

Лекция №3

**Дизайн А/В-тестов в
реальных условиях**

Елисова Ирина
ML Head



Зачем нам A/B-тестирование?

Почему нельзя просто делать
изменения, опираясь только на **свой
собственный опыт?**

Зачем нам A/B-тестирование?

Почему нельзя просто делать
и Потому что вывод на основе A/B - это
**единственный математически
верный** метод понять что изменение
ОК (или нет)

Зачем нам A/B-тестирование?

и Почему нельзя просто делать
Потому что вывод на основе A/B - это
единственный математически
верный

После анализа изменения
принимаются важные **решения о**
продукте, поэтому ошибка может стоить
дорого \$\$\$

Зачем нам A/B-тестирование?

Почему нельзя просто делать
и Потому что вывод на основе A/B - это
единственный математически

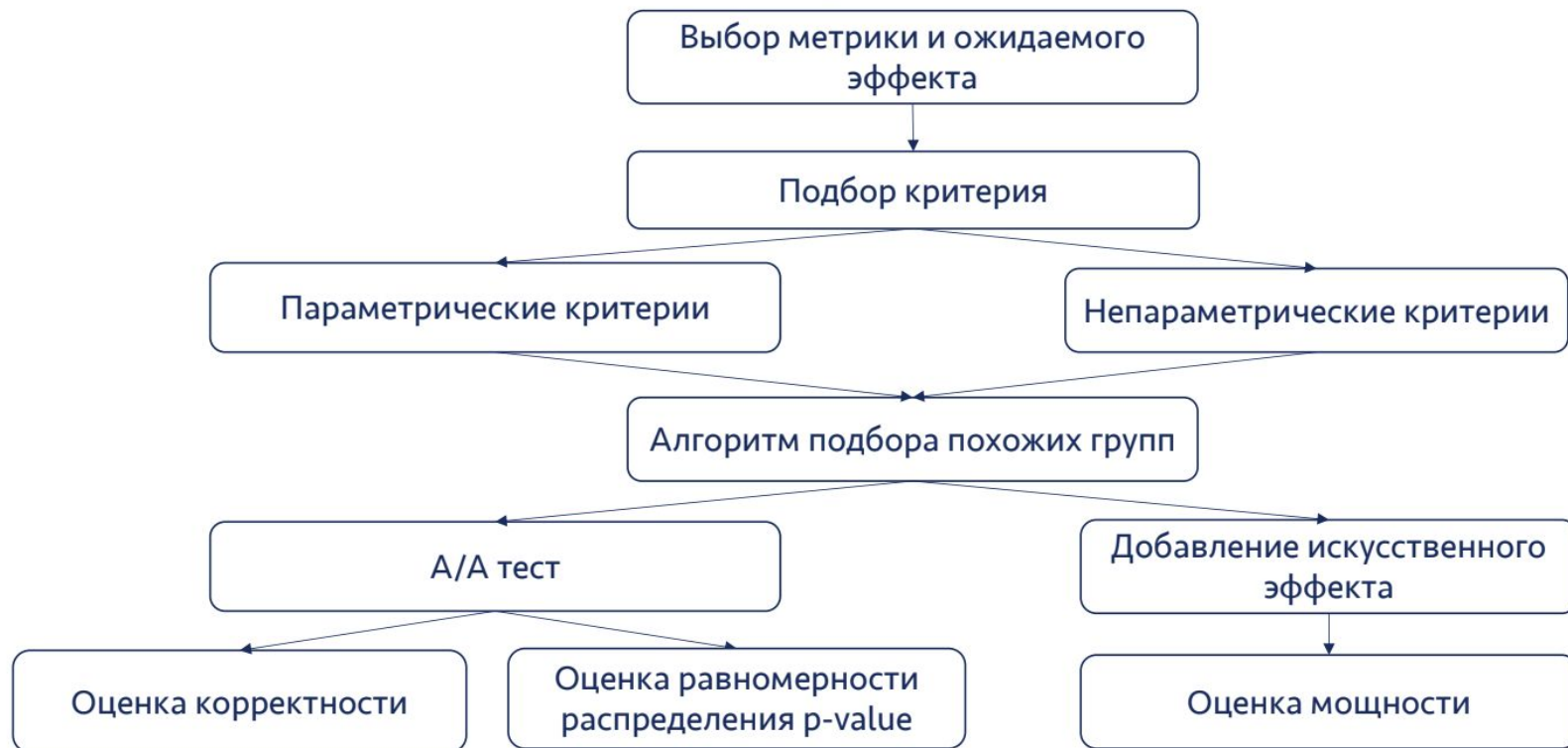
ве После анализа изменения

прод

Необходимо выделять рабочее
время на эксперименты с
дизайном

**Процесс
дизайна A/B-теста:
pov - real**

Ров: процесс дизайна



Real

Длительность
теста?

А что мы
измеряем?

Как выбрать из
100500 критериев

Там что-то говорили
про мощность

Надо выгрузить
данные из базы

команда



Решение: идем по шагам для создания дизайна

Бизнес
требования

Математика

Данные

Соединяем для создания дизайна

Бизнес
требования

Какие **метрики** интересуют бизнес?

Какое **увеличение метрики** имеет смысл
с точки зрения бизнеса?

Примеры метрик

Непрерывные = вектор чисел

Пример [11, 234, 1000, ... , 82]

- количество времени пользователя в приложении
- LTV = lifetime value
- ARPU = average revenue per user
- TVTu = total view time

Примеры метрик

Конверсии $[M/N] = [310 / 1000] = 0.031$

- основаны на факте целевого действия: клика, покупки
- **например, зашедшие в приложение (1000 чел.) -> затем кликнувшие на кнопку (310 чел.)**
- CTR = click through rate
- CR = conversion rate

Увеличение метрики

Пример:

на 0.05% или на 3% рост метрики в группе В

- **увеличение** метрики в A/B-тесте обычно **не очень большое** (до 5%)
- может быть больше, только если продукт новый и любое изменение в нем для пользователей - целое событие (например, выход на рынок LLM-моделей типа ChatGPT)

Соединяем для создания дизайна

Бизнес
требования

Как **часто** мы можем запускать тест?

Насколько **долго** мы можем держать тест?

Обсудите **ошибки первого и второго рода**

Проверим знания

Виды ошибок

https://docs.google.com/forms/d/e/1FAIpQLSdZHI65TtWP9gxq_k74XmHGUBj-kGFkcec_tDsDKzEKULZGLCA/viewform?usp=sf_link

Как узнать, что была ошибка стат. теста?

https://docs.google.com/forms/d/e/1FAIpQLSfGR_JH12fcgWR_rPa2HuTlwjBgei7w0IHwmiFkg8OaeudWg/viewform?usp=sf_link

Соединяем для создания дизайна

Знаем **какие критерии** бывают
и область их применимости

Знаем **какие** бывают
распределения (кроме
нормального)



Математика

Длительность теста

- Длительность теста измеряется в днях (неделях) или в количестве пользователей в группах А, В
- По сути это одно и то же, так как за 1 день приходит фиксированное кол-во пользователей
- Вы можете оценить эту величину по историческим данным в базе данных

