

# 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



# Что такое АБ-тест

- 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



# 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



# Типичные метрики

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

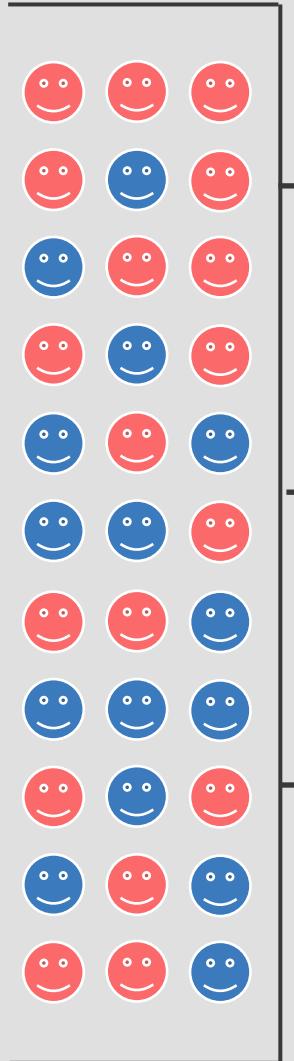


# Где используются АБ-тесты

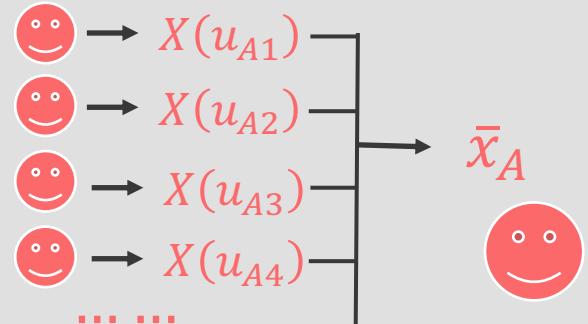
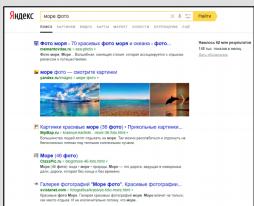
- Изменение дизайна на сайте
- Изменение функциональности в играх
- Работоспособность лекарств
- Выкатка нового алгоритма машинного обучения
- Изменения в онлайн-магазинах: смена порядка отделов, раскладки товаров, установка постаматов, промо-акции



# Процедура АБ-тестирования

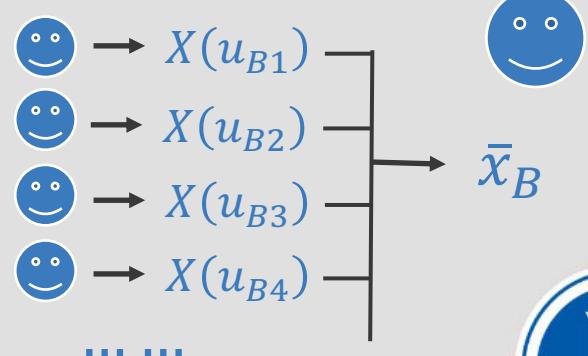
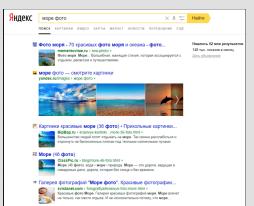


Вариант для А  
(продакшн-версия)



$X(u)$  – число поисковых  
сессий юзера  $u$

Вариант для В  
(изменённая версия)



# Процедура АБ-тестирования



$\bar{x}_A$

Статистика для  
группы А

Вычисляем  
критерий для  
оценки



$\bar{x}_B$

Статистика для  
группы В

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

Существенность

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

позитивные  
или  
негативные  
изменения

Принимаем  
решение

Значимость

Статистический  
тест

разница  
вызвана шумом  
или  
изменениями



# Значимость и существенность

- Интерес представляет не величина р-значения, а **размер эффекта** - степень отклонения данных от нулевой гипотезы
- На основе размера эффекта делается вывод о том, являются ли результаты **практически существенными**

**Значимо, но несущественно:** за три года женщины, упражнявшиеся не меньше часа в день, набрали значительно меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день. Разница в набранном весе составила 150 г.

