

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

Принимаем решение

Значимость

Статистический тест

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

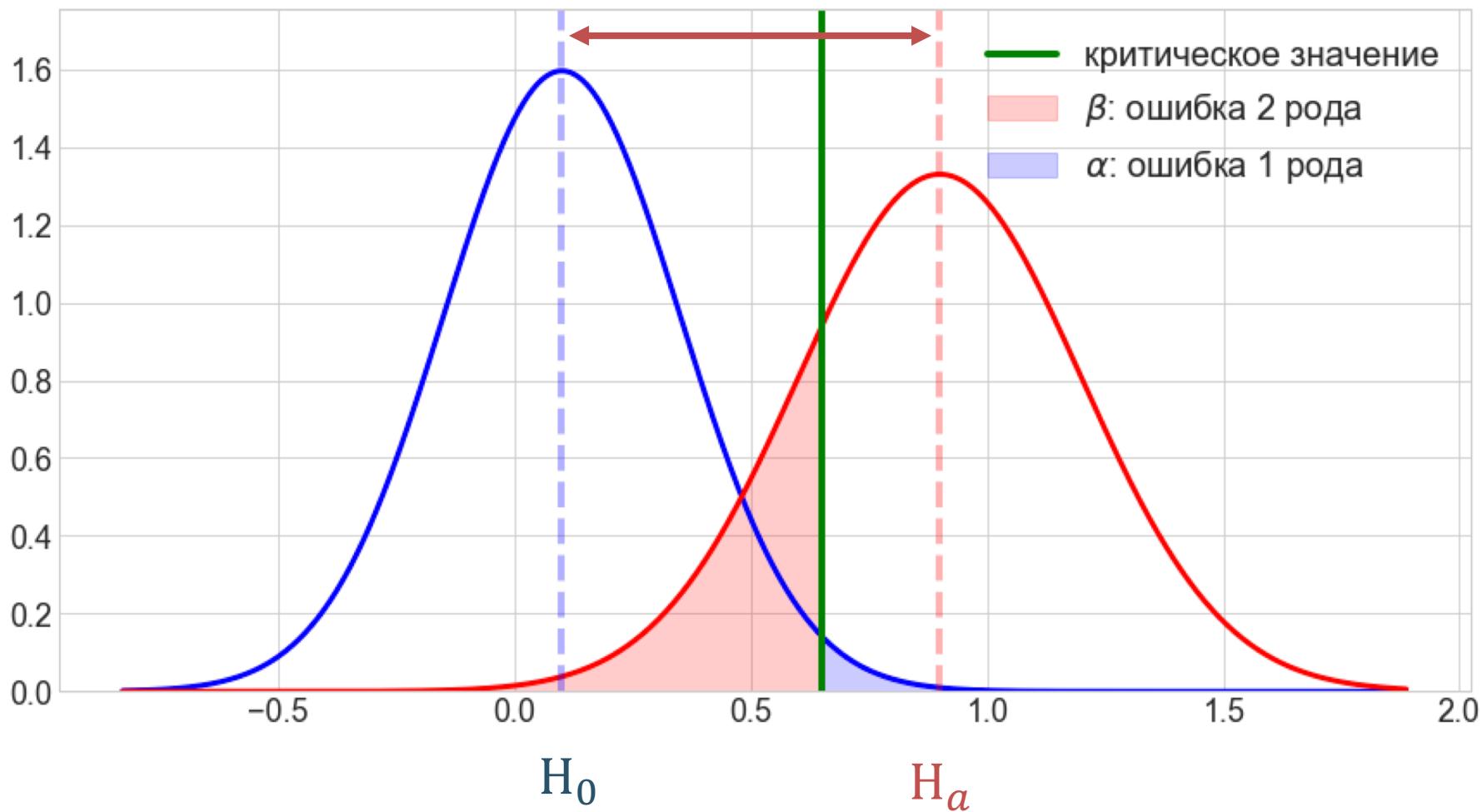
Значимость и существенность

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/dexvscov>

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:

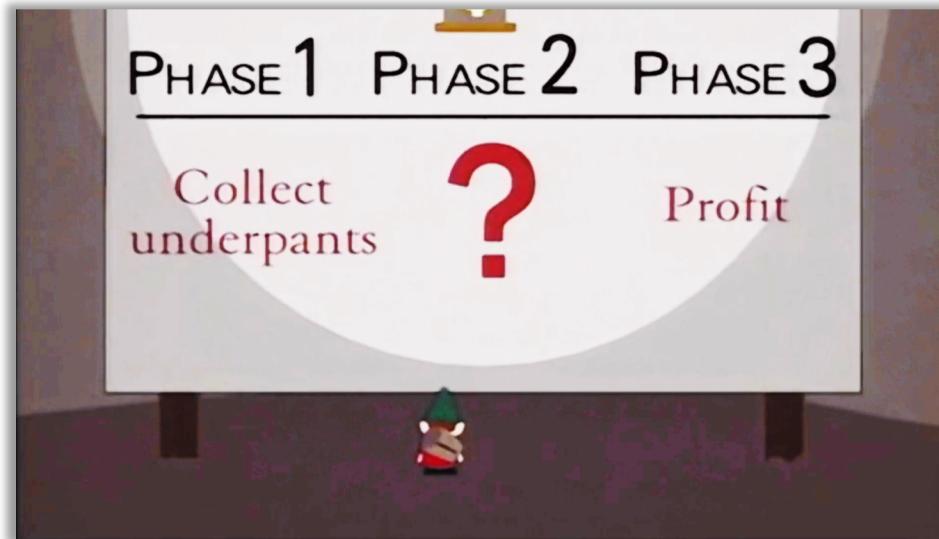
1. Что является целевой метрикой?
2. На какое увеличение мы рассчитываем?
3. Какой критерий мы используем для проверки результата на статистическую значимость?
4. Как должен выглядеть дизайн эксперимента, как разбить пользователей на группы?
5. Как долго должен идти эксперимент?

