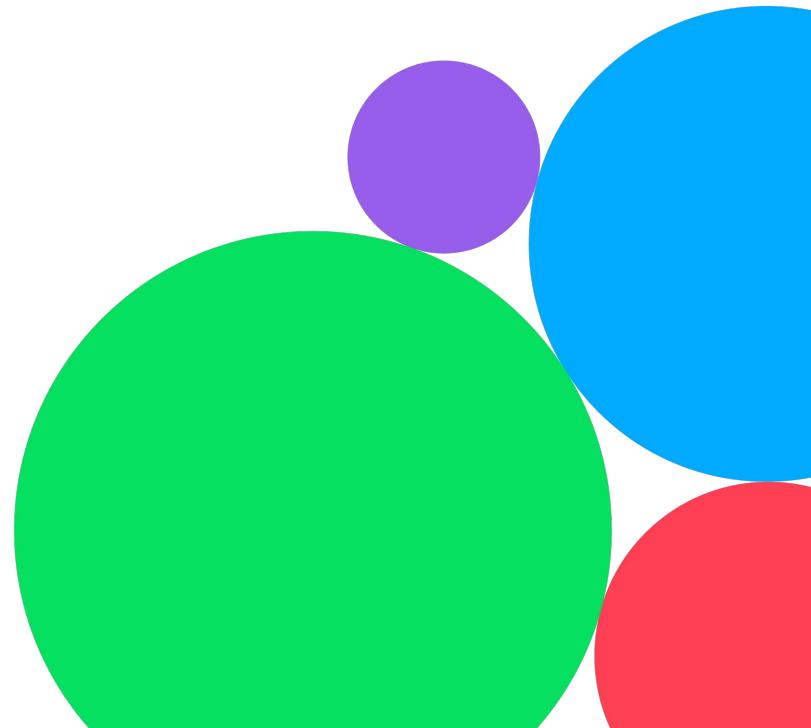


АБ-тесты: основа принятия решений в компании

Евгения Мурзаева
acting Team Lead at MNZ, Avito



Привет!

Мурзаева Женя
тг @jane_mur

- **acting Team Lead** в команде монетизации Авито
- Развиваю размещение и продвижение на площадке
- Закончила ВШЭ
- Провела **более 100 АБ-тестов**



О чём будет лекция?

Часть 1 — Бизнесовые исследования

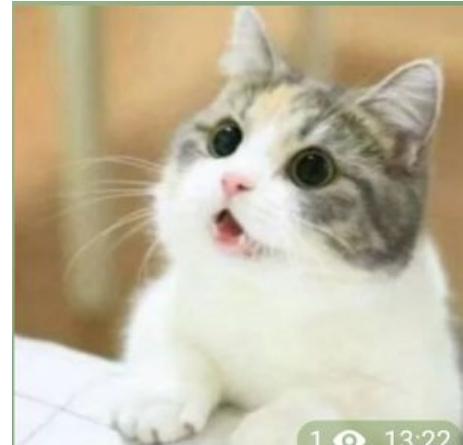
Часть 2 — Вводная про АБ-тесты

Часть 3 — Как запустить тест? Гипотезы, метрики, MDE

Часть 4 — Как интерпретировать результат? Статистика и продуктовое мышление

Часть 5 — Чуть больше про экспы

Часть 6 — Кейсы





Часть 1.

Бизнесовые исследования

Как X влияет на Y?



Как изменится метрика Y, если мы сделаем X?

Конверсия в регистрацию

Если сократить кол-во обязательных полей

Уровень подростковой преступности

Если ввести в школах курс по эмоциональному интеллекту

% завершенного курса

Если добавить ачивки и прогрессбар

Вероятность осложнений гриппа

Если сделать прививку от гриппа

Продолжительность жизни

Если ежедневно пить энергетики

Продолжительность сессии

Если ввести новый алгоритм рекомендаций

Как X влияет на Y?



Как изменится метрика Y, если мы сделаем X?

Конверсия в регистрацию

Если сократить кол-во обязательных полей

Уровень подростковой преступности

Если ввести в школах курс по эмоциональному интеллекту

% завершенного курса

Если добавить ачивки и прогрессбар

Вероятность осложнений гриппа

Если сделать прививку от гриппа

Продолжительность жизни

Если ежедневно пить энергетики

Продолжительность сессии

Если ввести новый алгоритм рекомендаций

Как отвечать на такие вопросы?

Как отвечать на такие вопросы?



Самый точный вариант:
две параллельные
вселенные

Как отвечать на такие вопросы?



Самый точный вариант:
две параллельные
вселенные

1

Минимум
0

Максимум
1

СГЕНЕРИРОВАТЬ

Share icon

A user interface element for generating random numbers. It features a large number '1' at the top left, followed by two input fields labeled 'Минимум' (Minimum) with '0' and 'Максимум' (Maximum) with '1'. Below these is a blue button labeled 'СГЕНЕРИРОВАТЬ' (Generate). In the bottom right corner of the interface, there is a small share icon.

Самый неточный вариант:
рандомайзер :)

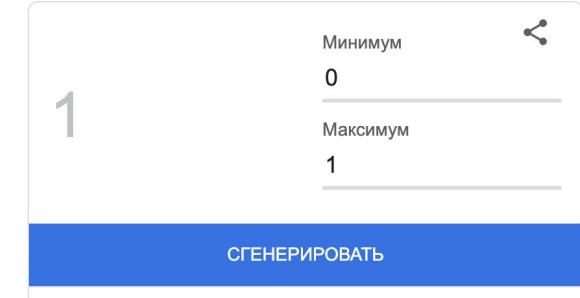
Как отвечать на такие вопросы?



- Опросы
- Найти похожих
- Сравнить во времени
- АБ-тесты

Самый точный вариант:
две параллельные
вселенные

Реалистичный вариант:
исследовательские методы



Самый неточный вариант:
рандомайзер :)

Опросы

Опросы в основном используются в двух случаях:

1. нам нужно ответить на вопрос “Почему”

Расскажите, чего вам не хватило на странице?

[Пройти опрос за 2 минуты](#)

Опросы

Опросы в основном используются в двух случаях:

1. нам нужно ответить на вопрос “Почему”
2. наша метрика – мнение пользователя



Дмитрий, здравствуйте.

Расскажите, чего вам не хватило на странице?

Поделитесь, пожалуйста, своими впечатлениями о YouDo. Это поможет нам понять, над чем стоит поработать.

Вы бы порекомендовали нас друзьям?

Выберите оценку от 0 до 10.

NPS



0 — не порекомендую

10 — обязательно порекомендую

Опросы – что может быть не так

Отражает мнение, а не реальное поведение

Почти все пользователи сказали, что подсказки – классные, они точно повысят retention. На практике retention не увеличился. Пользователи просто не хотели нас расстраивать

Опросы – что может быть не так

Отражает мнение, а не реальное поведение

Почти все пользователи сказали, что подсказки – классные, они точно повысят retention. На практике retention не увеличился. Пользователи просто не хотели нас расстраивать

Ошибка самоотбора

Опрос согласились пройти 12% из тех, кого мы попросили. Это оказались наиболее активные и лояльные юзеры, располагающие свободным временем. Результаты – нерепрезентативны.

Опросы – что может быть не так

Отражает мнение, а не реальное поведение

Почти все пользователи сказали, что подсказки – классные, они точно повысят retention. На практике retention не увеличился. Пользователи просто не хотели нас расстраивать

Ошибка самоотбора

Опрос согласились пройти 12% из тех, кого мы попросили. Это оказались наиболее активные и лояльные юзеры, располагающие свободным временем. Результаты – нерепрезентативны.

Мало данных

Мы получили данные по 1000 ответов.

Это несравненно с тем, что мы можем получить путем АБ-тестирования

Поиск похожих

Поиск похожих



Поиск похожих



Поиск похожих

Это могут быть:

- региональные тесты (для избежания сетевого эффекта)
- PSM (propensity score matching)

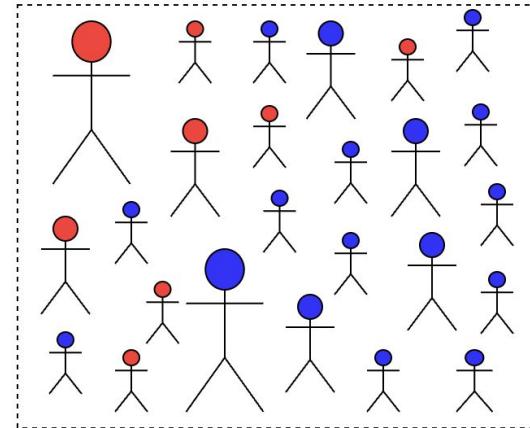


Поиск похожих

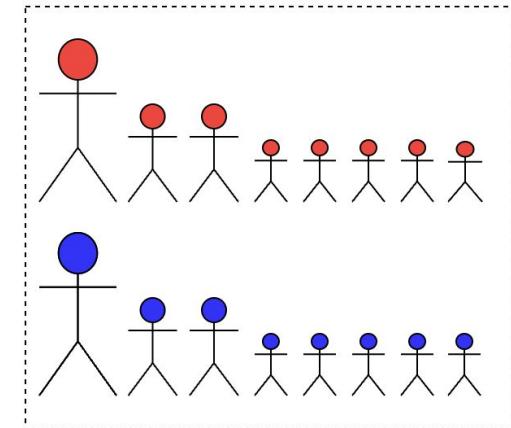
Это могут быть:

- региональные тесты
- **PSM (propensity score matching)**

Without Propensity Score Matching

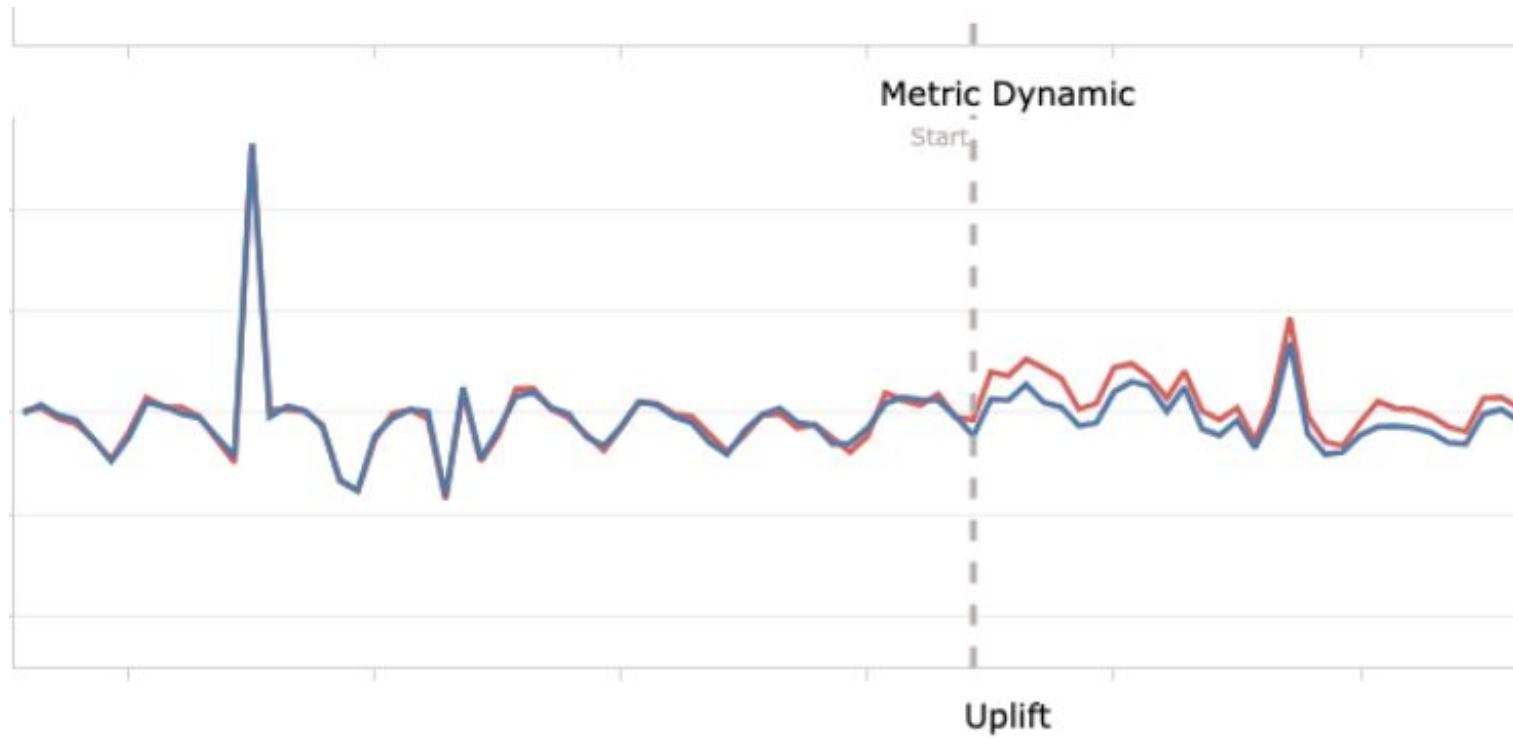


With Propensity Score Matching



● - Test ● - Control

Визуализация поиска похожих



Поиск похожих – что может пойти не так

Результаты зависят от фичей

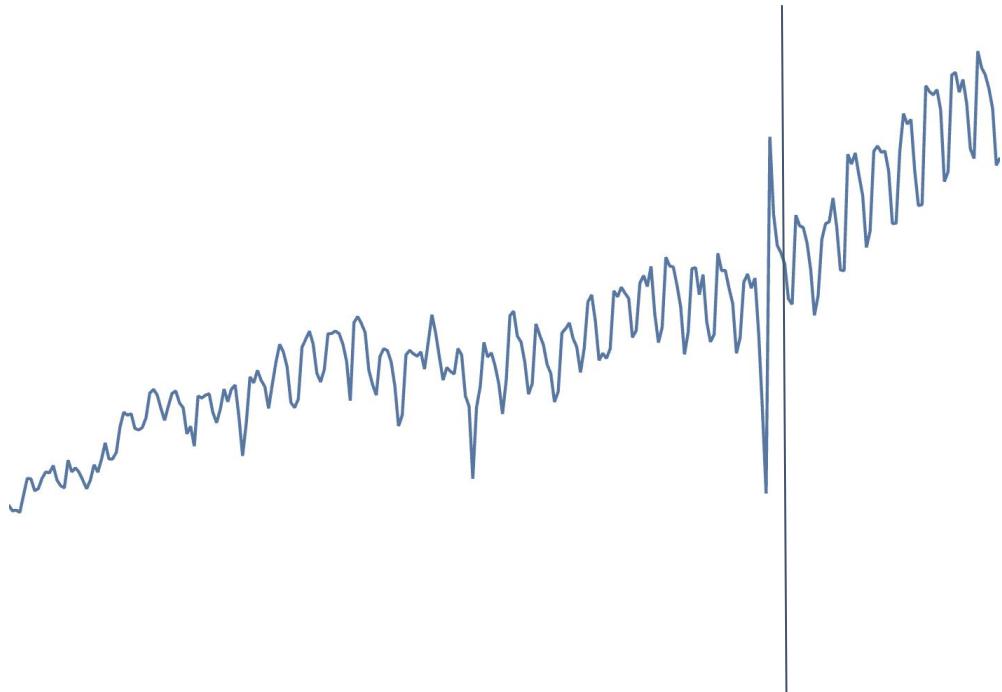
Мы выбрали в качестве фичей для поиска похожих: пол, возраст, дату регистрации. Получили эффект +4%. Потом добавили еще 3 фичи: ОС, был ли аккаунт ранее, % заполнения профиля. Получили результат +9%.

