

# АБ-тесты

# План

- Схема АБ-тестирования
- Проблемы, возникающие при проведении АБ-тестов и способы их решить
- АБ тесты в офлайне, естественные эксперименты
- Множественная проверка гипотез
- Как понять, сколько нужно наблюдений для проведения эксперимента
- Метрики для АБ-тестирования

# **Схема АБ-тестирования**

# Процедура АБ-тестирования



► <https://research.yandex.com/tutorials/online-evaluation/sigir-2019>

# Процедура АБ-тестирования

$\bar{x}_A$  

Статистика для  
группы А

Вычисляем  
критерий для  
оценки

$\bar{x}_B$  

Статистика для  
группы В

$$\rightarrow \Delta X = \bar{x}_B - \bar{x}_A$$

## Существенность

$$\Delta X \text{ vs } 0$$

позитивные  
или  
негативные  
изменения

## Значимость

Статистический  
тест

Принимаем  
решение

разница  
вызвана  
**шумом** или  
изменениями

► <https://research.yandex.com/tutorials/online-evaluation/sigir-2019>

# Типичные метрики

- Уникальные пользователи за сессию
- Клики на пользователя, клики на один запрос
- Среднее время пользователя на сайте
- Возвращаемость пользователя
- Средний чек
- Средний трафик
- Средняя разница между ценой товара и его себестоимостью (маржа)

# Где используются АБ-тесты

- Изменение дизайна на сайте
- Изменение функциональности в играх
- Работоспособность лекарств
- Выкатка нового алгоритма машинного обучения
- Изменения в онлайн-магазинах: смена порядка отделов, раскладки товаров, установка постоматов, промо-акции

