



команда роста на аутсорсе
для системного улучшения ключевых метрик

Фреймворк планирования и проверки A/B-теста

Пошаговое руководство для корректного проведения и анализа A/B-экспериментов

1. Формулировка гипотезы (SMART-гипотеза)

Если [изменение]	Например: 'Если я заменю заголовок на кнопке с "Отправить" на "Получить расчёт"'
то [метрика] изменится на [величина]	Например: 'то конверсия в заявку вырастет на 15%'
потому что [причина]	Например: 'потому что конкретный призыв снижает неопределённость у пользователя'
Пример полной гипотезы	Если мы изменим текст кнопки 'Купить' на 'Добавить в корзину' на странице товара, то количество добавлений в корзину увеличится на 5%, потому что формулировка 'Добавить в корзину' более точно отражает действие пользователя и снижает барьер к первому шагу.

2. Целевые метрики

☐ Primary-метрика

Единственный показатель, по которому принимается решение. Должна отражать бизнес-цель (например, конверсия в заявку, выручка). Не выбирайте две первичные метрики.

☐ Определение Primary-метрики

Например: 'Конверсия в покупку (Purchase Conversion Rate)'

☐ Единица измерения Primary-метрики

Например: 'Процент от уникальных посетителей'

☐ Guardrail-метрики

Показатели, которые не должны просесть при оптимизации основной метрики (3-7 метрик). Мониторятся вместе с Main-метрикой. Если CI guardrail пересекает '-5%', это подозрительно (kariernik.ru).

☐ Примеры Guardrail-метрик (с определением)

retention D7 (процент вернувшихся пользователей через 7 дней); Crashes per user (количество сбоев на пользователя); Load time (время загрузки страницы); NPS (индекс лояльности клиентов); Support tickets per user (количество обращений в поддержку на пользователя); Time spent browsing (время, проведенное на сайте без целевого действия); Revenue per user (выручка на пользователя) (kariernik.ru)

☐ Прокси-метрики (если применимо)

Используются при низком трафике для ускорения сбора данных. Обоснование: 'Вместо конверсии в покупку (долгосрочная метрика) используем 'добавление в корзину' как прокси, так как это первый шаг к покупке и имеет больший объем данных.'

3. Расчет выборки и длительности

Текущая конверсия (базовая)	Например: 5%
Минимальный детектируемый эффект (MDE, абсолютный)	Например: 0.5 п.п. (процентных пункта)
Минимальный детектируемый эффект (MDE, относительный)	Например: 10% (от 5% базовой конверсии)
Статистическая значимость (альфа)	0.05 (5%)
Статистическая мощность (бета)	0.8 (80%)
Расчетный размер выборки на группу	Рассчитывается с помощью калькуляторов A/B-тестов (например, Optimizely, Evan Miller). Зависит от базовой конверсии, MDE, альфы и беты.
Расчетная длительность теста	Зависит от размера выборки и ежедневного трафика. Например: '3 недели при 1000 посетителях в день и конверсии 1%.'
Обоснование длительности	Тест должен охватывать как минимум один полный недельный цикл, чтобы учесть разницу в поведении пользователей в будни и выходные. Учитывать сезонность и внешние факторы (например, распродажи, рекламные кампании).

4. Подготовка к сплитованию

☐ Метод рандомизации

Использование стабильного идентификатора (user_id или cookie) и хэширования. Хэш от идентификатора делится на 100, остаток определяет группу (например, <50 — контроль, >50 — эксперимент). Это гарантирует, что пользователь не 'перепрыгнет' из группы в группу.

☐ Проверка SRM (Sample Ratio Mismatch)

SRM — статистически значимое расхождение между заданным и фактическим соотношением групп. Проверяется с помощью хи-квадрат теста. Если p-value < 0.001 (часто используется более строгий порог, чем 0.05, для SRM), эксперимент невалиден (kariernik.ru, zasp/python.ru). SRM означает, что механизм распределения сломан, и результаты теста невалидны (microsoft.com, zasp/python.ru).

☐ A/A-тест

План проведения: Запустить тест, где обе группы получают одинаковый вариант. Ожидаемые результаты: Отсутствие статистически значимой разницы между группами. Если разница есть, это указывает на техническую проблему (утечка данных, боты, несимметричная загрузка) (kariernik.ru).

☐ Сегментация аудитории (если применимо)

Описание, как будут делиться сегменты (например, по источнику трафика, типу устройства). Убедиться, что тесты не пересекаются на одних и тех же пользователях, если проводятся одновременно.

☐ Исключение внутренних пользователей

План: Исключить сотрудников компании, разработчиков, QA-инженеров из выборки теста, чтобы их поведение не искажало результаты.

5. Проведение и мониторинг

☐ План QA

Тестирование варианта В перед запуском: кроссбраузерность, адаптивность, скорость загрузки, корректность работы трекинга событий, отсутствие ошибок в коде.

☐ Мониторинг в реальном времени

Список метрик: Количество пользователей в каждой группе (для SRM), основные технические метрики (ошибки сервера, время ответа), ключевые guardrail-метрики (например, количество сбоев). Пороговые значения аномалий: Например, 'более 1% ошибок сервера', 'отклонение SRM более 0.5% на больших выборках (10M+ пользователей)'. (kariernik.ru)

□ План действий при обнаружении аномалий

Остановка теста, анализ причин, исправление, перезапуск теста (при необходимости).

□ Коррекция множественных сравнений (если применимо)

Метод: Поправка Бонферрони, Холма-Бонферрони, FDR (False Discovery Rate). Применяется, если анализируется много метрик или гипотез одновременно, чтобы снизить вероятность ложноположительных результатов.

6. Анализ результатов

□ Инструменты анализа

Название: Например, 'SQL-запросы, Python (SciPy, Statsmodels), R (A/B testing package), Google Analytics, Amplitude, Tableau'.

□ Критерии принятия решения

p-value: <0.05 для статистической значимости. Доверительные интервалы: Не должны пересекаться или должны показывать значимый прирост. Практическая значимость: Минимальный эффект, который оправдывает внедрение с точки зрения бизнеса (например, прирост конверсии на 2% или увеличение выручки на \$X). Небольшой эффект стоит внедрять, если затраты минимальны, а масштаб трафика велик.

□ Оценка ошибок Type S и Type M (если применимо)

Type S (Sign Error): Ошибка в направлении эффекта (например, тест показал рост, а на самом деле падение). Type M (Magnitude Error): Ошибка в величине эффекта (например, тест показал 10% прирост, а на самом деле 2%). Важно для оценки рисков.

□ Сегментный анализ

План: Анализ результатов по ключевым сегментам аудитории (например, новые/возвращающиеся пользователи, мобильные/десктопные, по источникам трафика) для выявления скрытых эффектов.

7. Документирование и выводы

□ Место для хранения результатов

Ссылка: Например, 'Confluence, Notion, Google Docs/Sheets, внутренний репозиторий A/B-тестов'.

□ Формат записи выводов

Что сработало: Описание изменений и их влияние на Primary-метрику и Guardrail-метрики. Почему: Анализ причин наблюдаемых изменений. Следующие шаги: Решение о внедрении, дальнейшие тесты, масштабирование.

□ Уроки из теста

Какие новые знания получены о поведении пользователей, продукте или процессе тестирования. Что можно улучшить в будущих экспериментах.

Больше интересного в нашем блоге



развивайте бизнес-мышление с нашими исследованиями

Получить пользу



 rocketlab.bz

 [@rocketlab](https://t.me/rocketlab)