Результаты зависят от алгоритма

Мы использовали алгоритм Propensity Score Matching. Получили эффект +2%. Потом использовали алгоритм ближайшего соседа в пространстве. Получили результат +8%.

Сравнение во времени

Запуск продукта



Сравнение во времени

Невозможность отделить от внешних факторов

Мы увидели рост retention на 7%. Но в дни запуска фичи запустилась новогодняя распродажа. Непонятно, что привело к росту retention.

Сравнение во времени

Невозможность отделить от внешних факторов

Мы увидели рост retention на 7%. Но в дни запуска фичи запустилась новогодняя распродажа. Непонятно, что привело к росту retention.

Невозможность отделить эффект от дисперсии

Мы увидели рост на 4%. Но если взять случайны точки в прошлом и сравнить до-после, то увидим эффект в +3%, -2%, +5%, +1%. Непонятно, мы вырастили метрику или это естественный шум в данных

Сравнение во времени

Невозможность отделить от внешних факторов

Мы увидели рост retention на 7%. Но в дни запуска фичи запустилась новогодняя распродажа. Непонятно, что привело к росту retention.

Невозможность отделить эффект от дисперсии

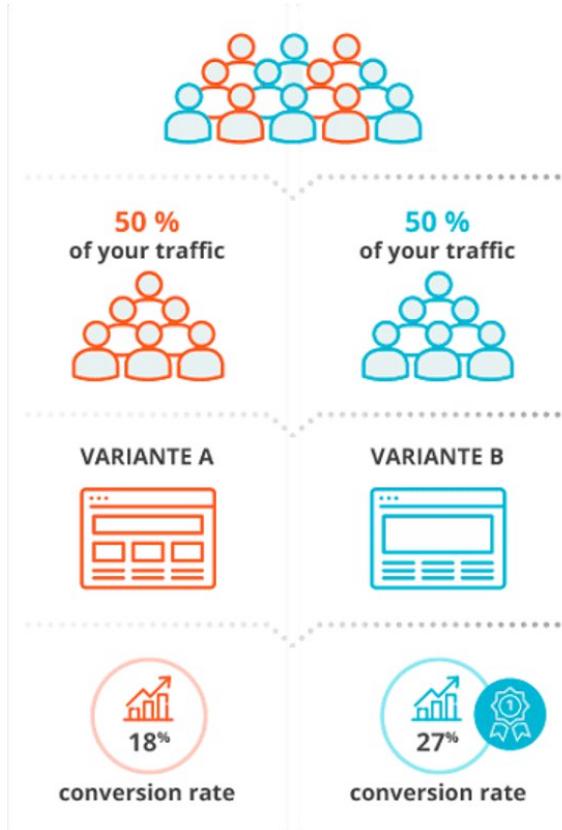
Мы увидели рост на 4%. Но если взять случайны точки в прошлом и сравнить до-после, то увидим эффект в +3%, -2%, +5%, +1%. Непонятно, мы вырастили метрику или это естественный шум в данных

Наличие тренда и сезонности

Мы увидели рост на 6%. Но на графике retention видно, что он и так рос весь последний год. Непонятно, мы повлияли или это следствие продолжающегося тренда.

Часть 2. АБ-тесты

Что такое АБ-тесты?



АБ-тесты – способ проверить гипотезу через параллельное тестирование текущей версии и альтернативной.

Для этого мы:

- Разбиваем выборку на группы
- Показываем каждой группе одну из вариаций
- Собираем результат и интерпретируем его

Зачем нам нужны АБ тесты?

- Убираем внешние факторы, так как эти факторы влияют на обе группы одинаково
- Можем определить точную выгоду от наших изменений
- Принимаем решения, основываясь на данных, а не наших предположениях

АБ-тесты – это простая идея, но сложная методология

В какой момент подвести итоги?

Как разделить на группы?

А если мой тест окажется неудачным?

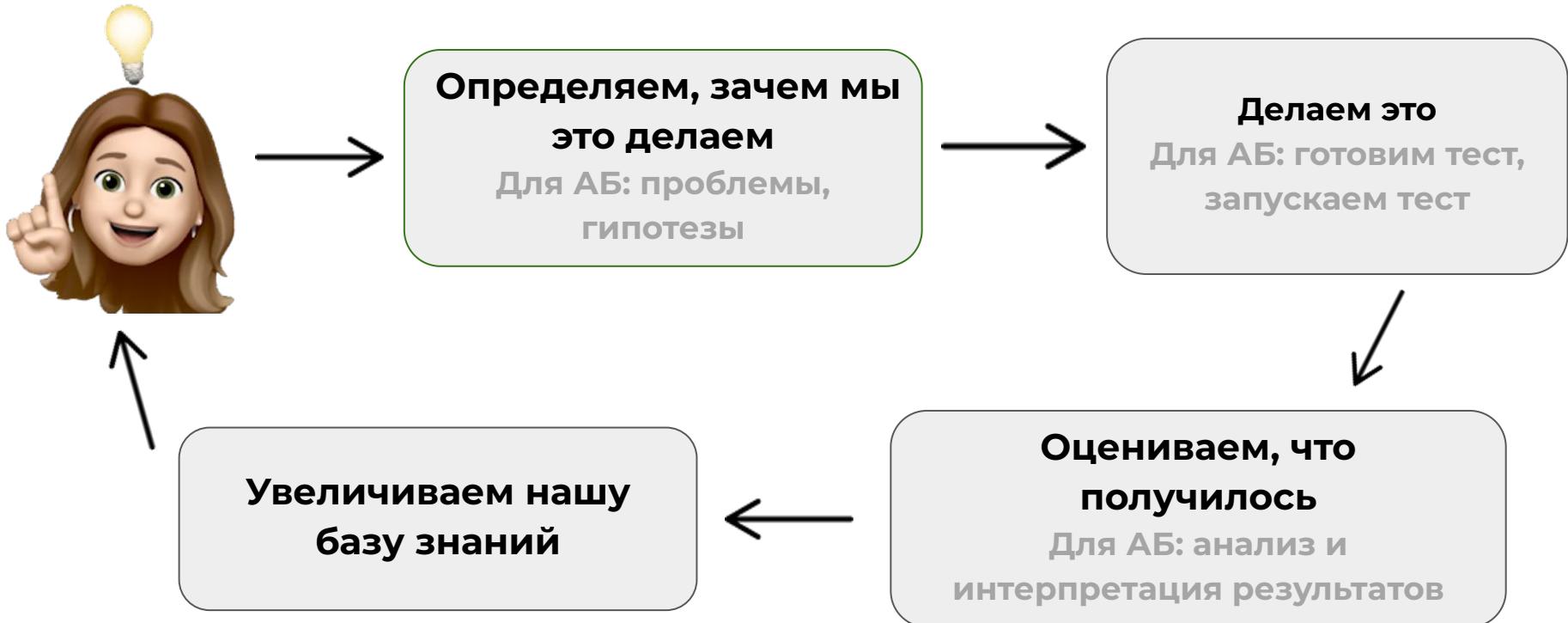


Сколько пользователей брать в тест?

А вдруг мы все сломали?

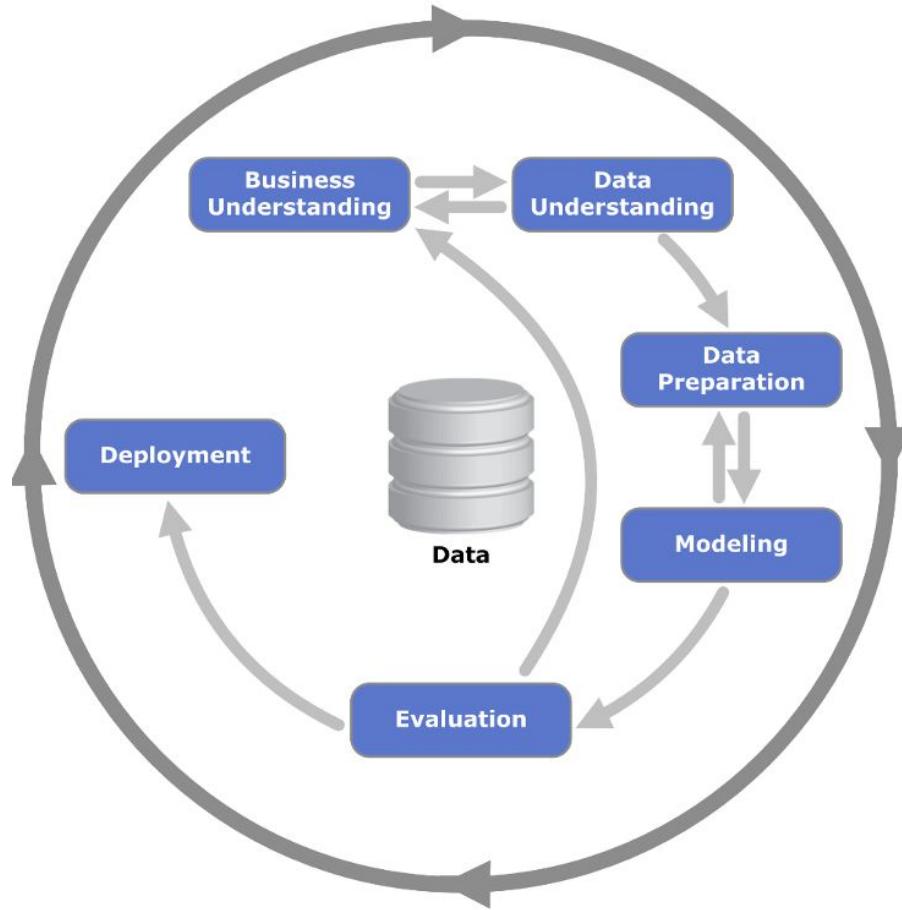
Как понять, есть ли результат?

Цикл АБ-теста



Цикл АБ-теста

CRISP-dm (Cross-Industry Standard Process for Data Mining)



Этапы АБ-тестов

Обнаружение идеи
или проблемы и
варианта ее решения

1

Выбор метрик

2-3

Формулировка
гипотез

2-3

Дизайн эксперимента

4

Запуск и мониторинг

5

Анализ и
интерпретация

6

АБ-тесты в Авито

В 2025 году запустили **6к** АБ-тестов

Процент успешных из них: ?

Все метрики							
		test vs control		Health OK		Exposed 3.1%	
Разрез	Metric	MoE	t-stat	Lift	Mean	Num	Den
total	i0	perf:fatal_app_errors:users_count	1.1%	4.7σ	1.5%	0.015	298.3K
		vas_user	0.6%	3.8σ	-0.7%	0.270	5.5M
	i1	user_vas	0.6%	3.8σ	-0.7%	0.270	5.5M
		user_vas_start_date	1.1%	6.3σ	-2.1%	0.078	1.6M
		vas_paying_user_ratio	0.6%	3.4σ	-0.7%	0.019	5.5M
		vas_transactions_performance_start_date	3.8%	3.5σ	-4%	0.087	1.8M
	i2	bbip_transactions	1.9%	3.7σ	-2.1%	0.289	5.9M
		bbip_users	0.8%	8.3σ	-1.9%	0.175	3.6M
		user_vas_performance	0.7%	7.9σ	-1.6%	0.226	4.6M
		user_vas_visual	1.2%	3.9σ	1.4%	0.095	1.9M
		user_vas_xl	1.2%	3.9σ	1.4%	0.086	1.7M
	other	user_lf_start_date	0.9%	3.9σ	1%	0.215	4.4M

АБ-тесты в Авито

В 2025 году запустили **6к** АБ-тестов
Процент успешных из них: 20%

Все метрики							
		test vs control				Health OK	Exposed 3.1%
Разрез	Metric	MoE	t-stat	Lift	Mean	Num	Den
total	i0	perf:fatal_app_errors:users_count	1.1%	4.7σ	1.5%	0.015	298.3K
		vas_user	0.6%	3.8σ	-0.7%	0.270	5.5M
	i1	user_vas	0.6%	3.8σ	-0.7%	0.270	5.5M
		user_vas_start_date	1.1%	6.3σ	-2.1%	0.078	20.4M
		vas_paying_user_ratio	0.6%	3.4σ	-0.7%	0.019	5.5M
		vas_transactions_performance_start_date	3.8%	3.5σ	-4%	0.087	288.9M
		bbip_transactions	1.9%	3.7σ	-2.1%	0.289	5.9M
	i2	bbip_users	0.8%	8.3σ	-1.9%	0.175	3.6M
		user_vas_performance	0.7%	7.9σ	-1.6%	0.226	4.6M
		user_vas_visual	1.2%	3.9σ	1.4%	0.095	1.9M
		user_vas_xl	1.2%	3.9σ	1.4%	0.086	1.7M
	other	user_if_start_date	0.9%	3.9σ	1%	0.215	4.4M

АБ-тесты в Авито

В 2025 году запустили **6к** АБ-тестов

Процент успешных из них: 20%

Процент масштабированных: ?

