Эксперимент в Яндекс Метрике: как провести A/B-тест и что учесть при подготовке

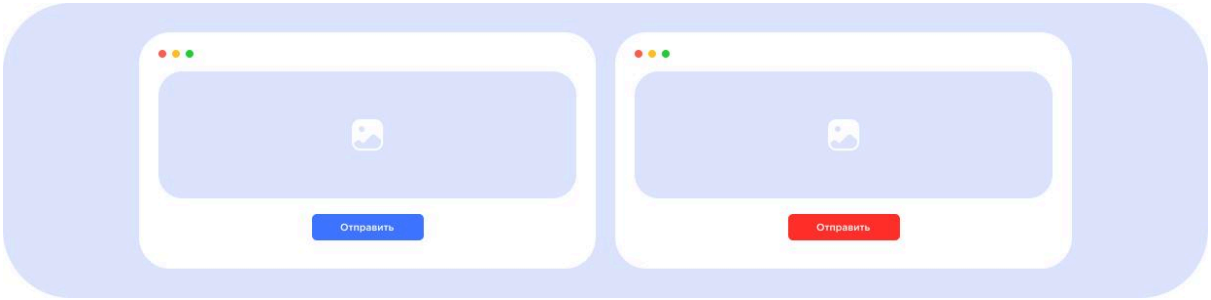
Что такое A/B-тестирование и как его провести с помощью инструмента «Эксперименты» в Яндекс Метрике.

Содержание

Что такое A/B-тестирование

A/B-тестирование проводится с целью сравнить две разные версии (вариант A и B) одного и того же продукта и выяснить, какая приносит лучший результат.

Например, тестируются два варианта кнопок на веб-странице. В ходе теста они случайным образом отображаются для разных пользователей в течение одного и того же периода времени. При этом пользователи разделены на равные группы: первая видит один вариант кнопки, вторая — другой.



Таким образом, тест позволяет проанализировать показатели на разных версиях сайта и принять решение, какой вариант оставить для улучшения взаимодействия пользователя с ресурсом.

Когда и зачем нужен A/B-тест

Предположим, у вас стоит задача — увеличить время пребывания пользователя на странице. Вы анализируете нужную страницу и выдвигаете гипотезы:

- если текст на странице разбить на разделы, пользователи будут проще его воспринимать и дольше находиться на странице;

- если на страницу добавить раздел с информацией о доставке, это тоже повысит вовлеченность;
- если добавить быструю навигацию по странице, пользователи смогут сразу переходить к нужному разделу, что увеличит их вовлеченность.

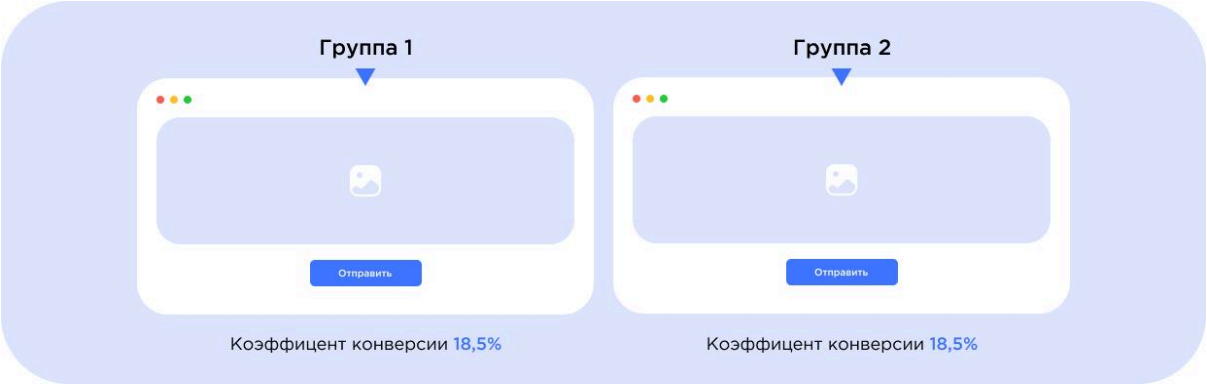
После того, как гипотезы сформулированы, их нужно или подтвердить, или опровергнуть. Для этого и проводится A/B-тестирование.