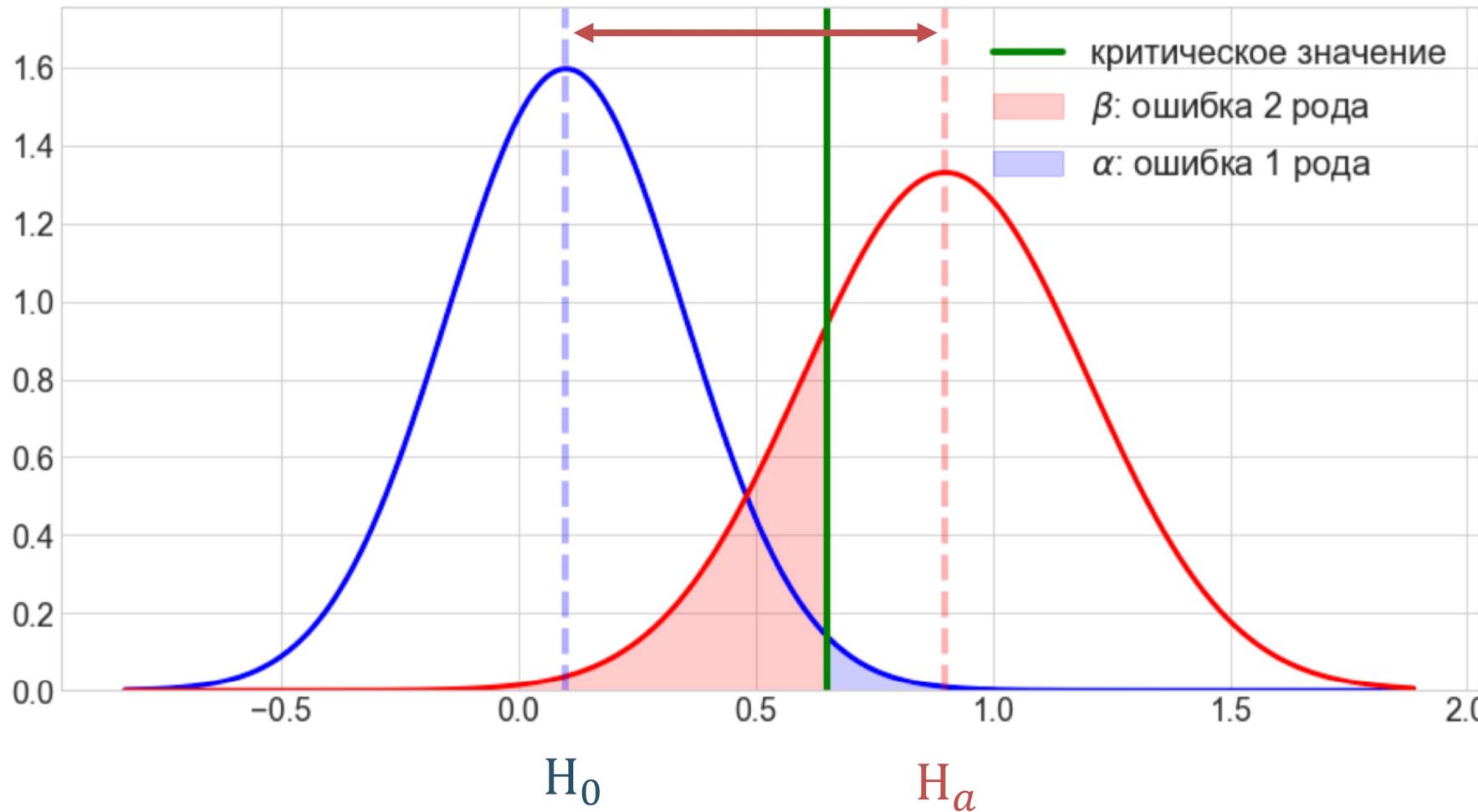
# Значимость и существенность

**Значимость** – статистический тест говорит нам, что изменения в метрике неслучайны

**Существенность** – насколько изменения большие по своей величине, насколько большой размер эффекта (изменение метрики), который мы ловим

# Размер эффекта

Размер  
эффекта



# Значимость и существенность

**Незначимо:** использование витамина D для борьбы с депрессией

- 18 тысяч человек, 5 лет
- Ежедневно принимают витамин D либо плацебо

**Результаты:**

- Тестовая группа (витамин): 609 с депрессией
- Контрольная группа (плацебо): 625 с депрессией

Разница оказалась незначима,  $pvalue = 0.62$

► <https://nplus1.ru/news/2020/08/04/vitamin-d-depression>

# Значимость и существенность

**Значимо, но несущественно:** польза позднего завтрака и раннего ужина для похудения

- 13 человек, 10 недель
- Завтракали на 1.5 часа позже и ужинали на 1.5 часа раньше обычного

**Результаты:**

- Содержание жира в экспериментальной группе снизилось на 1.9%, эффект значимый,  $pvalue = 0.047$
- Учёные отмечают, что это совсем небольшой эффект

► <https://nplus1.ru/news/2018/08/31/food-timing>

# Значимость и существенность

**Значимо и существенно:** дексаметазон снизил смертность пациентов с COVID-19 на ИВЛ (тяжёлая форма)

- 6.5 тысяч человек
- 2 тысячи в течение 10 дней получали 6 миллиграмм препарата раз в день

**Результаты:**

- Смертность больных снизилась на 30%, эффект значимый,  $pvalue = 0.0003$

► <https://nplus1.ru/news/2020/06/17/dexvsco>

# Резюме

- Во всех сферах бизнеса требуется улучшать показатели
- Идеи улучшения могут быть разными
- Хотим понять, какие из них будут работать, а какие нет
- Тестируем идеи на маленькой группе пользователей
- Проверяем гипотезу о значимости изменений

# Резюме

АБ-тест используется для проверки идей на группе пользователей. При проведении АБ-теста мы должны ответить на ряд вопросов:

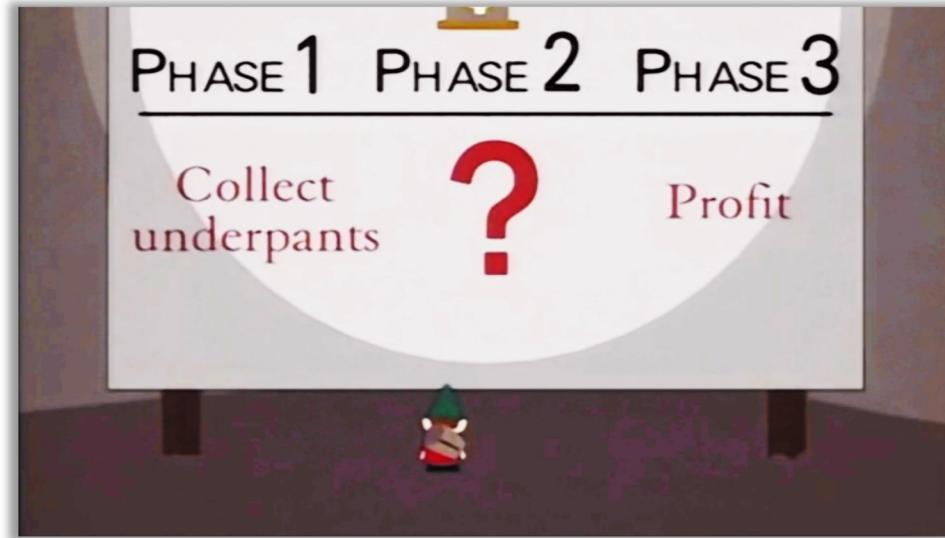
1. Что является целевой метрикой?
2. На какое увеличение мы рассчитываем?
3. Какой критерий мы используем для проверки результата на статистическую значимость?
4. Как должен выглядеть дизайн эксперимента, как разбить пользователей на группы?
5. Как долго должен идти эксперимент?

# Подводные камни АБ-тестирования

# Схема АБ-теста

**Фаза 1:** планирование эксперимента и его дизайна.

**Фаза 2:** сбор статистики и проверка гипотез



Кадр из мультипликационного фильма «Южный Парк».

Автор: Мэтт Стоун, Трей Паркер. Comedy Central

- ❗ Плохо спланированный дизайн эксперимента может привести к неверным выводам

# Кейс про кока-колу

- Как на продажи колы повлияет увеличение содержания сахара?
- Фокус-группа пробует напитки
- Обсчёт эксперимента показывает, что напиток с сахаром больше нравится людям
- Содержание сахара повышают ⇒ продажи падают



Что пошло не так?

# Кейс про кока-колу

- Исследование проходило не в тех условиях, в которых люди обычно пьют колу
- Если мы говорим про маленький стакан, то большее количество сахара нравится людям
- Если напиток постоянно употребляется в больших количествах, то большее количество сахара людям не нравится

**Мораль:** тестирование идей должно проходить в условиях максимально приближённых к реальности

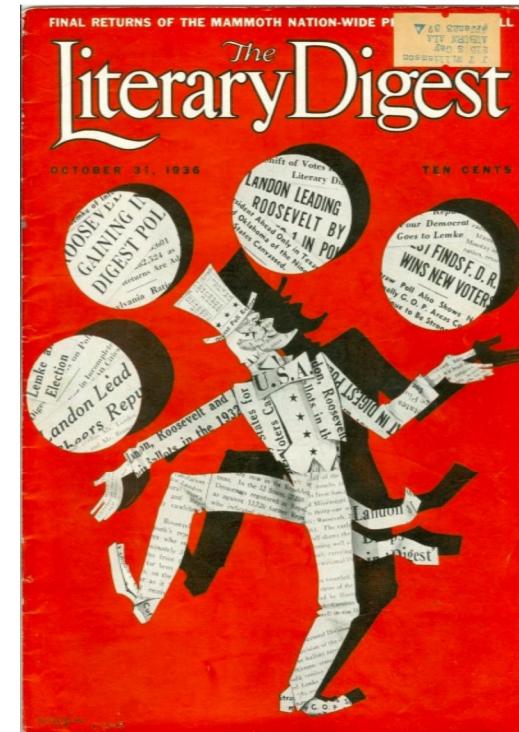
# Что **ещё** может пойти не так?



Что угодно!

# Репрезентативность выборки и выборы

- Выборы 1936 г. в США, журнал The Literary Digest опросил 10 млн. человек
- **Предсказание:** победит республиканец Альф Лэндон с результатом 60 на 40
- **Результат:** победа демократа Франклина Рузвельта 60 на 40



**Проблема:** выборка оказалась смещена. Журнал читали богатые, которые придерживались республиканской идеологии. Журнал попробовал скорректировать смещение телефонным обзвоном. Но телефон тоже был лишь у состоятельных граждан.