**Практическая значимость такого эффекта сомнительна.**

<https://jamanetwork.com/journals/jama/fullarticle/185585>



# Значимость и существенность

**Значимо и существенно:** в 2002 году клинические испытания гормонального препарата Премарин, облегчающего симптомы менопаузы, были досрочно прерваны.

Было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на **0.08%**, риска инсульта на **0.08%** и инфаркта на **0.07%**

Формально эффект крайне мал, но с учётом численности населения он превращается в тысячи дополнительных смертей.



# Значимость и существенность

**Незначимо и существенно:** лекарство, замедляющее ослабление интеллекта больных Альцгеймером.

При испытаниях выяснилось, что разница в IQ контрольной и тестовой групп составляет 13 пунктов. При этом она статистически незначима.

Возможно, выборка мала для детекции эффекта и изучение лекарства стоит продолжить.

<https://journals.sagepub.com/doi/abs/10.1177/0013164496056005002>



# 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?



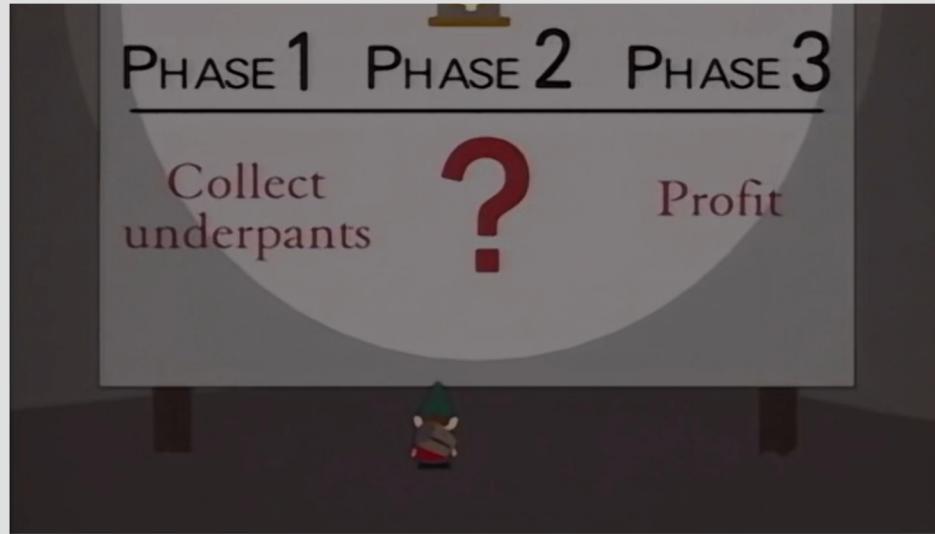
# Подводные камни АБ-тестирования



# Схема АБ-теста

Фаза 1: планирование эксперимента и его дизайна.

Фаза 2: сбор статистики и проверка гипотез



- ❗ Плохо спланированный дизайн эксперимента может привести к неверным выводам



# Кекс про кока-колу

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



Что пошло не так?



# Кекс про кока-колу

- Исследование проходило не в тех же самых 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?

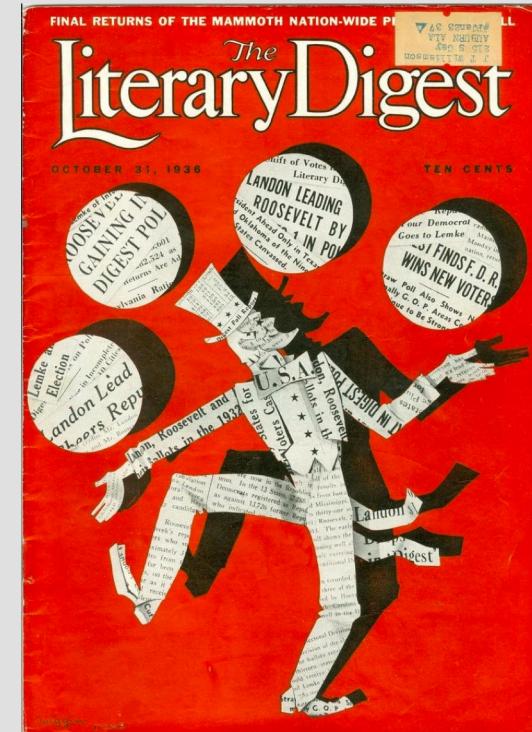


Что угодно!