Все метрики							
test vs control				Health OK		Exposed 3.1%	
Разрез	Metric	MoE	t-stat	Lift	Mean	Num	Den
total	i0 perf:fatal_app_errors:users_count	1.1%	4.7σ	1.5%	0.015	298.3K	20.4M
	vas_user	0.6%	3.8σ	-0.7%	0.270	5.5M	20.4M
	i1 user_vas	0.6%	3.8σ	-0.7%	0.270	5.5M	20.4M
		user_vas_start_date	1.1%	6.3σ	-2.1%	0.078	1.6M
		vas_paying_user_ratio	0.6%	3.4σ	-0.7%	0.019	5.5M
	i2 bbip_transactions	3.8%	3.5σ	-4%	0.087	1.8M	20.4M
		bbip_users	1.9%	3.7σ	-2.1%	0.289	5.9M
		user_vas_performance	0.8%	8.3σ	-1.9%	0.175	3.6M
		user_vas_visual	0.7%	7.9σ	-1.6%	0.226	4.6M
		user_vas_xl	1.2%	3.9σ	1.4%	0.095	1.9M
other	user_if_start_date	1.2%	3.9σ	1%	0.215	4.4M	20.4M

АБ-тесты в Авито

В 2025 году запустили **6к** АБ-тестов

Процент успешных из них: 20%

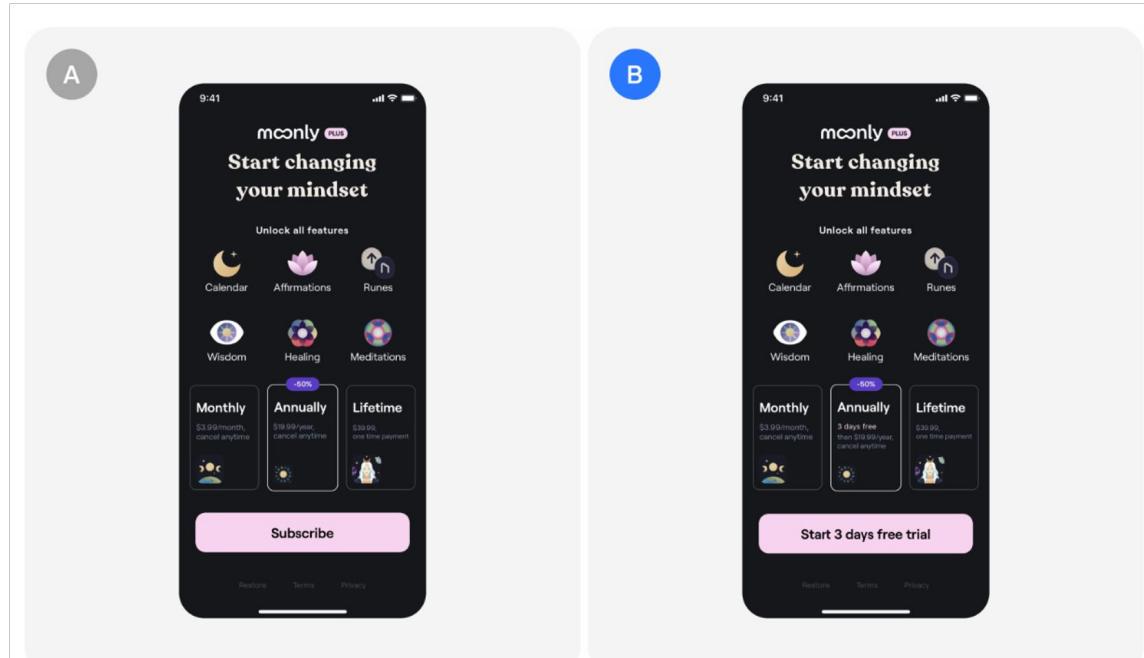
Процент масштабированных: 30%

Все метрики							
test vs control				Health OK		Exposed 3.1%	
Разрез	Metric	MoE	t-stat	Lift	Mean	Num	Den
total	i0 perf:fatal_app_errors:users_count	1.1%	4.7σ	1.5%	0.015	298.3K	20.4M
	vas_user	0.6%	3.8σ	-0.7%	0.270	5.5M	20.4M
	i1 user_vas	0.6%	3.8σ	-0.7%	0.270	5.5M	20.4M
		user_vas_start_date	1.1%	6.3σ	-2.1%	0.078	1.6M
		vas_paying_user_ratio	0.6%	3.4σ	-0.7%	0.019	5.5M
	i2 bbip_transactions	3.8%	3.5σ	-4%	0.087	1.8M	20.4M
		bbip_users	1.9%	3.7σ	-2.1%	0.289	5.9M
		user_vas_performance	0.8%	8.3σ	-1.9%	0.175	3.6M
		user_vas_visual	0.7%	7.9σ	-1.6%	0.226	4.6M
		user_vas_xl	1.2%	3.9σ	1.4%	0.095	1.9M
other	user_if_start_date	1.2%	3.9σ	1%	0.215	4.4M	20.4M

Что тестируют другие

Примеры ab в дизайне продукта

<https://abtest.design/>



Free trial only for annual subscription

Moonly • +39% conversion rate

А минусы будут?

Увеличивают time to market

Недели и даже месяцы!

А минусы будут?

Увеличивают time to market

Недели и даже месяцы!

Нетривиальная методология и подводные камни

А минусы будут?

Увеличивают time to market

Недели и даже месяцы!

Нетривиальная методология и подводные камни

Требуют ресурсов аналитика

Аналитики – дорогой ресурс
Это может стать бутылочным горлышком

А минусы будут?

Увеличивают time to market

Недели и даже месяцы!

Нетривиальная методология и подводные камни

Требуют ресурсов аналитика

Аналитики – дорогой ресурс
Это может стать бутылочным горлышком

Хорошо бизнесу не значит хорошо всем!

Участники эксперимента: пользователи могут считать эксперименты не этичными
Продукт оунеры: отсутствие эффекта в АБ не списать на проблемы с методологией и сложно сфальсифицировать

Когда АБ тесты не нужны? 1/4

Нужны качественные выводы

*Почему происходит X?
Как происходит X?*

**Почему байеры не пользуются
Автодоставкой?**

**Как покупатели ищут товары, если
не знают точного названия?**

Интервью/ открытые опросы

Наблюдение за поведением

**Почему 9% селлеров бросают
форму подачи объявления?**

**Как байер и селлер
договариваются о цене?**

Юзабилити-тесты/ интервью

Интервью/ анализ переписки

Когда АБ тесты не нужны? 2/4

Оценка эффекта не нужна

*Не требует значительных инвестиций
Является объективным улучшением
Нет альтернатив*

У пользователей на
андроид < 13.1 не
отправляются
сообщения

Новые требования
ЦБ: обязательная
верификация
арендодателей

Появилась новая марка
Авто, которая активно
выходит на российский
рынок

Когда АБ тесты не нужны? 3/4

Эффект огромен

Ожидаемый эффект кратно больше колебаний в прошлом периоде

одновременно не было других изменений, влияющих на метрику

До

После

01

28%

02

29%

03

28%

04

27%

05

42%

06

44%

Эффект огромен

Когда АБ тесты не нужны?

Невозможно провести
АБ-тест

*Физически невозможно
Этические причины
Финансовые риски*

Когда АБ тесты не нужны?

Невозможно провести
АБ-тест

*Физически невозможно
Этические причины
Финансовые риски*

Как увеличится ROI Sales-менеджеров, если увеличить их зарплату на 30%?

Сколько денег потеряет Авито, если сервис будет недоступен для пользователей в течении часа?

Как повлияет на траты факт нерешенной проблемы в поддержке более недели?

Как повлияло проведение ЧМ-2018 на выручку классифайда?



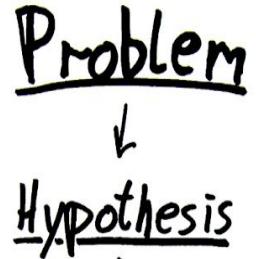
Часть 3.

Запускаем АБ-тест

С чего начинается АБ-тест?

С обозначения проблемы, которую мы решаем.

Далее мы формируем гипотезы, которые могут решить нашу проблему



С чего начинается АБ-тест?

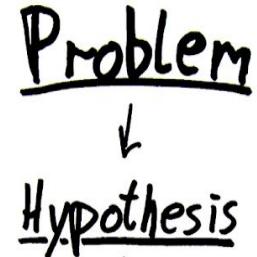
С обозначения проблемы, которую мы решаем.

Далее мы формируем гипотезы, которые могут решить нашу проблему

Почему так?

Понимание проблемы = понимание, зачем мы это делаем

С пропуском этого шага, увеличивается риск тестировать гипотезы с настроем “Ну а вдруг сработает”, что в лучшем случае может привести к потери времени, а в худшем – к падению важных метрик.



Примеры гипотез



Если мы сделаем рекомендации более привлекательными, станет лучше



Если мы в рекомендациях будем предлагать айтемы с более низкой ценой, они будут больше заинтересовывать пользователя, поэтому конверсия из визита в клик по айтему и в контакт с продавцом вырастет.

Зачем?

1. Стейкхолдеры сразу поймут вашу идею и смысл теста
2. Пока вы напишите подробную гипотезу, сами убедитесь еще раз, что понимаете, что делаете, зачем делаете и как сможете измерить :)

Где искать идеи для гипотез?

Количественные исследования

Анализ данных, АБ, опросы

Качественные исследования

UX-интервью, CustDev,
проблемные интервью

Идеи от команды

Продакты, аналитики,
инженеры

Бенчмаркинг

Исследования рынка,
конкурентов

Послушать

пользователей

Отзывы, обращения в
поддержку

ChatGPT

Что можно тестировать?

Клиентский интерфейс (Frontend)

- Механики заполнения форм
- Дизайн
- Контент

Бизнес - процессы

- Скрипты в call-центре
- Flow оформления продукта

Backend

- ML модели
- Скоринговые модели
- Скорость загрузки страницы /формы/приложения
- Проверяем, что не стало хуже

The screenshot shows the top navigation bar of the Avito website, featuring a red heart icon and the text "Помощь пострадавшим в Оренбургской". Below the bar are links for "Для бизнеса", "Карьера в Авито", "Помощь", "Каталоги", "Польза", and social media icons for heart and notifications. The main search bar has the placeholder "Поиск по объявлениям". Below the search bar is a grid of category icons: Авто, Недвижимость, Работа, Одежда, обувь, аксессуары, Хобби и отдых, Животные, Услуги, Электроника, Для дома и дачи, Запчасти, Товары для детей, and Путешествия. Further down are circular icons for "Оригиналы в Премиуме", "#яПомогаю", "Солнечные очки", "Устроить праздник", "Риелторы и юристы", "Помогаем «ЛизАлерт»", and "Зелень для дома". A section titled "Рекомендации для вас" displays three items: a yellow knitted hat and denim shorts labeled "Пакет женских вещей" for 1000 ₽; a portrait of a man in a suit labeled "Питер Ф. Друкер «Практика менеджмента»" for 1200 ₽; and a book cover labeled "Финансовый менеджмент и финансовая отчетность" for 1600 ₽.

AB x Guinness

Математик Вильям Госсет на производстве пива Guinness **использовал** в производстве продукции **разные виды ячменя, чтобы определить лучшее сочетание**, которое понравится потребителю больше всего.



Выбор метрик

Перед тестом определяем метрики, на которые будем смотреть.

Зачем?

1. Понимание на что мы повлияем и как это замерим
2. Настройка мониторинга, чтобы следить за ходом теста

Как правило есть **целевая** (главная) метрика и дополнительные

Контр-метрики – это метрики, которые не должны ухудшаться при росте целевой и дополнительных метрик. Например, не должны расти возвраты товара или количество брошенных корзин.

Выбор метрик

Кейс

Если мы в рекомендациях будем предлагать айтемы с более низкой ценой, они будут больше заинтересовывать пользователя, поэтому конверсия из визита в клик по айтему и в контакт с продавцом вырастет.

Метрики для теста:

- Конверсия из визита в контакт

...

Выбор метрик

Кейс

Если мы в рекомендациях будем предлагать айтемы с более низкой ценой, они будут больше заинтересовывать пользователя, поэтому конверсия из визита в клик по айтему и в контакт с продавцом вырастет.

Метрики для теста:

- Конверсия из визита в контакт
- Конверсия из визита в сделку/таргет контакт
- Количество контактов/кликов
- Количество обращений в поддержку
- Retention пользователей
- Денежные метрики (выручка, aov)

Определяем размер выборки

Сколько пользователей возьмем в наш тест?

Определяем размер выборки

Сколько пользователей возьмем в наш тест?

Двоих?



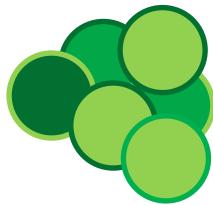
Определяем размер выборки

Сколько пользователей возьмем в наш тест?

Двоих?



Шестерых?



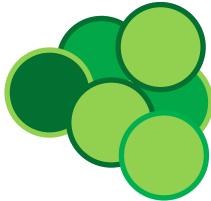
Определяем размер выборки

Сколько пользователей возьмем в наш тест?

Двоих?



Шестерых?



Всех на свете?



**Мы умеем отвечать на этот вопрос при
помощи статистики!**

**Мы умеем отвечать на этот вопрос при
помощи статистики!**

но сначала немного философии



вы сидите в комнате и не слышите посторонних
звуков

Значит ли это, что в комнате полная тишина?



Может быть да

А может посторонний звук
есть, но он такой тихий, что
мы его не слышим

очень
громкий звук

только после этого порога мы
что-то слышим

очень
тихий звук





Ты это слышишь?

очень
громкий звук

Нет, ты о чём?



наш порог слышимости

очень
тихий звук





Ты это слышишь?

Дааа



наш порог слышимости

То, что мы не слышим звук, не
всегда означает, что его нет,
вероятно мы просто не можем его
услышать

очень
громкий звук



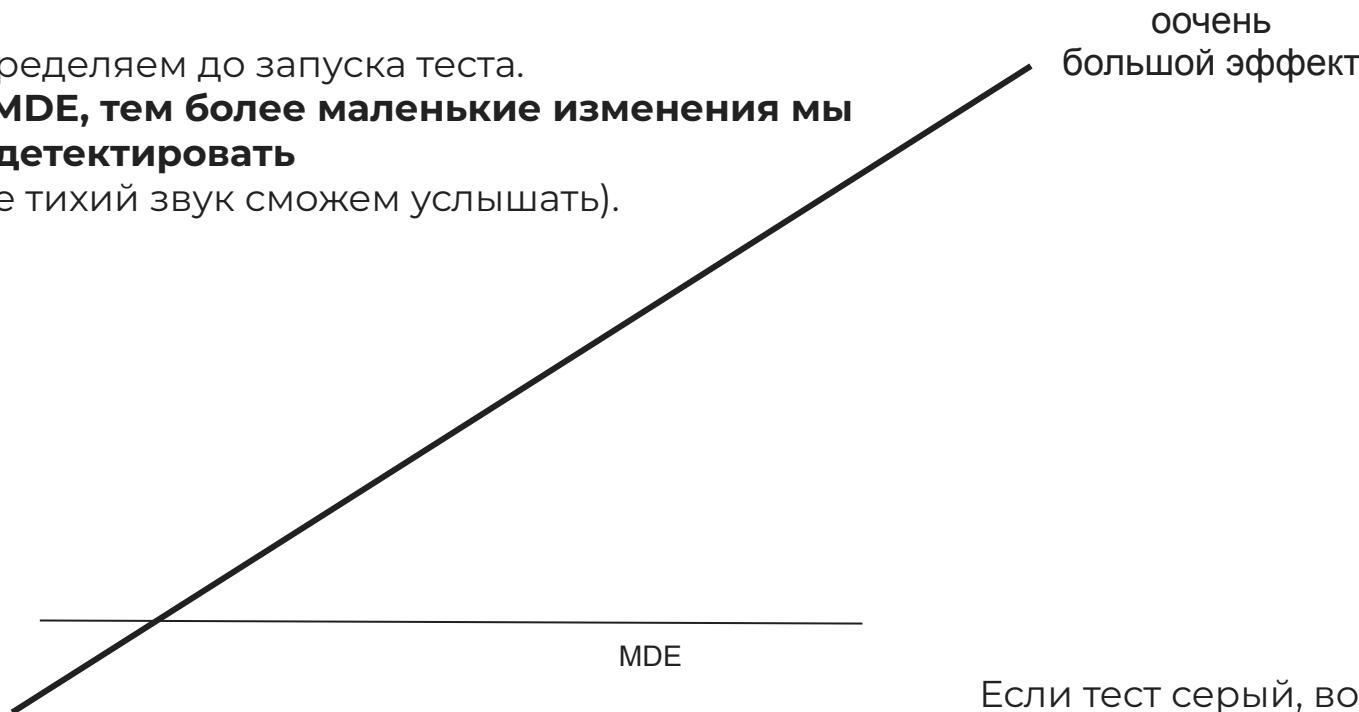
наш порог слышимости

очень
тихий звук

В статистике у нас тоже есть такой порог – MDE

MDE мы определяем до запуска теста.

