Вариации A/B-тестирования и его приложения

Почему A/B-тесты врут? Ошибки интерпретации метрик

Цель урока

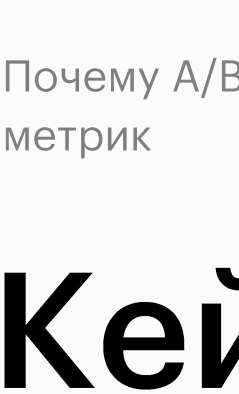
Обсудим, как можно ещё ошибиться при A/B-тестировании.

Задачи урока

- Поймём, почему нельзя завершать тестирование раньше запланированного срока
- Узнаем взаимосвязь регрессий и A/B-тестов
- Научимся использовать A/A-тест и инструменты для проверки тестирования

Почему A/B-тесты врут? Ошибки интерпретации метрик

Кейс:
преждевременное завершение тестирования



Хорошо, включаю тестирование — надо ждать две недели.

Стой!
Надо дождаться!

Давай проверим новый функционал.

Посмотрел метрики. Всё хорошо. Функционал оставляем!

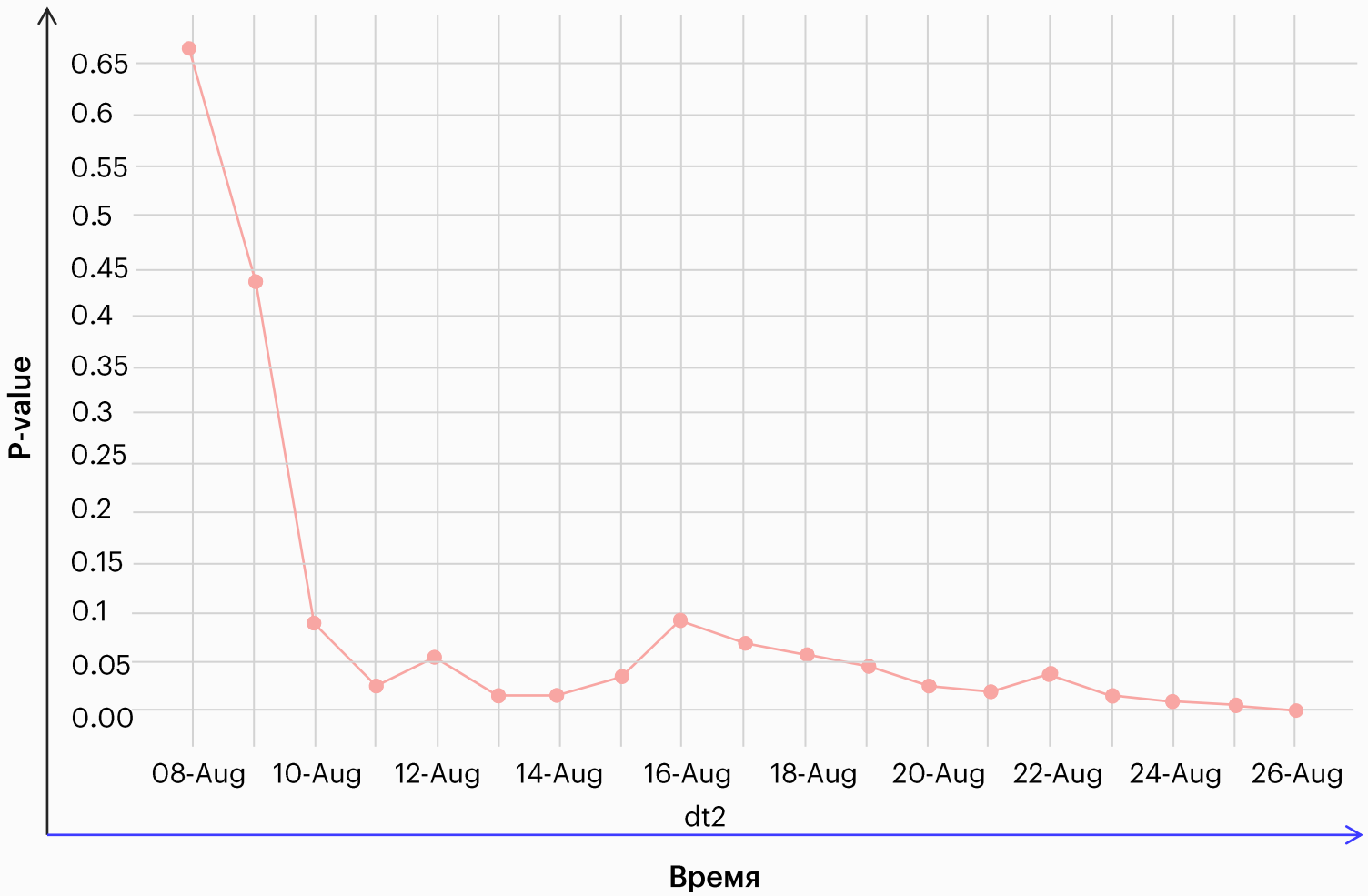
Но почему?