► <https://www.profmatt.com/statistics>

# Проблема самоотбора (selection bias)

- Часто можно увидеть разные анкеты и маркетинговые опросы
- Такие опросы подвержены проблеме самоотбора

Помогите нашему студенту в рисерче. Пожалуйста пройдите небольшой опрос, он займет менее 5 минут.

[https://docs.google.com/forms/d/e/1FAIpQLSfnp-8-jzV\\_Y..](https://docs.google.com/forms/d/e/1FAIpQLSfnp-8-jzV_Y..)

Влияние дополнительного образования на заработную плату.

Опрос проводится студентом экономического отделения РАНХиГС для дипломной работы.

\* Required

Заполните, пожалуйста, форму.

Пол \*

м

ж

Возраст (лет) \*

Влияние дополнительного образования на заработную плату.  
docs.google.com

**Проблема:** выборка окажется смещена из-за самоотбора.  
Люди принимают решение – участвовать в опросе или нет.  
Занятые люди с большой зарплатой явно не будут проходить этот опрос.

# Скрытые переменные

- Фермер, пшеница, эксперимент с новым удобрением
- Разделил поле на две части: на левую внёс удобрение, на правую нет



Кадр из фильма "Интерстеллар", Авторы: Кристофер Нолан, Джонатан Нолан. Legendary Pictures, Syncopy Films, Lynda Obst Productions

**Проблема:** скрытые переменные. Одна сторона поля может быть более солнечной, под ней может лежать геотермальный источник и т.п.

**Решение:** Разбиение на две части надо делать случайно, надо контролировать всевозможные сторонние факторы

# Связанные выборки

В эксперименте может быть важен порядок, в котором человеку показывают разные варианты



- В примере с колой, результат может зависеть от того, какой напиток человеку давали пробовать первым: сладкий или не сладкий
- **Решение:** рандомизировать порядок и давать напитки каждому человеку в случайном порядке

# Проблема подглядывания (peeking problem)

- Размер выборки для проведения АБ-теста должен быть определён заранее
  - **Нельзя** досрочно прерывать АБ-тест, при достижении значимости на более маленькой выборке
  - **Нельзя** продолжать АБ-тест, если за изначально запланированный период значимого результата получить не удалось
  - **Нельзя** менять метрики/критерии по результатам подглядывания
  - **Можно** запустить новый эксперимент, на новых выборках
- <http://varianceexplained.org/r/bayesian-ab-testing/>

# Проблема подглядывания (peeking problem)

Если мы подглядываем, мы отвечаем на вопрос

**Входит ли разница в диапазон неразличимости хотя бы раз за всё время тестирования?**

вместо

**Значима ли разница, когда вся выборка будет собрана?**

- Это завышает значение pvalue
  - **Решение:** дождаться конца теста либо использовать специальные методологии. Например, байесовских многоруких бандитов.
- <http://varianceexplained.org/r/bayesian-ab-testing/>

# Неправильная работа с метриками

- Неправильная интерпретация метрик

**Пример (эффект новизны):** метрики растут из-за того, что новизна привлекает пользователей, но со временем они упадут

- Неправильный выбор метрик

**Пример:** Британская Индия, проблема многочисленных кобр в Дели. Вознаграждение за каждую убитую змею.

- Люди начали разводить кобр, чтобы получить вознаграждение

# Смещение из-за оптимизма (Optimism bias)

- Мы недооцениваем вероятности плохих событий

Стать космонавтом

1 / 13,2 млн.



Быть насмерть  
покусанным собакой

1 / 700 000



Каковы шансы?

Выиграть в лотерею

1 / 14 млн.



Получить  
олимпийское золото

1 / 662 000



Умереть от алкогольного  
опьянения

1 / 820 000



# Смещение из-за оптимизма (Optimism bias)

- Когда метрика изменяется **в плохом направлении**, мы ищем проблемы

Стать космонавтом

1 / 13,2 млн.



Быть насмерть покусанным собакой

1 / 700 000



Каковы шансы?

Выиграть в лотерею

1 / 14 млн.



Получить олимпийское золото

1 / 662 000



Умереть от алкогольного опьянения

1 / 820 000

# Смещение из-за оптимизма (Optimism bias)

- Когда метрика изменяется в хорошем направлении, мы просто принимаем этот факт

Стать космонавтом

1 / 13,2 млн.



