Дизайн экспериментов

Алгоритм проведения АБ-теста

CR (conversion rate, %) в заказ продукта

Конверсия в группе А и в группе Б не отличается

3% абсолютного роста

Уровень значимости: 5%, мощность: 20%, MDE: 3%

На калькуляторе рассчитали, что нужно по 2000 наблюдений в каждой группе. Среднее количество пользователей в день – 400. Минимальная длительность – 5 дней.

Держим тест 5 дней и не вмешиваемся в его ход

Обнаружен ли эффект ≥3%?

Примеры A/B тестов
Кейс Bing

В 2012 сотрудник поисковой системы Microsoft Bing предложил изменить способ отображения заголовков: дописывать первую строку под заголовком в сам заголовок

Через несколько часов сработала система алерта об очень высокой выручке.

Выручка выросла на 12%, что принесло более $100M в год только на рынке США.

Примеры A/B тестов
Кейс Google:
41 shades of blue

В 2009 году Google протестировал 41 оттенок синего цвета на странице поиска.

Гугл не раскрыл точные цифры, но отметили, что это привело к существенному увеличению вовлеченности пользователей по их метрикам.

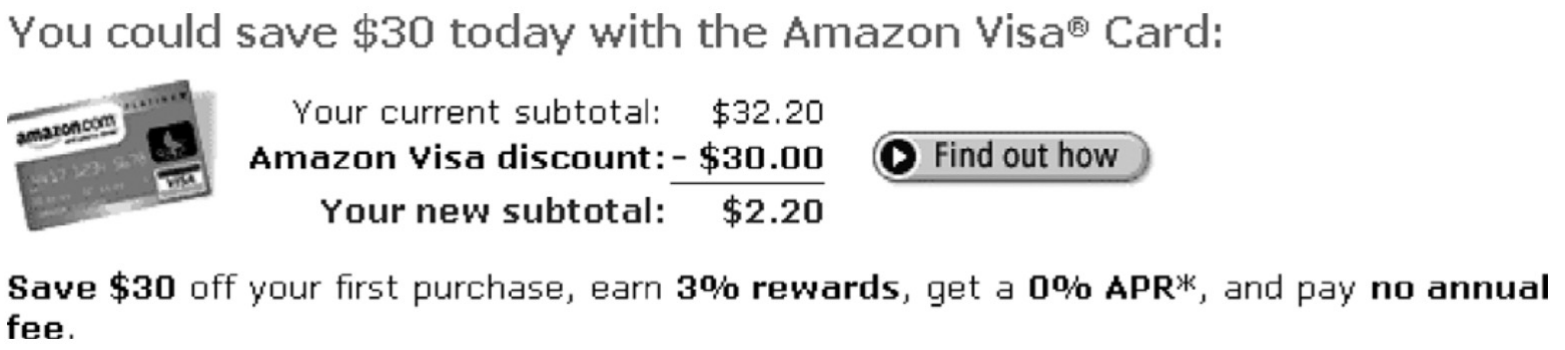
Microsoft Bing также показали, что цветовые настройки их поисковой системы увеличили годовую выручку на более чем $10M в США ежегодно, а также улучшилась метрика Time-to-success

Примеры A/B тестов
Кейс Amazon:
Making an offer at Right time

В 2004 году Amazon разместила на главной странице предложение кредитной карты.
Это было очень прибыльно, но имело очень низкий CTR

Команда переместила это предложение в корзину после добавления товара, а также начала показывать экономию с картой от каждой покупки.

