Конспект по теме "Подготовка к проведению A/B-теста"

AA-тест

Перед тем, как проводить A/B-тест, убедитесь, что:

- генеральная совокупность не содержит аномалий и выбросов;
- инструмент «деления» трафика работает безошибочно;
- данные отправляются в системы аналитики корректно.

Для этого проводят **A/A-тест**. Он похож на A/B-тест, только группам показывают не разные, а одинаковые версии страниц. Если трафик и инструмент проведения A/B-теста не подвели, различий в показателях не будет. Ещё A/A-тест помогает определить длительность теста и методику анализа данных.

Критерии успешного A/A-теста:

- Количество пользователей, попавших в различные группы, отличается не более, чем на 0.5%;
- Данные во всех группах отправляются в системы аналитики одинаково;
- Различие ключевых метрик по группам не более 1% и не имеет статистической значимости;
- Каждый посетитель, попавший в одну из групп теста, остаётся в этой группе до конца теста. Если пользователь видит разные вариации A/B-теста в течение одного исследования, неизвестно, какая из них повлияла на него. Значит, и результаты теста нельзя интерпретировать однозначно.

Степень различия ключевых метрик по группам может варьироваться — это зависит от необходимой чувствительности эксперимента.

Ошибки I и II рода при проверке гипотез. Мощность и значимость

Ошибкой первого рода (ложнопозитивным результатом статистического теста) называется ситуация, когда отвергается нулевая гипотеза H_0 несмотря на то, что она верна. То есть, различий между сравниваемыми группами нет, но тест показал *p-value* меньше уровня значимости. Вероятность ошибки первого рода равна уровню значимости `α` — вероятности случайно получить в реальном наблюдении значение, далёкое от предполагаемого в нулевой гипотезе.

Ошибка второго рода — ложнонегативный результат. Он указывает, что различия между группами есть, но тест показал *p-value* больше уровня значимости `α` и принять нужно H_0. Если обозначить вероятность ошибки второго рода как `β`, то параметр `1 - β` будет называться **мощностью статистического теста**. Раз `β` — вероятность ошибиться, то `1 - β` — вероятность *не ошибиться*, то есть правильно отклонить нулевую гипотезу, когда она неверна.

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка второго рода)
	H_1	H_0 неверно отвергнута (Ошибка первого рода)	H_0 верно отвергнута

Множественные сравнения: A/B vs A/B/n тесты

Часто одну и ту же гипотезу тестируют в разных вариациях. Несколько групп сравнивать с контрольной можно. Однако учитывайте увеличение вероятности ошибок, типы которых вы изучили в прошлом уроке.

Важная особенность **множественного теста** — нескольких сравнений, проводимых на одних и тех же данных — в том, что вероятность ошибки первого рода увеличивается с каждой новой проверкой гипотезы. Если каждый раз вероятность ошибиться равна равна `α`, то вероятность не ошибиться: `1-α`. Так, вероятность не ошибиться ни разу за `k` сравнений равна

$$(1-a)^k$$

В итоге, вероятность ошибиться хотя бы раз за `k` сравнений:

$$1-(1-a)^k$$

Чтобы снизить вероятность ложнопозитивного результата при множественном тестировании гипотез, применяют методы корректировки уровня значимости для снижения групповой вероятности ошибки первого рода, или **FWER** (групповой коэффициент ошибок).

FWER

-	Число принятых гипотез	Число отвергнутых гипотез	Всего
Число верных гипотез	U	V	m0
Число неверных гипотез	T	S	m1
Всего	W	R	m

Для контроля FWER и корректировки требуемых уровней значимости применяют методы Бонферрони, Холма и Шидака:

- Метод Бонферрони (поправка Бонферрони)

$$\alpha_1 = \ldots = \alpha_m = \alpha/m.$$

- Метод Холма

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{\alpha}{m-1},\ \ldots,\ \alpha_i = \frac{\alpha}{m-i+1},\ \ldots,\ \alpha_m = \alpha.$$

- Метод Шидака

$$\alpha_1 = \alpha_2 = \ldots = \alpha_m = 1 - (1-\alpha)^{\frac{1}{m}}\ .$$

Чаще всего применяют поправку Бонферрони из-за простоты решения. Легко поделить принятый уровень значимости на число сравнений, которые проводят на одних и тех же данных, без сбора новых наблюдений для каждого теста. Если вы собираете для каждой проверки гипотезы новые данные, проводите тест как обычно, выбрав нужное значение *p-value*, как это делали в курсе статистики.