Другими словами, A/B-тест помогает во всех ситуациях, когда вы предполагаете, как можно улучшить продукт. Если же воплощать в жизнь все идеи, исключая этап тестирования, в лучшем случае можно не заметить изменений, в худшем — потерять время на доработки, получить снижение показателей и потерять прибыль.

5 советов при подготовке и проведении A/B-теста

Перед запуском A/B-теста убедитесь в надежности инструмента для его проведения, а также в качестве трафика, который хотите использовать для исследования. Вот несколько советов перед стартом:

1. **Если ранее вы не проводили тест с помощью данного инструмента или не уверены в однородности трафика, для начала проведите A/A-тест.** В ходе теста трафик делится на две группы. Каждой показывается одна и та же версия продукта. Если трафик идентичен и инструмент тестирования в порядке, то при изучении поведения пользователей на одной и той же версии продукта, поведение группы 1 будет таким же, как поведение группы 2:



Таким образом, минимизируется вероятность того, что пользователи изначально чем-то отличаются и данные по предстоящему A/B-тесту будет некорректны.

2. **Учитывайте периоды активности пользователей.** Например, вы продаете канцтовары и проводите тест в августе. В конце августа спрос на канцтовары резко вырос, показатели изменились. И если вы выдвинете решение на основе полученных результатов, то, например, в октябре вы можете получить не тот результат, который ожидали.

3. **Перед тестированием проанализируйте веб-страницу на предмет ошибок и багов.** Если в ходе теста пользователи будут сталкиваться с какой-либо

ошибкой, а после теста проблема исчезнет, то вы рискуете принять решение по искаженному результату теста. Или тест нужно будет проводить снова.

4. **В ходе одного теста тестируйте одну гипотезу.** И не запускайте одновременно несколько тестов одной и той же страницы. Иначе вы не сможете понять, что именно привело к повышению или понижению целевого показателя

5. **Если на сайте мало трафика, скажем 2000 визитов и 5 заявок в месяц, то, скорее всего, пока нет смысла тратить время на тестирование.** Проведение A/B-теста на таких условиях может занять месяцы до получения статистически значимого результата.

Проведение A/B-теста с помощью Varioqub от Яндекс Метрики

Недавно мы проводили A/B-тест с помощью бесплатной версии Varioqub. Он доступен в интерфейсе Яндекс Метрики в разделе «Эксперименты». Поэтому на реальном примере расскажем, какие действия необходимо выполнить до, во время и после теста.

Подготовка к тесту

Шаг 1. Цель

Обозначьте цель, которую хотите достичь с помощью теста. Например, снизить показатель отказов или увеличить число просмотров видеоролика.

Нам нужно было понять, на какую промостраницу сайта лучше вести рекламную кампанию.

Шаг 2. Целевой показатель

Обозначьте показатель, на основании которого вы будете принимать решение, какой вариант тестирования сработал лучше. Если цель — увеличить число просмотров видеоролика, то показатель, соответственно, — число просмотров видеоролика или количество кликов на кнопку запуска видео и т. п.

В нашем случае ориентиром служил показатель конверсии по цели в Яндекс Метрике, которая собирала успешные отправки форм. Также в эксперименте Яндекс Метрики можно выбрать два дополнительных показателя, если хотите увидеть результаты не только по основному. Мы использовали эту возможность и в качестве дополнительных показателей обозначили конверсию по цели, собирающей звонки, и показатель отказов.

Шаг 3. Гипотеза

Изучите продукт и выдвиньте гипотезу. Если добавим элемент X, то произойдет увеличение показателя.

Наша гипотеза состояла в том, что конверсия на варианте A будет отличаться от конверсии на варианте B. Это поможет понять, куда лучше вести трафик.

Шаг 4. Аудитория

Определите, какой трафик хотите исследовать. Например, пользователей по определенной рекламной кампании, пользователей, которые уже были на вашем сайте, или пользователей только мобильных устройств. Если нужно, подготовьте рекламные кампании, добавьте utm-метки.

Также определите, сколько процентов трафика вы планируете использовать: 100% или, например, 70%.