В таком случае предложение показывалось в нужное время, и данный тест увеличил годовую прибыль Amazon на десятки миллионов долларов

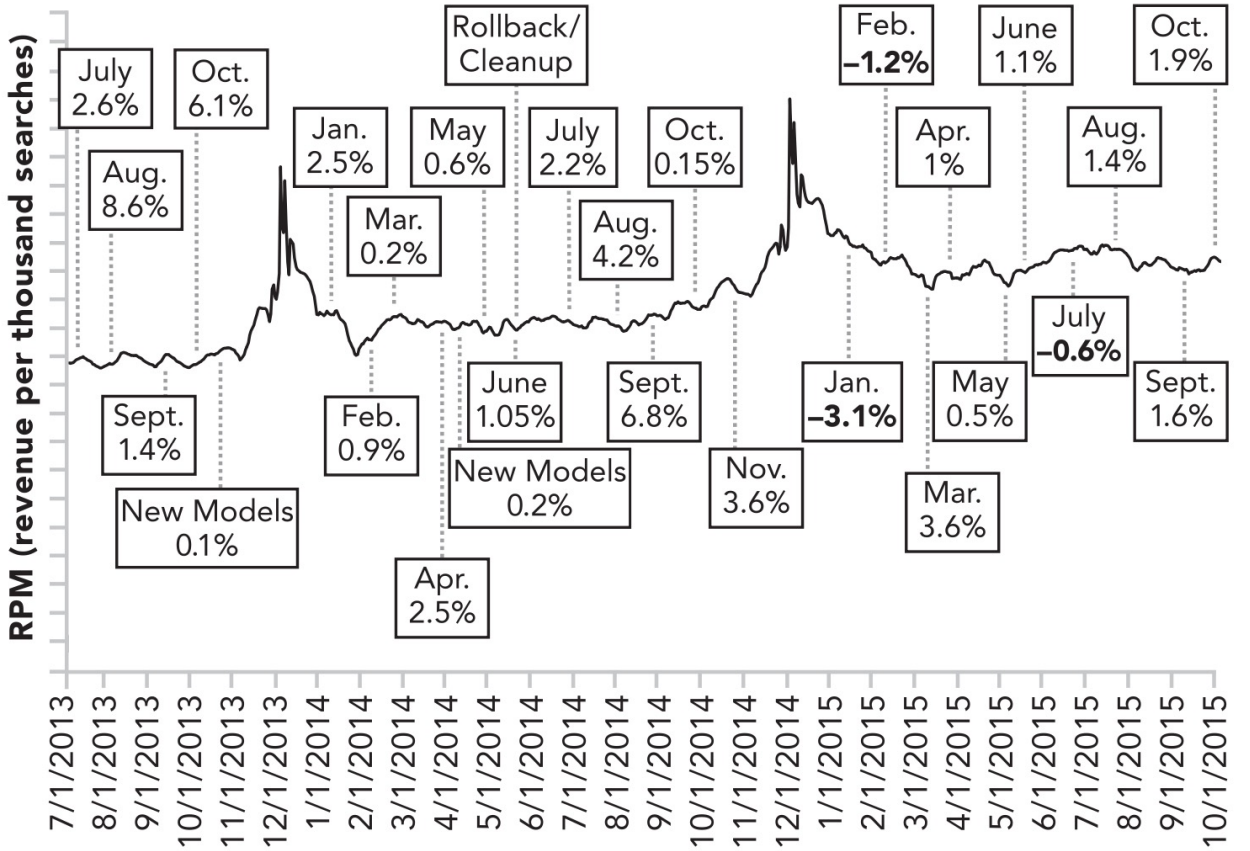


Примеры A/B тестов
Кейс Bing:
Ads team

Команда по рекламе в Bing последовательно увеличивала доход на 15–25 % с 2013-2025, как показано на рисунке справа.

Часто в больших компаниях проводится куча тестов, и улучшения от них небольшие (а иногда даже отрицательные)

Но именно это и есть правильная культура тестирования гипотез в компании:
маленькие изменения приводят в итоге к большим результатам



(*) Numbers have been perturbed for obvious reasons

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Amazon была система рекомендаций книг на основе следующих правил:
1) Рекомендуем новую книгу автора, если пользователь уже покупал книги данного автора
2) Программа рекомендаций на основе историй покупок
3) «Перекрестное опыление»: очень специфичные точечные рекомендации, отправляемые после покупок специальных категорий

Какую метрику подобрать под данную систему рекомендаций?