Чем ниже MDE, тем более маленькие изменения мы сможем задетектировать
(= тем более тихий звук сможем услышать).



очень
маленький эффект

MDE

очень
большой эффект

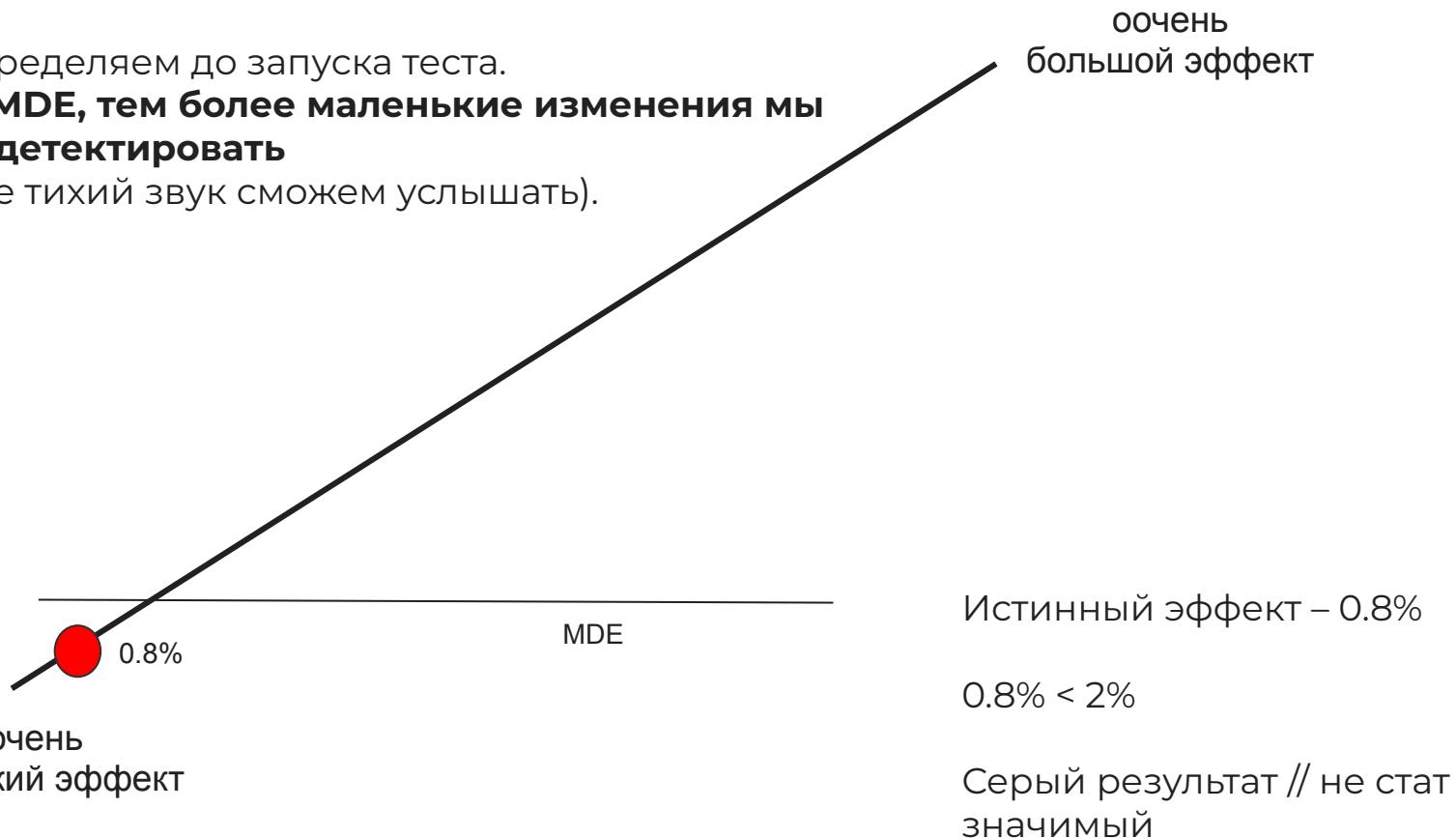
Если тест серый, возможно
эффект есть, но мы не можем
его задетектировать (как очень
тихий звук)

В статистике у нас тоже есть такой порог – MDE

MDE мы определяем до запуска теста.

Чем ниже MDE, тем более маленькие изменения мы сможем задетектировать

(= тем более тихий звук сможем услышать).

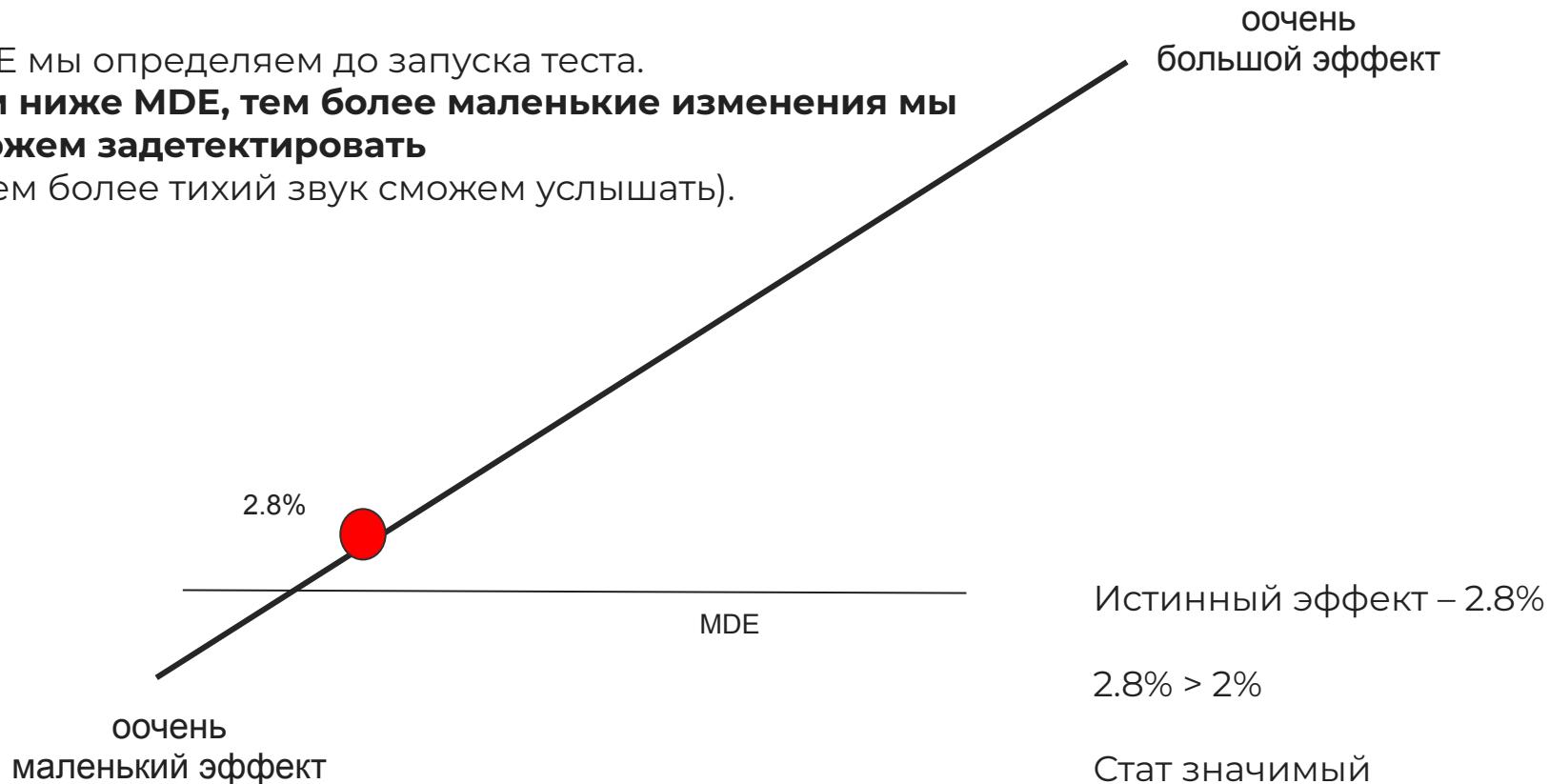


В статистике у нас тоже есть такой порог – MDE

MDE мы определяем до запуска теста.

Чем ниже MDE, тем более маленькие изменения мы сможем задетектировать

(= тем более тихий звук сможем услышать).



MDE – Minimum Detectable Effect

При MDE 4% – мы получим стат значимый результат, если в тесте значение метрики будет 4 и более %

При MDE 4% – мы получим не стат значимый результат, если в тесте значение метрики будет менее 4%

Вопросы!

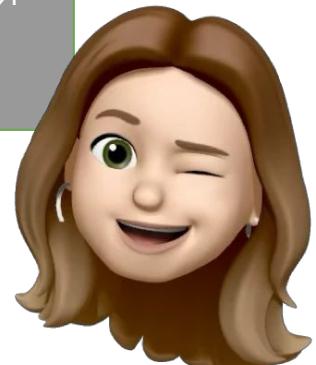


Так давайте брать самый маленький MDE, чтобы
детектить самые маленькие изменения!



Так давайте брать самый маленький MDE, чтобы
детектить самые маленькие изменения!

Чем меньше MDE – тем больше людей
нам понадобится, чтобы принять
решение



А от чего еще зависит MDE?

- **Разброс данных (дисперсия)** — если метрика «скачет» (например, конверсия колеблется от 5% до 20%), MDE будет большим.

А от чего еще зависит MDE?

- **Разброс данных (дисперсия)** — если метрика «скачет» (например, конверсия колеблется от 5% до 20%), MDE будет большим.



А от чего еще зависит MDE?

- **Разброс данных (дисперсия)** — если метрика «скачет» (например, конверсия колеблется от 5% до 20%), MDE будет большим.



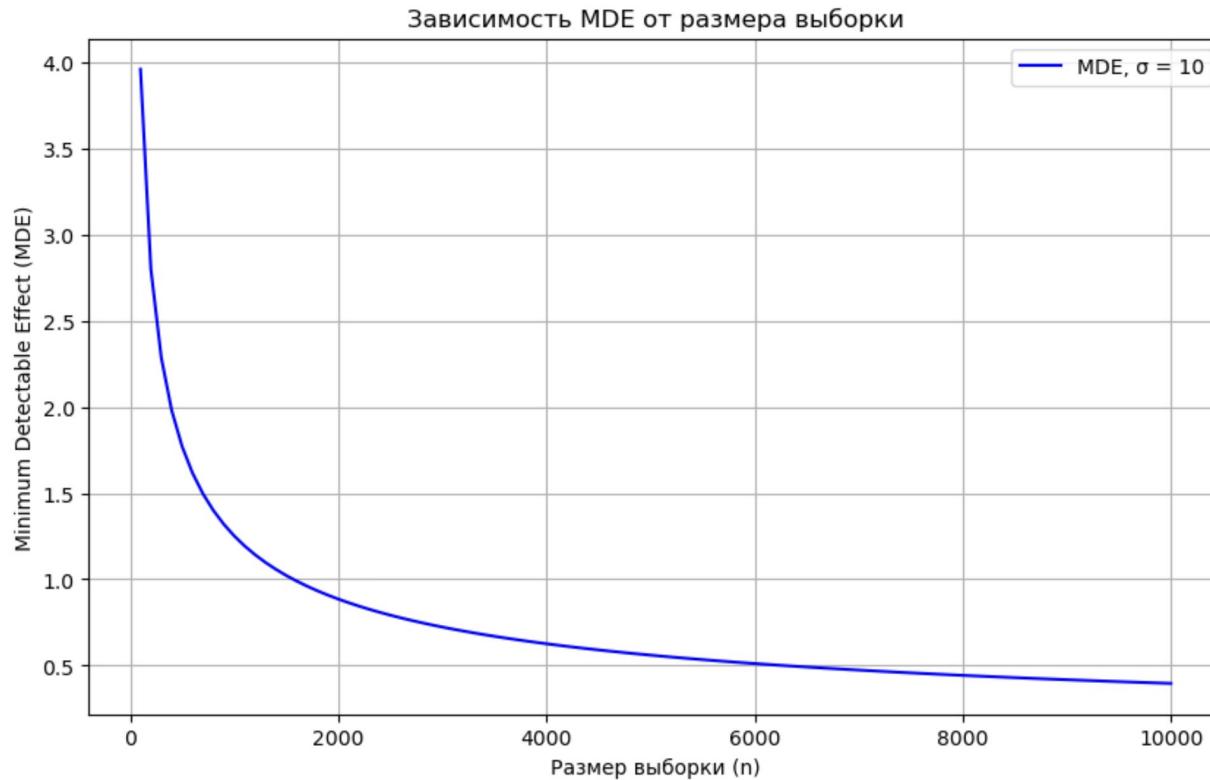
А от чего еще зависит MDE?

- **Разброс данных (дисперсия)** — если метрика «скачет» (например, конверсия колеблется от 5% до 20%), MDE будет большим.
- Размер выборки — чем больше людей в тесте, тем меньший эффект можно заметить.

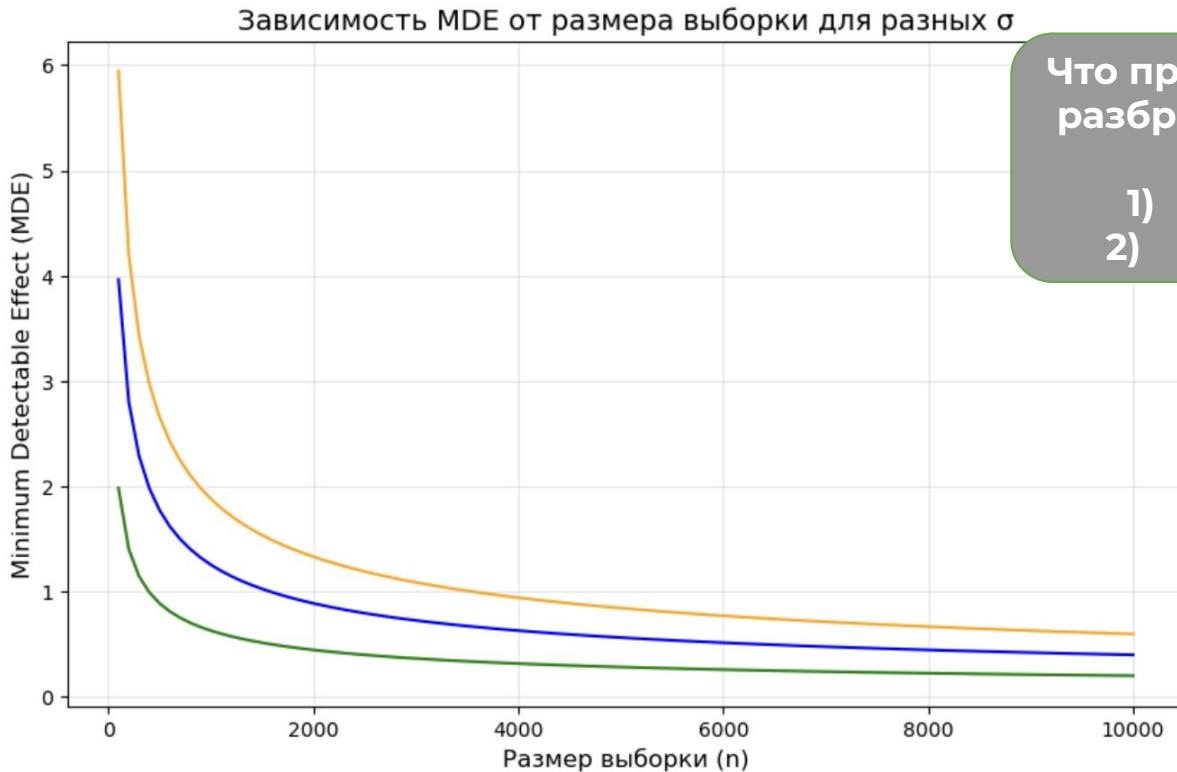
А от чего еще зависит MDE?