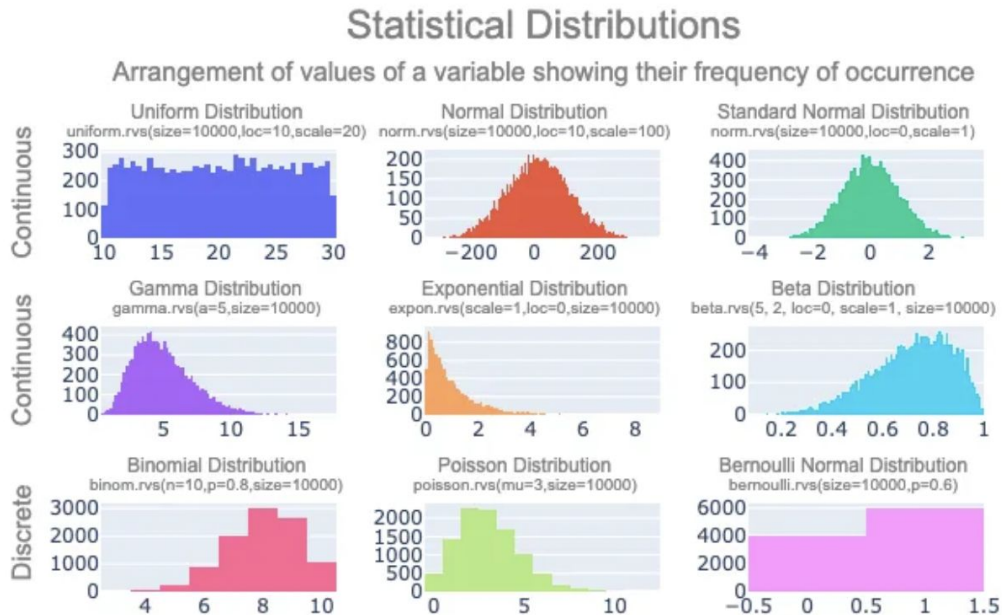
Распределение метрики

Выгружаем **исторические данные метрики** и строим гистограмму

- Это и будет распределением данных
- Данных должно быть достаточно ($>> 1000$ пользователей)
- Учитывайте сезонность (минимум 1 неделя)

Можно данные **смоделировать** из распределения

- Нужно знать теоретические распределения



[Источник](#)

Соединяем для создания дизайна

Знаем **какие критерии** бывают
и область их применимости

Знаем какие бывают
распределения (кроме
нормального)

А какую **метрику** мы **умеем считать**?

Какой **эффект** (MDE) мы можем
обнаружить с учетом пары (метрика +
критерий) за адекватное время?