Расчёт размера выборки и длительности теста

При проведении реальных A/B-тестов учитывают не только объём выборки, но и другие обстоятельства её получения. Например, сколько по времени длился тест, и характерна ли для него **проблема подглядывания.**

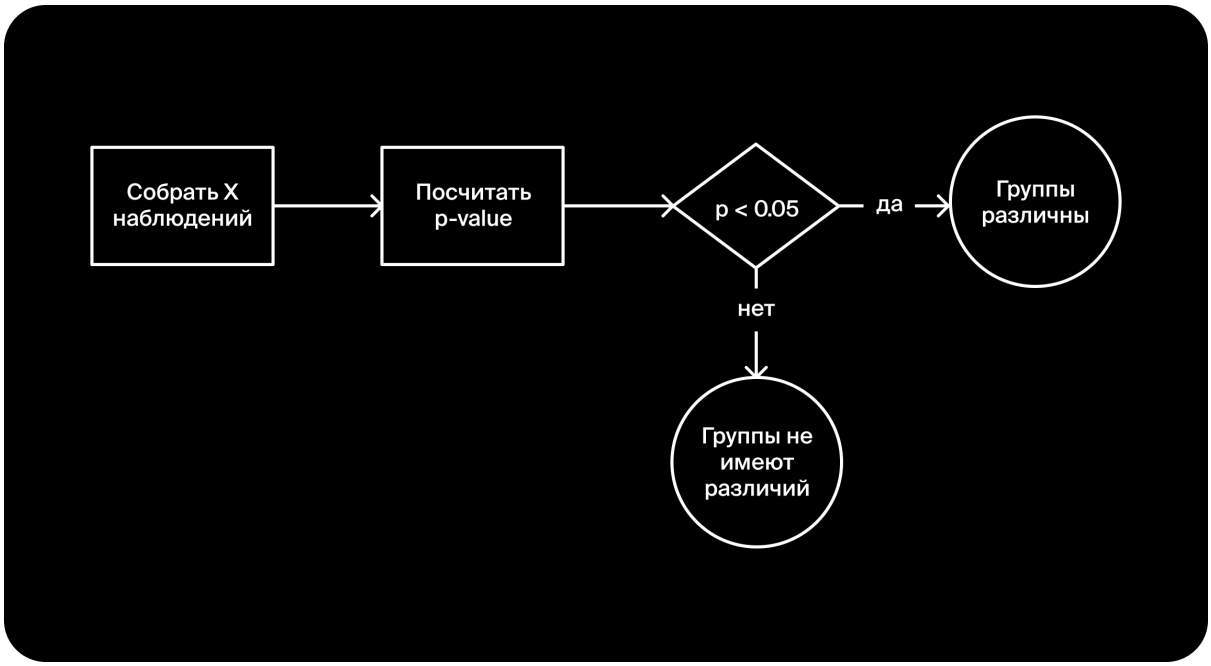
Принимая решение о длительности теста, учитывают циклические изменения трафика (ежедневные, еженедельные, ежемесячные) и то, как долго покупатель принимает решение о покупке исследуемых товаров или услуг.

Проблема подглядывания заключается в том, что в начале теста поступление новых данных значительно искажает общий результат. Каждый, даже небольшой фрагмент новых данных велик относительно уже накопленных — статистическая значимость достигается за короткий срок. Это одно из проявлений закона больших чисел. Если наблюдений мало, их разброс больше. Если их много, случайные выбросы успели компенсировать друг друга. Значит, если выборка слишком мала, легко увидеть различия. Это с точки зрения статистического теста будет