Кейс: преждевременное завершение тестирования

Когда вы завершаете тестирование раньше времени, считая, что метрика статистически значимо улучшилась, вы сразу же увеличиваете вероятность ошибки.

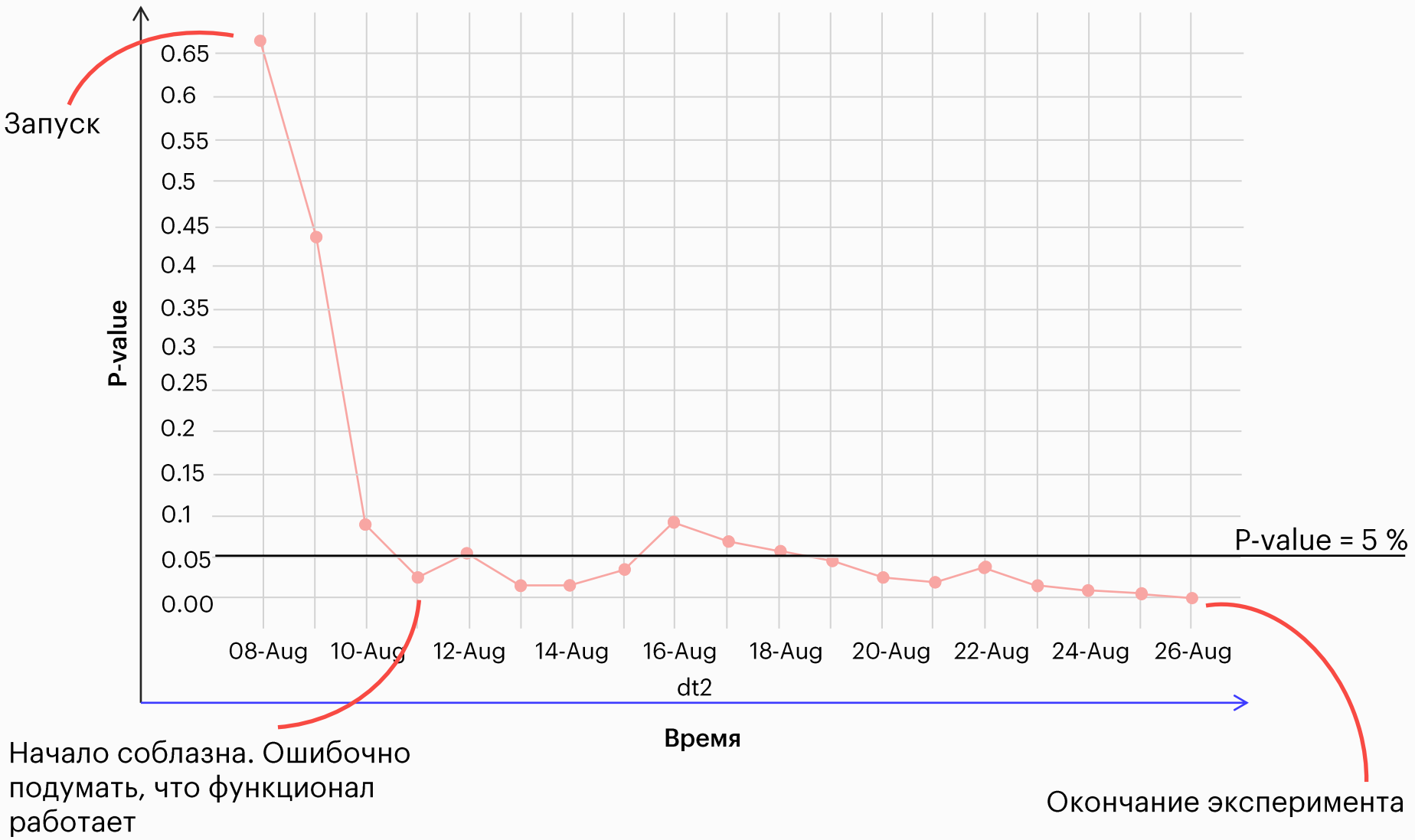
Рассмотрим причины этого неочевидного явления подробнее.

