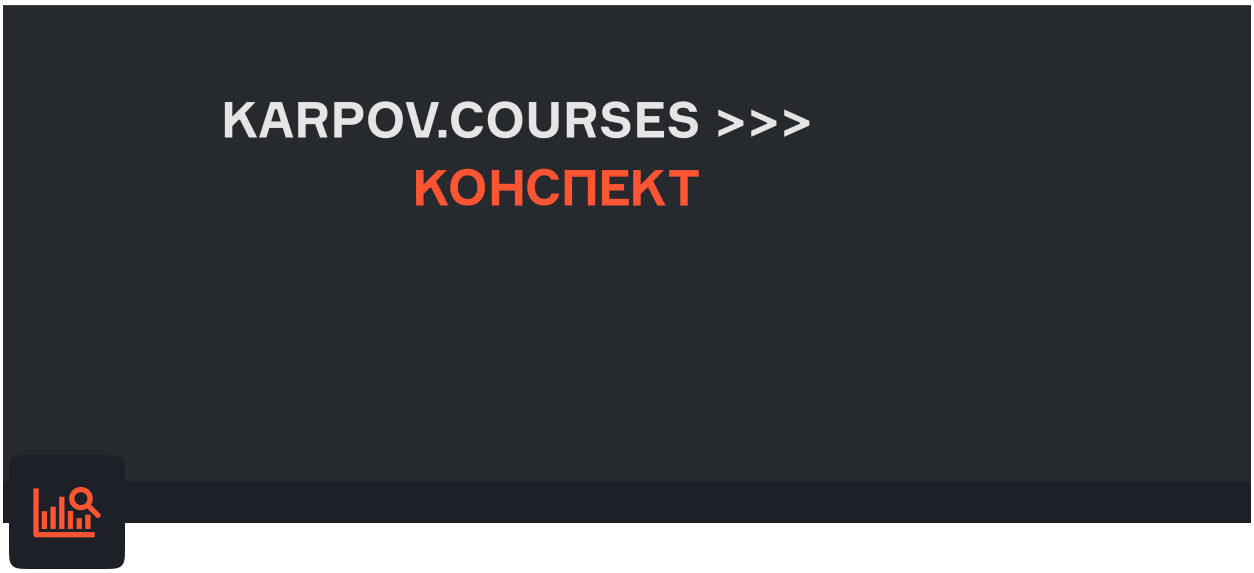


> Конспект > 8 урок > СТАТИСТИКА

> Оглавление

> Идеи и гипотезы

Всё начинается с идеи

Эксперимент – конечная точка, когда у нас уже сформулирована некоторая гипотеза, каким-то образом валидирована. Мы понимаем на какую часть продукта она влияет и какой метрикой что нужно проверить. Но перед тем, как формулируется гипотеза, откуда то берется идея. Откуда?

Идея – ещё не подтвержденная гипотеза. Гипотеза рождается на основе данных. Сами данные могут быть получены разными способами.
Например, путем изучения конкурентов, анализом потенциала рынка, или с использованием творческих механик (воркшопы, хакатоны).

Каждой идее – свой метод проверки

Помимо количественных методов анализа данных могут быть:

- **Глубинные интервью** – общаемся с пользователями, выясняем проблемы, рассуждаем
- **Опросы** – спрашиваем что-либо очно или по дистанционному каналу (напр. рассылка)
- **Юзабилити-тестирования**

Проверке гипотезы необходимы условия

- **Отсутствие влияющих факторов, кроме самого тестируемого изменения:** когда мы хотим сравнить две версии, должно влиять только изменение, никаких других внешних факторов.
- **Репрезентативная оценка:** репрезентативность – соответствие свойств выборки характеристикам генеральной совокупности, т.е. когда выборка отражает свойства генеральной совокупности. Пример в данном контексте: есть выборка А и выборка Б. В выборке А есть пользователи из Москвы. Если выборка репрезентативна, то и в А и в Б пользователей из Москвы будет по 30%. Если в А будет 25%, а в Б – 10%, то есть смещение и выборки не репрезентативны. Это критично для проведения честного A/B теста: если есть дисбаланс, то результат эксперимента может объясняться не тем, что у нас хорошая гипотеза, а тем, что нарушается репрезентативность (напр. другой покупательской способностью).
- **Точность оценки:** за точность оценки отвечают статистические критерии, плотность и количество данных, параметры распределения метрик. Точность

влияет на ошибки и на качество экспериментов.