Математика

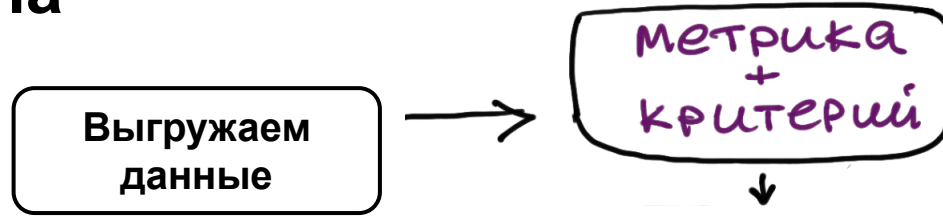
+

Данные

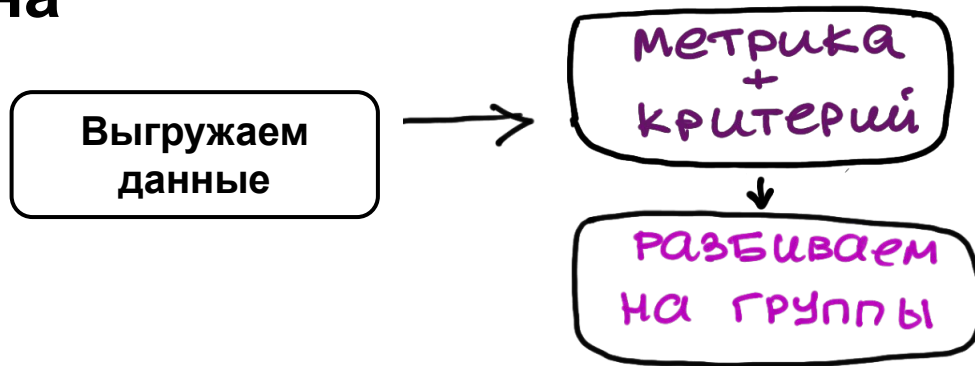
Шаги дизайна

Выгружаем
данные

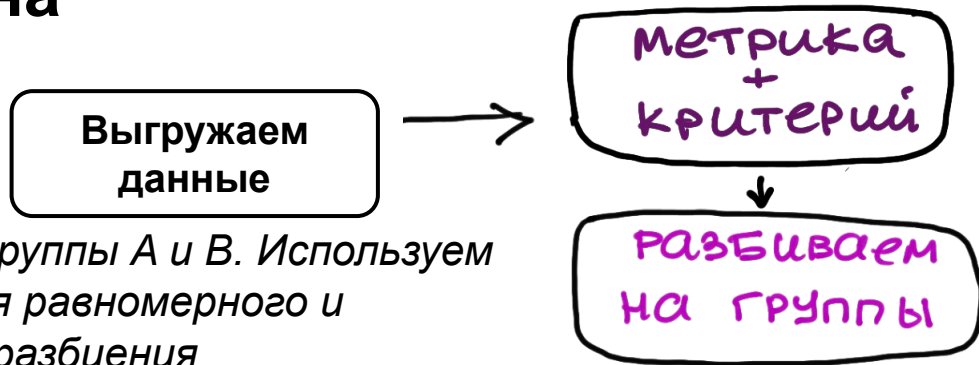
Шаги дизайна



Шаги дизайна

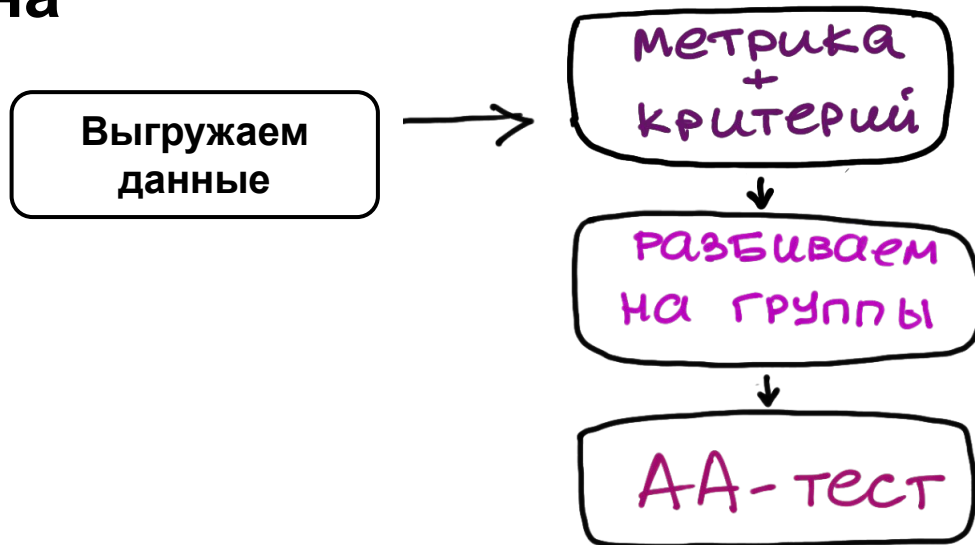


Шаги дизайна

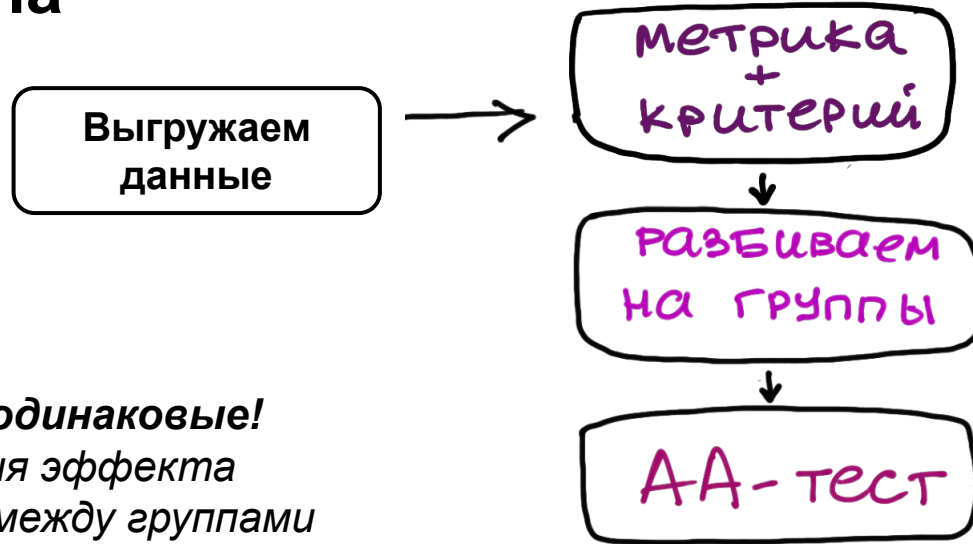


Разбиваем на две группы A и B. Используем хэш - функцию для равномерного и воспроизводимого разбиения

Шаги дизайна



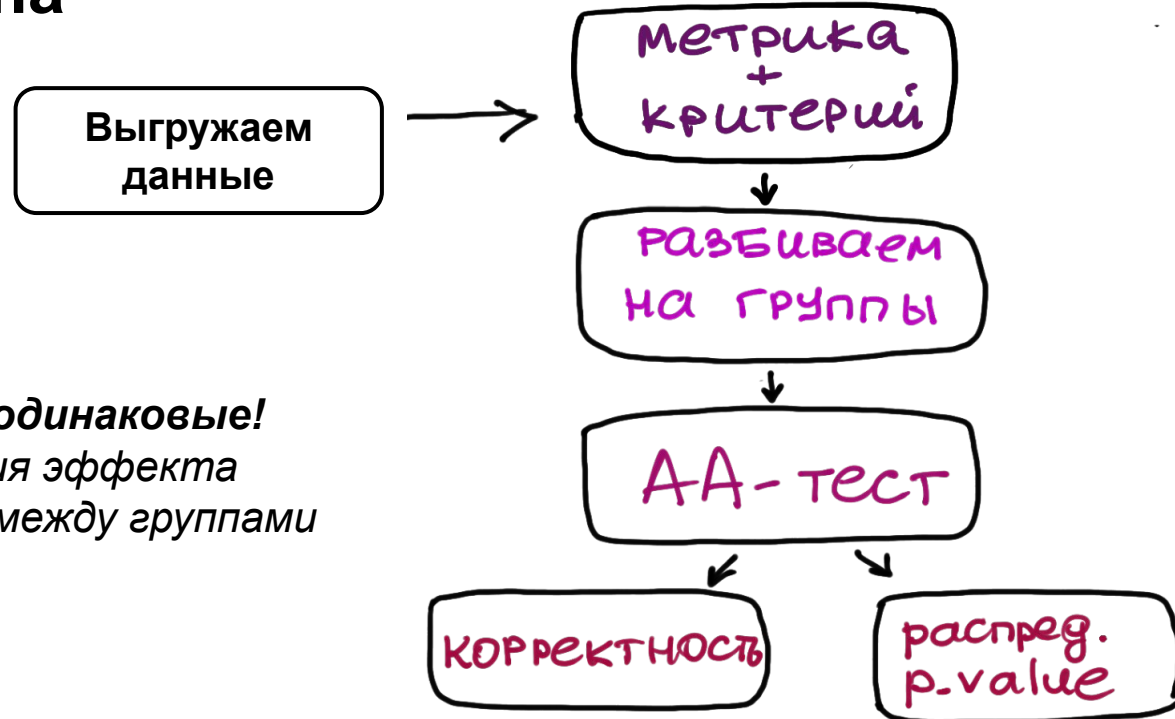
Шаги дизайна



Обе группы A и B - **одинаковые!**

- Нет добавления эффекта
- Нет различия между группами

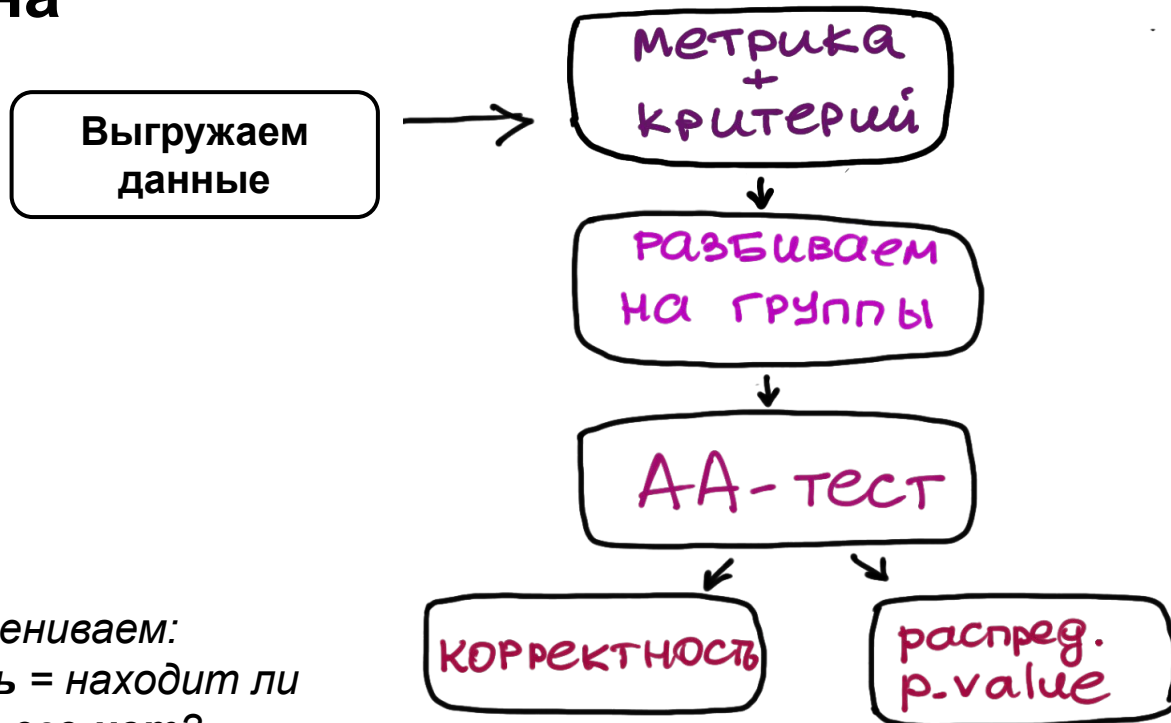
Шаги дизайна



Обе группы A и B - одинаковые!