Быть насмерть  
покусанным собакой

1 / 700 000

Выиграть в лотерею

1 / 14 млн.



Каковы шансы?



Получить  
олимпийское золото

1 / 662 000

Умереть от алкогольного  
опьянения

1 / 820 000

# Смещение из-за оптимизма (Optimism bias)

! У нас есть предрасположенность подвергать проверкам только неприятные выводы

Стать космонавтом

1 / 13,2 млн.



Быть насмерть покусанным собакой

1 / 700 000

Выиграть в лотерею

1 / 14 млн.



Каковы шансы?



Получить олимпийское золото

1 / 662 000

Умереть от алкогольного опьянения

1 / 820 000

# Ещё проблемы

- АБ-тест длится меньше недели: поведение пользователей различается в разные дни недели, присутствует сезонность
- Долгосрочные и краткосрочные эффекты
- Хочется проверять много идей сразу, один и тот же пользователь не должен попадать в несколько групп
- Проведение одновременно нескольких экспериментов: изменения могут взаимоуничтожать друг друга

# Ещё проблемы

- Неравномерный отбор пользователей в эксперимент искажает картину
- Не всегда ясно, как улучшить сервис, гораздо понятнее, как всё испортить

# АА-тест

- Иногда, чтобы понять насколько хорошим вышел дизайн эксперимента, проводят АА-тест
- Делим пользователей на две группы в соответствии с дизайном эксперимента
- Показываем обеим группам старый вариант
- Гипотеза о том, что метрики не изменились, должна не отвергаться, если она отвергается - с дизайном либо разбиением на группы что-то не так

# Резюме

- Эксперимент нужно аккуратно планировать, вести и анализировать
- Пользователей надо разбивать на группы случайно
- Дизайн эксперимента надо тщательно продумывать так, чтобы он соответствовал максимально приближённым к реальности условиям
- Нельзя жульничать: подглядывать, обрывать эксперимент раньше времени
- **Помощь:** АА-тесты, историческая база экспериментов

# **Оффлайн АБ-тестирование, естественные эксперименты**

# Офлайн АБ-тесты

АБ-тестирование в офлайне связано с большим количеством проблем. Реальный мир накладывает на нас довольно большое количество физических ограничений.

- ❗ Для онлайна таких проблем не возникает, так как мы чаще всего можем случайно разбить пользователей на группы

► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

## Пример: онлайн-ритейл

- Ограничение на количество магазинов
- Элементы выборки зависимы (чеки внутри магазина зависят друг от друга)
- Неоднородность магазинов: у каждого магазина своё среднее значение, свой размер и трафик
- Элементы выборки не из одного распределения, а из разных: Перекрёсток у бизнес-центра и в жилом районе – разные магазины

► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

# Офлайн АБ-тесты: ритейл

- Неоднородность по погоде: в разные погодные условия разный трафик
- Неоднородность по времени: в течение суток, по дням недели и времени года (праздники, сезонность, промоакции)