> Что такое A/B тесты

С чего все начиналось? Статистика как инструмент использовалась для клинических исследований. Благодаря этим задачам математическая статистика развивалась.

A/B тестирование – это где:

- проверяется два и более варианта (контроль и тест) с целью определения наиболее эффективного
- степень эффективности измеряется с помощью посчитанных вероятностей ложноположительных и ложноотрицательных случаев

Решения

1. **Ложноположительные** (False Positive, ошибка I рода): публикуем бесполезные изменения, которые на самом деле не работают.
2. **Ложноотрицательные** (False Negative, ошибка II рода): полезные изменения упускаются из виду. Например, зафиксировали результат как не значимый, хотя он значим.

Примеры:

1. **Ложноположительное решение**

Гипотеза: повышение недельной цены подписки с 1до 2

Итог: на первых 2 днях эксперимента был зафиксирован статистически значимый результат. Продакт и аналитик приняли решение принять результат как успешный. А после публикации изменения на всех пользователей – ключевой показатель изменился в худшую сторону. Так, приняли за истину то, что ей не является.

2. **Ложноотрицательное решение**

Проводили эксперимент 2 дня, не видели разницы, остановили

Итог: т.к. в эксперименте было охвачено только 2 дня, мы не учитываем поведение аудитории в остальные дни недели. Возможно, изменение имеет

отложенный эффект: пользователь в понедельник попал в тестовую группу, а в пятницу принял решение

> Метрики эксперимента: уровни

- **Целевые** – показатели, на которые направлено изменение (конверсия, средний чек)
- **Опережающие** – показатели, хорошо коррелируемые с целевыми, дающие предикт и полезны тогда, когда нет времени ждать основную метрику
- **Guardrail** – показатели, на которые направленно влияет изменение, но не являющиеся целевыми. Рекомендуется за ними наблюдать и на их основе в том числе принимать решение

Пример системы уровней

Пример: e-commerce, тест нового UI корзины

- *Целевые*: конверсия в покупку, средний чек, ARPU (средний доход на пользователя), ARPPU (средний доход на платящего пользователя)
- *Опережающие*: добавление товара в корзину на сессию, просмотры товаров на сессию, отток чекаута, ошибки на чекауте
- *Guardrail*: время от входа в корзину до ее прохождения, доля поисковых запросов из корзины, взаимодействие с рекомендательными блоками в корзине

Пример: образовательный продукт, тест нового образовательного контента

- *Целевые*: продление обучения, средний доход на платящего пользователя (ARPPU)
- *Опережающие*: интенсивность обучения, кол-во ошибок в момент обучения, частота обращений в службу поддержки, технические характеристики качества видео
- *Guardrail*: время проведенное за одним занятием, прерывание занятий, перемотка

> Метрики эксперимента: типы

Доли

– величина, которая берется из бинарного признака, когда есть две градации – 0 или 1: регистрации, удержание на 7 день [0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1]

Непрерывные

– среднее выборочное из некоторого распределения непрерывных величин: время в сек./мин./т.п., чек в рублях [1123.32, 324.4, 823.21, 924.91]

Отношения

– отношение двух случайных величин: поездок на водителя, кликов на сессии, цена за 1000 показов [10/123 = 0.081, 4129.2/12488 = 0.33, 1/100=0.01]

> Приоритет метрик

Система координат метрик – позволяет лучше и быстрее принимать решения, когда метрик много. Если нет понимания приоритета – принять решение достаточно сложно.