- Нет добавления эффекта
- Нет различия между группами

Шаги дизайна



После разбиения оцениваем:

- **Корректность** = находит ли эффект, когда его **нет**?
- **p_value** должно быть распределено **равномерно**

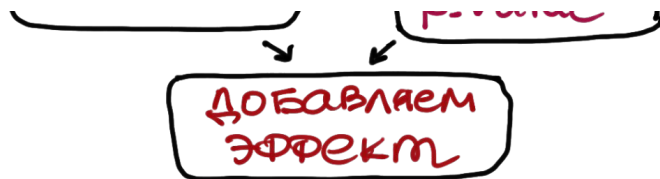
Шаги дизайна



Шаги дизайна

Теперь надо оценить, когда группы А и В - **НЕ** одинаковые!

- **Есть эффект**
- Мы его должны добавить сами, искусственно
- Какого размера?



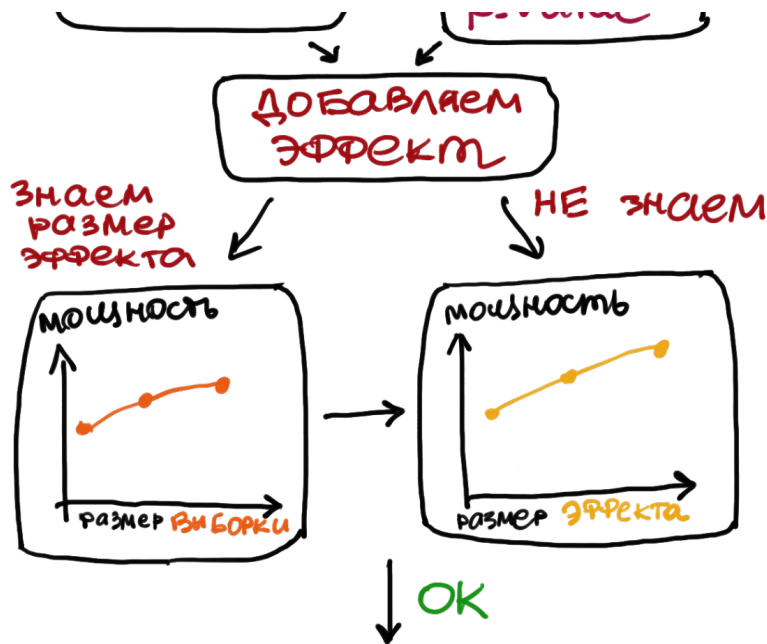
Шаги дизайна

Теперь надо оценить, когда группы А и В - **НЕ** одинаковые!

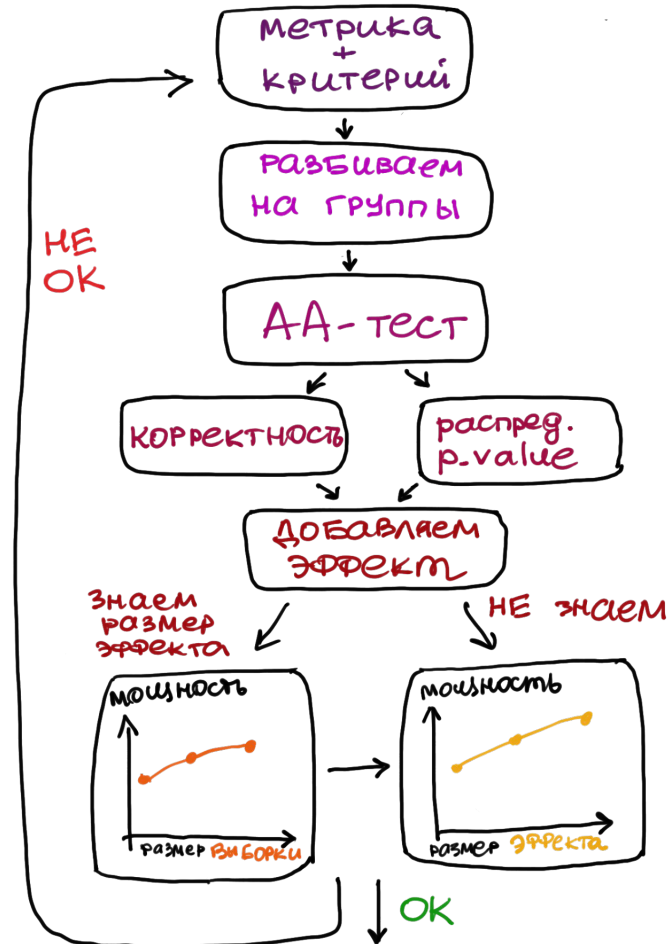
- **Есть эффект**
- Мы его должны добавить сами, искусственно

Проверяем **мощность** = **вероятность найти различие там, где оно ЕСТЬ**

Хотим **мощность больше 0.8**
(говорят “мощный” критерий)



Шаги дизайна



Когда дизайн ГОТОВ



Обязательно проводим A/A-
тест в продакшене !

АА-тест
на проде

↓ OK

АВ-тест
на проде

Подводим итоги теста

Выгружаем реальные
данные теста

Подводим итоги теста

Выгружаем реальные
данные теста



Считаем на них
выбранную метрику
(как на дизайне!)

Подводим итоги теста

Выгружаем реальные
данные теста

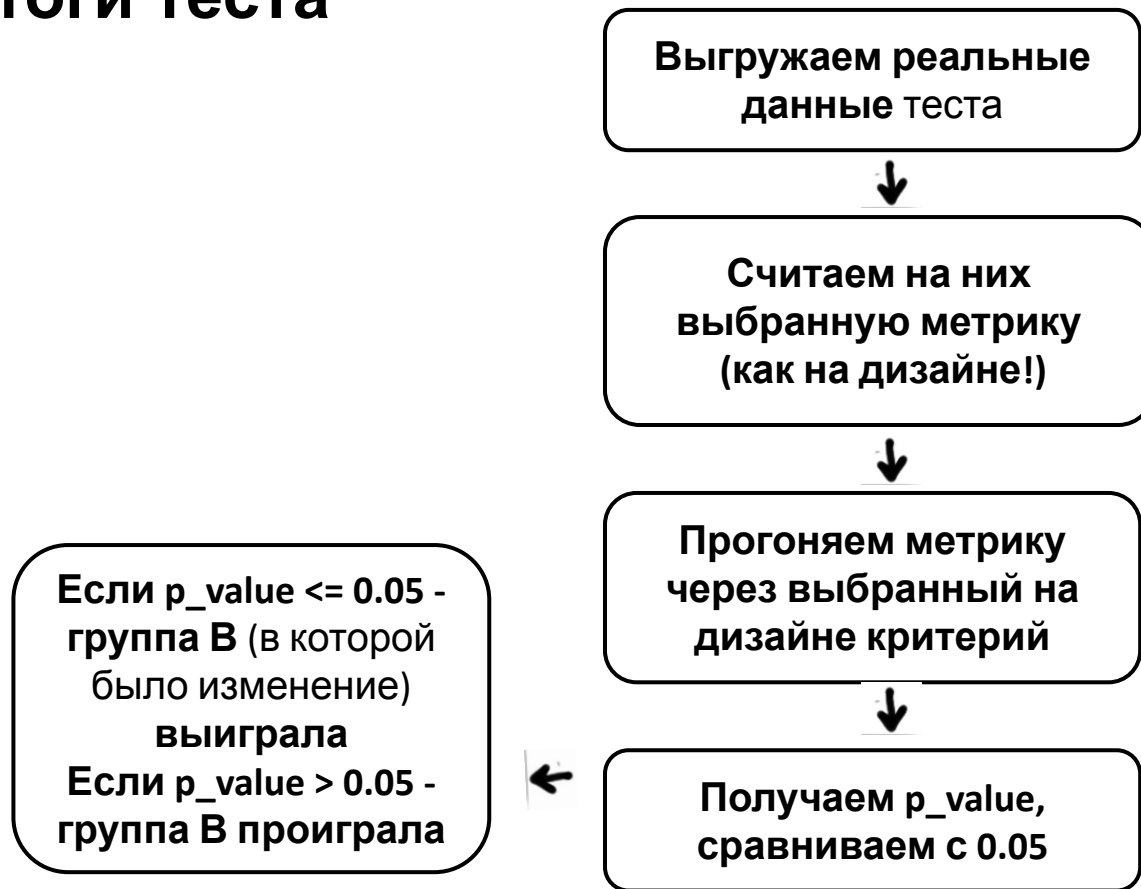


Считаем на них
выбранную метрику
(как на дизайне!)