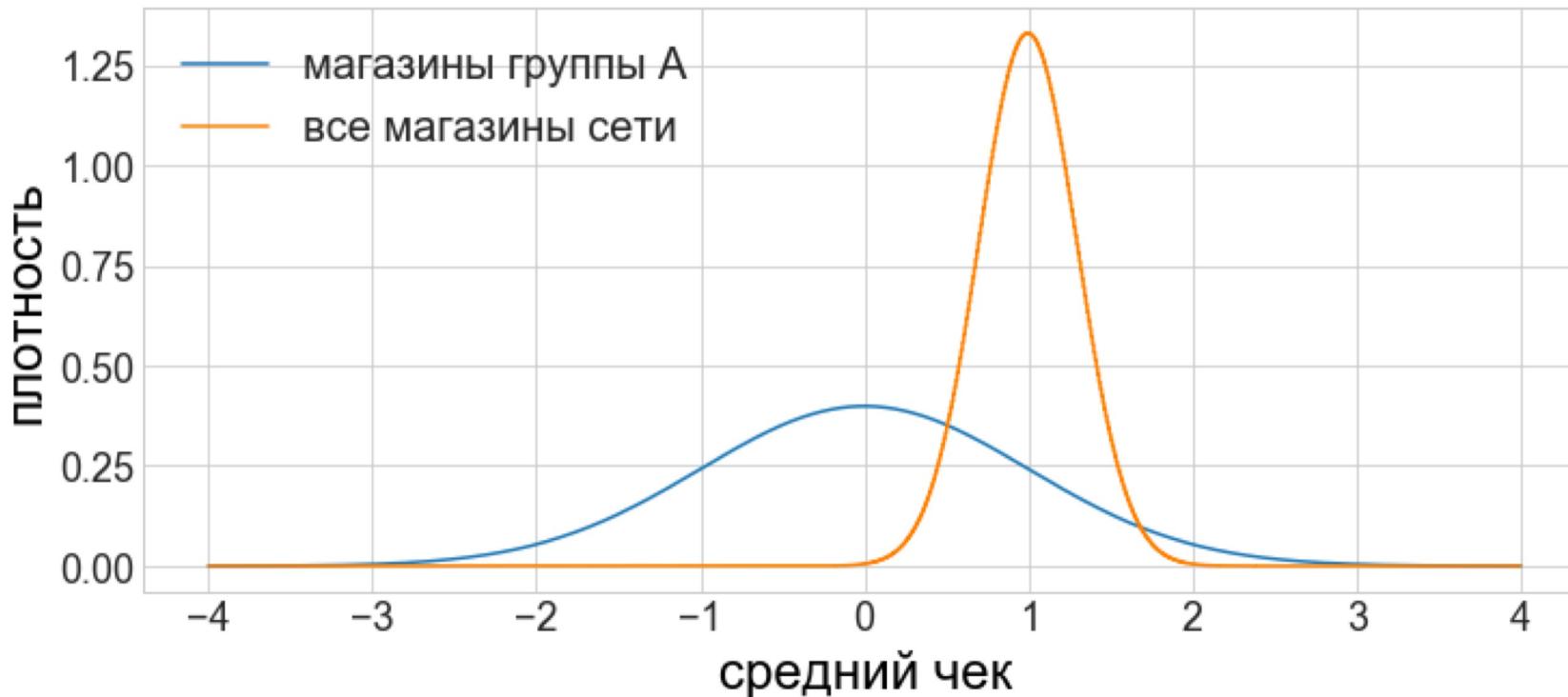


Неоднородность увеличивает дисперсии, тяжелее делать выводы, с ней нужно бороться

► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

# Офлайн АБ-тесты: ритейл

Часто сложно выделить тестовую группу так, чтобы она не отличалась по своим характеристикам от магазинов всей сети



► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

# Офлайн АБ-тесты: ритейл

⇒ нужны специальные приёмы, которые помогут повести АБ-тест корректно:

## Разбиение на группы:

- Каждый магазин описывается какими-то параметрами
- Можем посчитать между ними расстояния и найти похожие друг на друга магазины
- Если магазины были похожи до АБ-теста, то скорее всего и после него они останутся похожими
- Универсального способа нет

► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

# Офлайн АБ-тесты: ритейл

⇒ нужны специальные приёмы, которые помогут повести АБ-тест корректно:

## Проверка корректности:

- АА-тесты: правда ли, что в выделенных нами группах до эксперимента нет значимых различий
- Искусственный эффект: добавляем его в одну из групп и убеждаемся, что тест его находит
- Есть и другие методики валидации

► <https://habr.com/ru/company/X5RetailGroup/blog/466349/>

# Невозможность АБ-теста

Бывают ситуации, когда АБ-тест устроить невозможно

## Примеры:

- Вызывает ли курение рак? Нельзя поделить людей на две группы и заставить одну из них курить в течение всей жизни.
- Простимулирует ли экономику снижение налога? Нельзя поделить экономику страны на две части.

Многие эксперименты запрещает этика. Многие запрещает суровая реальность.



