Гайд по A/B-тестам для продакт-менеджеров

A/B-тест — эксперимент, где проверятся два или более варианта, чтобы определить наиболее эффективный с точки зрения метрик. Эффективность определяется с помощью вероятностей ложноположительных и ложноотрицательных случаев:

Ложноположительный случай или ошибка 1 рода

Результата на самом деле нет, но тест показывает, что он есть. Например, продакт-менеджер запустил эксперимент на повышение конверсии в первую покупку, который показал улучшение метрики и статистическую значимость. Поскольку была достигнута статистическая значимость, команда приняла решение опубликовать изменения, однако после публикации для всех пользователей целевая метрика упала.

Ложноотрицательный случай или ошибка 2 рода

Результат на самом деле есть, но тест его не показывает. Например, продакт-менеджер запустил эксперимент на повышение конверсии в первую покупку, однако улучшения метрик и статистической значимости добиться не получилось, поэтому изменения не были опубликованы. Такой тест не показал результата, хотя в реальности он был, однако эксперимент его не зафиксировал. Команда усилила продукт, но упустила этот момент.

Гайд по A/B-тестам для продакт-менеджеров

A/B-тест — эксперимент, где проверятся два или более варианта, чтобы определить наиболее эффективный с точки зрения метрик. Эффективность определяется с помощью вероятностей ложноположительных и ложноотрицательных случаев:

При организации A/B-теста важно предусмотреть:

- Каждая группа должна видеть только свой вариант — если пользователь попал в группу А, в течение всего эксперимента он видит вариант, предназначенный для группы А. Недопустимо, чтобы пользователь участвовал одновременно и в группе А, и в группе В, так как это влияет на его поведение и искажает результаты теста — пример интернет-магазина, где пользователь сначала находится в версии сайта без скидок, а потом попадает в версию сайта со скидками, после чего сразу бежит покупать.
- Группы независимы и берутся из одной генеральной совокупности — команда не сравнивает вариант А, который видели только пользователи из контекстной рекламы, с вариантом В, который видели только пользователи из таргетированной рекламы. Также важно предусмотреть, чтобы пользователи из двух группы не влияли на поведение друг друга — пример социальных сетей, где пользователи из контрольной и тестовой группы начинают обмениваться скриншотами о том, как у них поменялся дизайн сервиса.
- Распределение может быть неравномерным — можно разделять трафик в пропорции 60 на 40, 70 на 30 и так далее. Нет обязательного требования всегда делить 50 на 50 — пример почтового сервиса, который решил маркировать рекламу под входящие сообщения. Эта гипотеза ухудшала пользовательский опыт и провоцировала отрицательные отзывы, поэтому команда распределила трафик 90 на 10.

Гайд по A/B-тестам для продакт-менеджеров

Классификация метрик в A/B-тестах

В разных источниках метрикам дают разные названия, однако их суть всегда сохраняется:

Целевые метрики

Метрики, на которые команда планирует повлиять. Например, продакт-менеджер хочет поднять конверсию в первую покупку, поскольку повышение этой метрики позволяет вырастить продукт.
Если у команды получится это сделать, в компании увеличатся продажи. Целевые метрики обычно напрямую влияют на бизнес.

Прокси-метрики

Косвенные и опережающие метрики, которые обладают большей чувствительностью и изменчивостью, чем целевые. Например, прежде чем человек купит, пройдёт время, потому что пользователь должен оформить триал и погрузиться в бесплатный контент. Активность в триале и бесплатной части будет индикатором, по которым команда сможет заранее принять решение, как изменится целевая метрика. Допустим, если активности в триале или бесплатной части нет, команде не стоит ожидать высокой конверсии в первую покупку. Благодаря прокси-метрикам сокращают время эксперимента, их подбор — отдельная аналитическая задача.