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг на основе следующих правил:
1) Рекомендуем новую книгу автора, если пользователь уже покупал книги данного автора
2) Программа рекомендаций на основе историй покупок
3) «Перекрестное опыление»: очень специфичные точечные рекомендации, отправляемые после покупок специальных категорий

Какую метрику подобрать под данную систему рекомендаций?

Amazon выбрал метрику: выручка с кликов по эмейлам

В чем проблема такой метрики?

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг на основе следующих правил:
1) Рекомендуем новую книгу автора, если пользователь уже покупал книги данного автора
2) Программа рекомендаций на основе историй покупок
3) «Перекрестное опыление»: очень специфичные точечные рекомендации, отправляемые после покупок специальных категорий

Какую метрику подобрать под данную систему рекомендаций?

Amazon выбрал метрику: выручка с кликов по эмейлам

В чем проблема такой метрики?

Выручка монотонно увеличивается с увеличением количества писем и рекламных кампаний, что в конечном итоге приведет к спаму.

Что случится в долгосрочном итоге?

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг на основе следующих правил:
1) Рекомендуем новую книгу автора, если пользователь уже покупал книги данного автора
2) Программа рекомендаций на основе историй покупок
3) «Перекрестное опыление»: очень специфичные точечные рекомендации, отправляемые после покупок специальных категорий

Какую метрику подобрать под данную систему рекомендаций?

Amazon выбрал метрику: выручка с кликов по эмейлам

В чем проблема такой метрики?

Выручка монотонно увеличивается с увеличением количества писем и рекламных кампаний, что в конечном итоге приведет к спаму.

Что случится в долгосрочном итоге?
Произойдет отток пользователей

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг

Какую метрику подобрать под данную систему рекомендаций?
Amazon выбрал метрику: выручка с кликов по эмейлам

В чем проблема такой метрики?
Выручка монотонно увеличивается с увеличением количества писем и рекламных кампаний, что в конечном итоге приведет к спаму.

Что случится в долгосрочном итоге?
Произойдет отток пользователей

Первое решение:
Amazon сначала пытались решить эту проблему ограничениями на количество писем пользователю.
Но появилась проблема оптимизации:
какое письмо отправить пользователю каждые X дней, когда несколько программ рекомендаций хотят порекомендовать этому пользователю продукт?

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг

Какую метрику подобрать под данную систему рекомендаций?
Amazon выбрал метрику: выручка с кликов по эмейлам

В чем проблема такой метрики?
Выручка монотонно увеличивается с увеличением количества писем и рекламных кампаний, что в конечном итоге приведет к спаму.

Оптимальная метрика

$$\text{Metric } = (\sum_i Rev_i - s * \text{unsubscribe\_lifetime\_loss }) / n$$

- i – счетчик для получателей e-mail
- s – количество отписавшихся пользователей
- unsubscribe_lifetime_loss – ожидаемое потеря дохода из-за невозможности отправить ни одно письмо пользователю за все время
- n – количество пользователей

Как подобрать метрику?
Кейс Amazon: E-mail at Amazon

В Амазон была система рекомендаций книг

Какую метрику подобрать под данную систему рекомендаций?
Amazon выбрал метрику: выручка с кликов по эмейлам

Оптимальная метрика

$$\text{Metric} = (\sum_i Rev_i - s * \text{unsubscribe\_lifetime\_loss}\ )/n$$

- i – счетчик для получателей e-mail
- s – количество отписавшихся пользователей
- unsubscribe_lifetime_loss – ожидаемое потеря дохода из-за невозможности отправить ни одно письмо пользователю за все время
- n – количество пользователей

После внедрения такой метрики более половина кампаний показали отрицательное значение метрики.
Что еще более интересно, осознание того, что отказ от рассылок несет такие большие потери, привело к созданию другой страницы отказа от рассылок, где по умолчанию была отписка только от данной «кампании», а не от всех электронных писем Amazon, что резко снизило количество отписок от всех писем

Метрики

Нечувствительные

- Цена акции компании
- Доля пользователей, продливших подписки в годовых сервисах
- Total выручка

Чувствительные

Плохие

- CTR новой кнопки на главной странице
- Время, проведенное на сайте

Хорошие

- CTR в таргетное действие (покупка)
- Time-to-success целевого действия

Как собрать одну метрику?