В таких ситуациях приходится работать с наблюдаемыми данными

# Зефирный тест

Кладём перед ребёнком зефир, если он не съест его в полном одиночестве за 15 минут, то получит ещё одну

**1960-е:** первые тесты в Стендфорде

**1990-е:** изучение повзрослевших детей



# Зефирный тест

Кладём перед ребёнком зефир, если он не съест его в полном одиночестве за 15 минут, то получит ещё одну

**1960-е:** первые тесты в Стендфорде

**1990-е:** изучение повзрослевших детей

## Результаты:

- Тот, кто справился с искушением, показал себя успешнее, чем его сверстники
- Откладывание удовольствия приводит в жизни к успеху
- Зефирные тесты стали очень модными

# Зефирный тест

Кладём перед ребёнком зефир, если он не съест его в полном одиночестве за 15 минут, то получит ещё одну

**1960-е:** первые тесты в Стендфорде

**1990-е:** изучение повзрослевших детей

## Проблемы:

- 90 детей, все из детского сада при Стендфорде

# Зефирный тест

Кладём перед ребёнком зефир, если он не съест его в полном одиночестве за 15 минут, то получит ещё одну

**1960-е:** первые тесты в Стендфорде

**1990-е:** изучение повзрослевших детей

**Новое исследование:**

- 1000 детей из разных слоёв общества, контрольные переменные (демография, социальное положение и тп)
- Способность продержаться определяется финансовым положением, отсюда и будущая успешность детей

# Резюме

- В офлайне АБ-тесты делать сложнее, чем в онлайне
- На каждом этапе проведения эксперимента возникают проблемы
- Нужно быть аккуратнее с неоднородностью выборок, все результаты необходимо дополнительно валидировать
- Бывают ситуации, когда провести АБ-тест невозможно, но знать величину эффекта надо
- В таких ситуациях на помощь приходит эконометрика, которая помогает очистить эффект от влияния других переменных

# Сколько надо наблюдений

# Ошибки, что мы совершаем

	$H_0$ верна	$H_0$ неверна
$H_0$ не отвергается	<i>ok</i>	$\beta$
$H_0$ отвергается	$\alpha$	<i>ok</i>
ошибка 1 рода		

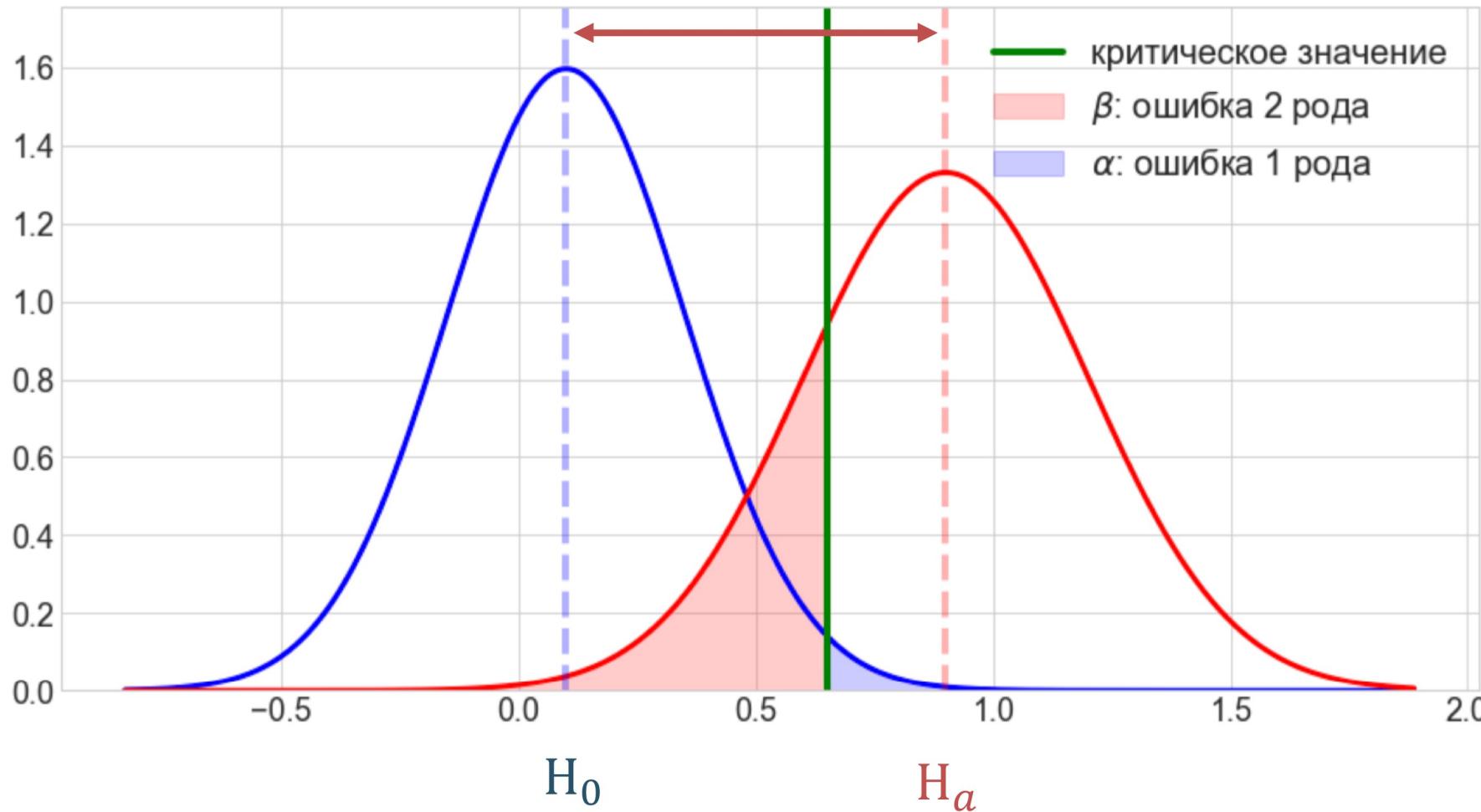
$$\alpha = \mathbb{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна})$$