Гайд по A/B-тестам для продакт-менеджеров

Классификация метрик в A/B-тестах

В разных источниках метрикам дают разные названия, однако их суть всегда сохраняется:

Debug-метрики

Метрики, которые используются для отслеживания технических моментов. Например, команда проводит эксперимент для повышения конверсии в первую покупку, поэтому ей важно проверить, что а) страница с платёжными инструментами быстро грузится и открывается, б) все способы оплаты работают, в) прочие технические моменты, которые указывают на работоспособность. Такие метрики помогают обнаружить ситуации, когда тест запущен, однако пользователи сталкиваются с багами, которые влияют на их поведение и соответственно результат эксперимента.

Guard-метрики

Оберегающие метрики, которые позволяют отслеживать, что в продукте не ломаются процессы или другие метрики. Например, команда работает над конверсией в первую покупку, однако тест может нарушить работу технической поддержки из-за того, что повышение конверсии сопровождалось увеличением количества обращений. Такие метрики позволяют отслеживать кейсы, когда повышение одного показателя может положительно или отрицательно влиять на другие части продукта.

Как применять метрики

Не во всех экспериментах нужны все 4 вида метрик — часто команды работают только с целевыми метриками, но если A/B-тест сложный и затрагивает множество пользовательских сценариев и операционных процессов, стоит задуматься и над прокси, и над debug, и над guard-метриками. Когда мы улучшаем один показатель, велика вероятность, что это положительно или отрицательно повлияет на другие метрики.

Гайд по A/B-тестам для продакт-менеджеров

Характеристики метрик в A/B-тестах

При работе с метриками и аналитике результатов важно учитывать следующие параметры:

Распределение метрики

Например, конверсия может принимать значение только от 0% до 100%, поэтому A/B-тесты на конверсию в покупку анализируют с помощью одних математических инструментов, а A/B-тесты на средний чек с помощью других, потому что средний чек может распределяться от $1 до бесконечности. В зависимости от типа и распределения метрики подбираются разные статистические критерии, которые определяют результаты анализа и которые мы обсудим дальше. Что важно запомнить — нельзя планировать и анализировать A/B-тесты на конверсию так же, как A/B-тесты на средний чек или другой тип метрики с отличающемся распределением.

Чувствительность метрики

Насколько быстро можно заметить изменения по метрике. Например, CTR рекламы обладает высокой чувствительностью — эксперимент можно провести за несколько недель и даже дней. В то же время LTV как правило обладает низкой чувствительностью, так как он меняется в течение нескольких месяцев или вообще лет. Если метрика обладает низкой чувствительностью, значит, команде необходимо закладывать дополнительное время на эксперимент, чтобы заметить изменения, либо прибегать к прокси-метрикам, которые коррелируют с целевой. Например, команда хочет измерить, как гипотеза повлияла на LTV 6 месяца, но столько ждать нет времени, поэтому для анализа берут прокси-метрику в виде LTV 2 месяца.

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

A/B-тест всегда начинается с формулирования гипотезы. В рамках эксперимента говорят про нулевую и альтернативную гипотезы:

Нулевая гипотеза

Предположение о том, что не существует взаимосвязи между выборками в эксперименте. Например, что команда запустит A/B-тест, по итогу которого разницы между контрольной и тестовой группы не будет. Допустим, что продакт-менеджер запускает эксперимент на CTR — нулевая гипотеза подразумевает, что изменения в продукте никак не повлияют CTR, что он останется тем же.

Альтернативная гипотеза

Между контрольной и тестовой группой разница есть, что изменения затронут и улучшат CTR. Если верна альтернативная гипотеза, значит, мы как продакт-менеджеры улучшили продукт.

По итогу если верна нулевая гипотеза, изменений не произошло. Но если альтернативная — мы добились результата.

Статистические критерии

Чтобы проверить нулевую и альтернативную гипотезы, применяют статистические критерии — математические правила, на основе которых принимают решение о верности той или иной гипотезы. Это язык, который помогает перевести продуктовые гипотезы в плоскость статистики, и основа для дальнейшей аналитики.

Если у нас в контрольной группе получился CTR 2%, а в тестовой 3%, мы не можем только на основе этих цифр утверждать, что улучшили продукт. Чтобы это сделать, нужно к полученным данным применить математический аппарат, после чего принимать решение.