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

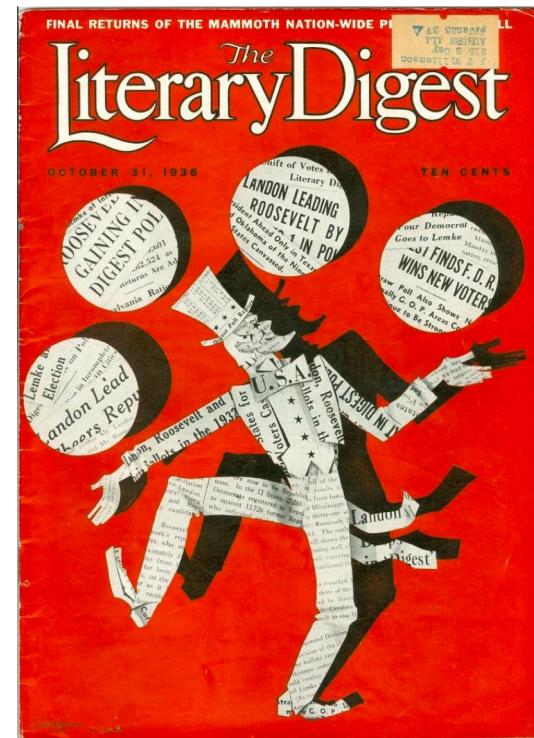
Что ещё может пойти не так?



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..

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



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)

- 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



Каковы шансы?

Выиграть в лотерею

1 / 14 млн.



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: офлайн-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

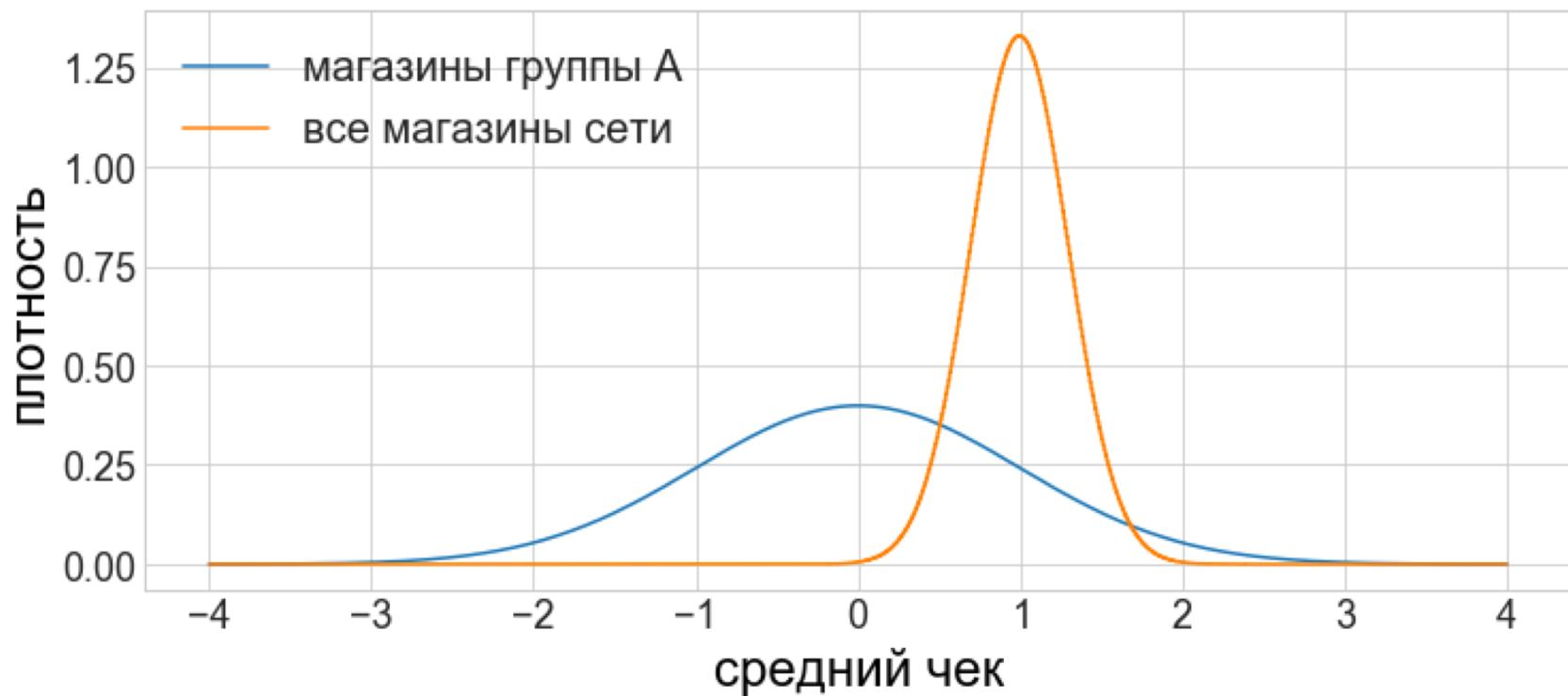
- Неоднородность по погоде: в разные погодные условия разный трафик
- Неоднородность по времени: в течение суток, по дням недели и времени года (праздники, сезонность, промо-акции)

! 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/>

Оффлайн 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

Естественный эксперимент

Задача: правда ли, что в более маленьких классах дети учатся лучше?