- **Разброс данных (дисперсия)** — если метрика «скачет» (например, конверсия колеблется от 5% до 20%), MDE будет большим.
- Размер выборки — чем больше людей в тесте, тем меньший эффект можно заметить.
- Статистические настройки — уровень значимости (α) и мощность ($1 - \beta$).

Посмотрим на графики!



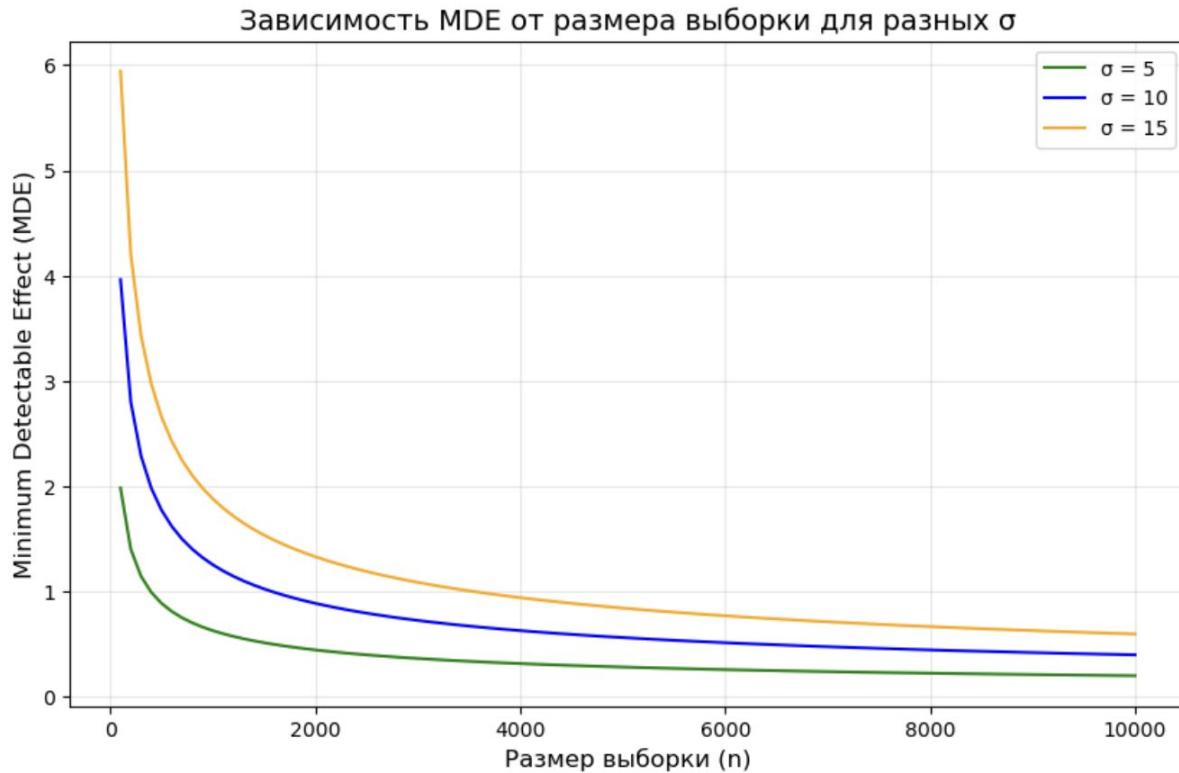
Посмотрим на графики!



Что произойдет с графиком, если разброс данных станет меньше?

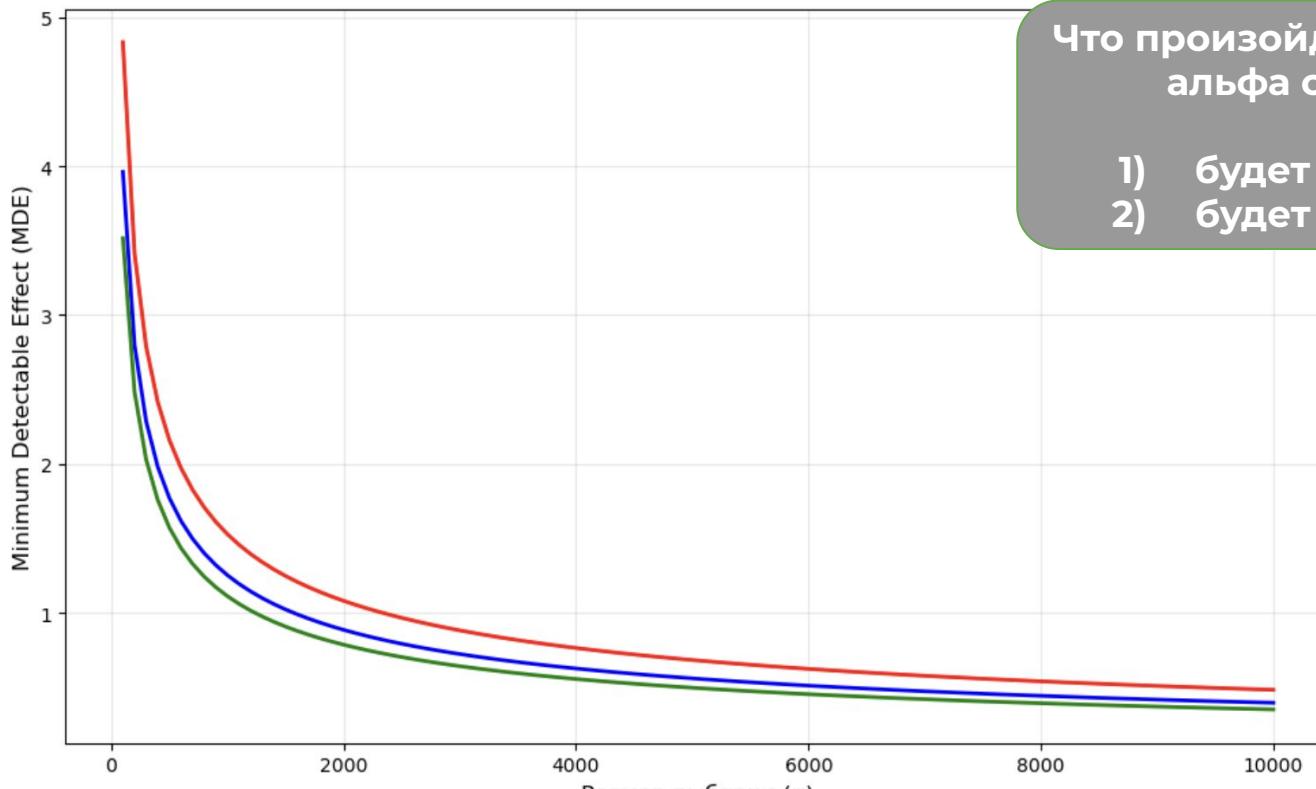
- 1) будет на месте рыжего
- 2) будет на месте зеленого

Посмотрим на графики!



Посмотрим на графики!*

Зависимость MDE от размера выборки при $\sigma = 10$
и разных уровнях значимости (α)

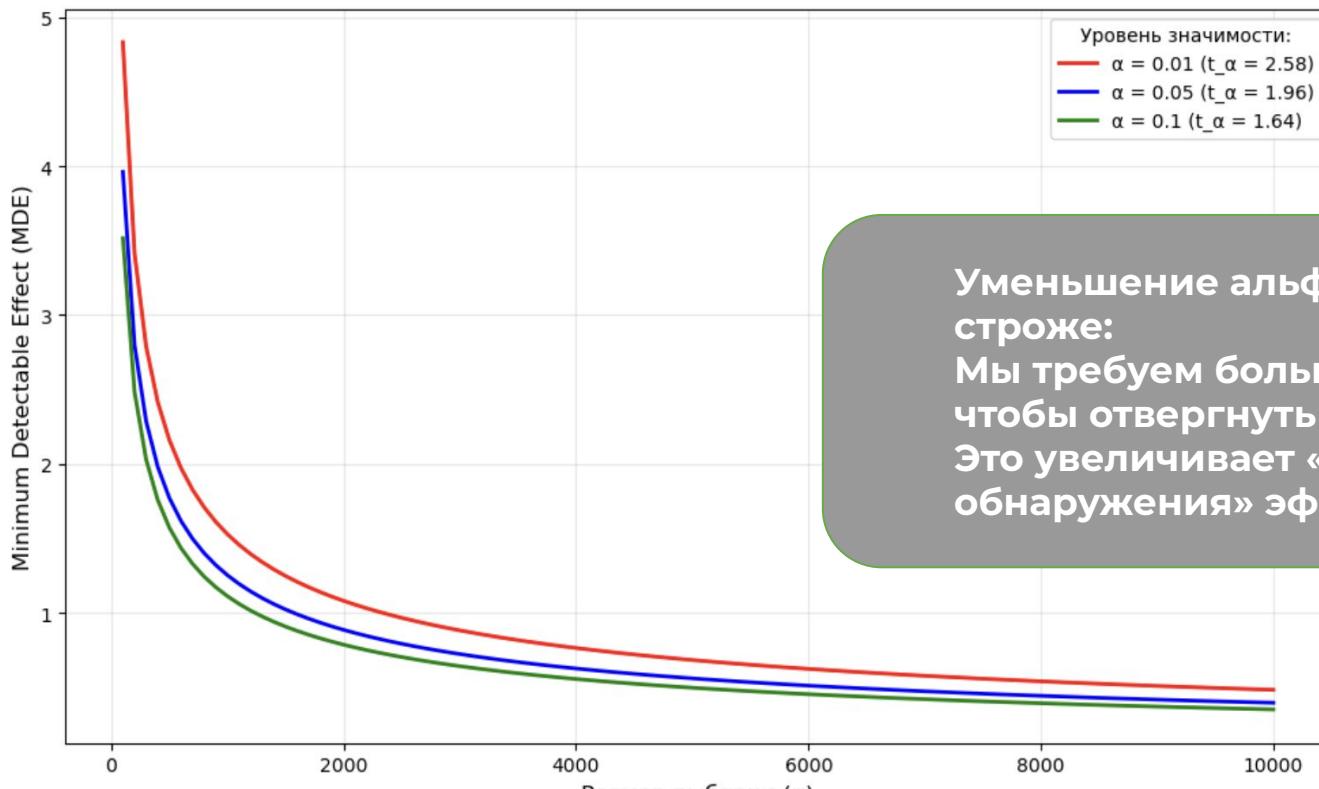


Что произойдет с графиком, если
альфа станет меньше?

- 1) будет на месте красного
- 2) будет на месте зеленого

Посмотрим на графики!*

Зависимость MDE от размера выборки при $\sigma = 10$
и разных уровнях значимости (α)



Уменьшение альфы делает критерий строже:
Мы требуем большей уверенности,
чтобы отвергнуть нулевую гипотезу.
Это увеличивает «порог
обнаружения» эффекта (MDE)

Как посчитать MDE?

Если вы не аналитик:

- попросите аналитика

Если вы аналитик:

- у команды есть код, который переиспользуется, чтобы посчитать MDE

Если вы любите формулы:

$$MDE = (t_{\alpha/2} + t_{\beta}) \cdot \sqrt{\frac{2 \cdot \sigma^2}{n}}$$

Подготовка к запуску АБ-теста

Чеклист, что сделать до запуска АБ:

-  Определить гипотезу и проблему, которую решаем
-  Определить метрики
-  Определить размер выборки / MDE

Подготовка к запуску АБ-теста

Чеклист, что сделать до запуска АБ:

-  Определить гипотезу и проблему, которую решаем
-  Определить метрики
-  Определить размер выборки / MDE

Что стоит еще сделать перед запуском теста?

Подготовка к запуску АБ-теста

Чеклист, что сделать до запуска АБ:

-  Определить гипотезу и проблему, которую решаем
-  Определить метрики
-  Определить размер выборки / MDE
-  Проверить, что смежные команды готовы к запуску (разработка, дизайн)
-  Проверить, что настроен треккинг
-  Предупредить всех стейкхолдеров о запуске

запускаем тест





Часть 4. Подводим итоги

Настройка мониторинга

А идеале делать до теста, но и сразу после запуска есть свои плюсы:

- будут данные
- Будет понятно, какие графика самые важные

Зачем нужен мониторинг?

