимого объёма данных и срока тестирования

Расчёт необходимого объёма
данных и срока тестирования

Цели урока

Настало время подробнее обсудить,
как перед началом запуска эксперимента
определить необходимое количество данных.
Мы разберём сегодня такие темы:

- какого объёма выборку взять?
- какие параметры влияют на размер выборки? А что влияет на время проведения эксперимента?
- формула для определения размера выборки
- примеры правильного и неправильного использования онлайн-калькуляторов

Расчёт необходимого объёма
данных и срока тестирования

Определение
оптимального размера выборки

Для каждого A/B-теста нужен определённый размер выборки, чтобы получить статистически значимый результат. Это важно, потому что без статистической значимости случайное совпадение можно ошибочно принять за успех варианта. Результат — неверное бизнес-решение.

Расчёт необходимого объёма
данных и срока тестирования

Определение
оптимального размера выборки

Для каждого A/B-теста нужен определённый размер выборки, чтобы получить статистически значимый результат. Это важно, потому что без статистической значимости случайное совпадение можно ошибочно принять за успех варианта. Результат — неверное бизнес-решение.

Например, доля открытых писем в рассылках составляет 20 %. Если хотите увеличить показатель на 25 % с помощью изменения, понадобится выборка минимум из 2 000 человек. Необходимый размер выборки рассчитывается с помощью калькулятора A/B-тестов.

Расчёт необходимого объёма
данных и срока тестирования

Осторожнее
с онлайн-калькуляторами!

Воспользоваться онлайн-калькуляторами можно, если вам надо очень быстро и примерно оценить объём необходимой выборки и при этом вы хорошо уверены в том, какие у вас будут данные!

Расчёт необходимого объёма
данных и срока тестирования

Оптимальный размер выборки

В онлайн-калькуляторах вы легко можете задать нужные вам **уровни значимости**, **мощность** для подходящего вам **процента изменений**.

Несмотря на простоту в использовании, важно знать, что «под капотом»!

Расчёт необходимого объёма
данных и срока тестирования

Как рассчитать срок A/B-тестирования?

Использовать методы, в которых полностью разбираетесь и которым доверяете. Для каждого случая есть свои критерии!

Распределение	Критерии	Примеры
Нормальное	Критерий Стьюдента, критерий Аспина — Уэлша	Средняя прибыль по пользователям
Бернуллиевское	Z-критерии, точный тест Фишера, критерий хи-квадрат	CTR
Пуассоновское	E-тест, C-тест	Количество транзакций на пользователя
Мультиноминальное	Критерий хи-квадрат	Количество приобретённых пользователем продуктов
Другое	Критерий Уилкоксона, Манна — Уитни, семплирование	

Расчёт необходимого объёма
данных и срока тестирования

Частые понятия: BCR и MDE

BCR (Basic Conversion Rate) —
разовый уровень конверсии.

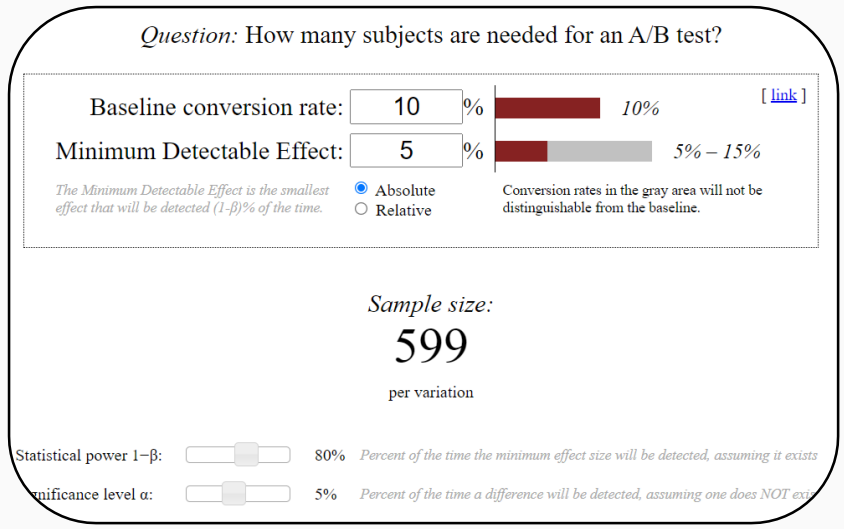
MDE (Minimum Detectable Effect,
абсолютный или относительный) —
минимальный размер эффекта.