Идеальный эксперимент:

- Случайным образом поделить всех детей и обучающих их учителей на классы разного размера
- Проводить регулярное тестирование и сравнивать различия в оценках

Проблемы:

- Родители хотят, чтобы их ребёнок учился в маленьком классе и мешают эксперименту
- Очень дорого: 4-летний проект STAR (~12 млн. \$)

Естественный эксперимент

Проект STAR (вторая половина 80-х): выборка результатов тестов в $n = 420$ школьных округах в Калифорнии

Переменные:

- средние результаты тестирования пятиклассников по округу (комбинация тестов по математике и чтению)
- соотношение учеников и учителей (число учеников в округе делённое на число учителей в округе)

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Линейная регрессия: статистический взгляд

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$

x_i – среднее значение размера класса в i -ом округе

y_i – среднее значение за тест в i -ом округе

ε_i – прочие факторы, влияющие на результаты обучения в i -ом округе

β_0, β_1 – коэффициенты линейной регрессии

Линейная регрессия: статистический взгляд

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$

Если:

- $\mathbb{E}(\varepsilon_i | x_i) = 0$,
- наблюдения независимы, одинаково распределены
- большие выбросы маловероятны

Тогда можно найти несмешённую, состоятельную и эффективную оценку $\hat{\beta}_1$ и проверить гипотезу о равенстве этого коэффициента нулю

Линейная регрессия: статистический взгляд

Задача: понять, как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$

Проблема:

- Есть ещё десяток признаков, которые могут влиять на успеваемость детей
- Если учесть их влияние, останется ли влияние размера класса значимым

Эконометрика

- Логическое продолжение статистики
- Пытается ответить на вопрос, как именно несколько переменных связаны между собой
- Пытается получить несмешённые, состоятельные и эффективные оценки для этих взаимосвязей
- Изучает методы оценки причинно-следственных связей и свойства этих методов

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

Множественное тестирование

История о зомби-лососе

- В 2012 году ряд авторов получил Шнобелевскую премию по нейробиологии
- Надо было протестировать аппарат МРТ
- Для этого обычно в него кладут шарик с маслом и сканируют его



- ▶ <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- ▶ <https://habr.com/ru/company/ods/blog/325416/>

История о зомби-лососе

- Это скучно, поэтому авторы решили купить на рынке мёртвого лосося и просканировать его мозг
- Лососю показывали фотографии людей и проверяли, есть ли у него в мозгу активность
- Оказалось, что активность есть



- ▶ <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- ▶ <https://habr.com/ru/company/ods/blog/325416/>

История о зомби-лососе

- Аппарат МРТ возвращает много данных
- Чтобы убедиться, что в мозгу нет реакции, надо проверить много гипотез об отсутствии активности на каждом маленьком участке мозга

Проблема множественного тестирования:
если мы проверяем несколько гипотез подряд,
уровень значимости выходит из-под контроля

Мы начинаем чаще отвергать верные гипотезы,
чем нам хотелось бы

- <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- <https://habr.com/ru/company/ods/blog/325416/>

Множественная проверка гипотез

Проверяем две гипотезы:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Каждую на уровне значимости α

Можно ошибиться сразу в двух местах:

$\mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$

$$= 1 - \mathbb{P}(\text{не ошибиться ни в одной}) = 1 - (1 - \alpha)^2$$

$$= 1 - (1 - 2\alpha + \alpha^2) = 2\alpha - \alpha^2 > \alpha$$

$$\alpha_i = 0.05 \Rightarrow \alpha = 0.1 - 0.025 = 0.075 > 0.5$$



Вероятность ошибки первого рода накапливается и выходит из-под контроля

Множественная проверка гипотез

Пример: показ на странице сервиса нескольких новых элементов

- Изменения взаимосвязаны и их можно протестировать только на одном временном промежутке
- В такой ситуации мы сталкиваемся с множественным тестированием
- С ростом числа гипотез, вероятность получить ошибку растёт экспоненциально: $1 - (1 - \alpha)^n$

! Нужно взять уровень значимости под контроль

Неравенство Бонферрони

- Нужно как-то скорректировать исходный уровень значимости, в этом помогает неравенство Бонферрони:

$$\mathbb{P}(A + B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

- То есть каждую гипотезу из двух надо проверять на уровне значимости $\frac{\alpha}{2}$

$$\alpha = \mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$$

$$\leq \mathbb{P}(\text{ош. в 1}) + \mathbb{P}(\text{ош. во 2}) = \frac{\alpha_i}{2} + \frac{\alpha_i}{2} = \alpha_i$$

- Если гипотез k , берём уровень значимости $\frac{\alpha}{k}$ для каждой

Неравенство Бонферрони

- Из-за коррекции уровня значимости возникают проблемы с мощностью тестов
- Чем больше гипотез проверяется, тем ниже шансы отклонить неверные гипотезы
- Более того, из-за презумпции нулевой гипотезы для более низкого уровня значимости нам нужно собрать большее число наблюдений, чтобы зафиксировать значимое отклонение от нулевой гипотезы

⇒ процедуру надо улучшить,
чтобы мощность стала выше

Матрица ошибок

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

- Неверно отклонили V гипотез, неверно не отклонили T гипотез
- На практике пытаются контролировать обобщения ошибки первого рода, например: FWER и FDR

Family-Wise Error Rate (FWER)

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

Групповая вероятность ошибки, FWER (Family-Wise Error Rate)

– это вероятность совершил хотя бы одну ошибку первого рода

$$FWER = \mathbb{P}(V > 0)$$

False Discovery Rate (FDR)

