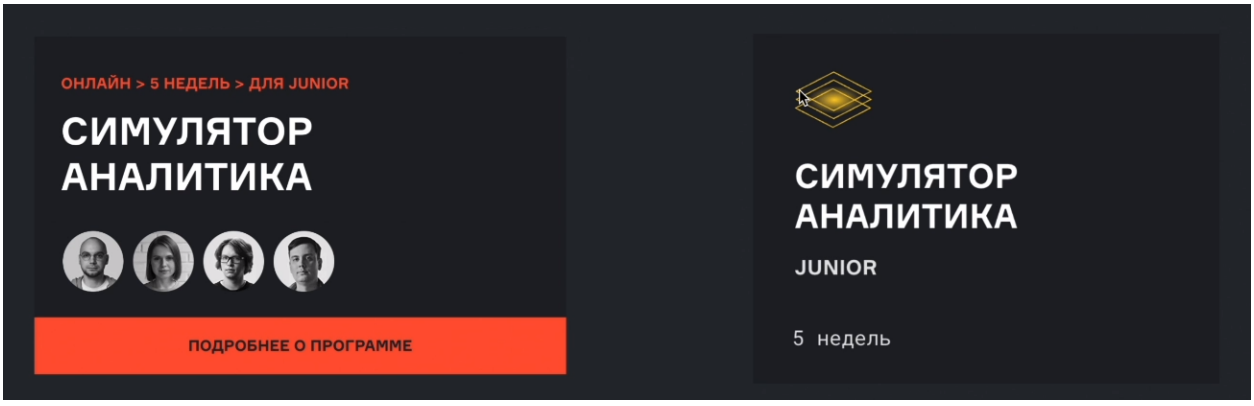
KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 10 урок > СТАТИСТИКА

> Суть АБ-тестирования

Представьте себе, что мы решили что-то поменять в своей жизни — во всяком случае, в жизни того продукта, над которым вы работаете. Например, систему рекомендаций или дизайн какого-то элемента вашего сайта:



Изменения — это здорово! Но не всегда они ведут к лучшему. Совершенно не исключено, что пользователи отреагируют на изменения совсем не так, как вы себе это представляете — и вместо прироста прибыли вы начнёте нести убытки.

Вопрос: а можем ли мы как-то заранее проверить, какой эффект окажет наше изменение?

Ответ: да, как раз с помощью А/Б-тестов!

Общая суть А/Б-тестов проста:

- делим пользователей на две группы;
- для одной из групп ничего не меняем (контрольная группа), вторая же видит изменения (тестовая группа);
- если тестовая группа ведёт себя лучше, чем контрольная — можно выкатывать изменения, в остальных случаях не стоит.

Реальность, конечно же, сложнее — для корректного А/Б-теста надо решить несколько важных вопросов:

- выбрать подходящую метрику;
- разбить пользователей на группы — две или больше;
- набрать необходимое количество данных;
- применить правильный статистический тест;
- сделать выводы, понятные заказчику.

ВНИМАНИЕ! По сути своей А/Б-тестирование — это разновидность большой научной области под названием **экспериментальный дизайн**. А/Б-тесты — это эксперименты, реализованные онлайн, и у них есть своя специфика. Однако для погружения в тему можно также читать материалы по тому, как проводят эксперименты в науках о жизни и социальных науках, например, в медицине и психологии. При всех различиях многие базовые идеи в конструировании экспериментов пересекаются между областями.

> Выбор метрики

Самое главное, с чем надо определиться на самых первых этапах А/Б-тестирования — что именно должно улучшиться, чтобы мы сделали вывод об эффективности воздействия. Нужен какой-то показатель, который можно измерить количественно — **метрика**.

В простейшем случае метрики можно делить на две основные категории:

- **Целевые** — то, что нас непосредственно интересует. В бизнесе практически все целевые метрики в итоге сводятся к прибыли, но сводиться они могут очень по-разному. **Пример**: число продаж курса.
- **Прокси-метрики** — показатели, которые непосредственно связаны с целевыми метриками, но сами ими не являются. Полезны, если по каким-то причинам целевая метрика на данный момент недоступна. **Пример:** число пользователей, перешедших на страницу курса (не все из них его купят, но рост ажиотажа может быть хорошим ранним сигналом качества ещё до появления статистики по продажам).

Можно встретить и другие категории метрик, о которых говорят специалисты — например, **guardrail-метрики** (дополнительные метрики, которые не должны измениться в худшую сторону при введении изменения) или **vanity-метрики** (бесполезные метрики, на которые просто приятно смотреть). Более обширную классификацию с примерами можно посмотреть в этой книге. Также про основные метрики продукта будет идти речь в блоке продуктовой аналитики.

> Деление на группы

Вы уже поняли, что мы должны поделить наших пользователей на группы, одна из которых не видит изменений, а другая — видит, после чего сравнить их поведение. Но не всякое деление на группы одинаково эффективно!

Например, представьте себе, что в контрольную группу у нас попали только старые пользователи, а в тестовую — только новые. Мы видим, что в тестовой группе больше продаж — можно ли заключить, что нововведение улучшило наши продажи? Необязательно: более разумно будет предположить, что **новые пользователи с большей вероятностью купят курс, чем старые** (которые могли уже и так всё купить).