Например, доля открытых писем в рассылках составляет **20 % (BCR)**. Если хотите увеличить показатель на **25 % (MDE)** с помощью изменения, понадобится выборка минимум из 2 000 человек.

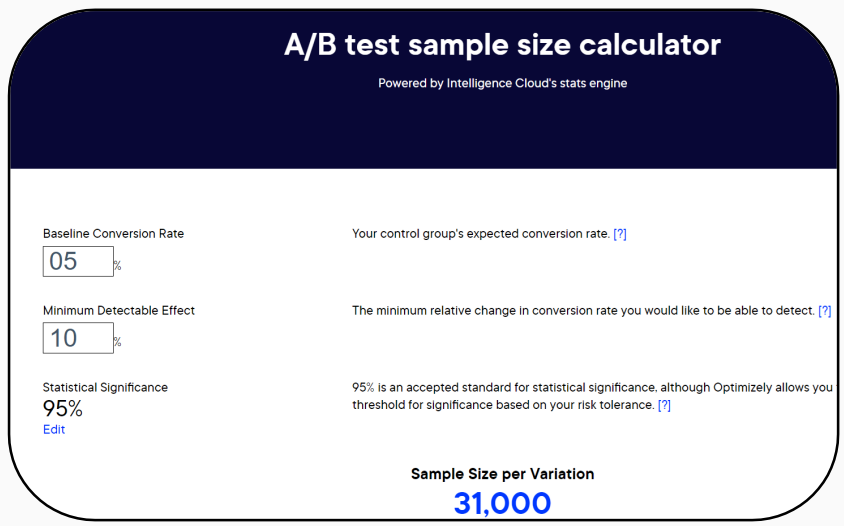
Расчёт необходимого объёма
данных и срока тестирования

Пример онлайн-калькуляторов

Evanmiller.org —
A/B-Test Sample Size Calculator



Optimizely.com —
A/B-Test Sample Size Calculator



Расчёт необходимого объёма
данных и срока тестирования

Длительность тестирования: кейс

Гипотеза: нужно убрать фильтр, но есть риск, что конверсия в покупку подписки упадёт.

Решение: сделать A/B-тестирование и за 2 недели задетектировать 1 % разницы в конверсиях. Собирая по 100 клиентов в день, успеем ли провести эксперимент?

Расчёт необходимого объёма
данных и срока тестирования

Сравнение калькуляторов

BCR (Basic Conversion Rate) — базовая конверсия.

MDE (Minimum Detectable Effect) — абсолютный или относительный.

EvanMiller	BCR — 20 % MDE — 5 %	1030 / 1030	Двусторонний тест
Optimizely		670 / 670	Байесовские методы
Unbounce		1024 / 1024	Двусторонний тест
VWO		1024 / 1024	Двусторонний тест
ABTasty		1030 / 1030	Двусторонний тест
Python		Зависит от реализации	Любой тест
R		Зависит от реализации	Любой тест

Расчёт необходимого объёма
данных и срока тестирования

Онлайн-калькуляторы и подводные камни

Как это всегда бывает, за удобством скрываются некоторые подводные камни, на которые можно незаметно наткнуться. Разберём все плюсы и минусы.

Плюсы:

- оценки получаются быстро и легко
- при этом не требуется глубоких знаний о ваших данных

Минусы:

- расчёт часто происходит из предположения, что данные имеют нормальное распределение. При этом калькуляторы не знают, какое распределение в данных именно у вас. А оно не всегда нормальное
- калькуляторы не учитывают наличие возможных выбросов в ваших данных

Расчёт необходимого объёма
данных и срока тестирования

Онлайн-калькуляторы и подводные камни

Чтобы не выстрелить себе в ноги при использовании удобных калькуляторов, есть несколько рекомендаций.