Рассмотрим случай, когда мы проверяем n гипотез

	верных H_{0i}	неверных H_{0i}
не отвергнутых H_{0i}	U	T
отвергнутых H_{0i}	V	S

Ожидаемая доля ложны отклонения, FDR (False Discovery Rate) – это математическое ожидание числа ошибок первого рода к общему числу отклонений нулевой гипотезы

$$FDR = \mathbb{E} \left(\frac{V}{V + S} \right)$$

Метод Холма

- Поправка Бонферрони пытается контролировать FWER (вероятность хотя бы одной ошибки 1 рода)
- **Бонферрони:** проверяем k гипотез на уровнях значимости

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = \frac{\alpha}{k}$$

- **Метод Холма** – улучшение поправки Бонферрони, обладает более высокой мощностью
- Проверяем k гипотез, но уровни значимости пытаемся выбирать разными

Метод Холма

- Отсортируем гипотезы по получившимся P -значениям по возрастанию: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$
- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{\alpha}{k-1}, \dots, \alpha_{(i)} = \frac{\alpha}{k-i+1}, \dots, \alpha_{(k)} = \alpha$$

- Если $p_{(1)} \geq \alpha_{(1)}$, все нулевые гипотезы не отвергаются, иначе отвергаем первую и продолжаем
- Если $p_{(2)} \geq \alpha_{(2)}$, все оставшиеся нулевые гипотезы не отвергаются, иначе отвергаем вторую и продолжаем
- Идём, пока не кончатся гипотезы

Метод Холма

- Метод Холма обеспечивает контроль $FWER$ на уровне α
- Метод Холма оказывается мощнее корректировки Бонферрони, так как его уровни значимости меньше

Метод Бенджамини-Хохберга

- Отсортируем гипотезы по получившимся P -значениям по возрастанию: $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$
- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{2\alpha}{k}, \dots, \alpha_{(i)} = \frac{i\alpha}{k}, \dots, \alpha_{(k)} = \alpha$$

- Если $p_{(k)} < \alpha_{(k)}$, отвергнуть все гипотезы, иначе не отвергнуть k – ую и продолжить
- Если $p_{(k-1)} < \alpha_{(k-1)}$, отвергнуть все оставшиеся гипотезы, иначе не отвергнуть $(k - 1)$ – ую и продолжать
- Идём, пока не кончатся гипотезы

Метод Бенджамини-Хохберга

- Для любой процедуры множественного тестирования гипотез $FDR \leq FWER$
- Метод Бенджамини-Хохберга обычно оказывается более мощным, чем методы контролирующие $FWER$
- Он отвергает не меньше гипотез с теми же α_i
- Это происходит за счёт того, что метод позволяет допустить большее число ошибок первого рода

Специальные тесты

Альтернатива для процедур множественного тестирования – разработка специальных тестов, которые проверяют гипотезы сразу о нескольких ограничениях

Примеры:

- Тест отношения правдоподобий (обсудим позже)
- ANOVA – равенство сразу же нескольких математических ожиданий
- Тест Бартлетта – равенство нескольких дисперсий

Резюме

- Если сделать поправку, мёртвый лосось остаётся мёртвым
- До 2010 около 40% статей по нейробиологии не использовали поправки при множественном тестировании гипотез
- Благодаря работе о лососе и Шнобелевской премии за неё удалось уменьшить число таких статей до 10%
- Корректировка уровня значимости помогает держать под контролем ложно-положительные результаты, это приводит к росту ложно-отрицательных результатов

Сколько надо наблюдений

Ошибки, что мы совершаем

	H_0 верна	H_0 неверна	
H_0 не отвергается	<i>ok</i>	β	ошибка 2 рода
H_0 отвергается	α	<i>ok</i>	ошибка 1 рода

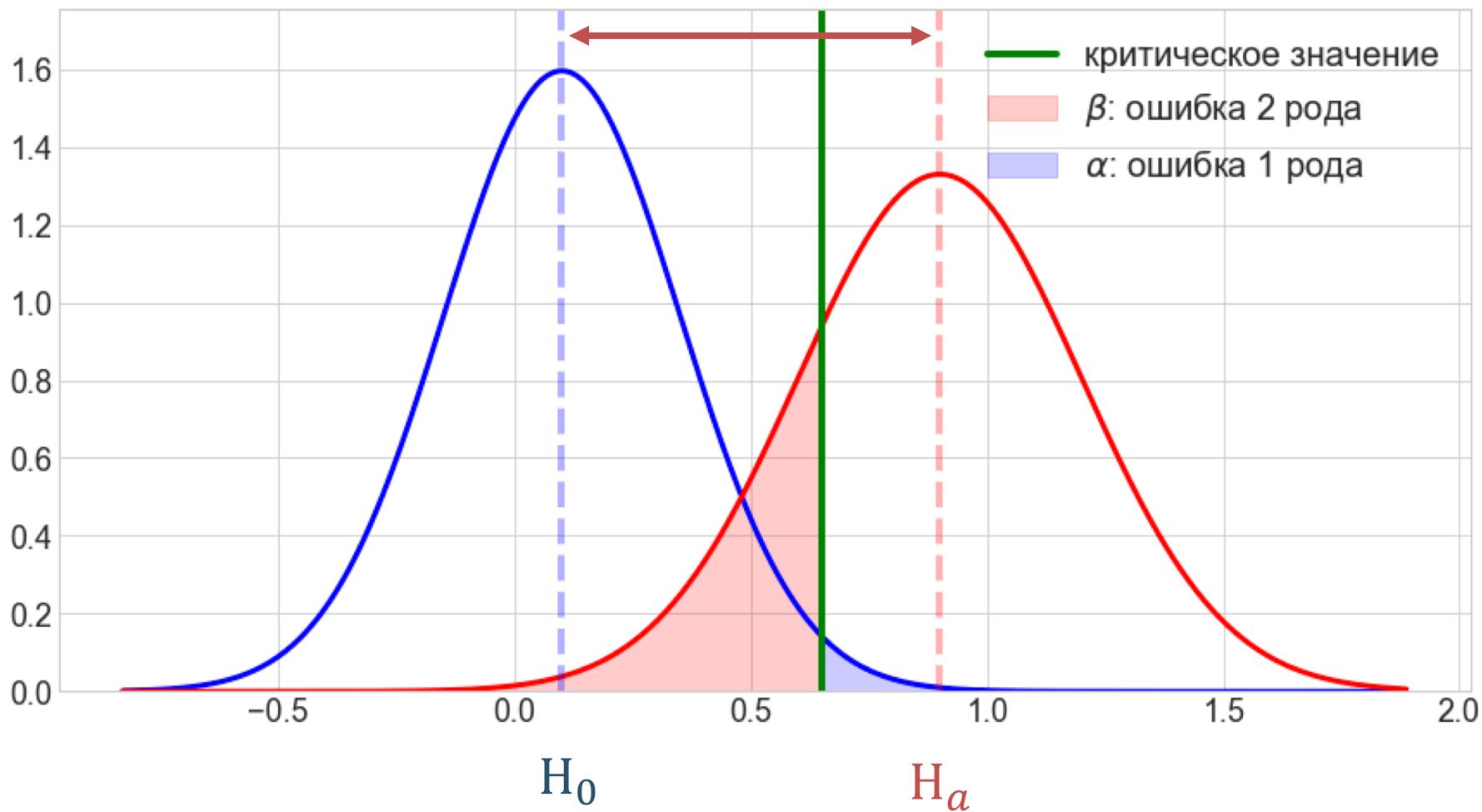
$$\alpha = \mathbb{P}(H_0 \text{ отвергнута} \mid H_0 \text{ верна})$$