На разных стадиях развития продукта система метрик может значительно меняться – это динамический артефакт, который должен постоянно версионировать.

долгие метрики vs. быстрые метрики

У каждой метрики есть "окно" – пользователь не сразу принимает решение о том, чтобы совершить желанное действие для продукта. Важно искать "быстрые" метрики, которые зависимы к "долгим" и на основе этого менять приоритет и всю иерархию.

> Типы экспериментов

A/B

Чем полезен: Измерить эффект от изменения

Ключевые особенности:

- каждая группа эксперимента видит свой вариант
- группы независимы
- группы взяты из одной ГС
- распределение может быть неравномерным

A/A

Чем полезен:

- проверить сплит-систему
- выбрать гомогенные группы

Ключевые особенности:

- группы независимы
- группы взяты из одной ГС
- часто используются для симуляций

A/B/C/...

Чем полезен: Тот же A/B, только проверяется от 2 и более изменений.

Ключевые особенности:

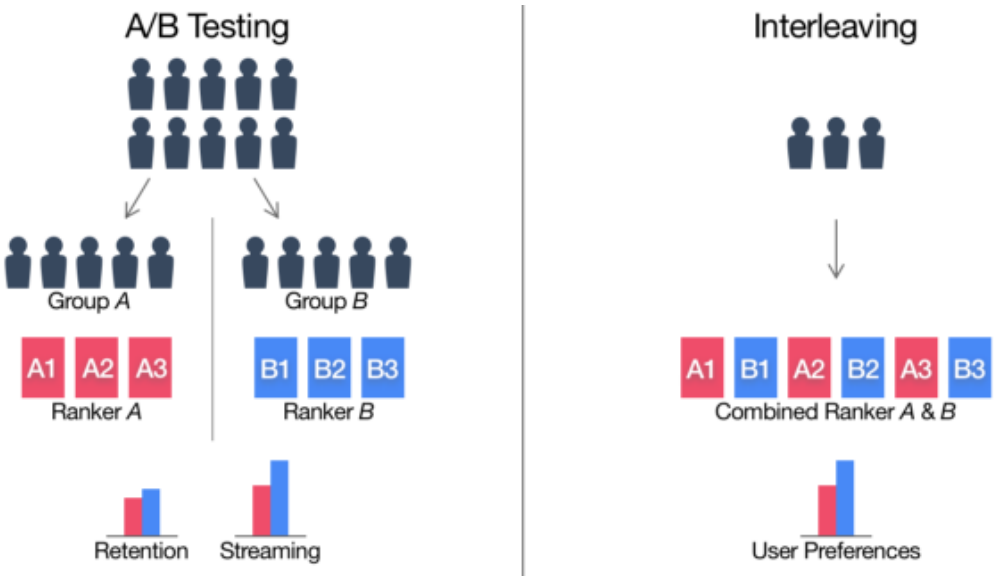
- группы независимы
- группы взяты из одной ГС
- сопряжена с проблемой множественной проверки гипотез

TDI (team draft interleaving)

Чем полезен: изменение в ранжированных списках

Ключевые особенности:

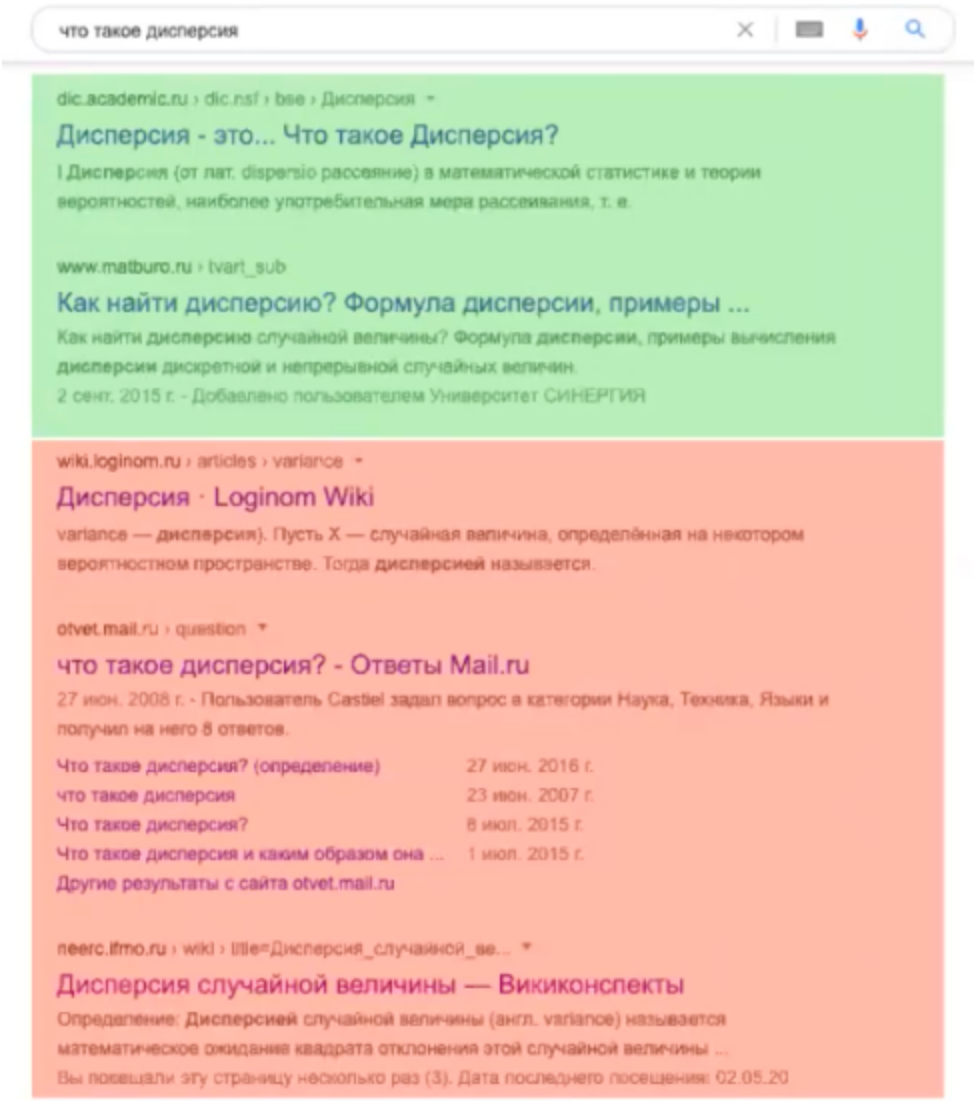
- один пользователь видит сразу несколько вариантов
- чаще всего используется в поиске и рекомендациях
- выборки зависимы – что накладывает особенности



Пример TDI:

Задача: протестировать два поисковых алгоритма

Метод TDI позволяет показать пользователю результат выдачи пользовательского запроса с применением двух алгоритмов. Пользователь решает кликом, какой же алгоритм дал релевантный ответ. Разделения на варианты здесь нет. Сами расположения ответов разных алгоритмов будут перемешиваться, чтобы результат эксперимента не был зависим от расположения.