Кейс:
преждевременное завершение тестирования



P-value — вероятность принять гипотезу при условии, что она неверна

Кейс:
преждевременное завершение тестирования



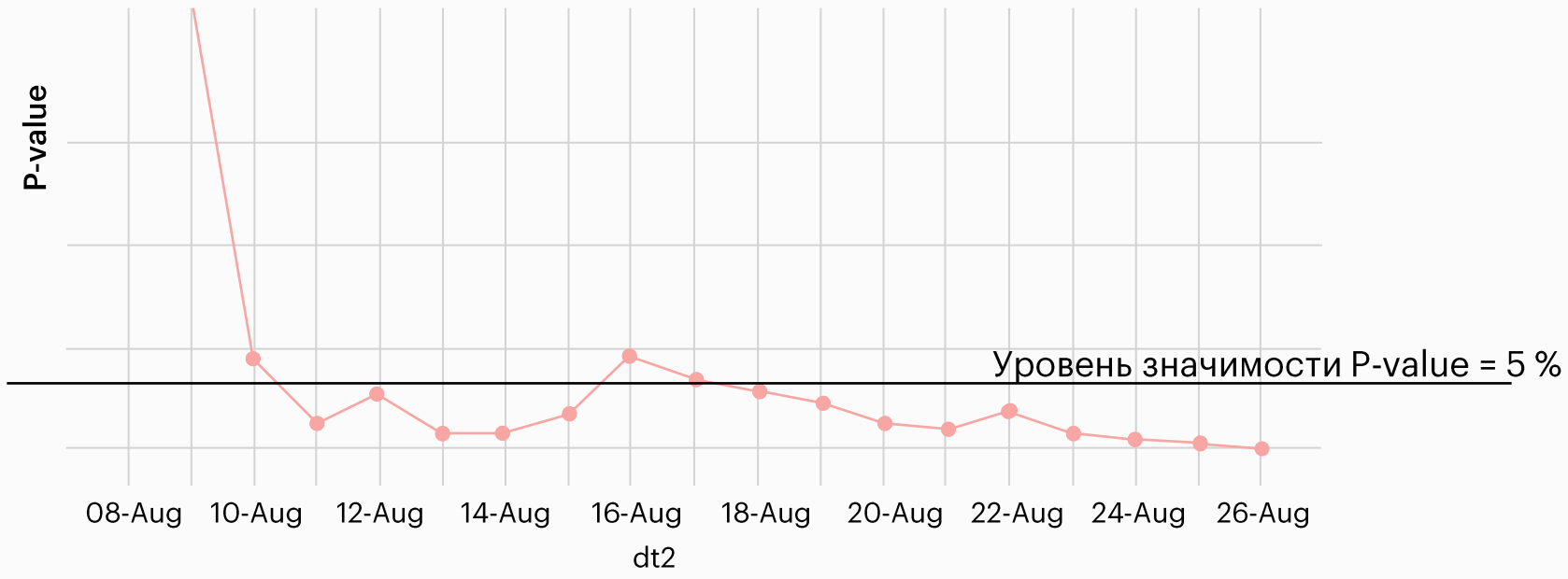
P-value — вероятность принять гипотезу при условии, что она неверна

Кейс:
преждевременное завершение тестирования

Можем ли принять решение, что тест успешный и один из вариантов выиграл? Ответ: нет.

Но, если мы будем продолжать рассчитывать P-value ежедневно, мы ведь будем повышать шансы поймать ложный эффект? Ответ: да.

Как же нам быть?

