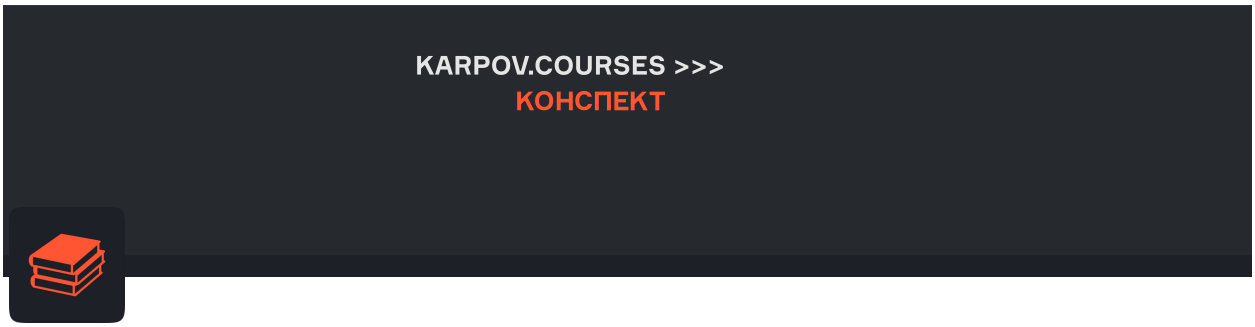

> Конспект > 9 урок > Ошибки при проведении A/B тестов

>Оглавление

> Зачем нужны A/B эксперименты

Когда мы развиваем продукт, этот процесс мы делаем итеративно: работаем над алгоритмами, видом приложения. И каждое такое изменение нужно проверять.

Проверку можно производить разными способами:

- Обсудить с коллегами
- Опросить пользователей

- Пообщаться подробно с группой пользователей

Но во всех этих случаях мы можем получить неполноценную картину, поэтому как раз приходится применять A/B эксперименты.

Основная идея состоит в тестировании двух версий приложения на двух различных группах пользователей. Мы ожидаем, что группы пользователей эквивалентны (из одной генеральной совокупности) - ведут себя одинаково.

Из-за того, что мы проводим эксперименты на разных пользователях и сравниваем метрики по разным пользователям, метрики в точности, как правило, не совпадают - пользователи никогда не будут себя вести одинаково в двух группах. Чтобы отличать шум от достоверных изменений, мы используем статистические критерии.

Процесс A/B тестирования состоит из четырех основных этапов:

1. Выбор метрик
2. Выбор размеров групп и продолжительности эксперимента
3. Запуск эксперимента и сбор метрик
4. Оценка результата

Но не всегда все проходит идеально, случаются и ошибки. Ошибки, как правило, бывают двух видов.

В некоторых случаях мы принимаем за улучшение то, что улучшением не являлось, а иногда - не замечает, что сделали какие-то хорошее изменение.

Если происходит одна такая ошибка - это не страшно. Проблемы начинаются, когда мы перестаем управлять этими ошибками.

> Проблемы при проведении A/B экспериментов

Подглядывание в результаты

Мы заранее смотрим, что происходит в нашем эксперименте. Например, мы запустили эксперимент на две недели. Метрики считаются автоматически, обновляются каждый день. Мы можем начать каждый день смотреть на метрики. И когда мы смотрим так каждый раз, мы ждем, когда прокрасится нужная нам метрика. И когда метрика прокрасилась, можем остановить эксперимент.

Неправильное разбиение

Разбить пользователей на две группы не всегда получается успешно. Сначала нужно ответить на вопрос: а, действительно, ли мы нигде ничего не забыли и разбиваем пользователей нормально? Представим, что у нас пользователи заходят на сайт, и при заходе на сайт мы их разбиваем на разные группы. Возникает вопрос: а что будет, если пользователь будет заходить на сайт несколько раз за период? В какую группу мы будем его заносить? Если мы случайно разбиваем пользователей на группы, в таком случае можем нарушить условие независимости разделения.

Т.е нам нужно разделять пользователей как-то консистентно. Поэтому нужно использовать айдишники пользователя, а не разбивать их случайно.

Нерепрезентативный период

Представим, что у нас много данных, и можно провести A/B эксперимент хоть за один день. Мы запустили эксперимент в будние дни, чтобы за неделю и запустить эксперимент, и подвести итоги. Но оказывается, что это не очень хорошо, потому что чаще всего пользователи ведут себя по-разному в выходные и будние дни.

С другой стороны, нужно понимать паттерны своего продукта, потому что могут проявляться какие-то циклы. Например, праздничные дни или дни зарплаты.

Большое количество метрик

Еще нужно быть аккуратными с большим количеством метрик. Для каждого эксперимента можно найти какой-нибудь хороший эффект - множественная проверка гипотез, но сами метрики скоррелированы между собой.

Особенности метрик

Также у конкретных метрик могут быть свои особенности. Чаще всего они связаны с распределением самих метрик. Они редко совпадают с идеальным распределением, на котором основаны статистические критерии. И из-за этого на каждой конкретной метрике статистический критерий может выглядеть как-то иначе.

Также в метриках бывают выбросы, и их лучше заранее фильтровать.

Воспроизводимость и срезы

Реализовывать разбиение пользователей можно разными способами. В некоторых случаях происходит так: там, где нам нужно провести A/B эксперимент, происходит разбиение на группы. В таком случае, если пользователь не дошел до этого места в приложении, то не поппадает ни в одну из групп.

Независимость наблюдений

Представим, что у нас один объект - это одна сессия пользователя, а целевое действие - это заказ в приложении. Считать будем метрику conversion to order. Проводить эксперимент будет две недели. За это время, конечно, найдутся пользователи, которые зайдут в приложение больше одного раза. В таком случае, даже если мы сплитование делали корректно, то все равно мы нарушаем условие независимости.

Сетевой эффект

Представим, что в приложении для доставки продуктов, мы сделали скидку на сезонные фрукты, Пользователи, которые получили скидку, начнут более активно заказывать товар, и быстрее его раскупят. Вышло так, что мы повлияли нашей тестовой группой на контрольную группу.