- Для перестраховки стоит увеличивать количество данных в 1,5–2 раза (если скорость сбора данных позволяет это делать)
- Чётко понимайте, что находится «под капотом калькулятора» и как оно работает
- Если результаты онлайн-калькулятора вызывают сомнения, воспроизведите результат в других версиях или собственноручно

Расчёт необходимого объёма
данных и срока тестирования

Длительность тестирования

Чекпоинты:

1. Рассчитать минимальный размер каждой группы, соответствующий заданным параметрам (вероятность ошибки 1-го рода, мощность, минимальный отслеживаемый эффект)

2. Если в продукте есть эффект сезонности, то держать тестирование 1–2 цикла

Расчёт необходимого объёма данных и срока тестирования

Длительность тестирования

Длительность тестирования сократится, если:

- увеличится вероятность ошибки 1-го рода
- уменьшится мощность тестирования
- увеличится минимальный отслеживаемый эффект

Лайфхаки для уменьшения длительности

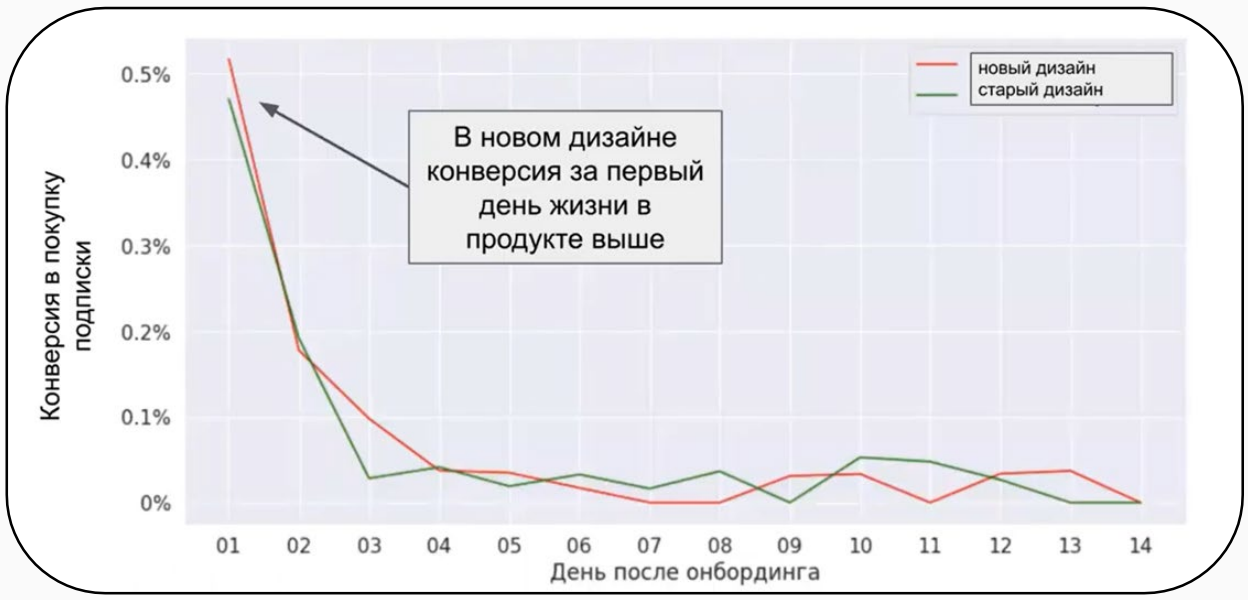
- Выбор более чувствительной метрики. Для определения статистически значимого изменения достаточно выборки меньшего размера
- Удаление выбросов. Экстремально большие или малые значения признаков

Расчёт необходимого объёма
данных и срока тестирования

Кейс: прокси-метрики

Факт: регистрация происходит только
в первый день пользования продуктом.