Прогоняем метрику
через выбранный на
дизайне критерий

Подводим итоги теста



Что такое нестатзначимый / непрокрашенные / серый результат?

$$p_value > 0.05$$

1. **p_value** (уровень значимости) выше установленного порога (обычно 0.05): **не можем** отвергнуть нулевую гипотезу о том, что группы **ОДИНАКОВЫ**
2. **Малый размер выборки**: Если выборка слишком мала, статистическая мощность теста может быть недостаточной для обнаружения даже существующих различий.
3. **Небольшой эффект**: Возможно, что разница между группами действительно есть, но она слишком мала, чтобы быть выявленной с текущей мощностью теста.

Типичные ошибки в дизайне

Окей, ~~google~~-chatgpt



What can I help with?

калькулятор длительности аб теста



Окей, ~~google~~-chatgpt



Калькуляторы работают
только для z-test / t-test
формулы

What can I help with?

$$\frac{2\sigma^2(\Phi^{-1}(\frac{\alpha}{2}) + \Phi^{-1}(\beta))^2}{MDE_{absolute}^2}$$

калькулятор длительности аб теста



Любитель митапов



*“В этом докладе я вам
расскажу, как мы
делаем АБ-тесты в
CompanyName”*

Любитель митапов



Ваши данные могут быть в 100 раз меньше, иметь другие сущности и т.д.



“В этом докладе я вам расскажу, как мы делаем АБ-тесты в `CompanyName`”

Готовая платформа а/б-тестов



OVERVIEW Performance Summary				
UNIQUE VISITORS	Variations	Visitors	Views	example
79,797	Original	19,942 25.0%	--- 10% (± 0.70)	10
DAYS RUNNING 131 Started: April 9, 2014 How long should I run my test?	Variation #1	19,899 25.0%	+20.0% 12% (± 0.70)	▲ + 12
	Variation #2	19,989 25.1%	+10.0% 11% (± 0.70)	▲ + 11
	Variation #3	19,967 24.9%	-10.0% 9% (± 0.70)	▼ - 9

Готовая платформа а/б-тестов



Подходит, когда дизайн аб эксперимента **УЖЕ продуман**

OVERVIEW Performance Summary				
<div>UNIQUE VISITORS</div> <div>79,797</div> <div>DAYS RUNNING</div> <div>131</div> <div>Started: April 9, 2014</div> <div>How long should I run my test?</div>	Variations	Visitors	Views	example
	Original	19,942 25.0%	--- 10% (±0.70)	10
	Variation #1	19,899 25.0%	+20.0% 12% (±0.70)	12
	Variation #2	19,989 25.1%	+10.0% 11% (±0.70)	11
	Variation #3	19,967 24.9%	-10.0% 9% (±0.70)	9

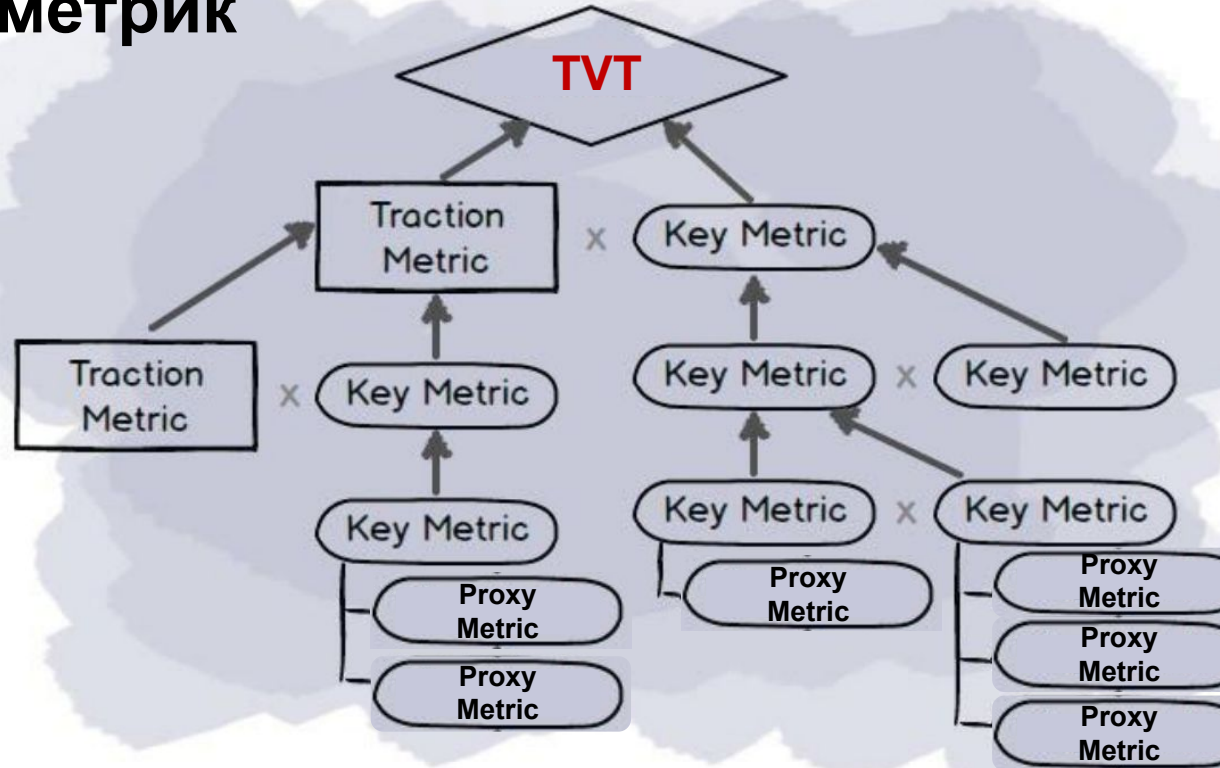
Выбранная метрика и тест на 6 месяцев

Часто эту метрику невозможно
прокрасить за адекватное
время

А кто захочет делать A/B-тест
длиной в полгода?