Diff-in-Diff

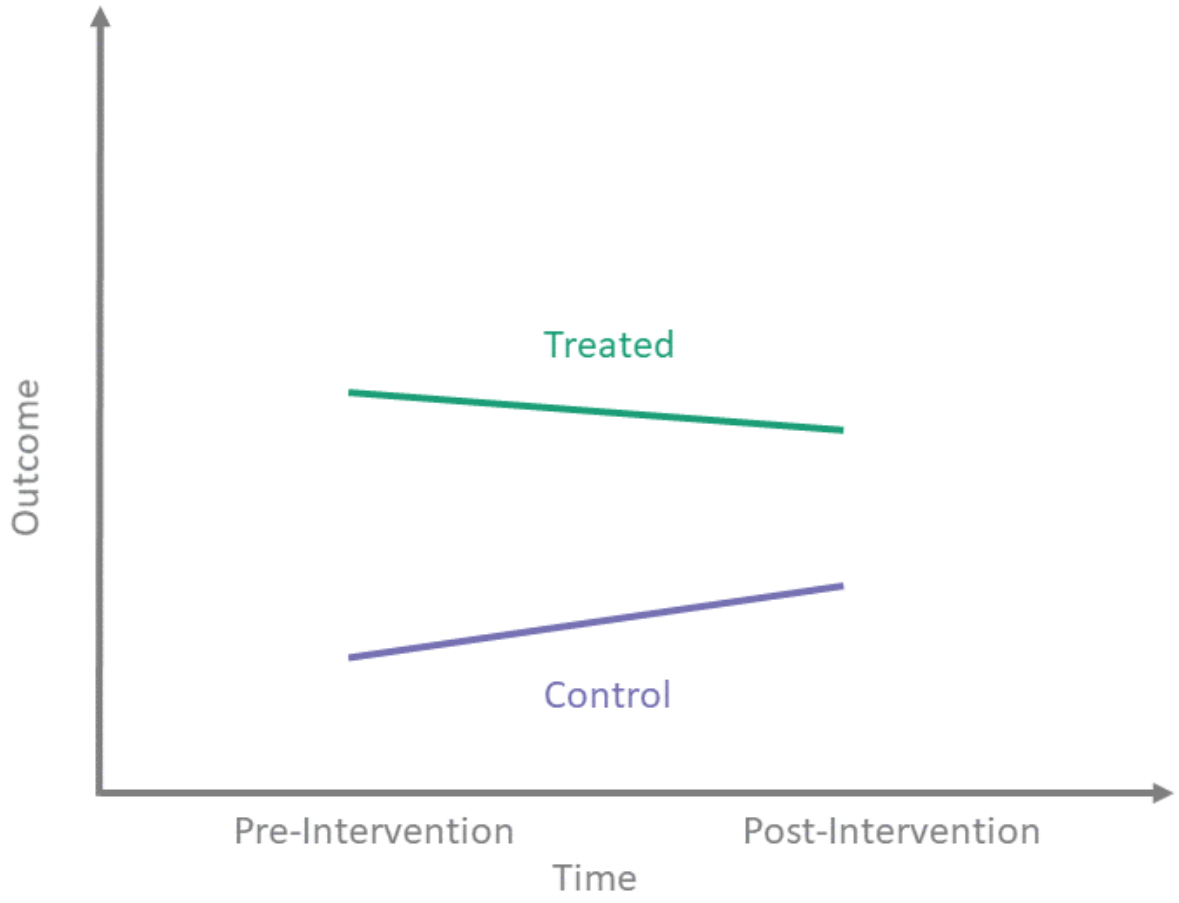
Чем полезен: когда нет возможности поделить пользователей на группы в один момент времени

Ключевые особенности:

- группы зависимы и разнесены во времени
- один из типов регрессий

Пример: экономическое изменение в государствеЕсть две группы (тест и контроль) и два периода – период до взаимодействия и после. Одна из групп подвержена воздействию, или участвует в некоторой программе, во втором

периоде, но не в первом. Вторая группа не подвержена воздействию ни в одном из периодов. Метод устраняет смещение при сравнении исходов в опытной и контрольной группах только во втором периоде, которое может быть следствием постоянных различий между этими группами.



Дополнительно: следующий вид эксперимента в лекции не упоминался, одна авторы просили нас включить и его описание, так как он бывает весьма полезен.