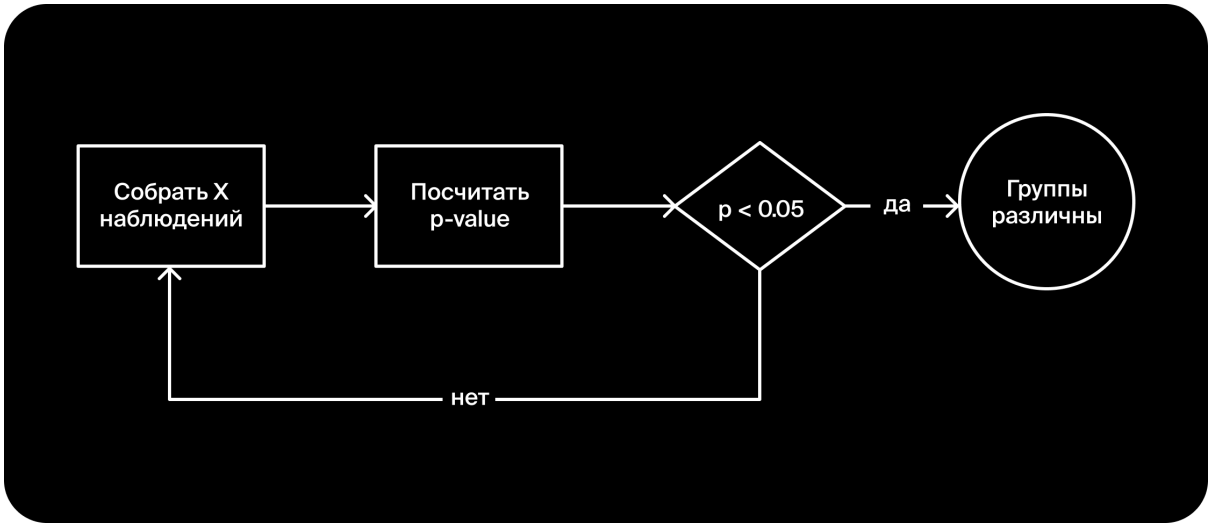
означать снижение *p-value* до значений, достаточно маленьких для отвержения нулевой гипотезы об отсутствии различий.

Чтобы исправить проблему подглядывания, размер выборки определяют до начала теста.

Вот правильная процедура A/B тестирования:



А вот так делать A/B тестирование не стоит:



Калькуляторы длительности теста и расчёта размера выборки

Простой способ найти размер выборки, не вдаваясь в детали — рассчитать в онлайн-калькуляторе.

Примеры калькуляторов:

- http://www.evanmiller.org/ab-testing/sample-size.html
- https://www.optimizely.com/sample-size-calculator/?conversion=20&effect=5&significance=95
- https://vwo.com/tools/ab-test-duration-calculator/

Эти сервисы хороши для оценки минимального необходимого объёма выборки, на которой будет заметно изменение показателя, если оно есть. Так можно оценить минимальную длительность теста.

Графический анализ метрик и определение предметной области

У применения калькуляторов, как и любого другого инструмента, есть плюсы и минусы.

Преимущества калькуляторов:

- Простота в случае подсчёта конверсии;
- Помогают избежать проблемы подглядывания;
- Помогают оценить минимальное необходимое время проведения теста.

Недостатки калькуляторов :

- Размер выборки — необходимый, но не достаточный критерий валидности теста.
- Калькуляторы считают, что конверсия и относительное изменение конверсии зафиксированы на всё время теста, что на практике не выполняется.
- Калькуляторы хорошо считают выборку только для конверсии. Есть калькуляторы и для других величин, но работать с ними гораздо сложнее.

Определение минимальной длительности теста в зависимости от предметной области

При определении длительности теста важно понимать, какие всплески активности бывают у вашей аудитории.

Причины всплесков разные:

- Будние или выходные дни;
- Праздники (увеличение спроса на «подарочные» товары);
- Распродажи, акции, маркетинговые активности (скидки увеличивают активность аудитории, меняя её покупательское поведение);
- Особые события (например, покупки товаров для школы в августе);
- Сезонность продукта (например, обогреватели);
- Деятельность конкурентов (конкуренты снизили цены на продукт, и активность ваших пользователей уменьшилась);
- Изменение в политической и экономической обстановке (кризис, рост цен, запрет на торговлю товаром, увеличение его стоимости из-за дополнительных пошлин).

Помимо всплесков активности следует понимать, какой **цикл реализации измеряемой метрики**? Чаще всего он связан с **циклом принятия решения о покупке** — временем, прошедшим от первой мысли о приобретении товара до его покупки.