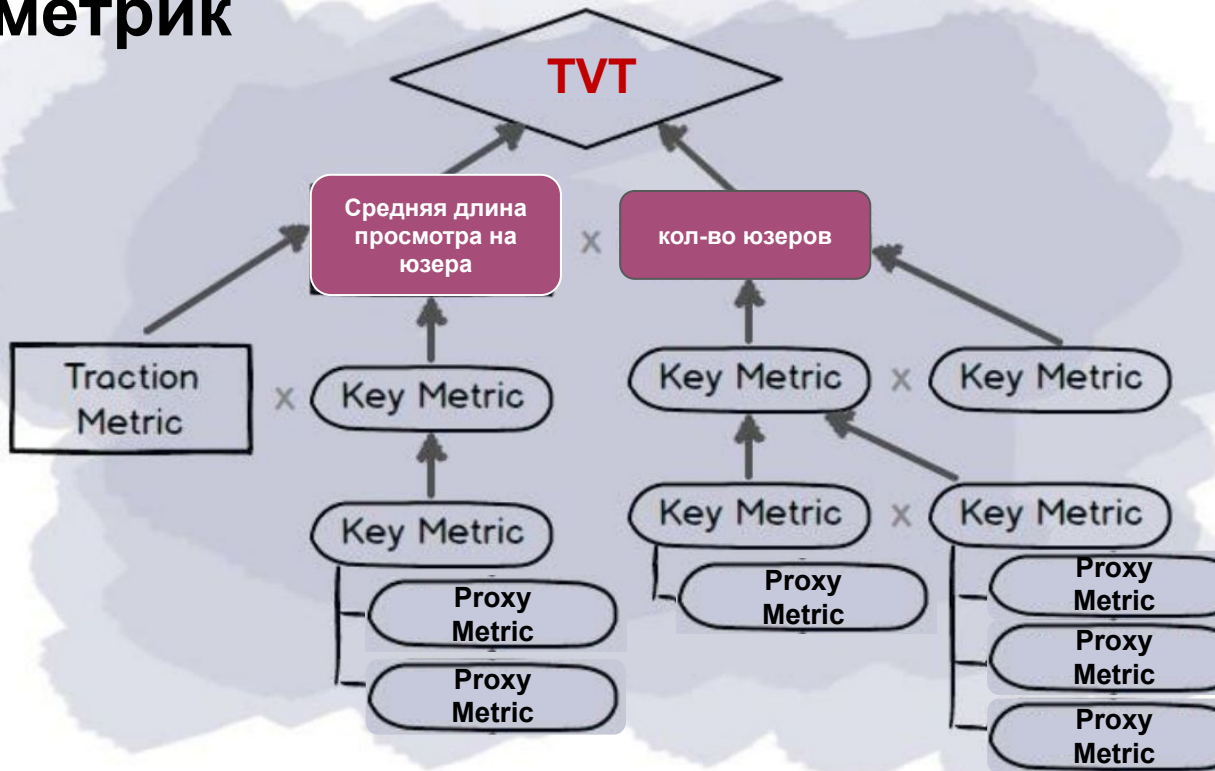
Решение: дерево метрик

Строим дерево метрик =
декомпозируем
 основную метрику
 проекта



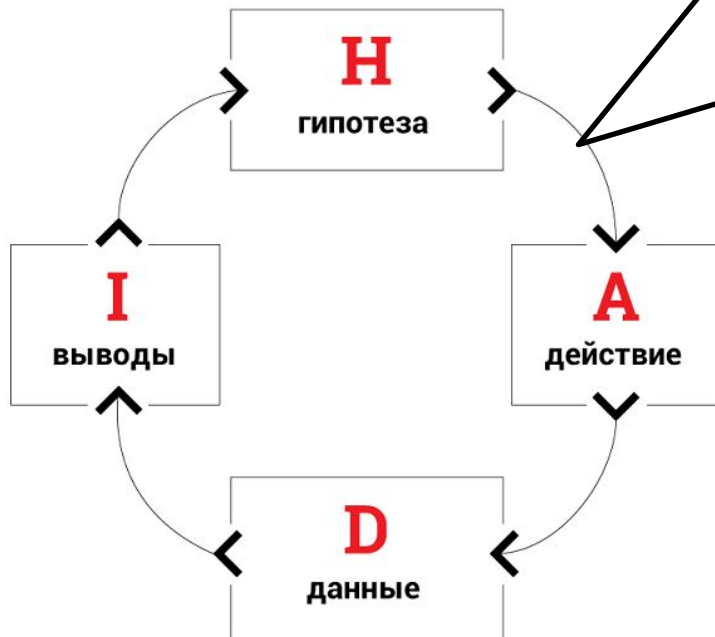
Решение: дерево метрик

метрики уровня ниже
(прокси-метрики)
оцениваем в **A/B-
тесте**



Длительность а/б-теста != цикл а/б-теста

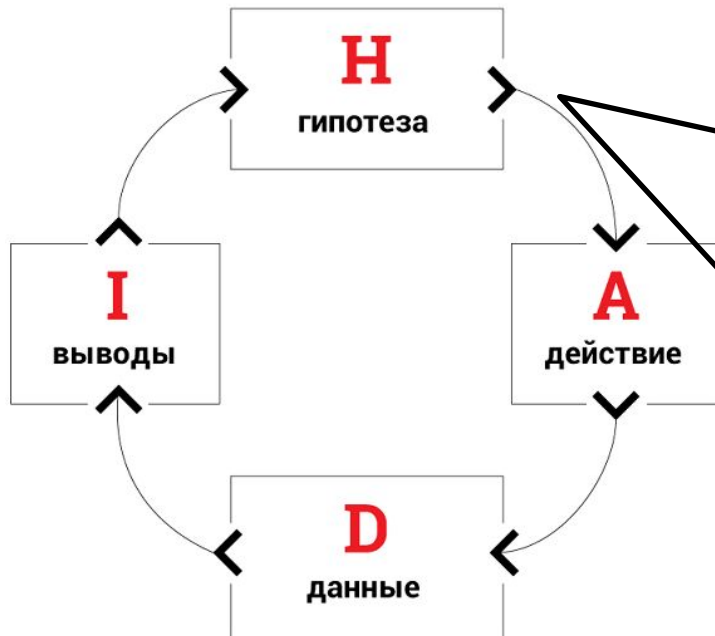
іТМО



Команда **продукта** работает над новой фичей = 1-2 недели

(например, поменять текст рекламной рассылки, перекрасить кнопку или улучшить модель кредитного скоринга)