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

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

		Ошибка 1/2 рода $\alpha = \beta$			
		0.1%	1%	5%	10%
размер эффекта	1%	МНОГО данных			
	1.5%				
	3%				
	5%				
	10%				мало данных

- ! 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:

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

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 Δ

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

Бутстрап

Бутстррап

- Не для всех описательных статистик можно найти распределение в аналитическом виде (медиана, эксцесс, куртосис)
- Бутстррап помогает решить эту проблему

Бутстррап

- **Идея метода:** имеющаяся выборка – это единственная информация об истинном распределении данных
- Давайте приблизим истинное распределение эмпирическим, то есть “сами себя вытащим”
- Предполагается, что бутстррап-распределение окажется похожим на реальное распределение

Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики \bar{x}

Выборки с повторениями:

$$1,4 \Rightarrow \bar{x} = 2.5$$

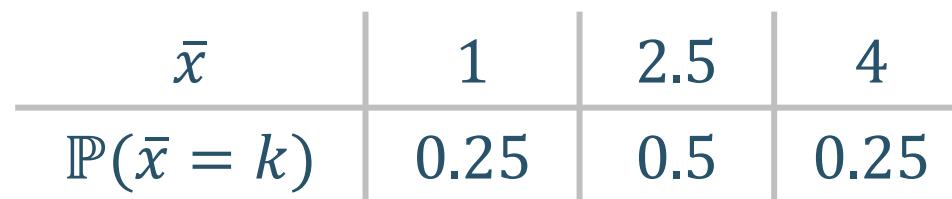
$$4,1 \Rightarrow \bar{x} = 2.5$$

$$1,1 \Rightarrow \bar{x} = 1$$

$$4,4 \Rightarrow \bar{x} = 4$$

Всего вариантов выборок:

$$n^n = 2^2 = 4$$



Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики \bar{x}

Всего вариантов выборок:

$$n^n = 2^2 = 4$$

- Строить такие распределения для больших выборок дорого
- Задача слишком сложная, а точность излишняя
- **Выход:** симуляции

Схема бутстрата

- Извлечение выборок из генеральной совокупности – сэмплирование из неизвестного распределения $F(x)$
- У нас есть оценка для $F(x) - \hat{F}_n(x)$
- Сэмплировать из такого распределения – то же самое, что брать выборки с повторениями объёма n