При аналитике A/B-тестов команда оперирует не на уровне утверждений «да» или «нет», у нас 1 или 0, а на уровне вероятностей. Ранее мы указывали, что эффективность A/B-тестов определяется на основе вероятности ложноположительных и ложноотрицательных случаев.

Обсудим каждый кейс и посмотрим, как здесь работают статистические критерии.

07/23

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

P-value (α или альфа)

За ложноположительное решение отвечает показатель P-value (α или альфа) — чаще всего берут 95% уровень значимости. Бытовое и практическое определение звучит так, что мы на 95% уверены в том, что не допустили ошибку первого рода или ложноположительное решение, когда результата нет, но нам A/B-тест показывает, что он есть. Если мы за основу берём 95%, в тесте должен получиться p-value меньше 0,05 (=100% - 95%), чтобы можно было отвергнуть нулевую гипотезу и принять за верную альтернативную.

P-value — уровень риска, на который готова пойти команда и бизнес при проведении A/B-тестов. Допустим, что команда провела 1000 A/B-тестов. Абсолютно во всех верной оказалась нулевая гипотеза, то есть целевая метрика в эксперименте не изменилась. При уровне значимости 95% команда по 950 тестам примет верное решение, что нельзя их раскатывать на всех пользователях. Однако по 50 тестам будет принято ложноположительное решение — они показывали результат там, где его на самом деле не было. Это бесполезные изменения, которые команда раскатит в продукте. В лучшем случае они никак не затронут метрики, а в худшем их поломают.

Чтобы рассчитать P-value, необходимо как раз применить статистический критерий. Статистических критериев существует большое количество. Их выбор и подбор под конкретный кейс и продукт — отдельная задача, которую обычно в продуктовой команде решают аналитики.

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

На практике чаще всего используют следующие критерии на основе распределения метрики:

Примеры метрик	Распределение метрики	Статистический критерий
• Среднее время звонка при работе с лидами • Длина сессии в продукте	Зависит от продукта, но чаще всего нормальное распределение *, поскольку продавцы работают по одному скрипту, который предполагает ± похожую длительность, например, когда большинство звонков попадут в диапазон 15–20 минут * — колоколообразный и симметричный график, который часто встречается в статистике. Примеры	T-критерий Стьюдента
• Средний чек в интернет-магазине • LTV	Зависит от продукта, но чаще всего распределение будет далеко от нормального из-за большого разброса. Например, одни пользователи совершают покупки с небольшим чеком в несколько десятков долларов, а другие могут делать дорогостоящие покупки за сотню и даже тысячу	U-критерий Манна-Уитни
• Конверсия в первую покупку • Retention	Конверсия имеет разные определения, но обычно под ней понимают долю. Например, в продукт пришло 100 новых пользователей, 3 сделало первую покупку. Тогда конверсия составит 3%	Z-критерий Фишера

09/23

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

Какие нюансы следует учесть при работе со статистическими критериями:

1. Критически относиться к калькуляторам и шаблонам, если неясно, какие статистические критерии в них используются, иначе это черный ящик. Предположим, что в калькулятор зашит T-критерий, а распределение метрики в A/B-тесте содержит выбросы, с которыми этот тест плохо справляется. Поэтому можно или поменять критерий, или трансформировать распределение, убрав выбросы, чтобы к нему можно было применить T-критерий.
2. Выбор статистического критерия для A/B-теста может быть непростой задачей, которая потребует аналитических компетенций, поэтому если продакт-менеджер чувствует, что их недостаёт, стоит или подтянуть знания, или позвать на помощь опытного аналитика.
   Если нет времени во всем разбираться, а решение нужно принять здесь сейчас, просто используйте статистические критерии из таблицы — с высокой вероятностью сработают именно они.

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

Рассмотрим самый распространенный вариант, когда продуктовой команде необходимо рассчитать P-value для A/B-теста, где тестировалась конверсия. Для решений этой задачи в подавляющем большинстве случаев используют Z-тест. Допустим, в эксперименте была контрольная и тестовая группа, где получились следующие показатели:

Метрики	Контрольная	Тестовая
Количество пользователей (наблюдений)	1500	1300
Количество новых клиентов	250	234
Конверсия в первую покупку	15%	18%

Z-тест = Разница долей / Стандартная ошибка = 0,03 / 0,01409 = 2,129

Формула разницы долей = p2 - p1 = 0,18 - 0,15 = 0,03, где p2 — конверсия по тестовой группе, а p1 — конверсия по контрольной группе.

