ТПОЭ - 23/24

Лекция 3

Нерсес Багиян

Мы сформулировали дизайн эксперимента

1. Формулировка гипотезы с сформулированным ожидаемым размером эффекта

   Добавление выделенного раздела для экологичных мыл на главной странице увеличит конверсию с главной страницы на страницы товаров на 15% и снизит отказы на главной странице на 10%.

2. Описание аудитории

   Посетители онлайн-магазина мыла, включая новых и возвращающихся пользователей.

3. Описание вариантов с размером каждой группы

   Контрольная группа (A): Главная страница без выделенного раздела для экологичных мыл. Размер группы – 50% посетителей.
   Экспериментальная группа (B): Главная страница с выделенным разделом для экологичных мыл. Размер группы – 50% посетителей.

4. Ожидаемые исходы и метрики

   Основная метрика: Увеличение конверсии с главной страницы на страницы товаров.
   Guardrail метрика: Конверсия из посещения главной страницы в активную сессию

5. Продолжительность

   Эксперимент продлится 4 недели, чтобы собрать достаточно данных для статистически значимых результатов, учитывая недельные колебания трафика и поведения покупателей.

6. Результаты

   TBD

Пример дизайн дока: https://docs.google.com/document/d/1y1uawhMsQ3tyZ8dI8W9jDmOWMICPnZhJnsnayt9UX5c/edit?usp=sharing

Как думаете что дальше?

Вопрос аудитории

В целом есть два подхода

Утром деньги вечером стулья

Подход 1

1. Выбираем группы
2. Подбираем размер

Подход 2

1. Подбираем размер
2. Выбираем группы

Рассмотрим ситуацию номер 1

Мы работаем в онлайн-магазине продажи мыла. В целом пользователей у нас немного, 1000 в неделю, что нам лучше делать?

Из-за того, что пользователей немного:

- У нас в целом не очень много событий
- Трафика на сайт тоже приходит немного
- Мы не можем тестировать больше 1 гипотезы

Рассмотрим ситуацию номер 2

Мы работаем на маркетплейсе. У нас миллионы пользователей, что нам лучше делать?

Из-за того, что пользователей много, что это значит:

- У нас много команд
- Много событий
- Мы проверяем очень-очень много гипотез

Разные компании - разные ситуации

Утром деньги вечером стулья

Smol

- Мало событий
- Размер группы всегда будет максимально возможный
- Поэтому лучше сразу начинать с похожих групп

Big

- Много событий
- Как и много гипотез
- Поэтому нужно сделать максимально универсальный алгоритм подбора групп

Вывод: всегда лучше фокусироваться на подборе групп, а затем тюнить размер

А что такое похожие группы?

Вопрос аудитории

Давайте разберемся на примере. Какие группы могли быть тут?

1. Формулировка гипотезы с сформулированным ожидаемым размером эффекта

   Добавление выделенного раздела для экологичных мыл на главной странице увеличит конверсию с главной страницы на страницы товаров на 15% и снизит отказы на главной странице на 10%.

2. Описание аудитории

   Посетители онлайн-магазина мыла, включая новых и возвращающихся пользователей.

3. Описание вариантов с размером каждой группы

   Контрольная группа (A): Главная страница без выделенного раздела для экологичных мыл. Размер группы – 50% посетителей.
   Экспериментальная группа (B): Главная страница с выделенным разделом для экологичных мыл. Размер группы – 50% посетителей.

4. Ожидаемые исходы и метрики

   Основная метрика: Увеличение конверсии с главной страницы на страницы товаров.
   Guardrail метрика: Конверсия из посещения главной страницы в активную сессию

5. Продолжительность

   Эксперимент продлится 4 недели, чтобы собрать достаточно данных для статистически значимых результатов, учитывая недельные колебания трафика и поведения покупателей.