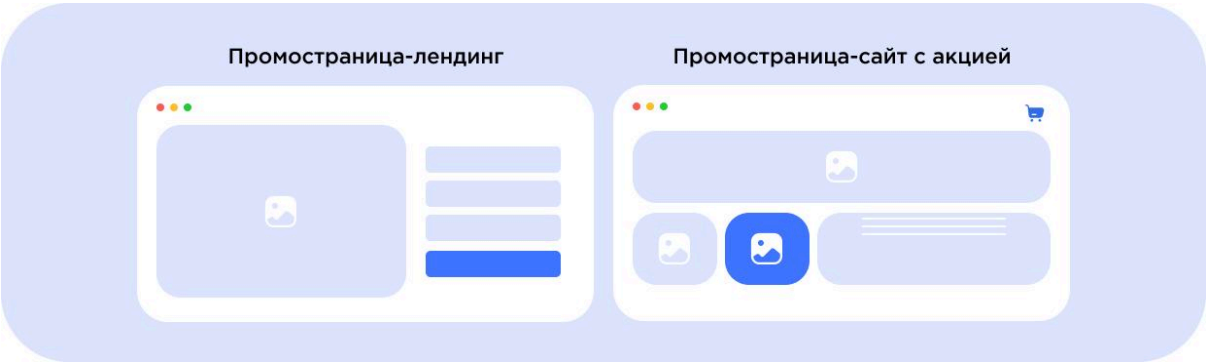
Допустим, целевой показатель — количество заполненных форм, и при этом вы выбираете 100% трафика. Если по одному варианту целевой показатель сильно ниже, то во время теста вы упустите возможные заявки от части аудитории, которые могли бы получить, если бы тест не проводился. А если будете использовать 70%, то упустите заявки только для 35% трафика. Но учтите: чем ниже доля тестируемого трафика, чем выше длительность теста (см. шаг 6).

Мы тестировали рекламный трафик с долей 100% по определенному региону с оптимизацией на цель, конверсия по которой используется как целевой показатель исследования.

Шаг 5. Варианты тестирования

Определите, что из себя будет представлять контрольный вариант (A) и сравниваемый вариант (B). Например, в варианте A кнопка зеленого цвета, в варианте B — красного. В Яндекс Метрике при создании эксперимента можно создавать варианты с простым изменением элементов сразу в интерфейсе (например, изменить текст на кнопке или ее цвет), не привлекая разработчика и не создавая отдельные страницы.

В нашем случае тестировались две действующие страницы сайта на одном домене, изменения не требовались. В контрольном варианте (A) мы использовали промостраницу-лендинг с акционным предложением и парой простых форм для заявок. В сравниваемом варианте (B) — промостраницу-сайт с акцией. Акционное предложение было таким же, как и в первом варианте, но эта страница представляла собой часть сайта из раздела «Акции».



Основное отличие было в том, что для заказа со страницы в сравниваемом варианте нужно было совершать заказ из корзины с прохождением стандартных этапов e-commerce воронки: корзина, оформление, покупка. В контрольном варианте для заказа можно было отправить быструю форму заявки.

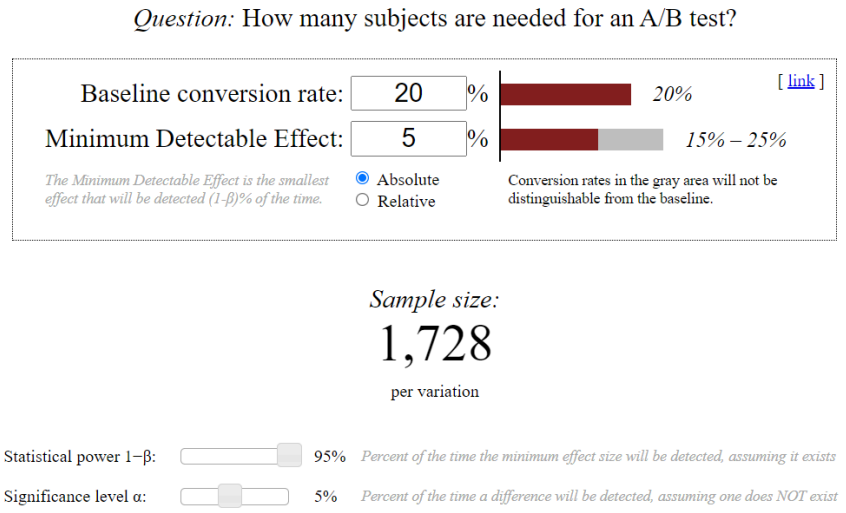
Если вы также планируете тестировать разные страницы сайта в эксперименте с типом «Ссылки для редиректа» Яндекс Метрики, учтите, что у тестируемых страниц должен быть одинаковый домен.