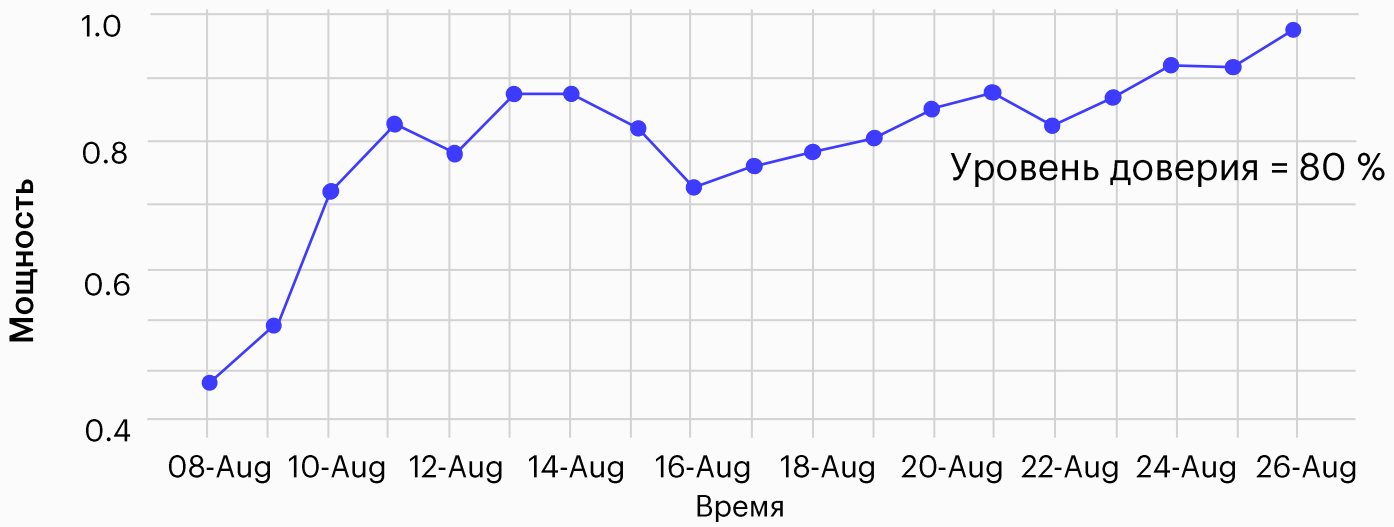


Кейс: преждевременное завершение тестирования

Просто не забывать, что статистическая успешность теста определяется 2 параметрами:

- significance level < 0,05
- power > 0,8

Значит, помимо динамики P-value, нам параллельно нужно также считать power теста. **Мощность** = 1 – β, где β — вероятность не найти различия там, где они есть.



Кейс 2: регрессия

Вы провели тестирование новой фичи, которая как-то связана с покупками. Через какое-то время вы провели повторное тестирование, и ваш успешный результат не воспроизвёлся.

Ошибочно: думать, что результат пропал.

Правильно: отнестись более критично и считать, что результата вообще не было. И в первый раз это был false-positive-случай.

Кейс 2: регрессия

Это хорошо известный феномен, который в статистике называется регрессией. Этот термин хорошо известен среди статистиков, но многие специалисты по A/B-тестированию едва ли о нем слышали.

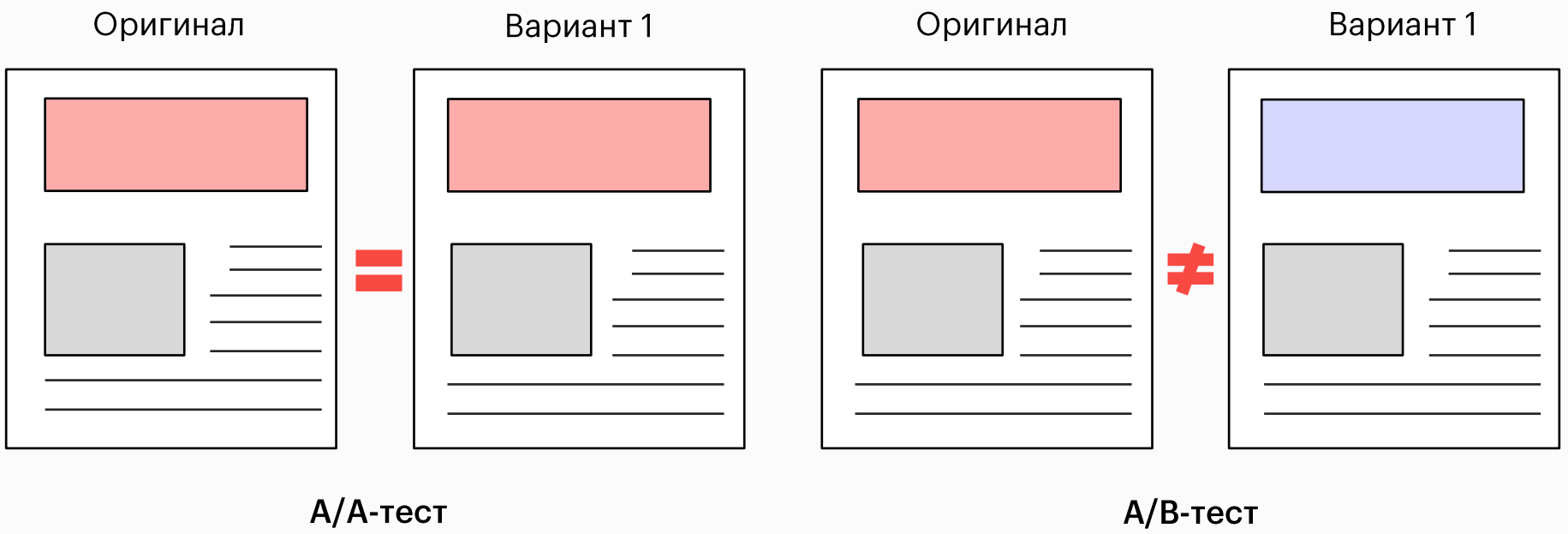
Как избегать таких ситуаций? Если результат важен, то лучше провести тестирование несколько раз. Если хотя бы раз результат не воспроизвёлся, то это повод задуматься, а есть ли он.