- Вовремя отследить, если тест запустила не как мы планировали (баги)
- Вовремя отследить, если тест идет сильно хуже, чем мы планировали и принять действия
- Отвечать на вопросы стейкхолдеров, как дела у теста

Parameters

is_revenue	avito_version	presence_type	product_subtype	logical_category	wave	Начало периода
false	ALL × +5 more	ALL × +5 more	performanc... ×	ALL × +44 more	ALL × +9 more	2023-06-19 ⏺

Конец периода

2023-09-01



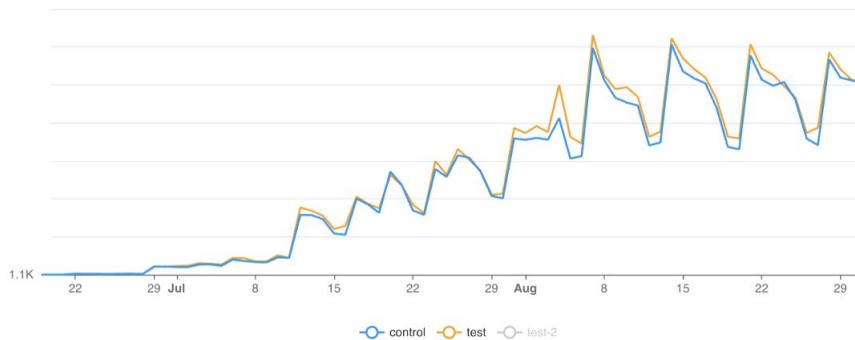
Карточка эксперимента в cf

- При is_revenue = true, все графики будут строиться по write-off. write-offs - признанная выручка VAS, распределенная по дням действия услуги продвижения
- При is_revenue = false, все графики будут строиться по reserves. reserves - метрика по покупкам, резервирование суммы, не является признанной выручкой

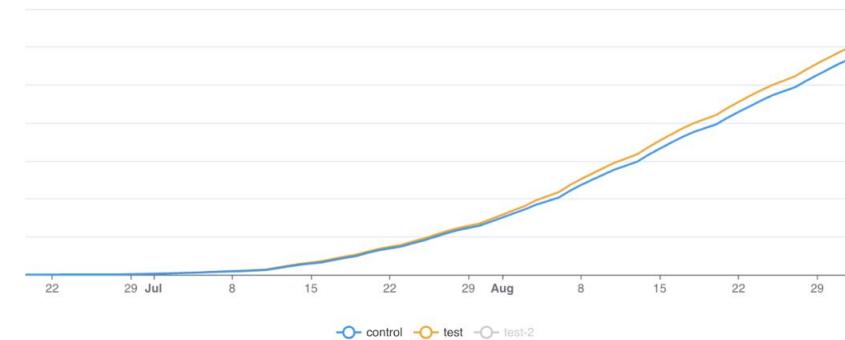
Графики:

- Выручка (amount_net) -- прибыль от покупок VAS
- Выручка (amount_net) накопительная -- суммарная прибыль от покупок VAS. Последний день показывает суммарное значение за все время теста.
- Платящие пользователи (paying_users) -- количество уникальных пользователей, купивших VAS
- Платящие пользователи (paying_users) накопительные -- Количество уникальных пользователей, купивших VAS за все время теста
- Транзакции -- количество VAS транзакций
- Транзакции накопительные-- количество VAS транзакций за все время
- ARPPU = sum(amount_net)/sum(paying_users)
- ARPU = sum(amount_net)/sum(active_users)
- PPU = sum(paying_users) / sum(active_users). Проникновение: процент пользователей, купивших VAS от всех пользователей в тесте. Расчитывается по дням
- TPPU = sum(transactions)/sum(paying_users)
- TPU = sum(transactions)/sum(active_users)
- AOV = sum(amount_net)/sum(transactions). Средний чек

Выручка (amount_net)



Выручка (amount_net) накопительная



MLP1 absolute uplift, ₽

5%

MLP1 relative uplift

В тесты можно подглядывать?



Можно и нужно следить за тем, как идет эксп. При этом – не вмешиваться в него



Мы не можем остановить тест, как только результаты будут нам «подходить»

В тесты можно подглядывать?



Можно и нужно следить за тем, как идет эксп. При этом – не вмешиваться в него



Мы не можем остановить тест, как только результаты будут нам «подходить»

Пример 1:

Мы увеличили цены на площадке: изначально вероятно мы увидим падение основных метрик, но со временем пользователи вернутся

В тесты можно подглядывать?



Можно и нужно следить за тем, как идет эксп. При этом – не вмешиваться в него



Мы не можем остановить тест, как только результаты будут нам «подходить»

Пример 1:

Мы увеличили цены на площадке: изначально вероятно мы увидим падение основных метрик, но со временем пользователи вернутся

Пример 2:

Мы добавили новый функционал: новый продукт продвижения, сначала люди попробуют что-то новое, но если продукт не оправдает их ожидания, мы увидим отток со временем

**THREE
WEEKS LATER**



Мы получили результат!

Теперь у нас есть данные по пользователям в двух группах:

1. В контрольной группе, где не было изменений **5600** пользователей из 98к совершили сделку
2. Во второй, где были изменения, **5790** пользователей из 98к совершили сделку

Мы получили результат!

Теперь у нас есть данные по пользователям в двух группах:

1. В контрольной группе, где не было изменений **5600** пользователей из 98к совершили сделку
2. Во второй, где были изменения, **5790** пользователей из 98к совершили сделку

Надо бы его интерпретировать, но как?

Для тестов с CR: калькулятор

AB Testguide

Is your test result significant? Does it have enough power?

Play with the controls and get a better feel for how a lower confidence level will boost the power or how an increase in test size can make a small CR-difference significant!

Pre-test calculation or post-test evaluation?

Pre-test analysis

Test evaluation

Test data

Visitors A	Conversions A
98000	5600
Visitors B	Conversions B
98000	5790

Apply changes

Settings

Hypothesis (?)

One-sided

Two-sided

Test result

save & share url

Significant test result!

Variation B's observed conversion rate (5.91%) was 3.39% higher than variation A's conversion rate (5.71%). You can be 95% confident that this result is a consequence of the changes you made and not a result of random chance.

A 5.71%

B 5.91%

The expected distributions of variation A and B.

Conversion Rate Control
Conversions A / Visitors A
5.71%

Conversion Rate B
Conversions B / Visitors B
5.91%

Relative uplift in Conversion Rate
 $CR_B - CR_A / CR_A$
3.39%

Для тестов с CR: калькулятор

Pre-test calculation or post-test evaluation?

Pre-test analysis

Test evaluation

Test data

Visitors A	Conversions A
98000	5600
Visitors B	Conversions B
98000	5790

Apply changes

Settings

Hypothesis (?)

One-sided

Two-sided

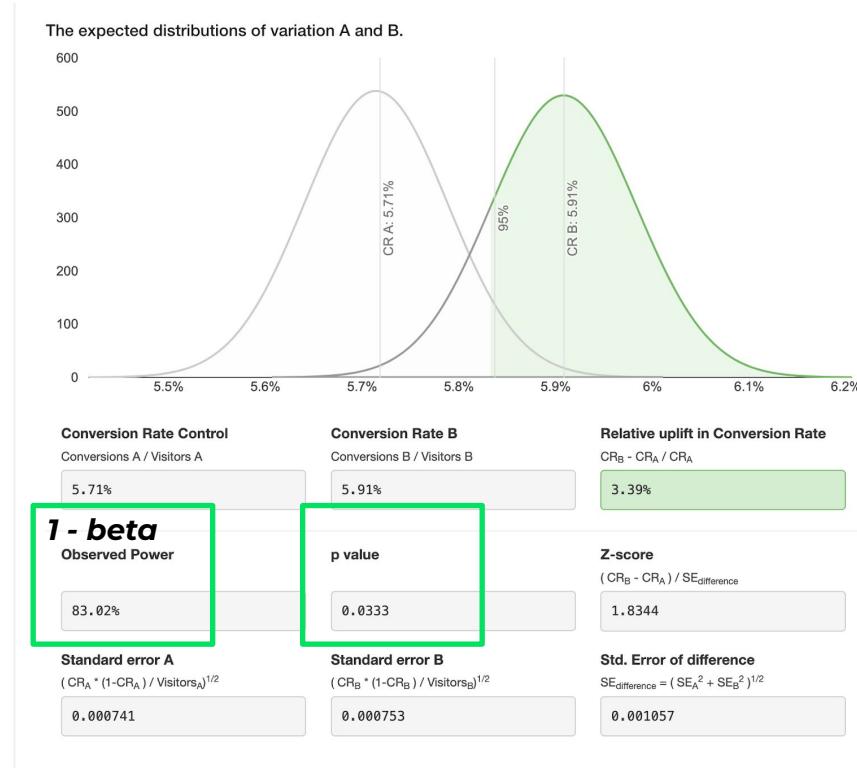
Confidence (?)

90%

95%

99%

alpha



Ошибки 1/2 рода: Гипотезы



H₀: Это мороженое

H₁: Это не мороженое

Критерий: будет вкусно

Тест: Было вкусно?

Результат: Нет

Итог: Отвергаем нулевую гипотезу, что это мороженое, принимаем альтернативную

Ошибки 1/2 рода: Гипотезы



H₀: Это мороженое

H₁: Это не мороженое

Критерий: будет вкусно

Тест: Было вкусно?

Результат: Нет

Итог: Отвергаем нулевую гипотезу, что это мороженое, принимаем альтернативную

H₀ – нулевая гипотеза, контроль

H₁ – альтернативная гипотеза, вариация

Ошибки 1/2 рода

Когда мы проводим тесты со стороны математики мы балансируем две ошибки.

Первая:

Ошибка совершил ложное
открытие = α



Масштабировали
результаты АБ теста, а на
самом деле контроль был
лучше

Вторая:

Ошибка пропустить
открытие = β



Не приняли результаты АБ
теста, хотя вариация была
лучше

Ошибки 1/2 рода

Когда мы проводим тесты со стороны математики мы балансируем две ошибки.

H₀ – нулевая гипотеза, контроль

H₁ – альтернативная гипотеза, вариация

α – Вероятность принять альтернативную гипотезу, когда она хуже нулевой. Ложное открытие

β – вероятность отклонить альтернативную гипотезу, когда она лучше нулевой. Пропуск открытие

$1 - \beta$ – мощность. Вероятность принять альтернативную гипотезу, когда она лучше нулевой. Вероятность не пропустить открытие

	H_0 отвергается, принимается H_1	H_0 принимается
H_0 верна	Ошибка 1-го рода, её вероятность $P_{H_0}(H_1) = \alpha$	Правильное решение, его вероятность $P_{H_0}(H_0) = 1 - \alpha$
H_0 не верна, т.е. верна H_1	Правильное решение, его вероятность $P_{H_1}(H_1) = 1 - \beta$	Ошибка 2-го рода, её вероятность $P_{H_1}(H_0) = \beta$

Ошибки 1/2 рода

Когда мы проводим тесты со стороны математики мы балансируем две ошибки.

H₀ – нулевая гипотеза, контроль

H₁ – альтернативная гипотеза, вариация

α – вероятность принять альтернативную гипотезу, когда она хуже нулевой. Ложное открытие

β – вероятность отклонить альтернативную гипотезу, когда она лучше нулевой. Пропуск открытие

$1 - \beta$ – мощность. Вероятность принять альтернативную гипотезу, когда она лучше нулевой. Вероятность не пропустить открытие

H_1 : есть беременность; H_0 : нет беременности

Истинный
позитив, верна
 H_1



Ложный
позитив,
ошибка I
рода,
ложная
тревога



Истинный
негатив,
верна H_0

Ошибки 1/2 рода

Когда мы проводим тесты со стороны математики мы балансируем две ошибки.

Н₀ – нулевая гипотеза, контроль

Н₁ – альтернативная гипотеза, вариация

α – вероятность принять альтернативную гипотезу, когда она хуже нулевой. Ложное открытие

β – вероятность отклонить альтернативную гипотезу, когда она лучше нулевой. Пропуск открытие

$1 - \beta$ – мощность. Вероятность принять альтернативную гипотезу, когда она лучше нулевой. Вероятность не пропустить открытие

P-value – вероятность получить значение, которое мы получили в экспе, если на самом деле верна Н₀

H_1 : есть беременность; H_0 : нет беременности

Истинный
позитив, верна
 H_1



Ложный
позитив,
ошибка I
рода,
ложная
тревога

Ложный
негатив,
ошибка II рода,
халатная
беспечность



Истинный
негатив,
верна H_0

Ошибки 1/2 рода



Про p-value непонятно, можно на примере?

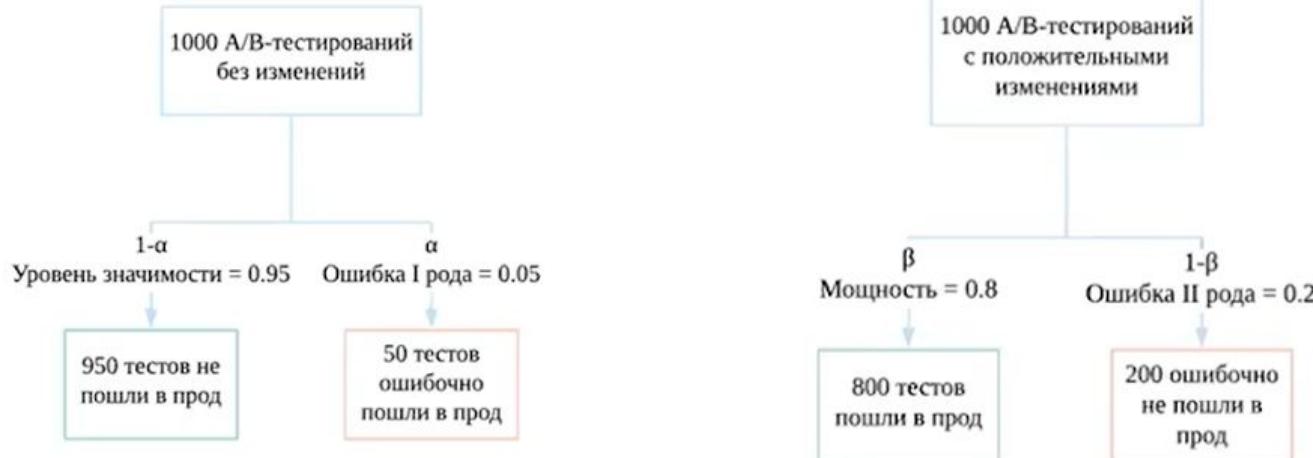
Да! Если мы провели эксперимент и его p-value 0.03 – это значит, что если мы принимаем альтернативную гипотезу (масштабируем вариацию), то будет шанс 3%, получить такие результаты, которые мы увидели, если верна H_0



Как выбрать параметры α и β ?

Стандарт индустрии: $\alpha = 5\%$, мощность = 80%

- 💡 Если выборка позволяет: уменьшают α до 1% или 0.5%
- 💡 Сделать хуже опаснее, чем пропустить возможность сделать лучше



Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

1. При включении пожарной тревоги?

Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

- При включении пожарной тревоги?

Ошибка первого рода (ложная тревога):

Сигнализация срабатывает, хотя пожара нет. Это приводит к эвакуации, панике и потере времени.

Ошибка второго рода (пропущенный пожар):

Сигнализация не срабатывает, хотя пожар есть. Это может привести к серьезным последствиям, включая материальный ущерб и угрозу жизни.

.

Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

1. При проведении медицинского теста?

Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

- При проведении медицинского теста?

Ошибка первого рода (ложноположительный результат): Тест показывает наличие заболевания, хотя пациент здоров. Это может привести к ненужному лечению, стрессу и тревоге.

Ошибка второго рода (ложноотрицательный результат): Тест показывает отсутствие заболевания, хотя пациент болен. Это может привести к задержке лечения и ухудшению состояния пациента.

Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

1. Отправка письма в спам?

Как выбрать параметры α и β ?

Какую ошибку **хуже допустить при:**

1. Отправка письма в спам?

Ошибка первого рода (ложное срабатывание):

Фильтр классифицирует важное письмо как спам. Пользователь может пропустить важную информацию.

Ошибка второго рода (пропущенный спам):

Фильтр пропускает спам в папку "Входящие". Пользователь может быть подвержен фишингу, вирусам или другим угрозам.

Для тестов с CR: калькулятор

Pre-test calculation or post-test evaluation?

Pre-test analysis

Test evaluation

Test data

Visitors A	Conversions A
98000	5600
Visitors B	Conversions B
98000	5790

Apply changes

Settings

Hypothesis (?)