Формула стандартной ошибки = √p1 * (1 - p1) / n1 + p2 * (1 - p1) / n2 = √0,15 * (1 - 0,15) / 1500 + 0,18 * (1 - 0,18) / 1300) = √0,000085 + 0,000113 = √0.000198 = 0,01409, где n2 — количество наблюдений в тестовой группе, а n1 — количество наблюдений в контрольной.

Значение 2,129 соответствует P-value = 0,033254. Полученный P-value < 0,05. Значит, продуктовая команда может отвергнуть нулевую гипотезу и принять за верную альтернативную — мы добились улучшения метрик. Часто здесь можно встретить выражение: «Мы получили статистически значимый результат». Статистическая значимость подразумевает, что команда получила уровень P-value меньше целевого значения 0,05.

Гайд по A/B-тестам для продакт-менеджеров

Проверка гипотез через A/B-тесты

Также встречаются расчёты через доверительные интервалы — диапазон значений, которые с определённой вероятностью принимают метрики в эксперименте. Для этого нужно преобразовать формулу статистического критерия. Но суть от этого преобразования никак не меняется — с помощью доверительных интервалов команда придёт к аналогичным выводам:

Доверительный интервал = Разница долей +- Коэффициент * Стандартная ошибка = (0,18 – 0,15) ± 1,96 * (√0,18 × (1 – 0,18) ÷ 1300 + 0,15 × (1 – 0,15) ÷ 1500) = 0,03 ± 1,96 * 0,01409 = 0,03 ± 0,027 = [0,003; 0,057]

1,96 — коэффициент, который соответствует уровню доверия 95%.
Какие еще используют коэффициенты:

Коэффициент	Уровень доверия
1,282	80%
1,440	85%
1,645	90%
1,960	95%
2,000	95,4%
2,576	99%
2,807	99,5%
3,291	99,9%

Если доверительный интервал не содержит 0, значит, результат признается статистически значимым. Если доверительный интервал включает 0, допустим, был получен интервал [-0,001; 0,061], результат признается статистически незначимым при заданном уровне доверия. Доверительные интервалы — альтернативный способ, но важно помнить, что их формула определяется статистическим критерием.

Гайд по A/B-тестам для продакт-менеджеров

Выборка и длительность A/B-теста

Представим, A/B-тест не показал статистически значимого результата

Значит ли это, что нулевая гипотеза оказалось верной, а эксперимент прошел зря? Вообще не факт, поэтому здесь важно поговорить про ложноотрицательные решения. Это ошибка 2 рода и ситуации, когда результат в эксперименте на самом деле есть, но тест его не показывает, поэтому команда признает результаты A/B-теста неудачными.

Если для ошибки 1 рода команда смотрит на P-value, то для ошибки 2 рода необходимо проанализировать мощность. **Мощность** — вероятность, что команда не пропустит полезные изменения, насколько верно была отклонена нулевая гипотеза. Допустим, что команда провела 1000 A/B-тестов, и абсолютно все оказались успешными. При классической мощности 80% или 0,8 в выборке 800 A/B-тестов будут опубликованы, так как они показали результат, однако 200 будут ошибочно выкинуты в корзину — на самом деле их нужно было тоже выкатывать.

Ошибка 2 рода считается по формуле 100% - 80% = 20% или 0,2. Чем выше уровень мощности, тем меньше вероятность пропустить сработавший эксперимент, поэтому при проведении A/B-тестов ее стараются всячески повышать, поскольку никто не хочет выбрасывать полезные изменения, которые сделала команда.

По итогу P-value — есть ли статистически значимое изменение в метрике или нет, какую из гипотез мы принимаем (нулевую или альтернативную). Мощность — насколько мы вообще верно мы отклоняем гипотезы и какой эффект получаем в метриках (конкретное изменение на 3%, 5%, 10% или другое значение). Упрощая можно сказать, что P-value — есть ли эффект, а мощность — правильно ли он зафиксирован и каким получился.