6. Результаты

   TBD

Пример дизайн дока: https://docs.google.com/document/d/1y1uawhMsQ3tyZ8dI8W9jDmOWMICPnZhJnsnayt9UX5c/edit?usp=sharing

Что вы могли назвать в качестве признаков?

Тут есть фантазия разгуляться:

1. Возраст
2. Пол
3. Девайсы
4. Время жизни
5. Количество покупок
6. Частота покупок
7. Средний чек
8. Конверсии

А как вообще сравнить, что две группы похожи?

Вопрос аудитории

Кто это и чем картинки связаны?

Гигант мысли, отец русской демократии

Уильям Госсет - британский ученый, работал на пивоварне Гиннесс над качеством пива. Любил публиковать различные статистические находки в журналах по математике. Одной из работы был подход к сравнению средней урожайности, в которой он использовал свою t статистику

Статистический тест - что это?

Абстрактные размышления о статистике

Если рассуждать абстрактно, статистический тест это способ получить ответ “да” или “нет” на некоторый вопрос сформулированный в виде гипотезы.

Пример: уголовное дело

Идет некоторый уголовный процесс, проводится заседание суда. Обвинитель пытается доказать, что обвиняемый виновен.

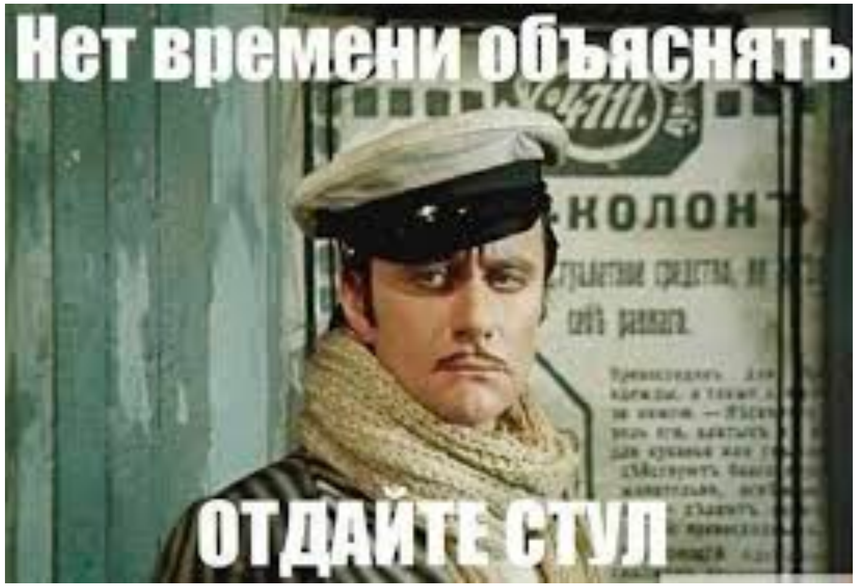
По дефолту предполагается, что обвиняемый невиновный и итог суда может быть одним из двух:

1. Доказали
2. Не доказали

Пример: ищем стул с бриллиантами

Представим себе ситуацию, у нас есть стулья и в одном из них есть бриллианты. А также есть эксперт, который видел кучу стульев без бриллиантов.

Мы приносим по очереди стулья и на основе каких-то своих критериев эксперт говорит: стул с бриллиантом или без



Статистический тест - что это?

Продолжаем менее абстрактные размышления о статистике

Первое, что делается в любом тесте это формулируется некоторая гипотеза, которую мы проверяем.

$$H_0 : \theta \in \Theta$$

$$H_1 : \theta \notin \Theta$$

H_0 : Стул без бриллианта

H_1 : Стул с бриллиантом

Статистический тест - что это?

Продолжаем менее абстрактные размышления о статистике

Откуда берется этот самый эксперт? Эксперт у нас бывалый, много всяких стульев в жизни видел, он точно знает и ведет себя как эксперт.

$$H_0 : \theta \in \Theta$$
$$H_1 : \theta \notin \Theta$$

H_0 : Стул без бриллианта

H_1 : Стул с бриллиантом

Нашим экспертом является некоторая статистика теста, которая “видела" много случаев, когда H_0 верна.