# 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.



# 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



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

**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



# 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 `p_value`
- **Решение:** дождаться конца теста либо использовать специальные методологии. Например, байесовских многоруких бандитов (их обсудим позже).

<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



# Optimism bias

- Мы недооцениваем вероятности плохих событий
- Когда метрика изменяется в **плохом направлении**, мы ищем проблемы
- Когда метрика изменяется в **хорошем направлении**, мы просто принимаем этот факт

What are the odds?



Source: Techjuice



У нас есть предрасположенность подвергать проверкам только неприятные выводы



## 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
- Неравномерный отбор пользователей в эксперимент искажает картину
- Пранки иногда выходят из-под контроля: 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
- 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.

## Пример: онлайн-ритейл

- Ограничение на количество магазинов
- Магазины сильно отличаются друг от друга: магазин Перекрёсток в спальном районе и у бизнес-центра - это совершенно разные магазины (элементы выборки не из одного распределения, а из разных)
- АБ нужно проводить во времени: праздники, сезонность, промо-акции и т.п.



# Оффлайн АБ-тесты: ритейл

- Элементы выборки зависимы (чеки внутри магазина зависят друг от друга)
- Неоднородность магазинов: у каждого магазина своё среднее значение, свой размер и трафик
- Неоднородность по погоде: в разные погодные условия разный трафик
- Неоднородность по времени: в течение суток, по дням недели и времени года



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



# Оффлайн АБ-тесты: ритейл

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



⇒ нужны специальные приёмы, которые помогут всё сделать корректно: KNN для подбора похожих групп, АА-тесты, разные симуляции и их проверка на значимость.

<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

$\varepsilon_i$  — прочие факторы, влияющие на результаты обучения в  $i$  —ом округе

$\beta_0$  — константа,  $\beta_1$  — коэффициент наклона



# 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 несмещённые, состоятельные и эффективные оценки для этих взаимосвязей
- Изучает методы оценки причинно-следственных связей и свойства этих методов



# Резюме

- В офлайне АБ-тесты делать сложнее, чем в онлайне
- На каждом этапе проведение 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



<https://habr.com/ru/company/ods/blog/325416/>

<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



<https://habr.com/ru/company/ods/blog/325416/>

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



# История о Зомби-лососе

- Аппарат МРТ возвращает кучу данных
- Один объект – воксель
- Чтобы убедиться, что в мозгу нет реакции, надо проверить гипотезу относительно каждого вокселя

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

<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  $\alpha$

Можно ошибиться сразу в двух местах:

$$\begin{aligned} & \mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0) \\ &= 1 - \mathbb{P}(\text{не ошибиться ни в одной}) = 1 - (1 - \alpha)^2 \\ &= 1 - (1 - 2\alpha + \alpha^2) = 2\alpha - \alpha^2 > \alpha \end{aligned}$$

$$\alpha_i = 0.05 \quad \Rightarrow \quad \alpha = 0.1 - 0.25 = 0.75 > 0.5$$



Вероятность ошибки первого рода  
накапливается и выходит из-под контроля



# Множественная проверка гипотез

- Требуется протестировать несколько изменений и их кумулятивное воздействие на продуктовые метрики.