Гайд по A/B-тестам для продакт-менеджеров

Выборка и длительность A/B-теста

Разберем, как обе ошибки работают вместе. Например, команда провела 1000 A/B-тестов при уровне значимости 95% и мощности 80% — допустим, в 700 A/B-тестах не было зафиксировано изменений, а в 300 были. Тогда:

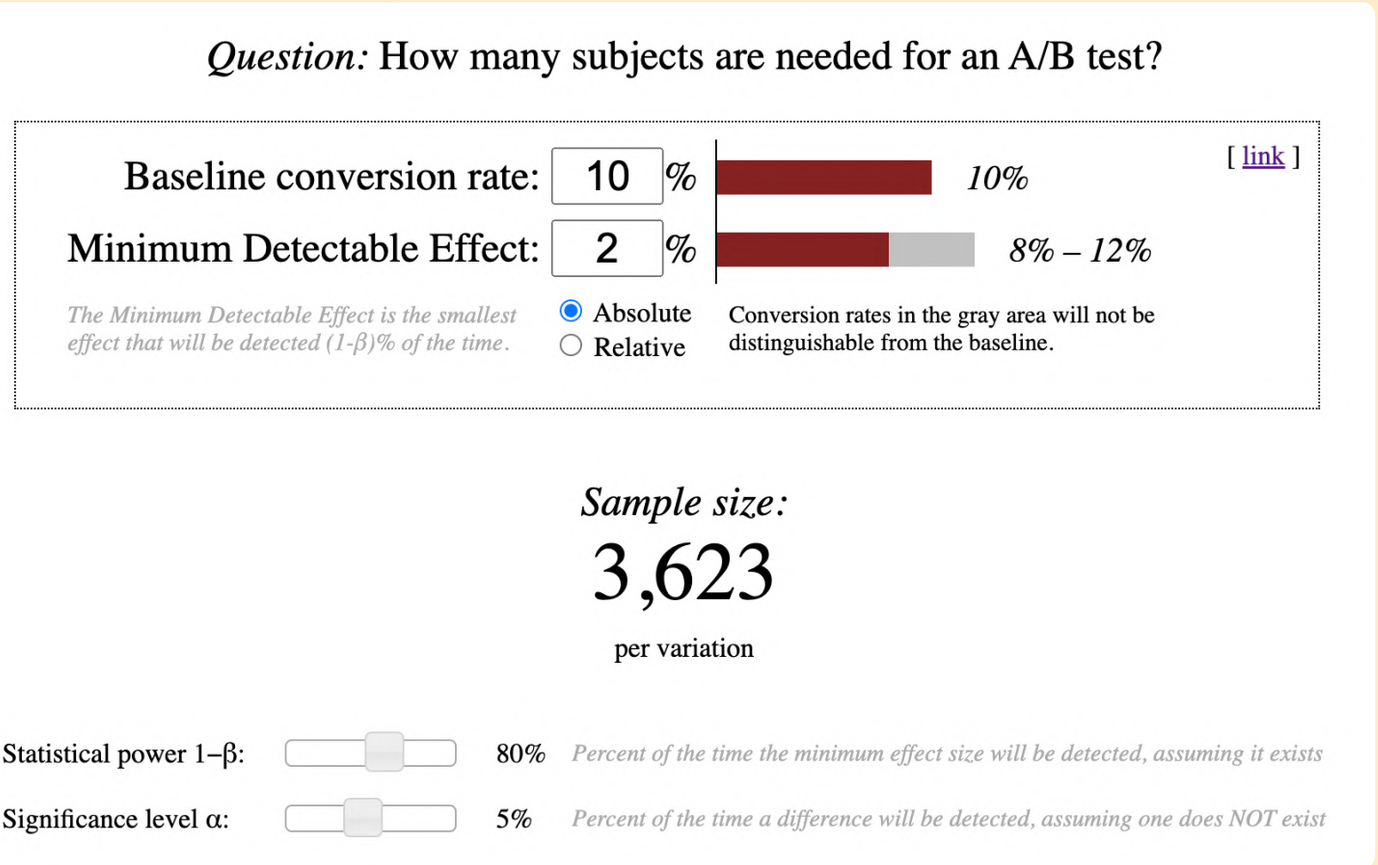
700 A/B-тестов без изменений	300 A/B-тестов с изменениями
• 665 A/B-тестов не статзначимы, поэтому не были опубликованы (95%) • 35 A/B-тестов статзначимы и были опубликованы, но сработала ошибка 1 рода, их на самом деле нельзя было публиковать (5%)	• 240 A/B-тестов статзначимы, поэтому были опубликованы (80%) • 60 A/B-тестов нестатзначимы и не были опубликованы, но сработала ошибка 2 рода, их на самом деле нужно было публиковать (20%)