Схема бутстрата

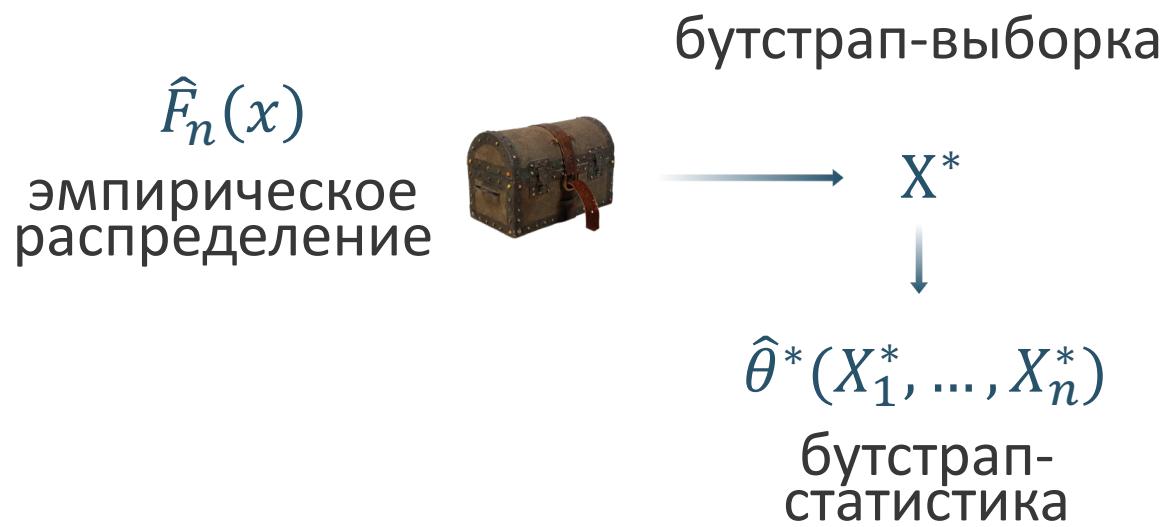


Схема бутстрата

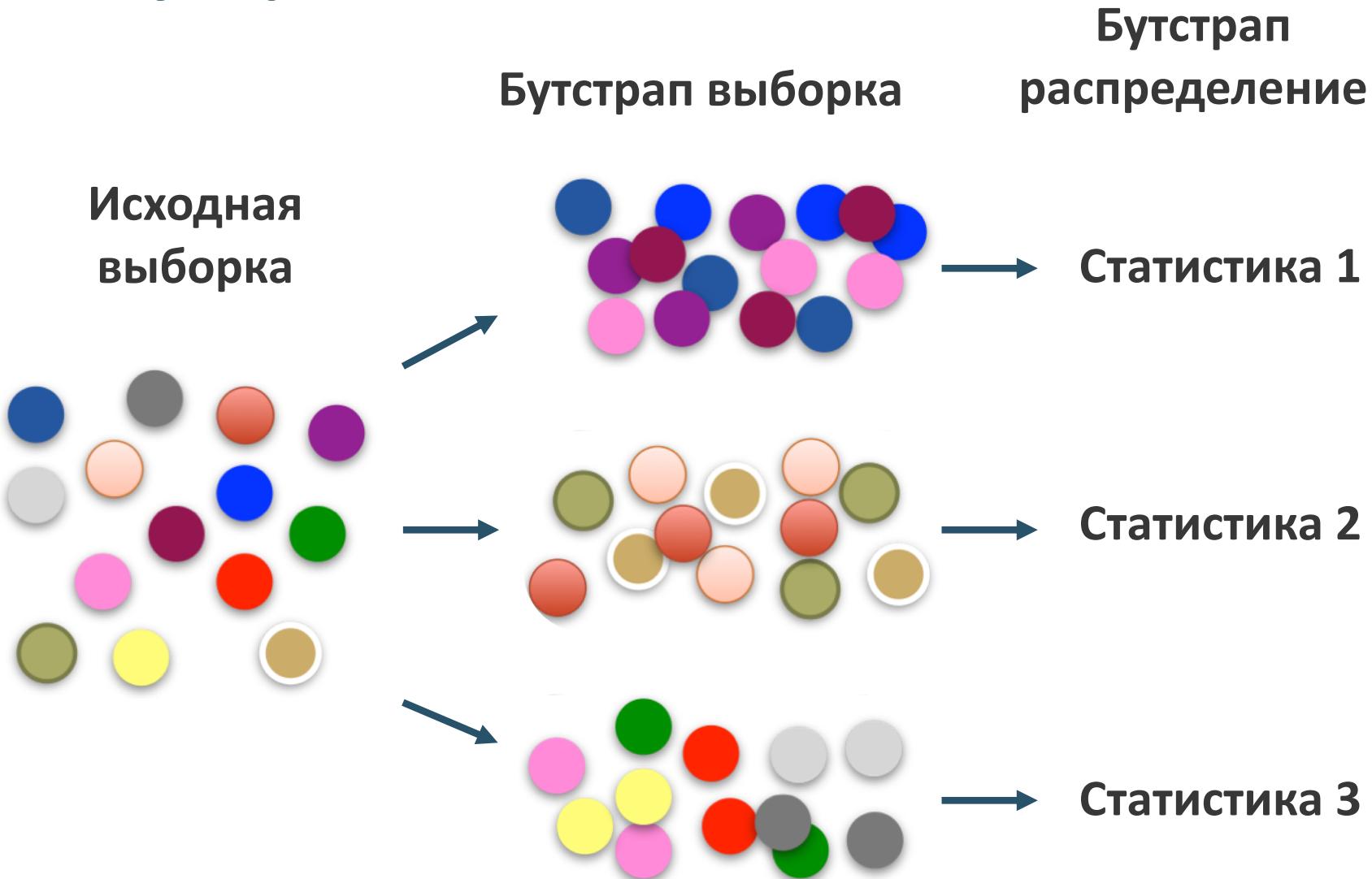


Схема бутстрата

- Генерируем B “псевдовыборок” с повторениями объёма n и оцениваем настоящее распределение “псевдоэмпирическим”
- По каждой выборке вычисляем интересующую нас статистику, получаем для неё бутстрат-распределение
- **Эфронов доверительный интервал:** находим выборочные квантили бутстрат-распределения
- В таком доверительном интервале возникает смещение \Rightarrow более сложные техники бутстрата

Доверительный интервал Эфона

Бутстранируем: оценку неизвестного параметра

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $\hat{\theta}^*$

Повторяем: B раз

Строим распределение: $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Интервал: $[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*]$

Доверительный интервал Эфона

Бутстранируем: оценку неизвестного параметра

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $\hat{\theta}^*$

Повторяем: B раз

Строим распределение: $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Интервал: $[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*]$

- ! Если распределение выборки несимметрично, такой доверительный интервал усиливает смещение, присущее изначальной выборке

Центрирование

- Чтобы не создавать искусственное смещение между $\hat{\theta}$ и θ , нужно бутстрартировать центрированную статистику
- Хотим бутстрартировать разность $\hat{\theta} - \theta$
- Бутстрарпируя $\hat{\theta}$, мы подсчитываем её каждый раз заново, но на бутстрарповских выборках, $\hat{\theta}^*$
- Бутстрарповским аналогом для θ будет $\hat{\theta}$, так как мы сэмплируем данные их эмпирического распределения
- Бутстрарповский аналог для $\hat{\theta} - \theta$ – это $\hat{\theta}^* - \hat{\theta}$