**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  $\frac{\alpha}{2}$

$$\begin{aligned}\alpha &= \mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0) \\ &\leq \mathbb{P}(\text{ош. в 1}) + \mathbb{P}(\text{ош. во 2}) = \frac{\alpha_i}{2} + \frac{\alpha_i}{2} = \alpha_i\end{aligned}$$

- 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)
- **Метод Холма** - улучшение поправки Бонферрони, обладает более высокой мощностью
- **Бонферрони:** проверяем  $k$  гипотез на уровнях значимости

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

- **Метод Холма:** 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 – **нисходящая процедура**, так как гипотезы проверяются по убыванию значимости
- Обеспечивает контроль  $FWER$  на уровне  $\alpha$
- Метод Холма оказывается мощнее корректировки Бонферрони, так как его уровни значимости меньше
- Для любой процедуры множественного тестирования гипотез  $FDR \leq FWER$



# Метод Бенджамини-Хохберга

- Метод Бенджамини-Хохберга – **восходящая процедура**, так как гипотезы проверяются по возрастанию значимости
- Восходящая процедура всегда отвергает не меньше гипотез, чем нисходящая с теми же  $\alpha_i$
- Метод Бенджамини-Хохберга обычно оказывается более мощным, чем методы контролирующие *FWER*
- Это происходит за счёт того, что метод позволяет допустить большее число ошибок первого рода



# 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



# 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	ok	$\beta$	ошибка 2 рода
$H_0$ отвергается	$\alpha$	ok	ошибка 1 рода

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

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

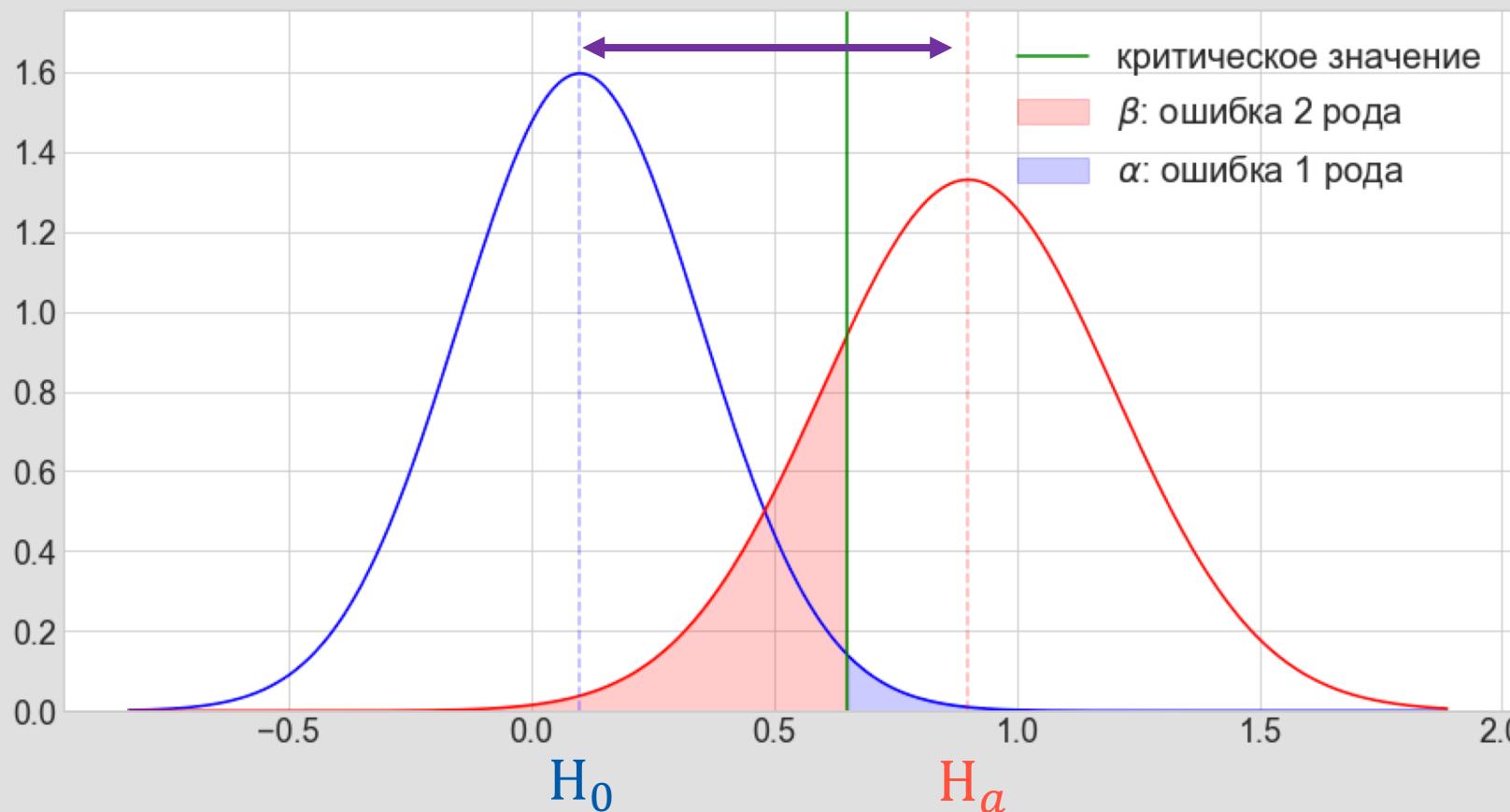
Величину  $1 - \beta$  называют **мощностью** критерия