Вывод: конверсия упадёт в первые дни,
если новый дизайн влияет на неё.

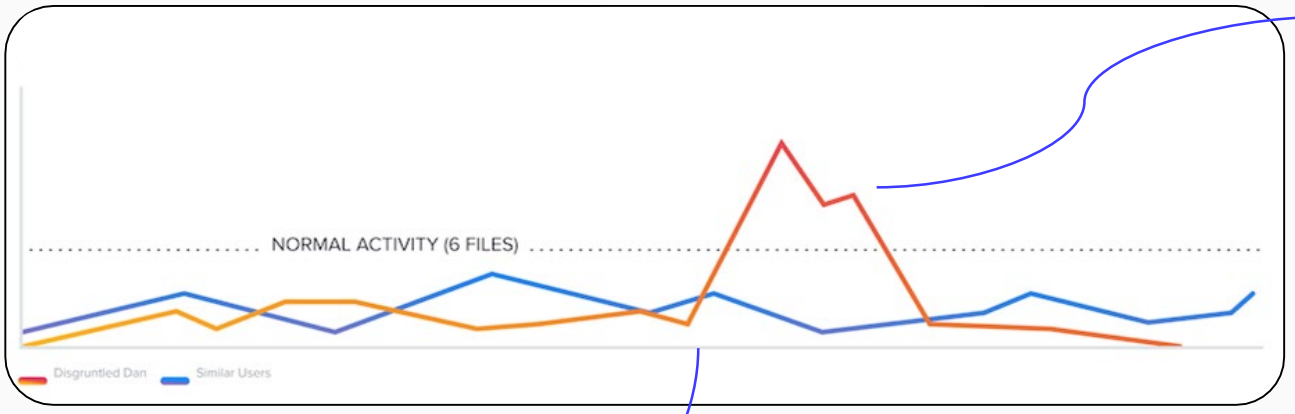


Конверсия не ухудшилась в первый день после регистрации. Значит, можем доверять результатам тестирования. Похож на **A/A-тест**, но отличие в том, что мы уже запустили тестирование.

Расчёт необходимого объёма
данных и срока тестирования

Кейс: стресс
пользователя от изменений

Сам факт изменений в сервисе может привести к аномальному поведению клиентов. Чтобы минимизировать эффект от такого поведения, почти всегда стоит выбрасывать из анализируемых данных части, собранные сразу после изменения.



Эти данные лучше удалить. Разница вызвана самим фактом изменений, а не эффективностью гипотезы

Тут произошло введение изменения

Расчёт необходимого объёма
данных и срока тестирования

Как рассчитывать
сроки A/B-тестирования?

Существует большое число калькуляторов,
но у всех есть свои нюансы:

- разные калькуляторы дают разные результаты
- поддерживают разбиение только 50 % на 50 %
- поддерживают только двустороннюю альтернативу
- калькулятор отталкивается от нормального распределения
- никто не отменял выбросы и случайный мусор в данных

Расчёт необходимого объёма
данных и срока тестирования

Как рассчитывать
сроки A/B-тестирования?

Существует большое число калькуляторов,
но у всех есть свои нюансы:

- для некоторых типов метрик калькуляторов
  не существует вовсе
- иногда имеет смысл давать «второй шанс»
  (если присутствует явно выраженный тренд,
  то подождать чуть дольше)
- Определяя примерное время (количество
  данных) тестирования, берите
  немного с запасом

Но универсального способа нет!

Итоги урока

Для быстрого определения необходимого количества экспериментов можно пользоваться уже готовыми онлайн-калькуляторами.

- Калькулятор — это быстро и просто, но есть риски ошибок. Так как калькуляторы не могут учесть абсолютно все случаи
- В случаях когда решение очень важное, лучше провести расчёт собственноручно
- Для простых случаев (таких, как конверсия) вы можете использовать онлайн-калькуляторы с высокой вероятностью, но для более сложных или ответственных случаев лучше считать результаты самим