Доверительный интервал Холла

Бутстранируем: Отклонение оценки от истинного значения

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $\hat{q}_i^* = \hat{\theta}^* - \hat{\theta}$

Повторяем: B раз

Строим распределение: $\hat{q}_1^*, \dots, \hat{q}_B^*$

Интервал: $[\hat{\theta} - \hat{q}_{1-\frac{\alpha}{2}}^*; \hat{\theta} - \hat{q}_{\frac{\alpha}{2}}^*]$

t -процентильный доверительный интервал

Бутстранируем: t - статистику

Сэмплируем: x_1^*, \dots, x_n^*

Считаем: $t_i^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$

Повторяем: B раз

Строим распределение: t_1^*, \dots, t_B^*

Интервал: $[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* se(\hat{\theta}); \hat{\theta} + t_{\frac{\alpha}{2}}^* se(\hat{\theta})]$

❗ Если бутстрарировать $|\hat{\theta}^* - \hat{\theta}|$, можно получить симметричный интервал

Проверка гипотез при помощи бутстрата

$$H_0: \theta = \theta_0$$

$$H_a: \theta > \theta_0$$

t-статистика:

$$t = \frac{(\hat{\theta} - \theta_0)}{se(\hat{\theta})}$$

Бутстррап-аналог:

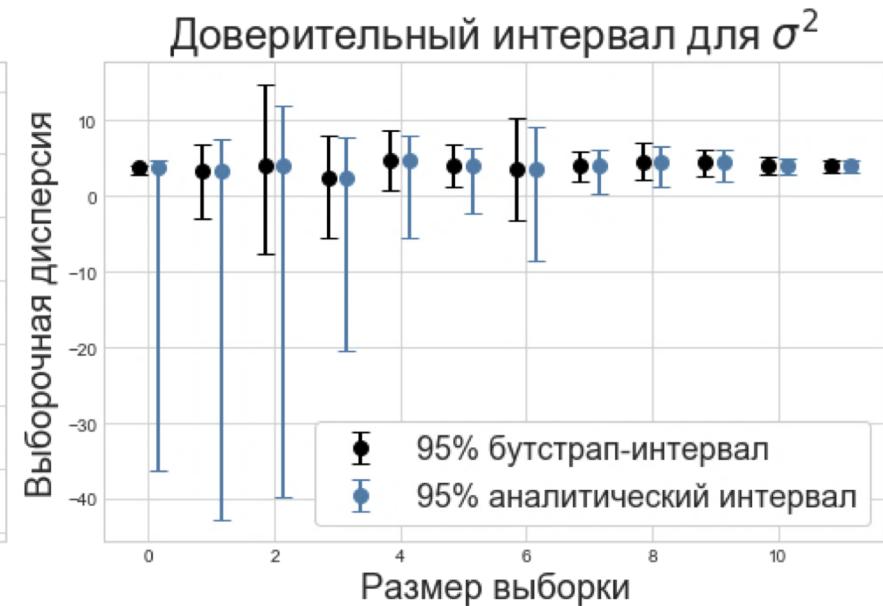
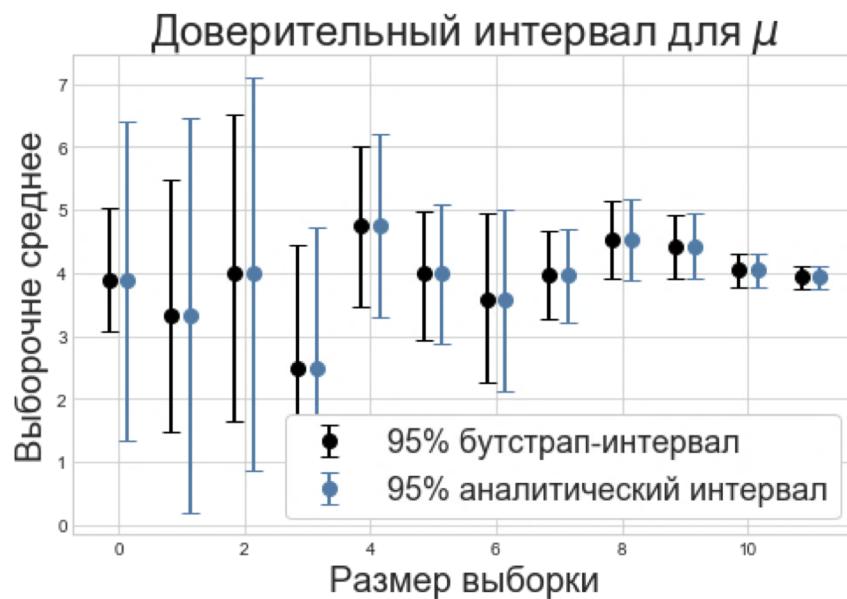
$$t^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$$

- Гипотеза отвергается, если $t_{obs} > t_{1-\alpha}^*$
- По аналогии можно проверять гипотезы против других альтернатив
- Для более сложных гипотез есть специальные бутстрраповские алгоритмы проверки

► <http://quantile.ru/03/03-SA.pdf>

Проблемы бутстрата

- Чтобы бутстрат сработал, выборка должна быть репрезентативной



- Если исходная выборка маленькая, бутстратовский доверительный интервал будет уже аналитического, так как в выборке недостаточно “неопределенности”

Проблемы бутстрата

- Если в данных есть структура (регрессия, временные ряды), бутстррап нужно устроить так, чтобы учитывать её ⇒ разные виды бутстрата
- Бутстррап ненадёжно работает в хвостах распределения из-за маленького числа наблюдений: мы можем хорошо оценить медиану, но не 99% квантиль
- Если у распределения тяжёлые хвосты, бутстррап может работать некорректно и в средиземье