- В результате из 1000 A/B-тестов по 95 экспериментам команда приняла ошибочное решение, так как сработали ложноположительные и ложноотрицательные вероятности. Чтобы снизить это количество и в целом балансировать обе ошибки, используют показатель MDE или minimum detectable effect — ожидания и минимальные изменения в метриках, которые команда хочет обнаружить в A/B-тесте с заданными порогами ошибок 1 и 2 рода. MDE используется для определения размера выборки в эксперименте.
- Расчеты MDE сложны с точки зрения математики, потому что в них также задействованы статистические критерии и прочие параметры, например, какое распределение мы закладываем в эксперимент (50 на 50, 60 на 40, 70 на 30 и и так далее), поэтому оставим их за скобками, однако разберем ключевые принципы.

Гайд по A/B-тестам для продакт-менеджеров

Выборка и длительность A/B-теста

Например, возьмем A/B-тест с тестом конверсии в первую покупку и допустим, что сейчас она равна 10%. MDE будет обозначать изменение метрики, которое команда хочет зафиксировать в эксперименте — предположим, что продакт-менеджеру важно, чтобы конверсия изменилась на 2%. Для демонстрации воспользуемся калькулятором EvanMiller.



При мощности 80% и уровне доверия 95% в A/B-тесте потребуется 3623 пользователей на один вариант или 7246 на оба. Это значит, что изменение конверсии в 2% можно заметить только на выборке из 7246 пользователей, поэтому так важно, чтобы эксперимент открутился именно на такой выборке Если в эксперименте примет участие меньшее число пользователей, прироста в 2% обнаружить не удастся — так работает MDE.

Гайд по A/B-тестам для продакт-менеджеров

Выборка и длительность A/B-теста

Как все параметры влияют друг на друга:

— Чем меньше MDE, тем больший размер выборки потребуется в эксперименте и наоборот. Если MDE станет 1%, а не 2%, нужно уже не 3623 пользователя, а 14313.

— Чем выше уровень доверия, тем больший размер выборки потребуется в эксперименте и наоборот. Если уровень доверия станет 99%, а не 95%, потребуется уже не 3623 пользователя, а 6076. Повышая уровень доверия, команда уменьшает вероятность ошибки 1 рода, но таким образом ей требуется больший объём выборки.

— Чем выше мощность, тем больший размер выборки потребуется в эксперименте и наоборот. Если мощность станет 95%, а не 80%, потребуется уже не 3623 пользователя, а 5366. Повышая мощность, команда уменьшает вероятность ошибки 2 рода, но таким образом ей вновь потребуется больший объём выборки.

— Если уменьшать MDE и одновременно повышать уровень доверия с мощностью, эксперимент будет требовать большего размера выборки и наоборот. При MDE 1%, уровне доверия 99% и мощности 95% потребуется уже не 3623 пользователя, 32610. Чем менее рискованные гипотезы команда тестирует при низких значениях ошибки 1 и 2 рода, тем больший объем выборки требуется и наоборот.

Гайд по A/B-тестам для продакт-менеджеров

Выборка и длительность A/B-теста

Допустим, что команда решила остановиться на выборке из 7246 пользователей, которые равномерно распределяться между вариантами A и B. Предположим, что в день сайт продукта посещает 500 пользователей — по итогу эксперимент займет 7246 / 500 = 15 дней.