# Размер эффекта

Чем больше размер эффекта, тем меньше наблюдений нужно, чтобы его уловить

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 tradeoff между минимальным размером эффекта, желаемой мощностью и объёмом выборки
- В выборе tradeoff помогает заказчик эксперимента, у него обычно есть ограничения, с которыми нам придётся работать (количество магазинов, длительность пилота и т.п.)



# Таблица эффекта-ошибки

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

❗ Совокупность этих трёх параметров (ошибка  $\frac{1}{2}$  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:

$$z = \frac{\hat{p}_a - \hat{p}_0}{\sqrt{P(1 - P) \cdot \left(\frac{1}{n} + \frac{1}{n}\right)}} \stackrel{\text{as}y}{\underset{H_0}{\sim}} N(0, 1)$$



# 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:

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



# 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



# Метрики для АБ



# 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



# Роли метрик

- **Метрики, отражающие ключевые цели сервиса** – должны обладать почти безусловной направленностью
- **Основные критерии оценки изменений** – должны быть согласованы с ключевыми метриками, должны быть чувствительными
- **Ограничительные метрики** – должны быть чувствительны и согласованы с тем, что нельзя ломать
- **Целевые метрики эксперимента** – должны выбираться в зависимости от смысла эксперимента



# Желательные свойства метрик

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



# Желательные свойства метрик

- **Чувствительность (sensitivity)** – способность метрики обнаруживать статистически значимую разницу, когда она есть.
- Чем выше чувствительность, тем меньшие изменения метрики могут быть обнаружены.

**Пример:** метрики, основанные на деньгах слабо реагируют на изменения



# Желательные свойства метрик

- **Стабильность** – должны быть чувствительны и согласованы с тем, что нельзя ломать
- Если у метрики высокая дисперсия, 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 юзера:

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

Имеют чёткую  
направленность

Хорошие предикторы  
для долгосрочного  
успеха продукта

Обладают слабой  
чувствительностью

## Активность юзера:

- Число кликов за сессию
- Длина пользовательской сессии

Обладают  
неоднозначной  
направленностью

Обладают сильной  
чувствительностью



# 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 кликбейтного контента.

- Метрики с чёткой интерпретацией часто обладают низкой чувствительностью
- Её повышение так, чтобы не испортить направленность – отдельная сложная задача



# 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 и тп
- Метрики, рассчитанные как средние могут быть подвержены выбросам, среднее может быть не самой важной величиной ⇒ можно использовать медианы (устойчивы к выбросам), квантили (когда надо следить за определённым сегментом)



# Математические хаки

Есть различные математические хаки, призванные улучшить свойства метрик:

- Сложные составные метрики с различными весами для составных частей
- CUPED
- Стратификация, разбиение пользователей на кагорты
- Различные трансформации данных



Разработка подобных метрик осуществляется под конкретный сервис



# Математические хаки

- После математических хаков, на АА-тестах метрика не должна краситься чаще, чем в  $\alpha$  процентах случаев, иначе мы сделали что-то странное
- Если преобразование оказалось успешным, оно должно быть проинтерпретировано