Сколько процентов выборки берем

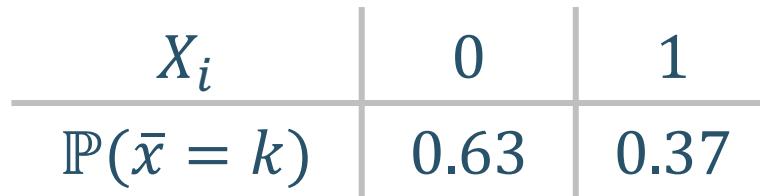
У Винни-Пуха есть 100 песенок (кричалок, вопелок, пыхтелок и сопелок). Каждый день он поёт одну равновероятно наугад. Одну и ту же песенку он может петь несколько раз. Сколько в среднем песенок не будут спеты ни разу за 100 дней?

- Вероятность конкретной песенки $\frac{1}{n}$
- В конкретный день не споёт эту песенку с вероятностью $1 - \frac{1}{n}$
- Не споёт песенку ни разу с вероятностью

$$\left(1 - \frac{1}{n}\right)^n \rightarrow e^{-1} = \frac{1}{e} \approx 0.37 \text{ при } n \rightarrow \infty$$

Пример

- Случайная величина $X_i = 1$, если песенка не была ни разу спета за n дней



- Случайная величина $Y = X_1 + \dots + X_n$ – число песенок, которое Винни-Пух ни разу не споёт за n дней

$$\mathbb{E}(Y) = \mathbb{E}(X_1 + \dots + X_n) = n \cdot \mathbb{E}(X_1) = 0.37 \cdot n$$

! В одной бутстрраповской выборке будет в среднем оказываться 63% наблюдений

Резюме

- Бутстреп – это метод получения критических значений статистики
- Процедура может требовать много времени для оценки
- При некоторых ограничениях бутстреп даёт состоятельные оценки, но не в общем случае
- Плохо работает для статистик, значение которых зависит от небольшого числа элементов выборки

Метрики для АБ-тестов

Метрики

- Показатель для улучшения – метрика
- Метрики бывают разными, они конструируются в зависимости от бизнес-задачи
- Иногда метрики привязаны к деньгам
- Чаще всего денежные метрики грубые (слабо реагируют на изменения либо, надо очень много времени, чтобы их измерить)
- Из-за этого чистым денежным метрикам предпочтительнее промежуточные метрики

Пример: Сайт с арендой квартир: число посетителей за день, число уникальных посетителей и тп.

Желательные свойства метрик

- **Согласованность** – метрика должна быть согласована с целями сервиса и его ключевыми метриками
- **Направленность** – если значение метрики изменилось, должна быть чёткая интерпретация этого изменения (хорошо это или плохо)

Желательные свойства метрик

- **Чувствительность** (sensitivity) – способность метрики отражать статистически значимую разницу между контрольной и тестовой группами, когда она есть
- Чем выше чувствительность, тем меньше данных нужно, чтобы обнаружить статистически-значимые изменения

Пример: метрики, основанные на деньгах слабо реагируют на изменения

Желательные свойства метрик

- **Стабильность** – метрика должна быть чувствительной и согласованной с тем, что нельзя ломать
- Если у метрики высокая дисперсия, то для того, чтобы уловить значимый эффект, надо собирать много данных

Пример: розничный торговый оборот магазина может колебаться в очень широких диапазонах. Чтобы уменьшить его дисперсию, обычно смотрят торговый оборот отдельных отделов.

Желательные свойства метрик

Лояльность пользователя:

- Число пользовательских сессий
- Время, которое юзер проводит в сервисе

Имеют чёткую направленность

Хорошие предикторы для долгосрочного успеха продукта

Обладают слабой чувствительностью

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

Желательные свойства метрик

Активность пользователя:

- Число кликов за сессию
- Длина пользовательской сессии

Обладают сильной чувствительностью

Обладают неоднозначной направленностью

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

Желательные свойства метрик

Пример: клики пользователей в рекомендательной системе отражают как позитивные, так и негативные сигналы

- С одной стороны, они говорят, что пользователю нравится пользоваться продуктом
- С другой, они говорят, что у нас много кликбайтного контента
- Метрики с чёткой интерпретацией часто обладают низкой чувствительностью

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

Как изучать свойства метрик

- Надо понимать особенности тех метрик, которые используются
- Замерять их характеристики
- Находить модификации, которые улучшают эти характеристики
- Нужен большой пул полезных исторических экспериментов

Примеры свойств

В метриках могут быть тренд и сезонность, их необходимо от них очищать:

- Линейная регрессия
- Взятие 12-ой разности

Среднее не единственная метрика, которую можно считать:

- Средние чувствительны к выбросам
- Медианы устойчивы к выбросам
- Квантили помогают следить за определённым сегментом

Математические трюки

Есть различные математические трюки, призванные улучшить свойства метрик:

- Сложные составные метрики с различными весами для составных частей
- CUPED (техника для увеличения чувствительности)
- Стратификация, разбиение пользователей на когорты
- Различные трансформации данных

! Разработка подобных метрик осуществляется под конкретный сервис

Математические трюки

- После математических трюков, на АА-тестах метрика не должна показывать значимые изменения чаще, чем в α процентах случаев, иначе мы сделали что-то странное
- Если преобразование оказалось успешным, оно должно быть проинтерпретировано

Резюме

- Метрики используются, чтобы понять, что изменилось в нашем сервисе
- Метрики обладают различными свойствами
- Нужно аккуратно подбирать их для проведения эксперимента