Соответственно: $t \sim F \mid H_0$ is true

Если ~~Остап~~ эксперт, считает, что стул **похож**: “Да я таких стульев кучу видел, отвечаю вам, что дело в шляпе”

Если эксперт, считает, что стул **не похож**: “Ну я чет хз... не похоже”

Вернемся к чуть более формальному языку

Статистический тест - правило, по которому принимается решение о том принимать или не принимать нулевую гипотезу. Это правило строится на предположении о том, что статистика теста имеет некоторое распределение, если нулевая гипотеза верна. Разберемся на картинке:

p-value

Магическая цифра, которая говорит, что нам делать

p-value - вероятность получить, такое или более экстремальное значение статистики при условии, что нулевая гипотеза верна

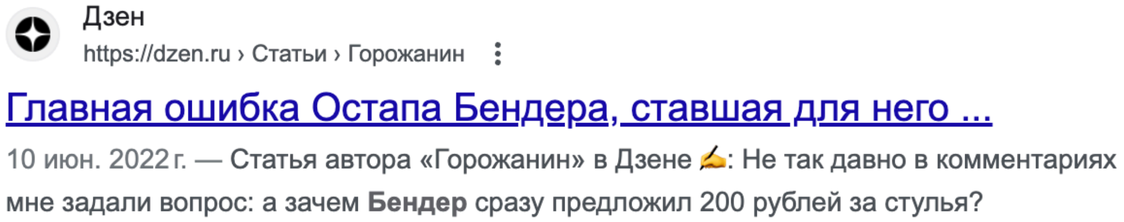
Вернемся к стульям:
p-стула - а с какой вероятностью я получу такой же или еще более не похожий стул при условии, что я видел только стулья без бриллиантов

$$\text{p} = P(t >= T \,|\, H_0 \text{ is true})$$

Как думаете такой эксперт всегда говорит правду?

Вопрос аудитории

Конечно нет, он может ошибиться



Как это может случиться:

- Сказать, что в стуле есть бриллиант, хотя его нет (ошибка 1-ого рода)
- Сказать, что в стуле нет бриллианта, хотя он есть (ошибка 2-ого рода)

Ошибки 1-ого и 2-ого рода

Почему они так называются

Ошибка первого рода — ситуация, когда отвергнута верная нулевая гипотеза. Она первого рода, потому что мы предполагаем, что по дефолту нулевая гипотеза верна

Ошибка второго рода — ситуация, когда принята неверная нулевая гипотеза. Она второго рода, потому что мы не предполагаем, что по дефолту альтернативная гипотеза верна (ТАК НЕЛЬЗЯ ГОВОРИТЬ, но мне можно)

здесь должен быть баянистый мем, но я передумал

Обычно работают с вероятностями ошибок

Вероятность ошибки первого рода — обозначается как α. Показывает вероятность отвергнуть нулевую гипотезу, если она верна. **Отвечает на вопрос:** как часто эксперт скажет, что стул с бриллиантом, если его там на самом деле нет?

Мощность (1 - вер-ть ошибки второго рода) — обозначается, как $(1 - \beta)$. Показывает вероятность отвергнуть нулевую гипотезу, если она не верна. **Отвечает на вопрос:** как часто эксперт скажет, что стул с бриллиантом, если он там есть?

Чем вероятность ошибки первого рода отличается от p-value?

Вопрос аудитории

Вероятность ошибки первого рода — обозначается как α. Показывает вероятность отвергнуть нулевую гипотезу, если она верна. **Отвечает на вопрос:** как часто эксперт скажет, что стул с бриллиантом, если его там на самом деле нет?

p-value - вероятность получить, такое или более экстремальное значение статистики при условии, что нулевая гипотеза верна

Вернемся к нашему дизайну эксперимента

1. Формулировка гипотезы с сформулированным ожидаемым размером эффекта

   Добавление выделенного раздела для экологичных мыл на главной странице увеличит конверсию с главной страницы на страницы товаров на 15% и снизит отказы на главной странице на 10%.