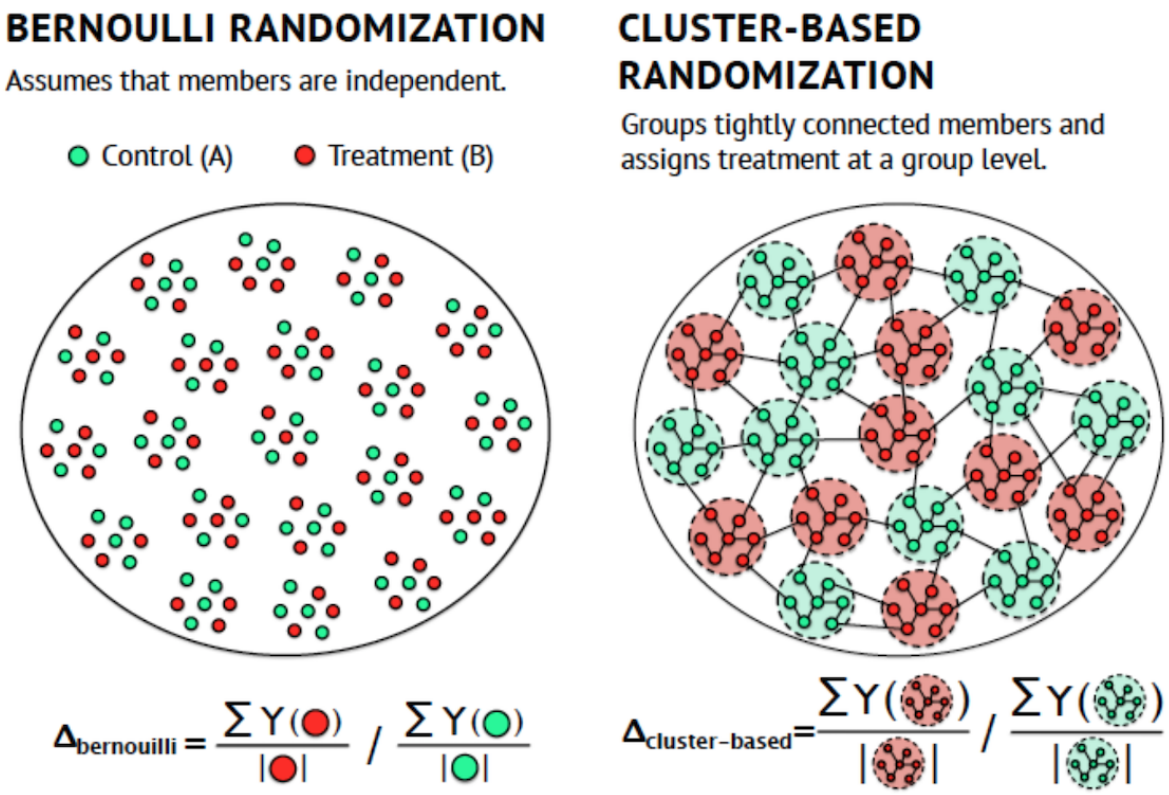
Synthetic control

Чем полезен: не чувствителен к социальным (сетевым) эффектам

Ключевые особенности:

- группы отделены друг от друга географически или физически
- группы схожи по описательным статистикам, но находятся далеко друг от друга
- контроль регулярно версионирует

Пример: Мы хотим провести эксперимент в социальной сети – дать возможность пользователям отправить анимированные смайлики в сообщениях. В подобных продуктах есть большая особенность – пользователи общаются между собой. И общение пользователей из групп A и B могут оказывать сильное влияние на исход всего эксперимента, т.к. в процессе кто-то может, например, отправить смайлик тому, кто в ответ этот смайлик отправить не сможет.



Самое важное в synthetic control – корректно выбрать группы для его формирования. Важно, чтобы группы до получения "влияния" не отличались друг от друга по описательным статистикам и были репрезентативны друг другу.

Формирование групп не является разовым процессом, скорее перманентным поиском близких выборок для проведения эксперимента.

> Как устроено A/B-тестирование в продуктовых командах

Каждый большой продукт имеет свои особенности как с точки зрения бизнеса, так и с точки зрения метрик. Эти особенности накладывают определенные ограничения и дают творческий простор для развития методологии экспериментов.

Примеры будут в презентации :)

> Дополнительные материалы

- Материалы по математической статистике