Вообще очень сложно.
Бизнес постоянно должен оптимизировать разные вещи (вовлеченность пользователя, деньги и тд)

Но есть такая темка:

1. Берем несколько метрик, которые хотим оптимизировать
2. Нормализуем их в заданный диапазон (например, 0-1)
3. Даем вес каждой метрике
4. Итоговая метрика – взвешенная сумма отнормированных показателей

Metric = 0.5 * CTR + 0.5 * (1 / Time-to-success) -> max

Важно: если не можете сделать 1 метрику, то делайте их меньше 5!
Если нулевая гипотеза верна (нет изменений), то P(p-value < 0.05) = 0.05, т.е. O1P = 5% для 1 метрики

Если у вас k независимых метрик, то вероятность того, что хотя бы одна метрика будет иметь p-value < 0.05:

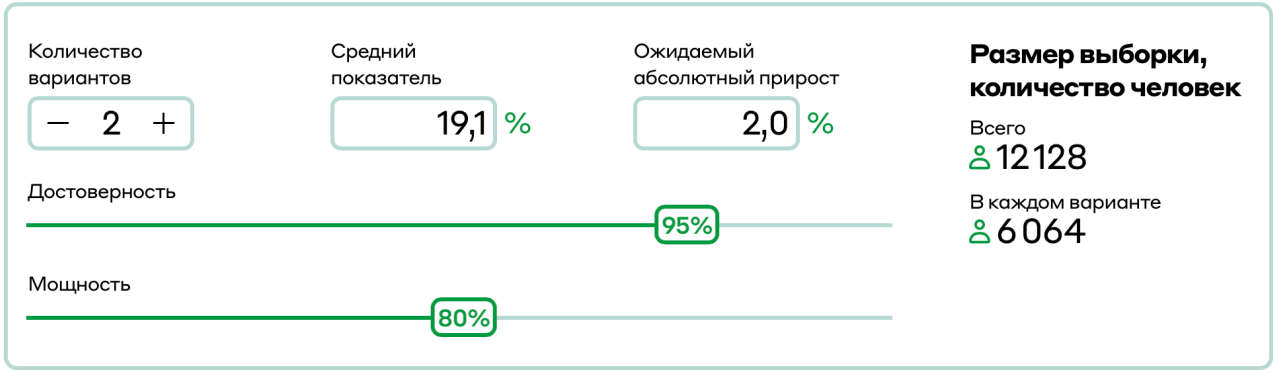
$P(\text{p-value} < 0.05) = 1 - (1 - 0.05)^k$
Для k = 5 вероятность того, что одна из метрик будет статистически значима = 23%
Для k=10 будет 40%

Типичные ошибки в A/B тестировании

- Множественное тестирование
- Преждевременная остановка теста (подглядывание)
- Неверный выбор чувствительности критериев (MDE)
- Оценка эффекта по зависимым выборкам
- Оценка эффекта по части выборки
- Незафиксированные уровень значимости и мощность

Множественное тестирование

Описание теста
Тест на увеличение конверсии в заказ продукта.
A группа: контрольная
B группа: чекбоксы с дополнительными продуктами
C группа: картинки с дополнительными продуктами