2. Описание аудитории

   Посетители онлайн-магазина мыла, включая новых и возвращающихся пользователей.

3. Описание вариантов с размером каждой группы

   Контрольная группа (A): Главная страница без выделенного раздела для экологичных мыл. Размер группы – 50% посетителей.
   Экспериментальная группа (B): Главная страница с выделенным разделом для экологичных мыл. Размер группы – 50% посетителей.

4. Ожидаемые исходы и метрики

   Основная метрика: Увеличение конверсии с главной страницы на страницы товаров.
   Guardrail метрика: Конверсия из посещения главной страницы в активную сессию

5. Продолжительность

   Эксперимент продлится 4 недели, чтобы собрать достаточно данных для статистически значимых результатов, учитывая недельные колебания трафика и поведения покупателей.

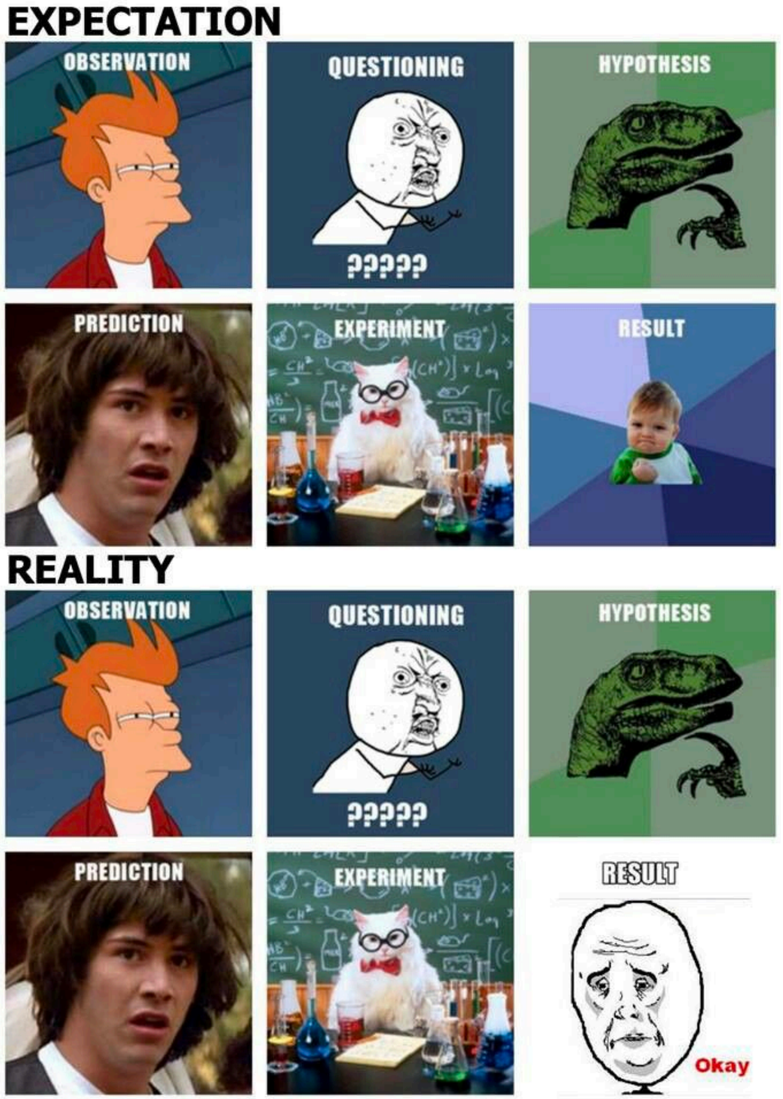
6. Результаты

   TBD

Пример дизайн дока: https://docs.google.com/document/d/1y1uawhMsQ3tyZ8dI8W9jDmOWMICPnZhJnsnayt9UX5c/edit?usp=sharing

