



команда роста на аутсорсе
для системного улучшения ключевых метрик

Чек-лист для аудита A/B-теста: от подготовки до анализа

Детальный чек-лист, который поможет провести аудит A/B-теста на всех этапах, помогая избежать распространённых ошибок.

Раздел 1: Подготовка и планирование

☐ Формулировка гипотезы: SMART-критерии

Гипотеза конкретна, измерима, достижима, релевантна, ограничена по времени.

☐ Формулировка гипотезы: Одна переменная

Меняется только один элемент, чтобы однозначно определить причину изменения (Mindbox).

☐ Формулировка гипотезы: Нулевая и альтернативная гипотезы

Чётко сформулированы H0 (изменения не влияют) и H1 (изменения влияют).

☐ Формулировка гипотезы: Основная метрика (OEC)

Выбрана одна ключевая метрика, по которой будет приниматься решение (например, конверсия, CTR).

☐ Формулировка гипотезы: Защитные метрики

Определены метрики для контроля негативного влияния (например, время на сайте, отказы, средний чек).

☐ Формулировка гипотезы: Ожидаемый эффект

Определён минимальный детектируемый эффект (MDE), который вы хотите зафиксировать.

☐ Расчет выборки: Минимальный детектируемый эффект (MDE)

Определен минимальный процентный рост, который тест должен обнаружить (например, 10% роста конверсии с 3% до 3.3%).

☐ Расчет выборки: Статистическая мощность

Установлена статистическая мощность (обычно 80%), вероятность обнаружить эффект, если он существует.

☐ Расчет выборки: Уровень значимости (p-value)

Установлен уровень значимости (обычно 5%, $p\text{-value} < 0.05$), вероятность ложноположительного результата.

☐ Расчет выборки: Длительность теста

Минимум 7 дней для учета недельного цикла, рекомендуется 14 дней. Не более 4-6 недель, чтобы данные не устарели.

☐ Расчет выборки: Использование калькуляторов

Применены калькуляторы статистической значимости (например, Яндекс.Метрика, калькулятор Эвана Миллера).

☐ Техническая подготовка: A/A-тест

Проведен A/A-тест для проверки корректности системы сплитования и сбора данных до основного A/B-теста.

☐ Техническая подготовка: Скорость загрузки

Обе версии (A и B) загружаются с одинаковой скоростью, чтобы избежать искажения результатов.

☐ Техническая подготовка: Анти-мерцающий код (anti-flicker)

Реализован код для предотвращения «мерцания» страницы при загрузке варианта B.

☐ Техническая подготовка: Корректность трекинга

Все необходимые события и метрики отслеживаются корректно в обеих группах.

❑ Техническая подготовка: Правовые аспекты

Проверены правовые основания для обработки персональных данных пользователей.

❑ Сегментация аудитории: Равномерность распределения

Трафик равномерно распределяется между группами (например, 50/50).

❑ Сегментация аудитории: Репрезентативность

Группы репрезентативны и сопоставимы по ключевым характеристикам (новые/возвращающиеся пользователи, устройства).

❑ Сегментация аудитории: Выделение трафика

Под эксперимент выделено не более 5-10% трафика для снижения рисков.

Раздел 2: Проведение эксперимента

❑ Мониторинг: Sample Ratio Mismatch (SRM)

Ежедневный контроль равномерности распределения пользователей между группами. Если ожидалось 50:50, а получилось 45:55, это проблема (Habr).

❑ Мониторинг: Технические ошибки

Проверка отсутствия ошибок в коде, проблем с кешированием, блокировкой событий аналитики.

❑ Мониторинг: Внешние факторы

Учет праздников, акций конкурентов, изменений в рекламных кампаниях, которые могут повлиять на результаты.

❑ Мониторинг: 'Ошибка подглядывания' (peeking)

Избегать преждевременного просмотра и остановки теста до набора необходимой выборки.

❑ Мониторинг: Взаимодействие с другими тестами

Проверка отсутствия пересечений с другими активными A/B-тестами.

❑ Мониторинг: Динамика метрик

Отслеживание стабильности метрик в обеих группах на протяжении всего теста.

❑ Мониторинг: Негативные эффекты

Быстрое реагирование и остановка теста, если наблюдается значительное ухудшение ключевых метрик или пользовательского опыта.

❑ Контроль качества данных: Дубликаты

Проверка отсутствия дублирующихся записей или событий.

❑ Контроль качества данных: Аномалии

Выявление и анализ аномальных значений или всплесков в данных.

❑ Контроль качества данных: Целостность данных

Убедиться, что все данные собираются и передаются корректно, без потерь.

❑ Контроль качества данных: Отсутствие ботов/спама

Фильтрация нерелевантного трафика, который может исказить результаты.

❑ Контроль качества данных: Соответствие трекингу

Сравнение собираемых данных с ожидаемым трекингом, описанным на этапе планирования.

Раздел 3: Анализ результатов и принятие решений

❑ Статистический анализ: p-value

Проверка p-value (обычно < 0.05) для подтверждения статистической значимости различий.

❑ Статистический анализ: Доверительные интервалы

Оценка диапазона, в котором находится истинный эффект.

❑ Статистический анализ: Практическая значимость

Оценка, является ли статистически значимое изменение достаточно большим для бизнеса (например, +0.5% конверсии может не стоить усилий).

❑ Статистический анализ: Анализ защитных метрик

Проверка, не ухудшились ли другие важные метрики (например, средний чек, время на сайте).

❑ Статистический анализ: Отсутствие ложных результатов

Убедиться, что результат не является случайным колебанием (Mindbox: только треть гипотез, считающихся 'очевидно верными', подтверждается).

❑ Сегментный анализ: По устройствам

Анализ результатов отдельно для мобильных, десктопных устройств.

❑ Сегментный анализ: По источникам трафика

Анализ результатов для разных каналов привлечения пользователей.

❑ Сегментный анализ: Парадокс Симпсона

Проверка, не меняется ли направление эффекта при агрегации данных по сегментам.

❑ Документирование и выводы: База знаний

Все результаты, гипотезы, настройки и выводы теста занесены в общую базу знаний.

❑ Документирование и выводы: Воспроизводимость

Описание шагов и условий, позволяющих повторить эксперимент и получить схожие результаты.

❑ Общие рекомендации: Закон Тваймана

Если результат кажется слишком хорошим, скорее всего, это ошибка в данных или расчетах.

❑ Общие рекомендации: CUPED (Controlled-experiment Using Pre-Experiment Data)

Рассмотреть применение CUPED для снижения дисперсии и повышения чувствительности теста.

❑ Общие рекомендации: ОЕС (Overall Evaluation Criterion)

Всегда фокусироваться на одной главной метрике для принятия решения.

Больше интересного в нашем блоге



развивайте бизнес-мышление с нашими исследованиями

Получить пользу