Длительность а/б-теста != цикл а/б-теста



ITMO

+ Команда
аналитики
продумывает, как
измерять
результат

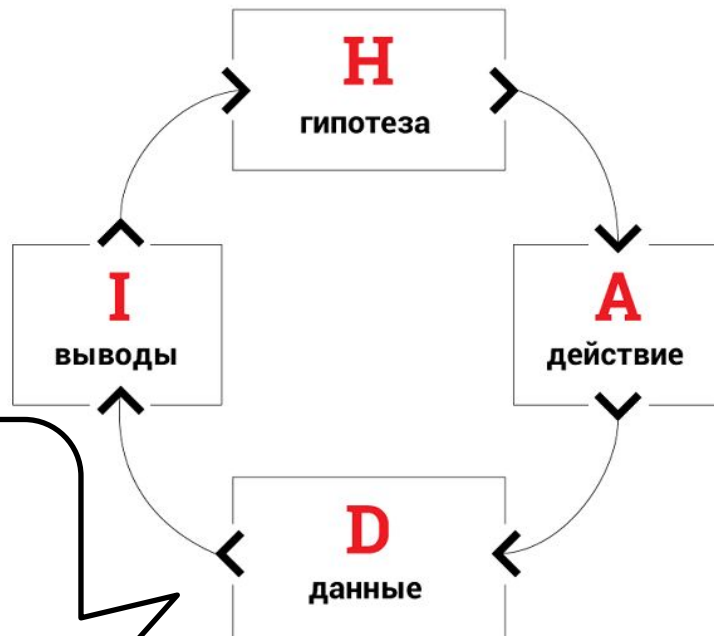
= 1 неделя

*какой стат.
критерий,*

*сколько по
времени будет
длиться тест*

на каких данных

Длительность а/б-теста != цикл а/б-теста



Идет А/В-тест = 1-2-3
недели

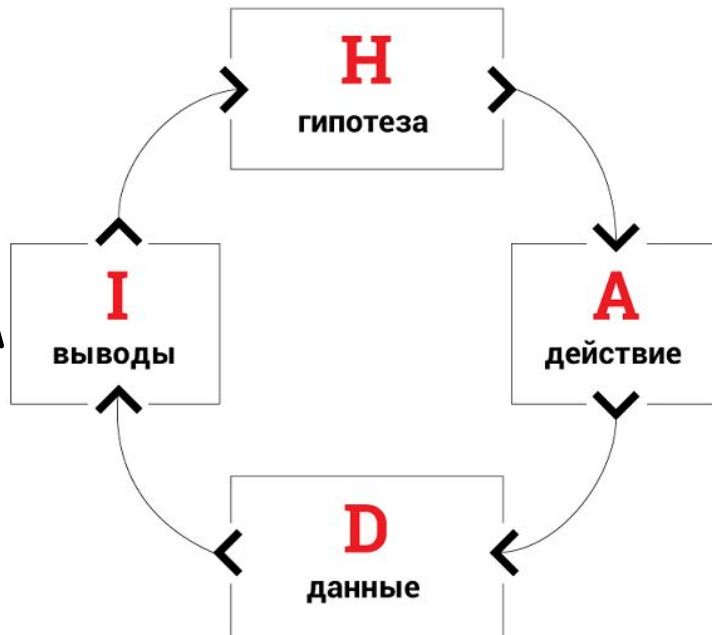
*изменение становится
видно части клиентов в
приложении*

Длительность а/б-теста != цикл а/б-теста

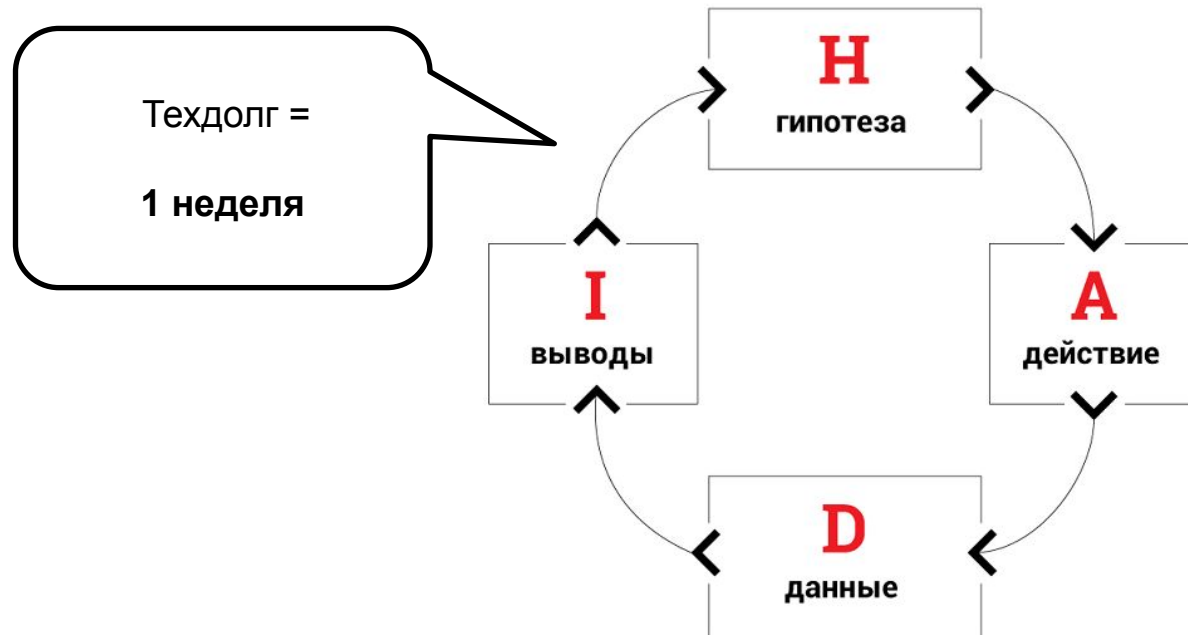
Результаты теста =
до 1 недели

Команда
аналитиков
собирает данные
по результатам
теста

считает для стат.
критерия p -value и
делает вывод о
том что изменение
было полезное или
нет



Длительность а/б-теста != цикл а/б-теста



Дизайн: выводы

Это важный этап всего пайплайна А/В-тестирования

Выделяем отдельно время на дизайн (минимум неделя)

Обязательно проводим А/А-тест на продакшн, если дизайн новый

Понимаем что именно бизнес хочет измерять

Метрика + размер эффекта

Дизайн: выводы

Это важный этап всего пайплайна A/B-тестирования

Выделяем отдельно время на дизайн (минимум неделя)

Обязательно проводим A/A-тест на продакшн, если дизайн новый

Понимаем что именно бизнес хочет измерять

Метрика + размер эффекта

Хорошо знаем свои данные

Каким распределением они моделируются?

Сколько у нас данных?

Понимаем сколько времени (= пользователей) надо на тест

Понимаем какие метрики можем считать

Дерево, прокси метрики

Истории из практики

Cool story: о добавлении эффекта

Есть метрика “время пользователя в секундах на карточке товара” в виде вектора, в котором **есть нули**

время пользователя в секундах = **[0, 4, 45, 900, 0, 56, ..., 0, 23]**

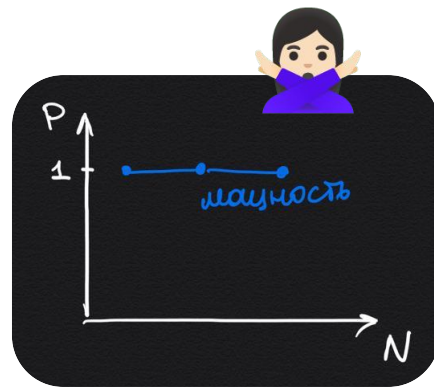
Хотим искусственно добавить эффект в 3% в группу В для подсчета мощности

Cool story: о добавлении эффекта

1. Приплюсуем 3% ко всей группе В
2. Домножим группу В на $(1 + 3\%)$
3. Распределение бы

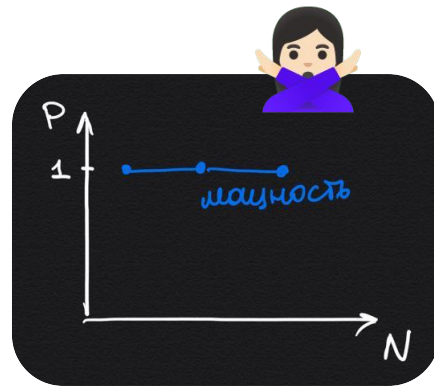
Cool story: о добавлении эффекта


1. Приплюсуем 3% ко всей группе В
2. Домножим группу В на $(1 + 3\%)$
3. Распределение бы



Cool story: о добавлении эффекта

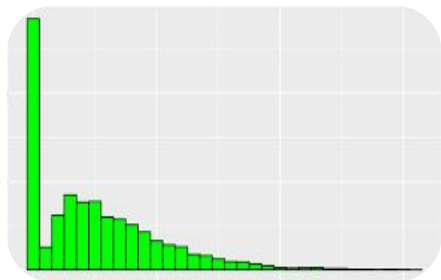
1. Приплюсуем 3% ко всей группе В
2. **Домножим группу В на $(1 + 3\%)$**
3. Распределение бы