Напоминание: зачем нужен эксперимент и его дизайн

- Гипотеза — это предположение, которое описывает феномен (факт, событие, явление, etc), и может быть проверено в ходе эксперимента
  - Важно отличие гипотезы от идеи заключается в том, что в основе гипотезы лежат наблюдения
- Эксперимент - научный метод проверить эту гипотезу и найти потенциальные противоречия



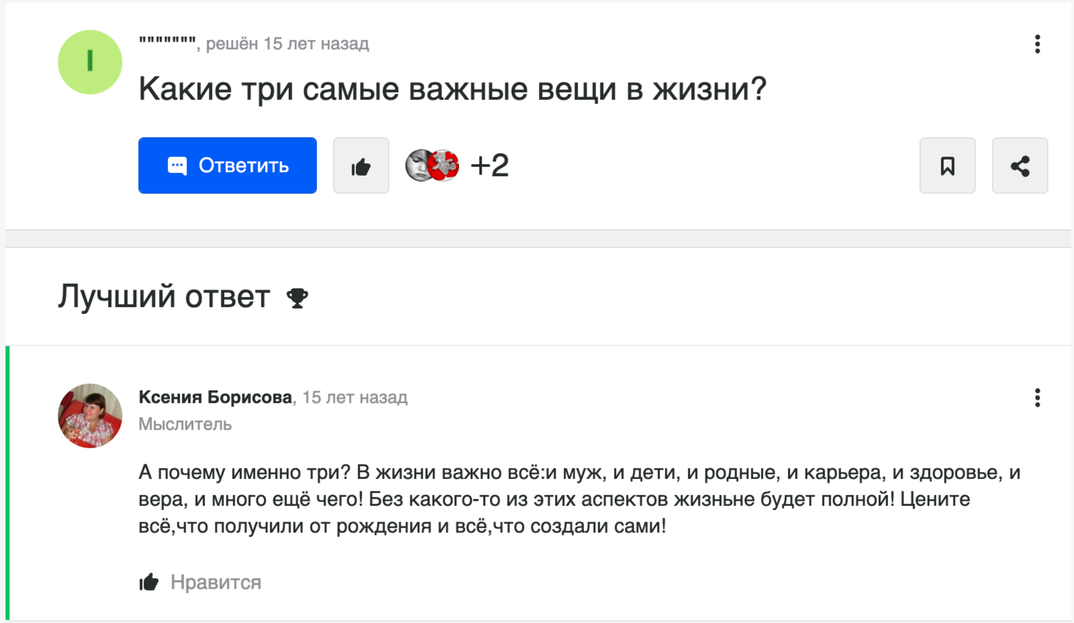
А что нам тогда важно в эксперименте?

Вопрос аудитории

Важны две вещи: мощность и ошибка 1-ого рода

1. **Вероятность ошибки 1-ого рода** - мы не хотим краснеть перед продактом, когда скажем, что в эксперименте был эффект, хотя после лесенки на графике не было

2. **Мощность** - при этом мы хотим по-максимуму находить все успешные гипотезы, которые были



Как сделать так, чтобы ошибка первого рода была нулевой?

Вопрос аудитории

А мощность 100%?

Вопрос аудитории

Промежуточные итоги: что узнали

Нужны две похожие группы

Чтобы понять группы похожи или нет нужно выбрать метрику

Статистический тест

Нужен, чтобы сравнить две группы между собой

У теста есть два параметра

Вер-ть ошибки первого рода и мощность

Как это применимо к гипотезам?

Мы не хотим краснеть перед начальством (type 1 error), но при этом хотим по максимуму находить рабочие гипотезы

Type I error

Фиксируем на приемлемом уровне для нас

Power

Стараемся максимизировать при таком уровне ошибки первого рода

Сплит система

Ключевая вещь, которая отвечает за подбор групп в А/Б

Сплит система - маленький заводик по производству похожих групп. Ее задача производить максимально похожие группы по заданным входным параметрам.