Другой вариант — представьте себе, что мы оцениваем влияние нововведения на средний чек пользователей. Так вышло, что контрольная группа у нас состоит из жителей Москвы, а тестовая — из жителей Саранска. Очень вероятно, что по результатам А/Б-теста мы увидим ощутимые различия между группами, но их будет легко объяснить **различиями в средней зарплате этих двух городов**.

Ну и как тогда правильно?

Рандомизация

Пользователей нужно распределять таким образом, чтобы у каждого из них была **одинаковая вероятность попасть в любую из групп**. Так мы создаём ситуацию, при которой две группы отличаются друг от друга лишь тем, видят ли они старую версию или уже нововведения. Все остальные различия будут несистематическими — соответственно, мы сможем с гораздо большей уверенностью утверждать, что различия между группами обусловлены именно нововведениями, а не чем-то ещё.

Само собой, руками это делать не стоит: это довольно ресурсозатратно, к тому же из людей выходят очень плохие рандомизаторы. Поэтому этим в компаниях занимается **система сплитования** — некоторый алгоритм, который случайно присваивает каждому пользователю метку, в какой группе он находится. Её настройкой обычно занимаются уже инженеры данных в тесном сотрудничестве с аналитиком.

Перед тем, как начать пользоваться системой сплитования, стоит проверить её на адекватность. Это делается посредством **A/A-тестирования** — по сути, это то же А/Б-тестирование, только обе группы контрольные. Соответственно, если пользователи действительно распределились по группам случайно, то их сравнение не должно дать статистически значимых различий. Если же группы различаются — это тревожный звоночек, означающий либо дефект системы сплитования, либо появление какого-то непредвиденного фактора.

Мы не будем останавливаться на методологии A/A-тестирования, но почитать про неё можно, например, тут.

> Размер выборки

Мы уже знаем, что чем больше собранная нами выборка, тем лучше наши выводы. Однако такая ориентировка не позволяет понять, в какой момент её размер становится *достаточным*. Для ответа на этот вопрос аналитику может помочь **анализ статистической мощности (power analysis)**.

Что такое мощность? Это (1 - вероятность ошибки II рода) — иначе говоря, **вероятность обнаружить различия, когда они реально есть**. Об этом показателе часто говорят, когда речь заходит о чувствительности теста — в каком-то смысле чувствительность и статистическая мощность являются синонимами.

Посчитать мощность для различных статистических тестов можно через pingouin. Все местные функции для анализа мощности имеют 4 аргумента. Нам нужно указать 3 из них, в результате посчитается четвёртый. Что это за аргументы?

- **Размер выборки** — сколько именно пользователей должно попасть в АБ-тест
- **Вероятность ошибки I рода** — порог p-value, с которым мы работали всё это время (по умолчанию 0.05)
- **Мощность** — вероятность обнаружить различия, если они есть
- **Размер эффекта** — как правило, это величина различий. При выборе этого показателя стоит ответить на следующий вопрос: **какое минимальное различие между группами может быть интересно бизнесу?**

Как правило, нас интересует размер выборки, поэтому для такого расчёта нам надо задать вероятность ошибки I рода, мощность и размер эффекта. Но можно считать и другие показатели — например, если мы укажем размер выборки, вероятность ошибки I рода и мощность, то мы получим **минимальный детектируемый эффект (minimum detectable effect, MDE**) — то есть самый маленький размер эффекта, который можно получить с таким размером выборки, уровнем значимости и мощностью.

Общая взаимосвязь между MDE и размером выборки проста: чем больше выборка, тем более маленькое различие мы можем обнаружить. Соответственно, чем более мелкие различия интересны бизнесу, тем большая выборка нам понадобится.

Что ещё посмотреть и почитать на тему:

- Вебинар Толи

- Выступление Толи
- Статьи из блога expf: раз и два

> Выбор теста и выводы

Теме выбора теста в каком-то смысле были посвящены все уроки статистики, начиная с 4 🙂

Толя предлагает следующую эвристику:

- если метрика похожа на конверсию — используем хи-квадрат (см. урок “Аналитика категориальных переменных”)
- если величина непрерывная или дискретная — используем t-тест (если не уверены - Манна-Уитни, но с этим поаккуратнее)

Дополним конспект двумя дополнительными эвристиками:

- если групп больше 2 — используем дисперсионный анализ
- если хотим сравнить что-то хитрее среднего — используем бутстрап

Также можно опираться на:

- вот этот гайд от VK
- деревья выбора метода в pingouin
- всё, что найдёте по запросу “схема выбора статистического теста”, “how to choose statistical test” и любым подобным — в интернете огромное количество подобных схем

Последний пункт краток, но его важно упомянуть: **самый важный результат работы аналитика — это выводы**. Причём для выводов недостаточно просто привести ноутбук с вычислениями — большей части ваших работодателей такой формат не будет понятен. Важно учиться объяснять результаты своей работы простыми словами и давать интерпретацию этим результатам. В этом вы попрактикуетесь в двух заданиях этого урока.

> Дополнительные ссылки

- Интервью с Никитой Маршалкиным — обратите внимание на полезные ссылки в описании, некоторые из них уже фигурировали в конспекте
- Книжка про причинно-следственный вывод — первые две главы объясняют, почему рандомизация в экспериментах вообще эффективна, остальные посвящены случаям, когда А/Б-тесты невозможны. Про случаи с невозможностью А/Б-тестов также можно почитать вот это.
- Для более глубокого и математизированного погружения в вопрос “как обосновать причинно-следственный вывод” — вот эта книга.