One-sided

Two-sided

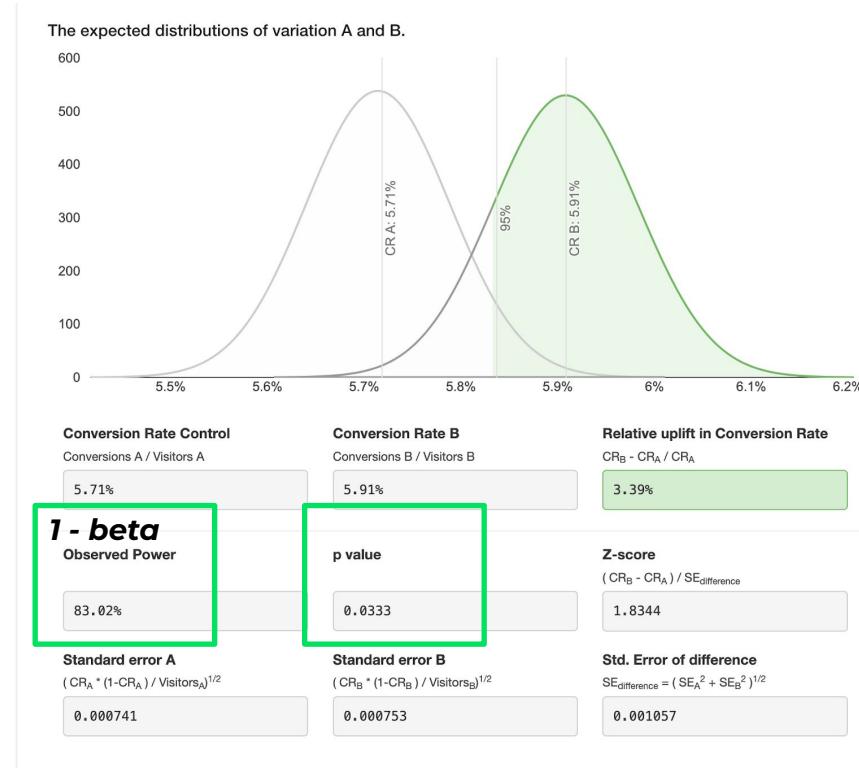
Confidence (?)

90%

95%

99%

alpha



Выбор стат критерия

Критерии помогают нам в тестах, когда мы смотрим не на конверсию, а например на средний чек или количество айтемов. Самые распространенные:

- **t-критерий Стьюдента** – если выборка распределена нормально (параметрические критерий)
- **U критерий Манна-Уитни** – если выборка не распределена нормально (непараметрический критерий)

```
from scipy.stats import ttest_ind, norm, expon
import numpy as np
np.random.seed(42)
X = expon(scale=1100).rvs(1000)
Y = norm(loc=980, scale=30).rvs(1000)
ttest_ind(X, Y, equal_var=False, alternative='two-sided')
Ttest_indResult(statistic=2.5645688722251325, pvalue=0.010475352713690184)
```

Принятие решения

$(P\text{-value} < \alpha)$
и набрана достаточная выборка?



Есть стат значимый результат!

Подведение итогов

Красный

Серый

Зеленый

Подведение итогов

Красный

- Еще раз проверяем, что тест работал корректно
- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Проводим UX
- Собираем знания и перезапускаем АБ или отвергаем гипотезу

Серый

Зеленый

Подведение итогов

Красный

- Еще раз проверяем, что тест работал корректно
- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Проводим UX
- Собираем знания и перезапускаем АБ или отвергаем гипотезу

Серый

- Еще раз проверяем, что тест работал корректно
- Проверяем, что держали достаточно дней
- Смотрим, в какую сторону сторону тест серый
- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Собираем знания, принимаем решения
- Иногда, мы и хотели получить серый тест
-> масштабируем

Зеленый

Подведение итогов

Красный

- Еще раз проверяем, что тест работал корректно
- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Проводим UX
- Собираем знания и перезапускаем АБ или отвергаем гипотезу

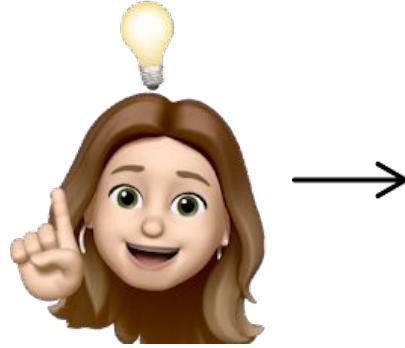
Серый

- Еще раз проверяем, что тест работал корректно
- Проверяем, что держали достаточно дней
- Смотрим, в какую сторону сторону тест серый
- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Собираем знания, принимаем решения
- Иногда, мы и хотели получить серый тест
-> масштабируем

Зеленый

- Исследуем разные сегменты
- Исследуем метрики на разных этапах
- Принимаем решение о масштабировании

Цикл АБ-теста





Часть 5.

Виды экспериментов

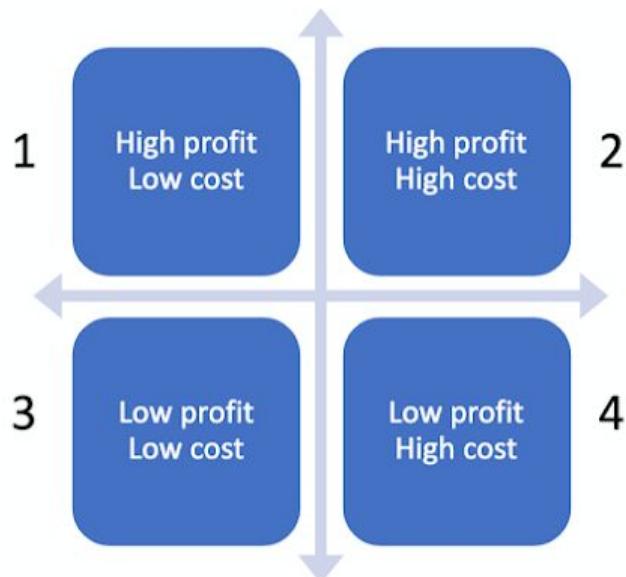
А давайте запустим сразу несколько тестов?

Можно, если:

1. Сегменты, на которых проходит тест не пересекаются
2. Заводить на тест не 2 вариации, а 2^{λ} (кол-во проверяемых гипотез)

Во всех остальных случаях мы не поймем какая именно гипотеза повлияла на результат

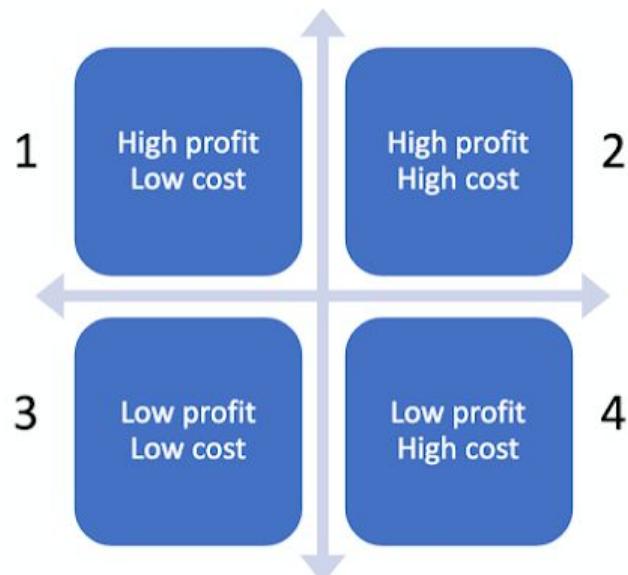
А как тогда приоритизировать?



В первую очередь запускаем тесты, которые могут принести нам большой профит, и которые просто реализовать.

В последнюю – запускаем тесты, которые скорее всего принести нам небольшой профит, и которые будет дорого реализовать.

А как тогда приоритизировать?



В первую очередь запускаем тесты, которые могут принести нам большой профит, и которые просто реализовать.

В последнюю – запускаем тесты, которые скорее всего принести нам небольшой профит, и которые будет дорого реализовать.

Дорогой тест – это какой?

Тест, для которого, нужно много времени разработчиков/дизайнеров/контента/аналитики/других команд

Какие тесты еще бывают?

АА-тесты

A/A – то же самое, что A/B, только контроль и вариация не отличаются

Цель – проверить, что сплит системы проведения АБ-тестов работают корректно и распределяют пользователей между вариациями равномерно. Если в конце эксперимента показатели конверсии обеих страниц совпадают, можно запускать А/В тест.

Проводим, когда появилась новая система сплита для АБ-теста или начинаем работать с новой метрикой и хотим убедиться, что результатам АБ можно доверять

Какие тесты еще бывают?

ABC-тесты

A/B/C – тест, в котором две тестовые вариации и контроль.

Цель – проверить сразу несколько вариантов.

Важно: для определения результата теста с несколькими вариациями, важно помнить про ошибку первого рода, которая возрастает. Чтобы с ней справиться, нам будет нужна поправка Бонферрони.

Для этого $p\text{-value}$ должно быть меньше альфа/ m , где m – кол-во вариаций

 [Статья](#) про A/B/n тестирования

Какие тесты еще бывают?

Ухудшающие тесты

Умышленное отключение/ухудшение функционала, чтобы узнать пользу от него – когда сломать проще, чем сделать.

Пример: мы хотим инвестировать время в то, чтобы сделать сайт быстрее на 0.1 сек, чтобы оценить, возможный импакт от задачи, мы можем искусственно замедлить сайт на 0.1 сек и проверить, ухудшатся ли метрики

Часть 6.

Кейсы

Кейс 1

Контроль

Услуги продвижения

 Продвижение
Чем больше бюджет, тем больше получите просмотров

1680 ₽

Дневной бюджет 240 ₽
 120 ₽ 1678 ₽

Срок 7 дней
 1 30

Ваш выбор

 Игра настольная Время
Валеры




Прирост просмотров ~ 19-477 Итого за 3 услуги 2 010 ₽

Вариация

Заметный внешний вид

 Выделить цену цветом на 7 дней
Привлечёт внимание к цене объявления

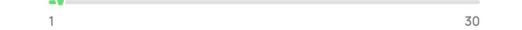
180 ₽

Услуги продвижения

 Продвижение
Чем больше бюджет, тем больше получите просмотров

120 ₽

Дневной бюджет 120 ₽
 120 ₽ 1678 ₽

Срок 1 день
 1 30

Ваш выбор

 Игра настольная Время
Валеры




Прирост просмотров ~ 0-36 Итого за 3 услуги 450 ₽

Заметный внешний вид

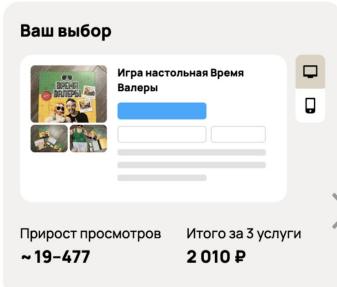
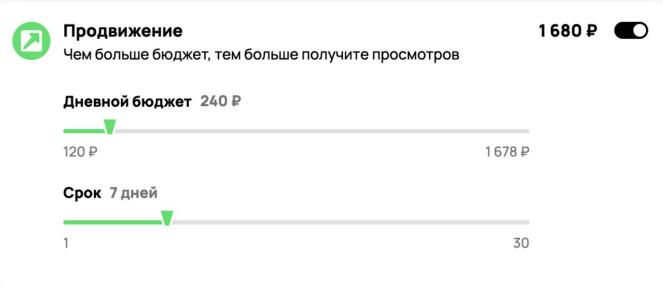
 Выделить цену цветом на 7 дней

180 ₽

Кейс 1

Контроль

Услуги продвижения



Вариация

Заметный внешний вид



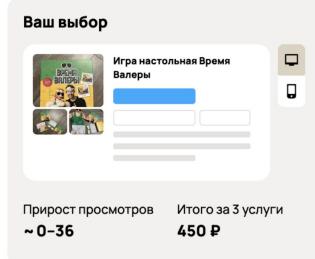
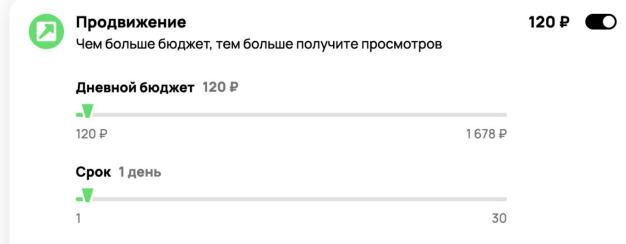
Результат:

Paying Users: +10%

AOV: -15%

Revenue: -5%

Услуги продвижения

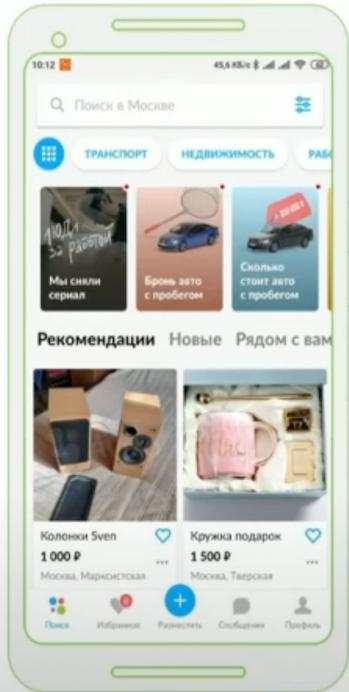


Заметный внешний вид



Кейс 2

Контроль



Вариация 1

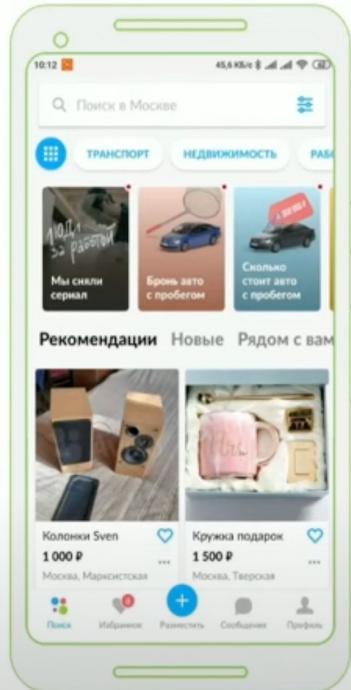


Вариация 2



Кейс 2

Контроль



Вариация 1

Buyers с вкладки: +2%



Вариация 2

Buyers с вкладки: -2%



Кейс 3

Обратный эксп, в котором скрыли блок «Похожие объявления»

и когтеточка на 5+. Приезжайте знакомиться. Смотрите наши объявления, у нас есть и белые котята. Могу отправить в другой город.