Пример самой простой сплит системы?

Вопрос аудитории

Пример: А/Б в поиске

Вы аналитик, работаете в некоторой компании в уездном городе N и у вас много людей, которые генерируют гипотезы и постоянно их хотят проверять. Что вам делать?

В целом есть два ~~стула~~ концептуальных варианта:

- Забить всю рабочую неделю задачами по подбору групп
- Сделать универсальный алгоритм по подбору, который бы это делал за вас (сплит систему)

Как проверить, что сплит система работает нормально?

Вопрос аудитории

Сплит система

Ключевая вещь, которая отвечает за подбор групп в А/Б

Сплит система - маленький заводик по производству похожих групп. Ее задача производить максимально похожие группы по заданным входным параметрам.

Качества хорошей сплит системы:

- Генерирует стабильные во времени одинаковые группы (с точки зрения ошибки 1-ого рода)
- Генерирует стабильные одинаковые группы при бесконечном количестве разбиений (с точки зрения ошибки 1-ого рода)

Как проверить качество сплит системы?

Критерий на основе распределения p-value

Пусть есть некоторая тест статистика T, при верности нулевой гипотезы она имеет распределение $F(t)$, как мы уже выяснили $p_{val} = F(T)$. Нам нужно понять, а какое распределение имеет $F_p(p_{val}) = P(p_{val} < p)$

$$P(p_{val} < p) = P(F(T) < p) = P(T < F^{-1}(p))$$

Заметим, что это по определению является $F(t) = P(T < t)$

$$P(T < F^{-1}(p)) = F(F^{-1}(p)) = p$$

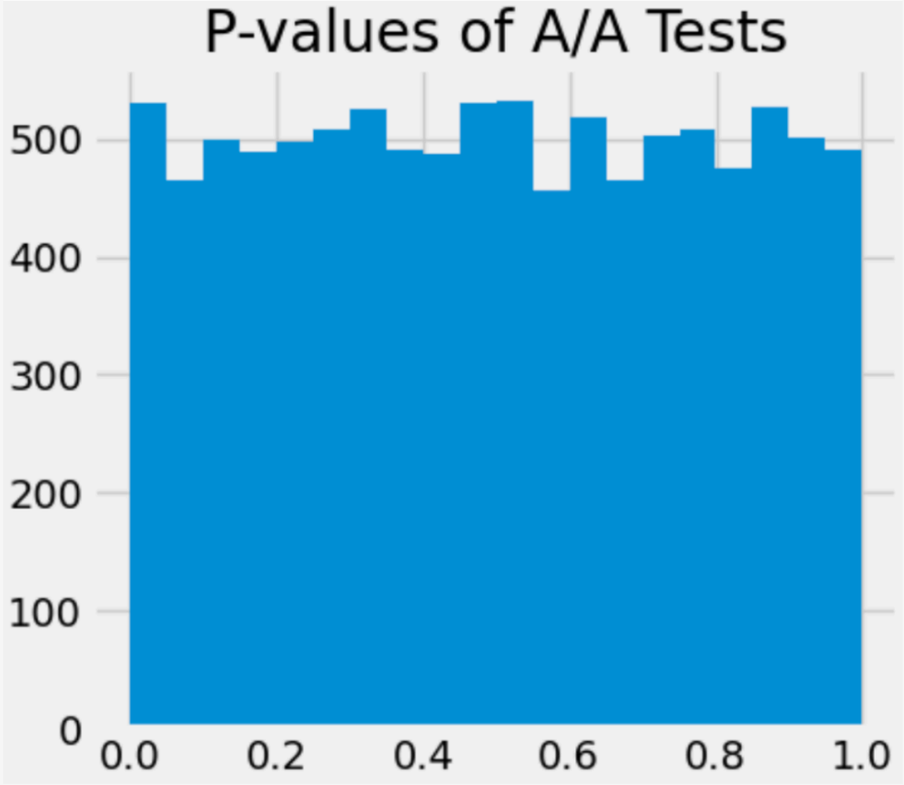
Что является равномерным распределением по определению

A/A тест

Способ проверить, что группы подбираются нормально

Пусть сплит система на основе исторических данных выдала две группы, которые не участвовали в эксперименте. Если мы сравним их между собой - мы проведем A/A тест. Тест, где между двумя группами нет изменений (нулевая гипотеза верна).