Какие ещё параметры влияют на длительность эксперимента:

1. **Сезонность** — в определенные дни, недели и месяцы пользователи более активны, чем в другие. В разные периоды может быть разное количество пользователей, что будет влиять на время проведения эксперимента.
2. **Метрика** — разные метрики имеют разную чувствительность. Например, конверсия способна меняться быстро, а средний доход с пользователя скорее всего потребует времени, поэтому тест на средний доход продлится гораздо дольше.
3. **Окно закрытия** — например, если команда тестирует продление месячной подписки, эксперимент должен продлится больше месяца, поскольку именно в течение этого срока клиенты сделают целевое действие, и накопятся данные об их поведении.

Гайд по A/B-тестам для продакт-менеджеров

Процессы в A/B-тесте

Как выглядит работа с экспериментом:

Шаг	Пример
1. Определить цель и гипотезу для проверки	Если убрать из формы заявки почту, оставив там только имя и телефон, это увеличит конверсию в лид на 1%.
2. Определить переменную для проверки	Для проведения эксперимента необходимо изменить один элемент на сайте — форму заявки, убрав оттуда поле с почтой.
3. Определить метрики для проверки	Ключевой метрикой для проверки станет конверсия в лид. В качестве guard-метрик выступят процент дозвонов до лида и конверсия в оплату.
4. Разработать контрольный и проверочный вариант	Контрольный вариант — текущая форма с тремя полями (имя, почта, телефон), проверочный вариант — форма с двумя полями (имя, телефон).
5. Определить размер выборки и поделить пользователей на группы	Для проведения эксперимента понадобится 220 тысяч пользователей, которые распределятся по группам 50 на 50.
6. Проверить параллельные тесты и запустить эксперимент	Уточнить, какие тесты сейчас проходят, удостовериться, что нет пересечений, и запустить A/B-тест.
7. Дождаться сбора данных	Помнить про проблему подглядывания и не делать преждевременных выводов — эксперимент должен открутиться на 220 тысяч пользователей.
8. Проанализировать данные и принять решение	Посчитать статистическую значимость, оценить успешность A/B-теста и определить следующие шаги.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Ситуация №1.

«Запустим A/B-тест и посмотрим, какие метрики получаются — конверсия в первую покупку и средний чек первой покупки. Соберутся данные по двум метрикам, а дальше посчитаем по ним данные — одним выстрелом двух зайцев».

Главные ошибки и риски:

1. Отсутствует дизайн эксперимента и гипотезы — команда бежит сразу делать, а не планирует A/B. Когда мы работаем с A/B-тестом, важно сформулировать нулевую и альтернативную гипотезу, которые уйдут в проверку.
2. Некорректная работа с метриками, которые требуют разного подхода. Например, для конверсии в первую покупку будет один расчёт MDE, а для среднего чека другой. С высокой вероятностью получатся разные значения, которые сформируют разный размер выборки для теста конверсии и среднего чека. Кроме того, для подсчёта P-value потребуются разные статистические критерии, поскольку конверсия в покупку распределяется одним образом, а средний чек другим.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Как избежать ошибок:

1. Создавать дизайн эксперимента, где нужно формулировать нулевую и альтернативную гипотезу. Например, что если упростить форму регистрации, убрав оттуда 3 поля, конверсия в первую покупку вырастет на 1%. Нулевая гипотеза — разницы между тестовой и контрольной группы нет. Альтернативная — разница есть. Не «запустим и посмотрим, что будет», а понимание, что хотим сделать.
2. Закладывать в эксперимент в качестве целевой одну метрику, например, что раз гипотеза предполагает увеличение конверсии в первую покупку, значит, весь процесс проведения и аналитики будет выстраиваться вокруг неё. Если хочется взять средний чек, его можно рассмотреть в качестве guard-метрики, что рост конверсии не должен сопровождаться падением среднего чека в первой покупке.
3. Изучать, как работать с каждом типом метрики при проведении A/B-тестов. Например, что для расчёта P-value конверсии в первую покупку будет использоваться один статистический критерий. Поскольку часто это доля, поэтому команда возьмёт Z-тест Фишера. Для расчёта P-value среднего чека первой покупки будет использовать другой статистический критерий, например, U-критерий Манна-Уитни.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Ситуация №2.