№ 3911006222 · сегодня в 06:56 · 48 просмотров (+21 сегодня)

Пожаловаться

15 000 ₽

... ❤

или [предложите свою цену](#)

Отвечает около часа

Позвонить через Авито
8 958 XXX-XX-XX

Написать сообщение

Похожие объявления



Британские плюшевые

котята

15 000 ₽

Москва, б-р Маршала Рокоссовского, Бульвар Рокоссовского 20 апреля 10:26



Британские котята

Цена не указана

Москва, пр. Воскресенские Ворота, Охотный ряд 21 апреля 15:22



Котята снежные

10 000 ₽

Москва, Южнобутовская ул., 2, Бульвар адмирала Ушакова 19 апреля 11:58



Клубные британские

котята

Цена не указана

Москва, Северо-Восточный административный округ, район Марьина Роща, жилой комплекс Марьина Роща, Марьина Роща 3 апреля 23:59



Британские котята

10 000 ₽

Москва, Серпуховско-Тимирязевская линия, Тульская 16 апреля 10:54



Котёнок

8 500 ₽

Москва, Центральный административный округ, Пресненский район, Московский международный деловой центр Москва-Сити, Выставочная 12 апреля 07:36

H

Наталья

4.8 ★★★★☆ 21 отзыв

Частное лицо

На Авито с января 2017

3 объявления

Подписаться

Эковклад: -223 кг CO₂

Спросите у продавца

Здравствуйте!

Когда можно посмотреть?

Ещё продаёт?

Позвоните мне?

Торг уместен?

Пришлите видео?

Кейс 3

Обратный эксп, в котором скрыли блок «Похожие объявления»

и когтеточка на 5+. Приезжайте знакомиться. Смотрите наши объявления, у нас есть и белые котята. Могу отправить в другой город.

№ 3911006222 · сегодня в 06:56 · 48 просмотров (+21 сегодня)

Пожаловаться

Похожие объявления



Британские плюшевые котята

15 000 ₽
Москва, б-р Маршала Рокоссовского, Бульвар Рокоссовского
20 апреля 10:26



Британские котята

Цена не указана
Москва, пр. Воскресенские Ворота, Охотный ряд
21 апреля 15:22



Котята снежные

10 000 ₽
Москва, Южнобутовская ул., 2, Бульвар адмирала Ушакова
19 апреля 11:58



Клубные британские котята

Цена не указана
Москва, Северо-Восточный административный округ, район Марьина Роща, жилой комплекс Марьина Роща, Марьина Роща
3 апреля 23:39



Британские котята

10 000 ₽
Москва, Серпуховско-Тимирязевская линия, Тульская
16 апреля 10:54



Котёнок

8 500 ₽
Москва, Центральный административный округ, Пресненский район, Московский международный деловой центр Москва-Сити, Выставочная
12 апреля 07:36

15 000 ₽

... ❤

или предложите свою цену

Отвечает около часа

Результат:

Buyers: +2%

Позвонить через Авито
8 958 XXX-XX-XX

Написать сообщение

H

Наталья

4.8 ★★★★★ 21 отзыв

Частное лицо
На Авито с января 2017
3 объявления
Подписаться

ЭкоКлад: -223 кг CO₂

Спросите у продавца

Здравствуйте!

Когда можно посмотреть?

Ещё продаёт?

Позвоните мне?

Торг уместен?

Пришлите видео?

**Что может
пойти не так?**

Что может пойти не так? Кейс 1

Проводили обратный тест.

Планировали не увидеть изменения в продукте.

Сделали мониторинги.

Запустили тест.

Ничего не увидели.

Что может пойти не так?

Что может пойти не так? Кейс 1

Проводили обратный тест.

Планировали не увидеть изменения в продукте.

Сделали мониторинги.

Запустили тест.

Ничего не увидели.

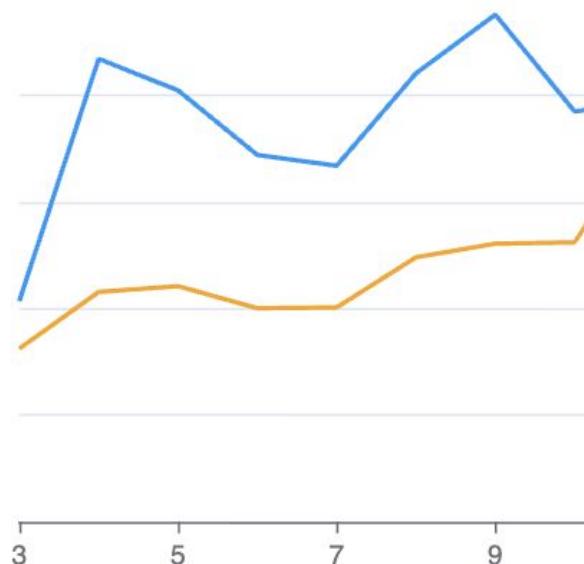
Что может пойти не так?

Тест не запустился, поэтому и не было изменений

Что может пойти не так? Кейс 2

Запустили тест, который уронил выручку
в два раза :)

Что пошло не так?

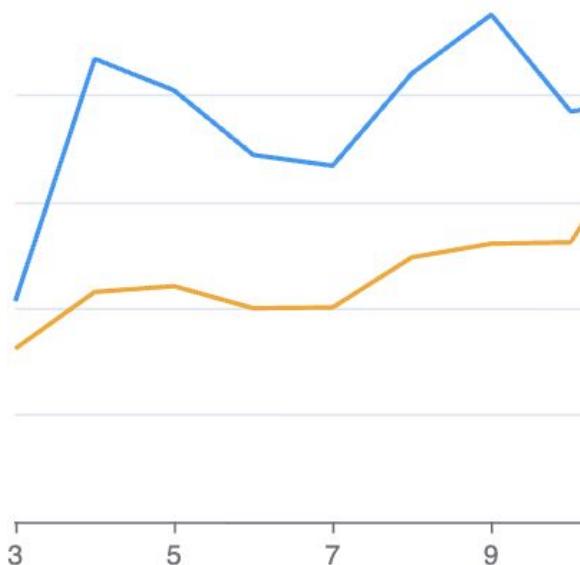


Что может пойти не так? Кейс 2

Запустили тест, который уронил выручку
в два раза :)

Что пошло не так?

Запуск теста сломал разметку сегмента,
на котором он запускался, часть тестовой
аудитории оказалась в другом сегменте.
На самом деле, выручка не пострадала



Что может пойти не так? Кейс 3

Чтобы увидеть изменения, нужно обновить версию iOS. В тесте стали предлагать обновиться и считать, что тестовые пользователи только те, кто обновился. В контроле была вся контрольная группа.

Что может пойти не так?

Что может пойти не так? Кейс 3

Чтобы увидеть изменения, нужно обновить версию iOS. В тесте стали предлагать обновиться и считать, что тестовые пользователи только те, кто обновился. В контроле была вся контрольная группа.

Что может пойти не так?

Шли обновляться уже более лояльные люди, они были более открыты к нашим новинкам, чем общая совокупность. Так, результаты в тесте были сильно завышены

Что может пойти не так? Кейс 4

Собрали в 6 вечера костыль, чтобы запустить тест и не откладывать его. В 8 вечера на телефон звонят, что мы сломали сайт :)

Что пошло не так?

Что может пойти не так? Кейс 4

Собрали в 6 вечера костыль, чтобы запустить тест и не откладывать его. В 8 вечера на телефон звонят, что мы сломали сайт :)

Что пошло не так?

Костыль не костылил)) Мы никого не предупредили, ну и запускать тесты вечером – к работе ночью

Что может пойти не так? Кейс 5

Другая команда также запустила АБ-тест на наш контроль. На вариацию не запустила.

Что пошло не так?

Что может пойти не так? Кейс 5

Другая команда также запустила АБ-тест на наш контроль. На вариацию не запустила.

Что пошло не так?

Тест оказался успешным и контроль был сильно лучше на фоне тестовой категории, тк 50% от контроля начало приносить на 20% больше выручки

Что может пойти не так? Кейс 6

Запустили региональный тест, при анализе получили огромный плюс метрики. Через год эффекта не обнаружилось.

Что пошло не так?

Что может пойти не так? Кейс 6

Запустили региональный тест, при анализе получили огромный плюс метрики. Через год эффекта не обнаружилось.

Что пошло не так?

Эффект связан с сезонностью, а не нашими изменениями. Каждый апрель выручка растет и уменьшается в августе

Что может пойти не так? Кейс 7

Включили эксперимент с новым интерфейсом на тест и контроль. Тест не сработал, мы сделали доработки и заново запустили тест. Получили кучу негатива

Что пошло не так?

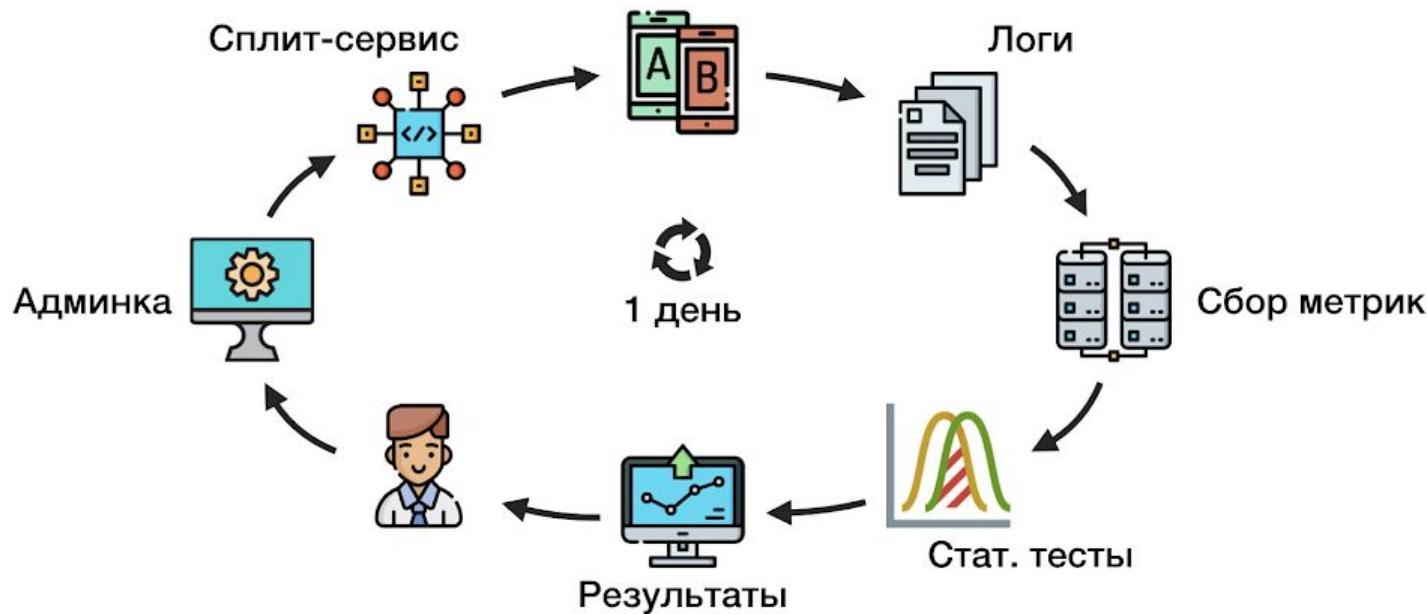
Что может пойти не так? Кейс 7

Включили эксперимент с новым интерфейсом на тест и контроль. Тест не сработал, мы заново запустили тест. Получили кучу негатива

Что пошло не так?

Часть пользователей, у кого был тест, стал контроль (пропал функционал продукта). Часть пользователей уже привыкла к изменениям и не заметила наших доработок.

А как запустить тест?

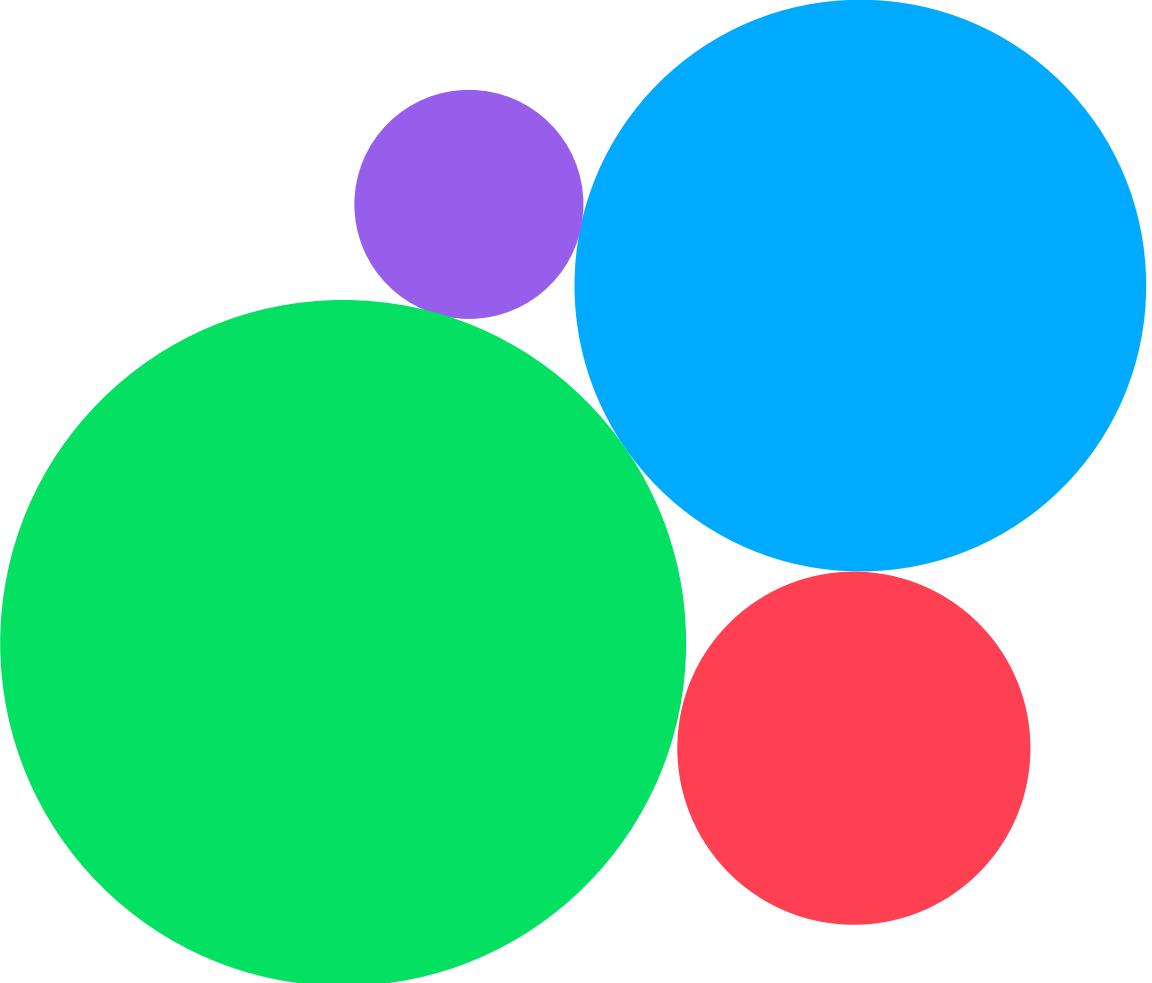


А как запустить тест?

У Авито есть свой in-house инструмент -- TriSigma

А если у компании нет in-house инструмента?

- Поискать открытые решения, есть у Google и Яндекс
- Подумать, точно ли маленькой компании нужны АБ-тесты и хватит ли размера выборки



**Спасибо
за
внимание**