Шаг 6. Длительность тестирования

Длительность тестирования напрямую зависит от необходимого объема выборки, поскольку по итогу важно получить не просто результат, а статистически значимый результат. Результат не должен быть случайным из-за неоднородности выборки. Важно получить высокую вероятность того, что итоги теста достоверные.

Для расчета длительности тестирования можно использовать онлайн-калькулятор. Например, в калькуляторе, представленном ниже, для расчёта необходимо:

- понимать значение целевого показателя в среднем (Baseline conversion rate);
- определить минимальное изменение показателя, которое хотите зафиксировать (Minimum Detectable Effect). Например, в среднем значение показателя — 20%. Мы ожидаем, что в тестируемом варианте показатель вырастет до 25%. Тогда MDE в абсолютном выражении — 5%. Чем ниже показатель MDE, тем больше должна быть выборка, тем больше будет длительность теста.



Калькулятор выдает размер выборки для каждого варианта. Остается только умножить это значение на 2 (т. к. мы тестируем два варианта) и разделить на количество визитов исследуемого трафика, которое у нас регистрируется в среднем за день. Если в день 100 визитов, то длительность эксперимента составит 35 дней.

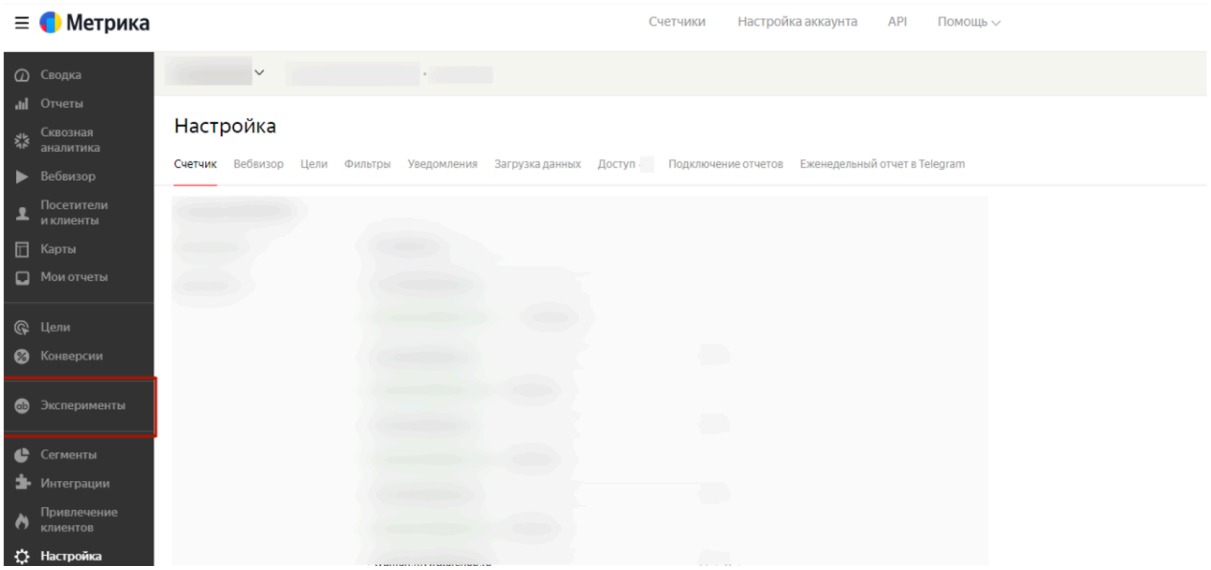
Также калькулятор можно использовать и в ходе теста, чтобы определить, когда его завершить. Мы как раз выбрали этот вариант, поскольку у нас было недостаточно данных для расчета длительности перед запуском теста. Подробнее опишем наш расчет в разделе «Тестирование». А на старте мы задали условную длительность — два месяца: если потребуется период меньше, эксперимент можно будет остановить.