«Запустим A/B-тест и по ходу эксперимента посмотрим, как ведут себя наши изменения. Попробуем покрутить эксперимент на небольшой выборке, чтобы собрать минимальную статистику и решить, что делать».

Главные ошибки и риски:

1. Отсутствует дизайн исследования — команда вместо планирования A/B-тест пытается разработать упрощённые правила, который ей якобы помогут принять решения. Кроме того, непонятно, на какой эффект на метрики планируется получить, из-за этого ей трудно спланировать ход A/B-теста, в частности объём требуемой выборки.
2. Проблема подглядывания — команда планирует смотреть, когда P-value достигнет нужного значения, чтобы остановить эксперимент. Обычно P-value в первое время ведёт себя нестабильно, резко меняя свои значения, но по мере увеличения выборки в эксперименте стабилизируется, поэтому так важно, чтобы A/B-тест открутился на всей выборке.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Как избежать ошибок:

1. Заранее продумывать, какой эффект на метрику команда ожидает получить в эксперименте. Это поможет определить MDE и рассчитать размер выборки, который необходим для тестирования. Если не считать MDE, значения ошибок и размер выборки, принятие решений в эксперименте превратится в угадайку и лотерею. Проработка этих нюансов также поможет при проверке гипотез.
2. Если команда понимает, что у неё не получится провести A/B-тест, возможно стоит подумать над другими методами исследований. Например, вместо A/B-теста воспользоваться опросами или вообще провести качественное исследование. Следует помнить, что хоть A/B-тест считается эталоном научного эксперимента, однако часто существуют более быстрые и дешёвые способы проверить гипотезу — не на каждую гипотезу нужен A/B.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Ситуация №3.

«Запустим сразу много A/B-тестов, чтобы проверить как можно большее количество гипотез, которые нагенерирует команда. Чем больше гипотез проверим, тем быстрее будем принимать решения и развивать продукт».

Главные ошибки и риски:

1. С высокой вероятностью у команды отсутствует стратегия и discovery-процессы по тому, как следует развивать продукт, поэтому возникает желание тестировать все подряд без понимания, зачем нужны такие тесты. Кроме того, опасен сам подход, когда каждая идея будет тестироваться через A/B-тест, что скорее всего сильно удлинит цикл принятия решений.
2. Если запустить сразу много A/B-тестов, есть риск создать пересечения между экспериментами.
   Так как все эксперименты влияют друг на друга, что пользователь был сначала в одном A/B-тесте, а затем сразу в другом, это с высокой вероятностью повлияет на их поведение и исказит результаты в обоих тестах.

Гайд по A/B-тестам для продакт-менеджеров

Популярные ошибки

Как избежать ошибок:

1. При работе с исследованиями и в частности с A/B-тестами, важно в первую очередь проанализировать стратегию и собрать план исследований, которые помогут её реализовать. Смысл не столько в том, что сделать множество экспериментов, а выбрать самые важные, которые позволят команде собрать данные для принятия ключевых решений и приблизиться к цели.
2. Хороший A/B-тест — часто долгий и затратный процесс, начиная с этапа проработки дизайна эксперимента и заканчивая аналитикой и проверкой гипотез. Если продакт-менеджер хочет тестировать все через A/B-тесты, потому что «гипотезы нужно проверять только на основе цифр, ведь без них все субъективно», он увеличивает цикл принятия решений, что может тормозить развитие продукта и скорость команды.
3. При запуске экспериментов стараться проектировать их таким образом, чтобы между ними не было пересечений. Например, один из популярных способов — разделить всю аудиторию на группы, допустим, 50. Когда команда проектирует A/B, она выбирает 2 группы из 50 — одну для контрольной а другую для тестовой. Таким образом, можно одновременно провести сразу 25 A/B-тестов с минимальным риском пересечений.