Если провести много-много таких тестов, то распределение будет равномерным и это значит, что сплит система работает правильно с точки зрения ошибки 1-ого рода



Как мощность проверить?

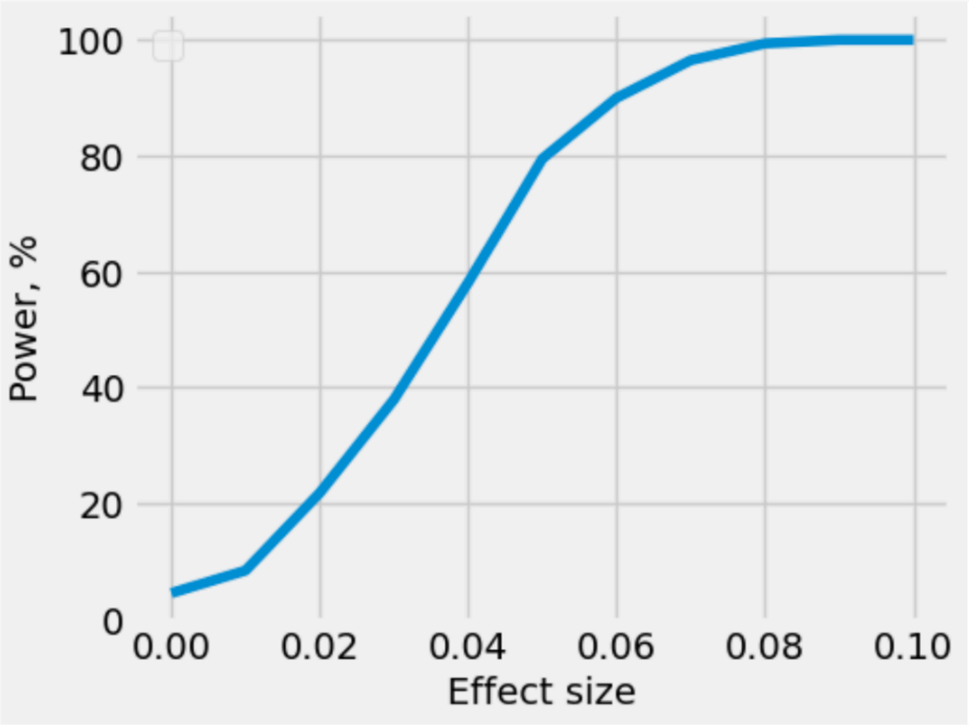
Вопрос аудитории

Синтетический A/B тест

Способ смоделировать реальный эксперимент

Алгоритм:

1. Берем две группы из сплит системы
2. Генерируем шум, чтобы добавить одной из групп
3. Считаем метрики, натравливаем тест
4. Повторяем много-много раз с разными значениями среднего у шума (это будет размер эффекта)
5. Получаем картинку как справа



Итого

1. После составления дизайна и выбора метрик нам надо подобрать одинаковые группы
2. За подбор групп отвечает сплит система
3. Главные параметры сплит системы - вероятность ошибки первого рода и мощность
4. Мы фиксируем вероятность ошибки первого рода и стараемся максимизировать мощность
5. Чтобы экспериментировать с этим можно и нужно использовать алгоритмы валидации: AA тест и синтетический A/B
6. Если все окей и вы удовлетворили требованиям, то можем двигаться дальше по дизайну