Как тестировали?
Сравнили конверсии
A vs B
A vs C

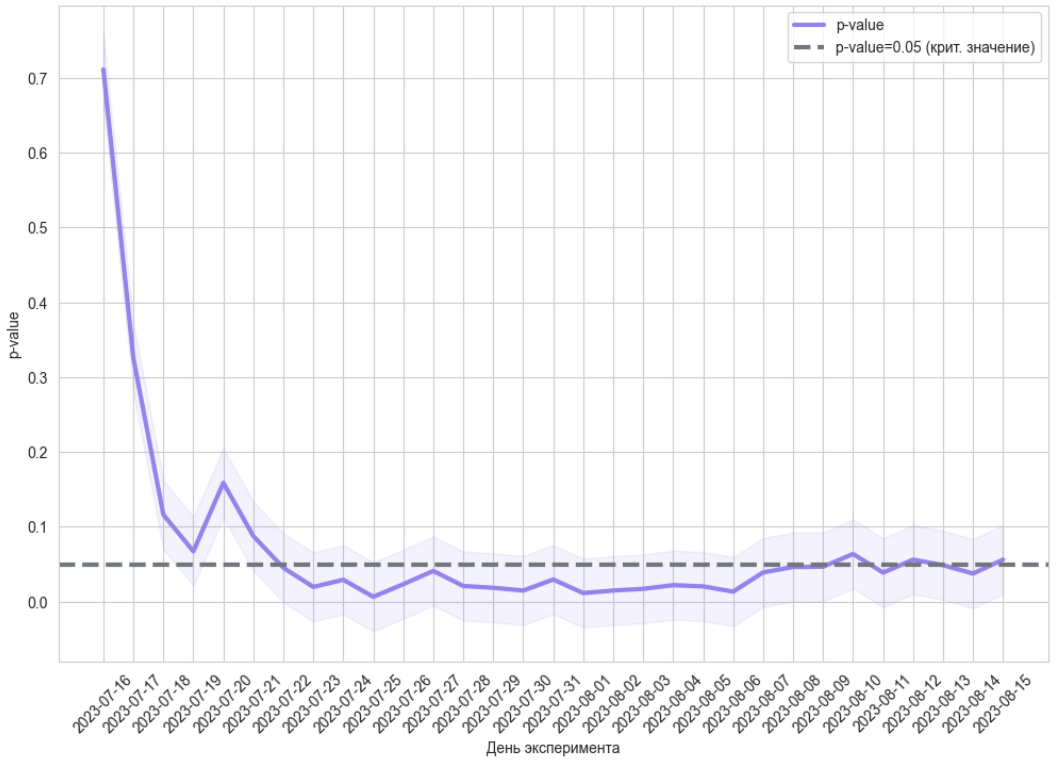
Важно: нельзя использовать одну и ту же контрольную выборку дважды. Снижается достоверность теста.