Кейс 2: регрессия и A/A-тест

Кроме повторной проверки результатов можно воспользоваться A/A-тестированием. Об этом мы как раз сейчас и поговорим.

A/A-тест

A/A-тест — это вариация A/B-теста, особенность которой понятна из названия. Если при A/B-тесте сравниваются разные варианты сайта или письма, то при A/A-тесте оригинал сопоставляется сам с собой.



Зачем нужен A/A-тест? Проверить тестирование

Возможно, у вас в голове сразу возник вопрос: «А зачем это нужно, если разницы в результатах не будет?»

В этом и заключается суть такого эксперимента!

По данным Instapage, около 80 % результатов A/B-тестов не подтверждаются на практике. Чтобы решить эту проблему, и были придуманы A/A-тесты.

Зачем нужен A/A-тест? Проверить тестирование

Главная цель A/A-теста — показать, можно ли доверять результатам эксперимента, который будет запущен в тех же условиях, но уже с разными вариантами страницы. Если в ходе A/A-теста победителя выявить не удалось, можно запускать A/B-тест. В противном случае придётся проверить настройки сервиса и однородность выборки.
Таким образом, A/A-тест предоставляет контрольные данные для проверки точности A/B-теста.

Зачем нужен A/A-тест? Диапазон изменения

С помощью A/A-теста можно определить доверительный интервал, в рамках которого изменения конверсии могут быть случайными и не зависеть от изменений на странице. Например, в ходе A/A-теста одна и та же страница показала конверсию 2 % и 3 %. Значит, если при A/B-тесте конверсия попадёт в диапазон от 2 % до 3 %, вносить изменения на страницу не стоит, так как они не повлияют на результат.

Зачем нужен A/A-тест — диапазон изменения

С помощью A/A-теста можно определить доверительный интервал, в рамках которого изменения конверсии могут быть случайными и не зависеть от изменений на странице. Например, в ходе A/A-теста одна и та же страница показала конверсию 2 % и 3 %. Значит, если при A/B-тесте конверсия попадёт в диапазон от 2 % до 3 %, вносить изменения на страницу не стоит, так как они не повлияют на результат.

Так, в компании Avast после проведения A/A-теста приняли внутренние для себя рекомендации по тестированию: любые изменения конверсии в диапазоне 5 % могут быть вызваны случайно и не заслуживают внимания по внедрению изменений.

Итоги и выводы урока

При A/B-тестировании всё ещё масса мест, где можно выстрелить себе в ногу, но мы уже умеем не допускать самые распространённые ошибки.

- Помните, что надо ждать конца эксперимента, иначе вы имеете высокую вероятность принять неверную гипотезу (допустить ошибку 1-го рода)
- Стоит крайне скептично относиться к наиболее успешным результатам. Для уверенности необходимо пытаться воспроизвести результаты
- A/A-тестирование используется, когда хотим получить дополнительные доказательства того, что тестирование работает правильно
- Также A/A-тест полезен, чтобы определить диапазон случайных отклонений, на который не стоит реагировать