 уже
 неплохо, нули
 не теряем

Cool story: о добавлении эффекта

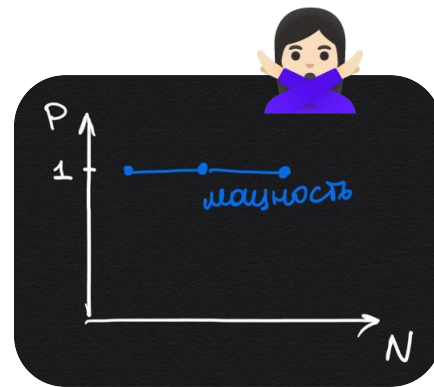
1. Приплюсуем 3% ко всей группе В
2. Домножим группу В на $(1 + 3\%)$
3. **Распределение бы**



Tweedie распределение

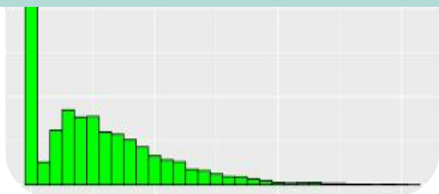
- ✓ Задаем количество нулей
- ✓ Задаем mean

✓ уже
неплохо, нули
не теряем



Cool story: о добавлении эффекта

Обращайте внимание на способ, которым вы добавляете эффект



Tweedie распределение

- ✓ Задаем количество нулей
- ✓ Задаем mean

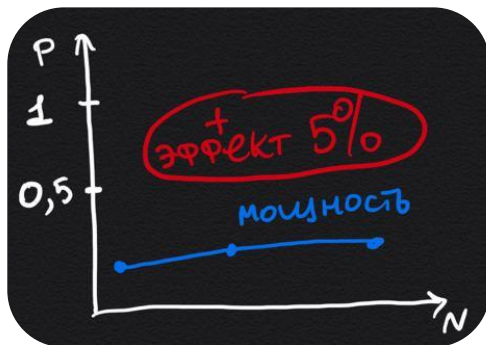
✓ уже
неплохо, нули
не теряем



Cool story: о логарифмировании метрики

Проблема:

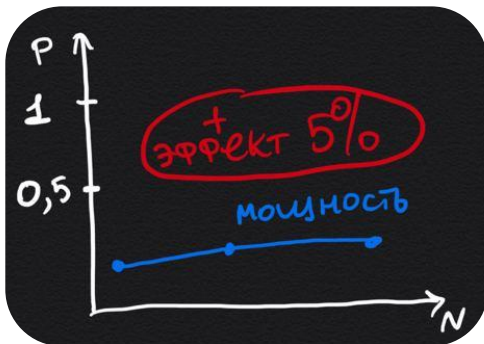
метрика нечувствительная,
мощность $\ll 0.8$ на доступном
нам количестве пользователей



Cool story: о логарифмировании метрики

Проблема:

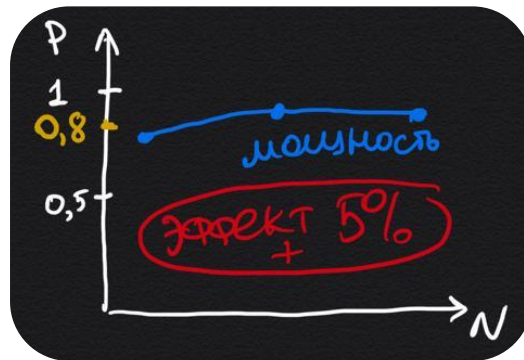
метрика нечувствительная,
мощность $\ll 0.8$ на доступном
нам количестве пользователей



А давайте возьмем

логарифм от метрики!

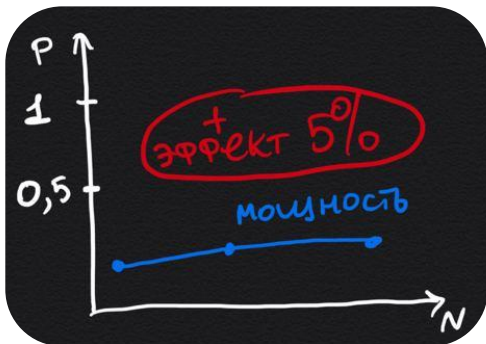
Добавили 5%, пересчитали,
выглядит неплохо



Cool story: о логарифмировании метрики

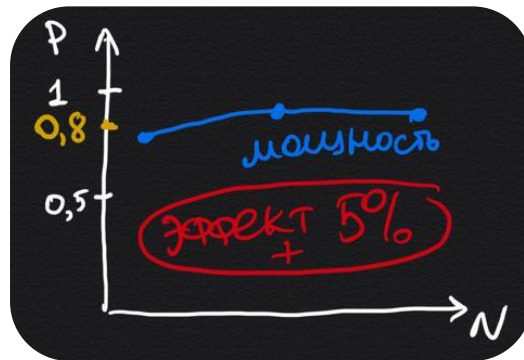
Проблема:

метрика нечувствительная,
мощность $\ll 0.8$ на доступном
нам количестве пользователей



Что здесь может
быть не так?

А давайте возьмем
логарифм от метрики!
Добавили 5%, пересчитали,
выглядит неплохо



1. Все так, выкатываю тест
2. Эффект нужно было взять больше
3. Эффект нужно было взять меньше

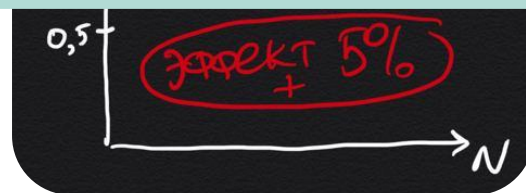
Cool story: о логарифмировании метрики

✗ **+5%** $\log(\text{метрики}) = \textbf{+30\%}$ метрики
[недостижимый результат]

✗ **+5%** метрики = **+0.01%** $\log(\text{метрики})$
[не отловишь такой MDE в реальном тесте]




Что здесь может
быть не так?



1. Все так, выкатываю тест
2. Эффект нужно было взять больше
3. **Эффект нужно было взять меньше**

Cool story: о группах А и В на проде

API не умеет на лету раздавать юзеру группу (А или В)

 Поэтому новым пользователям **группу раздаем оффлайн (записываем в таблицу)** каждый день

Бывают **незарегистрированные** пользователи, носящие один и тот же **дефолтный user_id = guest**

Свой **уникальный логин** этот пользователь получает **на следующий день** (если регистрируется)

Cool story: о группах А и В на проде

Сколько групп у нас будет в тесте?

1. Две
2. Три
3. Четыре

Бывают незарегистрированные пользователи,
носящие один и тот же дефолтный `user_id = guest`

Свой уникальный логин этот
пользователь получает на следующий
день (если регистрируется)

Cool story: о группах А и В на проде

Сколько групп у нас будет в тесте?

1. Две
2. Три
3. Четыре

- 🎯 Новый пользователь зашел в приложение под guest (группа А)
- 🎯 Впечатлился, зарегистрировался и на следующий день **опять зашел**
- 🎯 Сменил id с guest на user_1234 и внезапно попал в другую группу (В)

Итог: создаем группу “сменивших группу” А -> В

Полезные ссылки

Советы по А/Б тестированию от аналитиков Avito ([раз](#), [два](#))

[Пример построения дерева метрик \(видео\) от ex CEO Skyeng](#)

[Валерий Бабушкин | Метрики: от офлайна до иерархии](#)

[Пример онлайн калькулятора тестов от Mindbox \(CTR only\)](#)