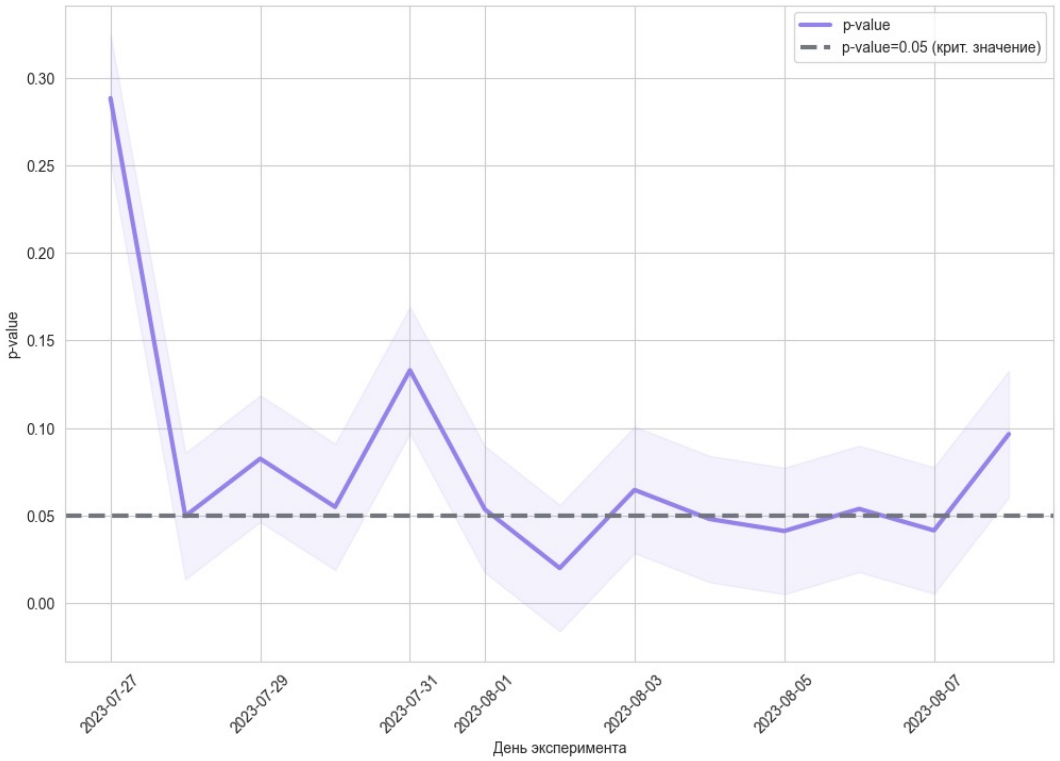
Решение:
1) в этом тесте нужно удвоить контрольную выборку. В подобных тестах можем делить трафик в соотношении:
- 50% - контрольная группа
- 25% - чекбоксы
- 25% - баннеры

2) Множественное тестирование

Преждевременная остановка теста (подглядывание)

Если остановить тест раньше рассчитанного времени, можно получить ложную прокраску:

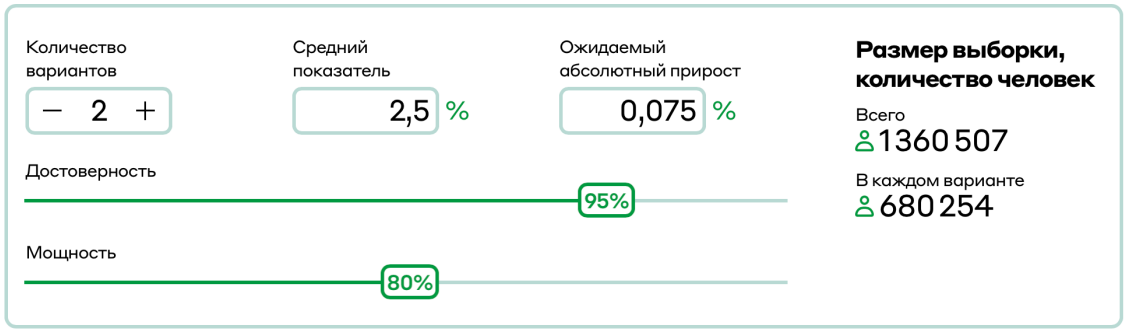




Решение:
1) Использовать последовательное тестирование, если есть вероятность, что тест придется остановить раньше или если рассчитанная длительность слишком большая
2) Ждать столько времени, сколько нужно для fixed-horizon тестирования

Выбор чувствительности критериев (MDE)

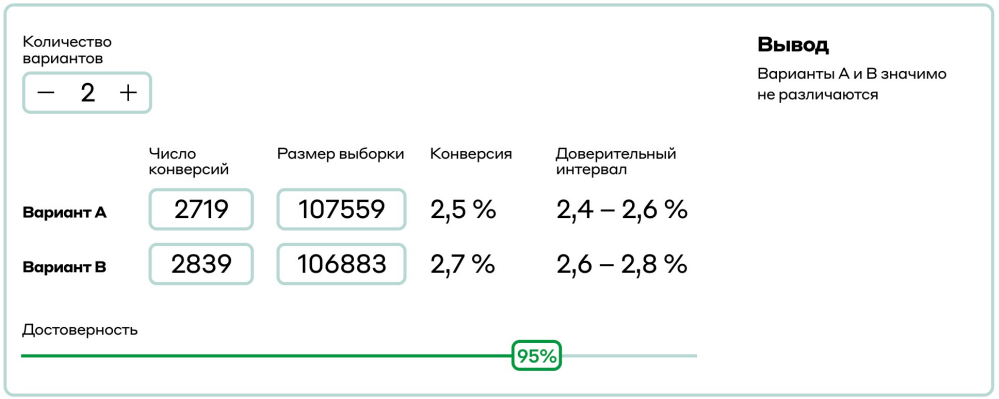
В тесте хотим детектить **относительный** рост конверсии на 3%: то есть при средней конверсии 2,5% нужно заметить **абсолютный** прирост 0,075%. Тогда необходимый размер выборки:



Это очень много. Как решить эту проблему:

- Выбирать больший MDE для тестов с низкой базовой конверсией. В этом случае мы не сможем уловить совсем небольшие изменения, но зато тест будет корректен.
- Ориентироваться на абсолютный прирост, чтобы понимать масштаб детектируемого эффекта.

Оценка эффекта по зависимым выборкам



Одному пользователю может принадлежать несколько визитов. Такие визиты будут зависимыми друг от друга. Тогда предпосылки нашего теста не соблюдаются, и результат нельзя интерпретировать. Самое простое решение – всегда использовать **поюзерные** метрики.

Оценка эффекта по сегменту выборки

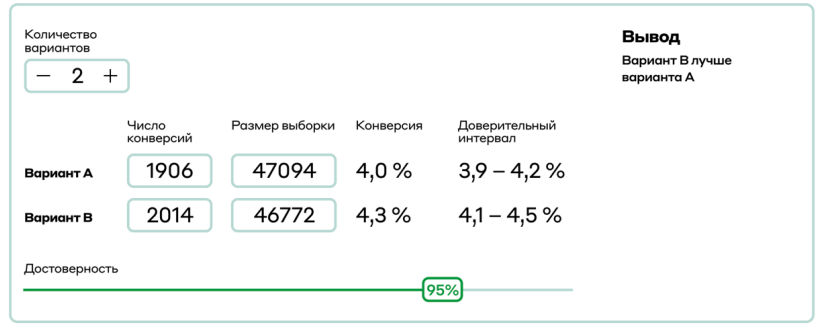
- Проверка значимости разницы между контролем и экспериментом может проводиться только по достижении заранее рассчитанного размера выборки.

- Если тестировать на сегменте выборки – например, только на Мобайле – может возникнуть проблема подглядывания.

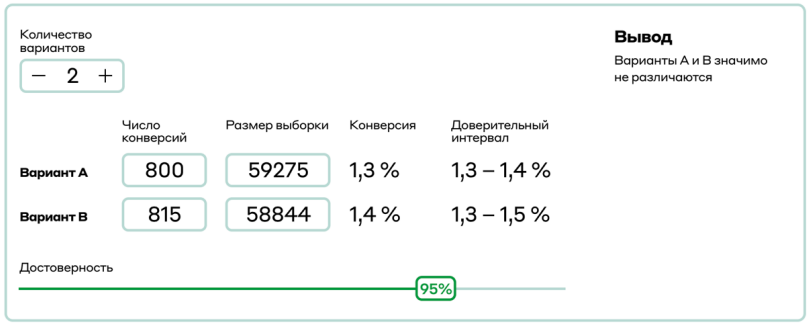
Решение:
1) Делать выводы только на основе всех данных (все устройства)

2) делать стратификацию по сегментам – тогда размер выборки не сократится, а сегменты с большим весом привнесут больший вклад в среднее по выборке.

Мобайл:



Декстоп:



Последствия ошибок

Ошибки в дизайне АБ-тестов приводят к некорректным результатам, которые нельзя интерпретировать: нам могло повезти и тест прокрасился, хотя на самом деле эффекта нет; и наоборот, тест мог не прокраситься, хотя эффект есть.