$$\beta = \mathbb{P}(H_0 \text{ не отвергнута} \mid H_0 \text{ не верна})$$

Величину  $1 - \beta$  называют **мощностью** критерия

# Размер эффекта

Размер  
эффекта



# Сколько нужно наблюдений

- Необходимое количество наблюдений зависит от размеров ошибок первого и второго рода, а также от размера эффекта
- Фиксируем уровень значимости (ошибку 1 рода), на которую мы согласны
- Подбираем соотношение между минимальным размером эффекта, желаемой мощностью и объёмом выборки
- В выборе соотношении помогает заказчик эксперимента, у него обычно есть ограничения, с которыми нам придётся работать (количество магазинов, длительность АБ-теста и т.п.)

# Таблица эффекта-ошибки



- ! Совокупность этих трёх параметров (ошибка 1/2 рода, размер эффекта) позволяют рассчитать необходимый для эксперимента объём выборки.

# Сколько нужно наблюдений

**Пример:** проверяем равенство конверсий до и после нововведений

$$H_0: p_0 = p_a$$

$$H_a: p_0 \neq p_a$$

Используем асимптотически-нормальный тест:

$$z = \frac{p_a - p_0}{\sqrt{P(1 - P) \cdot \left(\frac{1}{n} + \frac{1}{n}\right)}} \stackrel{\substack{asy \\ H_0}}{\approx} N(0, 1)$$

размер  
эффекта

# Сколько нужно наблюдений

Ошибка второго рода:

$$\beta = \Phi \left( \frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_a(1-p_a)}} \cdot z_{1-\alpha} + \frac{p_0 - p_a}{\sqrt{\frac{p_a(1-p_a)}{n}}} \right)$$

Число наблюдений:

$$n = \left( \frac{z_{1-\alpha} \cdot \sqrt{p_0(1-p_0)} + z_{1-\beta} \cdot \sqrt{p_a(1-p_a)}}{p_a - p_0} \right)^2$$

размер  
эффекта

# Анализ мощности

## До эксперимента:

- Какой нужен объём выборки, чтобы найти различия с разумной степенью уверенности
- Различия какой величины мы можем найти, если известен объём выборки

## После эксперимента:

- смогли бы мы найти различия с помощью нашего эксперимента, если бы величина эффекта была равна  $\Delta$

# Резюме

- Для многих критериев можно вывести формулу для расчёта необходимого числа наблюдений
- Число наблюдений зависит от ошибок  $\frac{1}{2}$  рода и минимального размера эффекта, который мы хотим уловить
- Перед экспериментом необходимое число наблюдений определяют исходя из пожеланий заказчика и физических возможностей