Запуск теста

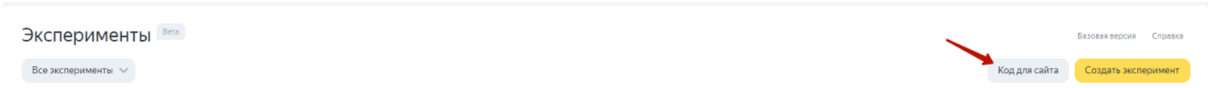
Шаг 1. Чтобы использовать инструмент от Яндекс Метрики, у вас должен быть создан и установлен на сайт счетчик Яндекс Метрики. Также нужен доступ к счетчику на уровне «Редактирование».

Если ранее вы не работали с этой системой аналитики, воспользуйтесь инструкцией для создания счетчика.

Находясь в интерфейсе, перейдите в раздел «Эксперименты»:

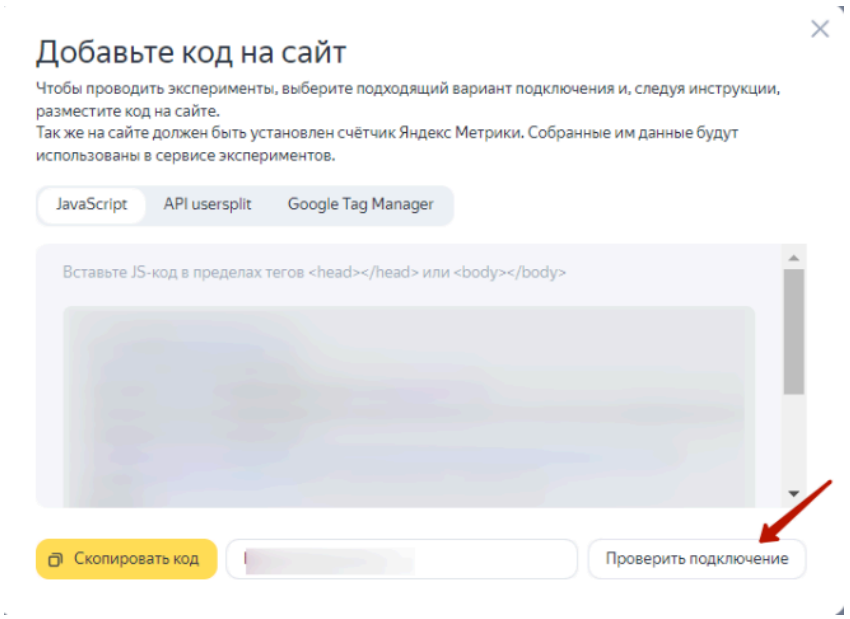


Шаг 2. Чтобы эксперимент работал, установите код эксперимента в код сайта. Для этого на панели эксперимента нажмите «Код для сайта» и установите его любым удобным способом, как будет описано в открывшемся окне:

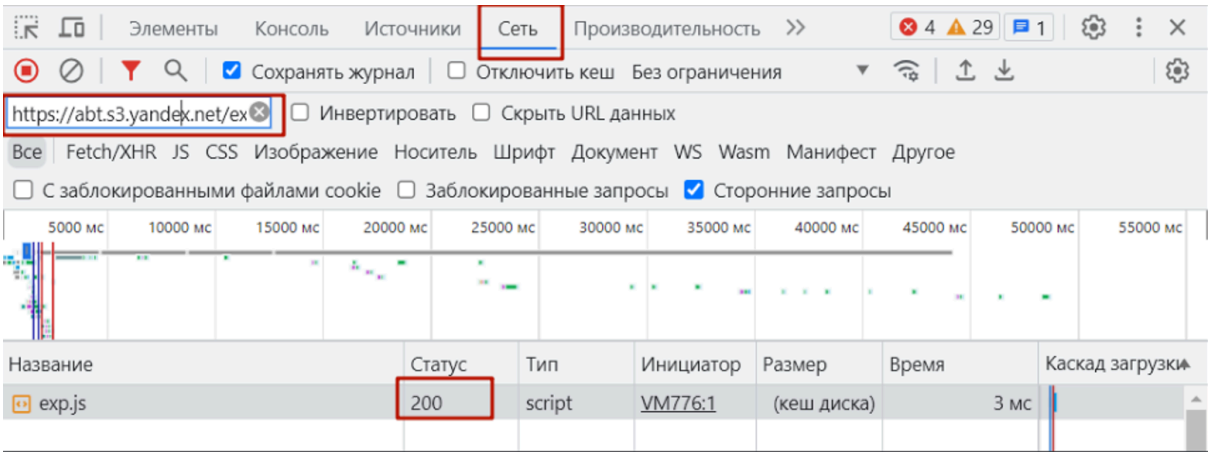


Шаг 3. Убедитесь, что код эксперимента установлен на сайте. Сделать это можно двумя способами:

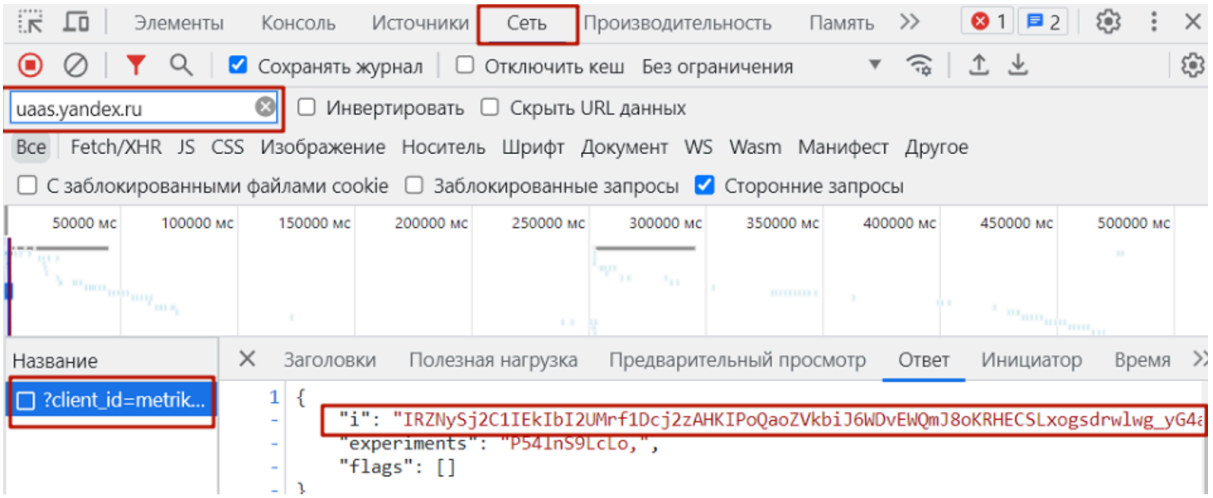
1. После установки кода эксперимента откройте снова окно с кодом эксперимента в интерфейсе Яндекс Метрики. Внизу будет поле для ввода ссылки на сайт/страницу для проверки и кнопка «Проверить подключение»:



2. После установки кода эксперимента перейдите на сайт/страницу, где этот код установлен. Откройте панель разработчика. Во вкладке **Network (Сеть)** отфильтруйте выдачу, указав https://abt.s3.yandex.net/expjs/latest/exp.js. Если код установлен, вы увидите успешно выполненный (статус **200**) запрос «exp»:



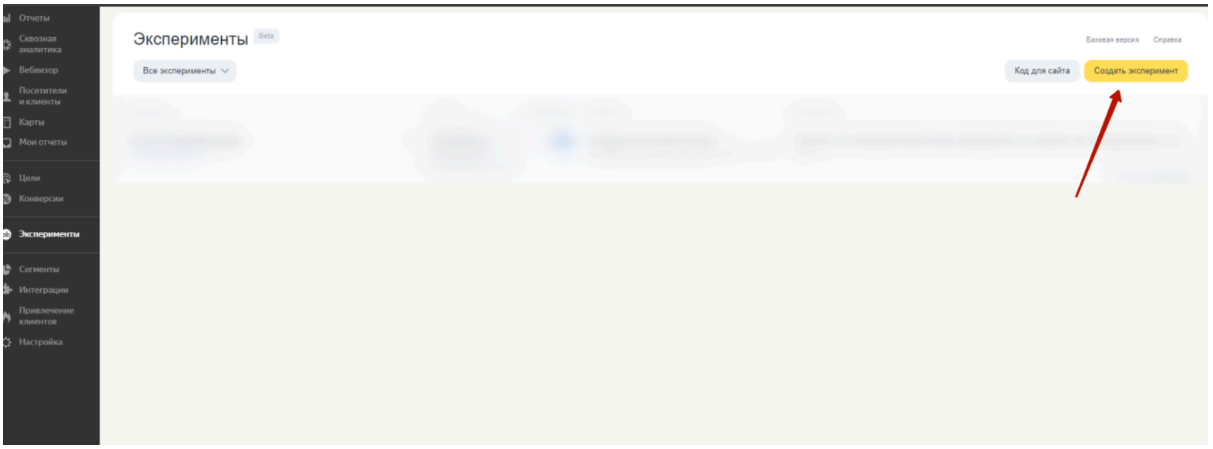
Также при фильтрации по домену **uaas.yandex.ru** во вкладке «Ответ» («Response») по запросу должен быть ответ сервера и значение у параметра «l»:



За информацию по проверке спасибо статье Якова Осипенкова.

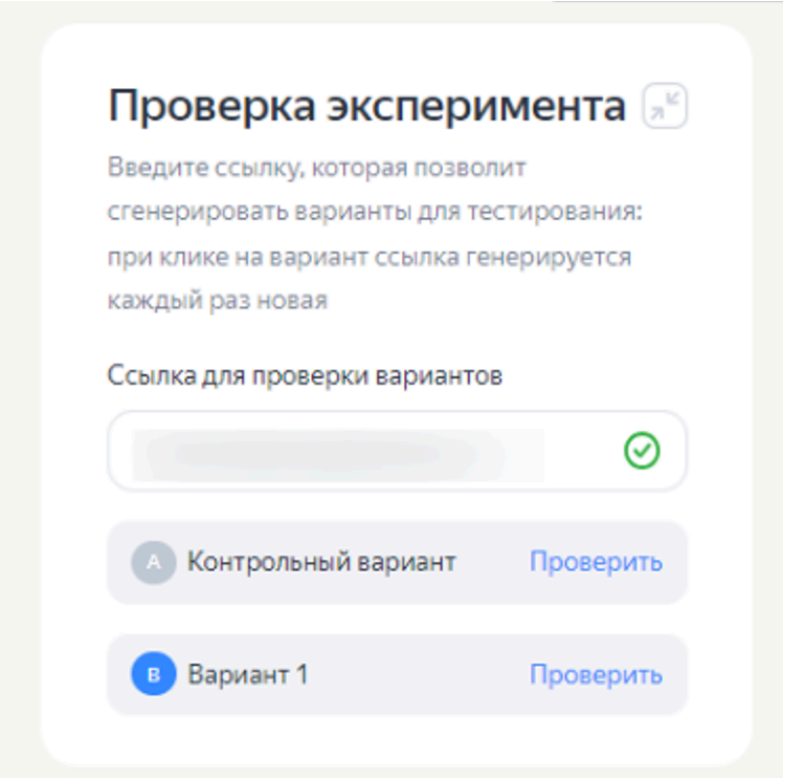
Шаг 4. Можно приступать к созданию эксперимента.

В разделе «Эксперименты» нажимаем «Создать эксперимент»:



И заполняем все необходимые поля, используя информацию, полученную в процессе этапов «Подготовки к тесту». Также можно воспользоваться документацией к созданию эксперимента, если возникнут вопросы.

Шаг 5. После заполнения полей эксперимента можете протестировать, как будут отображаться разные варианты:



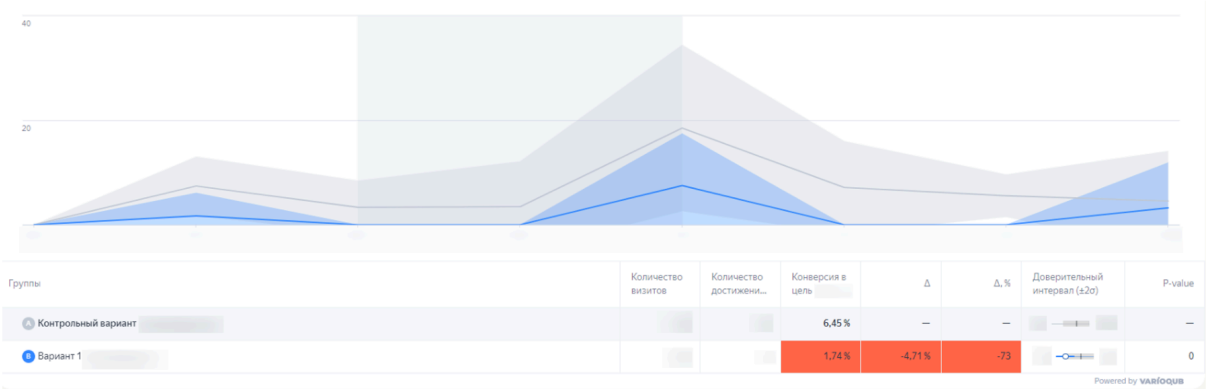
Если все в порядке, запускайте тестирование.

Тестирование

В ходе тестирования важно не поспешить и не остановить тест преждевременно, даже если вы видите, что исследуемые показатели имеют окрас. В Яндекс Метрике, если выводимый показатель статистически значимый, он окрашивается в красный или зеленый цвет.

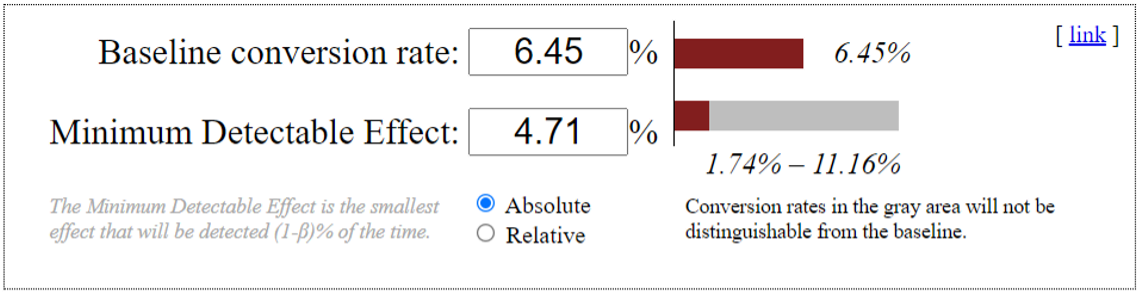
Если вам удалось определить длительность эксперимента на этапе подготовки к тестированию, то просто дождитесь автоматического завершения теста и переходите к интерпретации результатов.

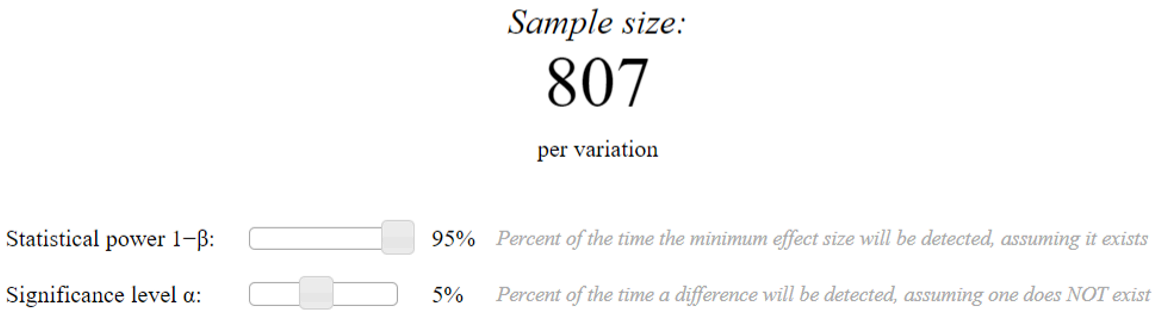
В нашем кейсе мы не смогли точно определить длительность теста перед запуском, поэтому следили за показателями в ходе эксперимента. Спустя неделю увидели окрас по основному целевому показателю:



Далее для каждого используемого в эксперименте целевого показателя (напомним, у нас один основной показатель и два дополнительных) решили рассчитать, сколько еще нужно ждать, используя калькулятор (представлен в шаге 6 в подготовке к тесту).

Пример расчета для основного показателя:





Несмотря на имеющийся окрас, мы не остановили тестирование: по расчету для одного из дополнительных показателей тестирование нужно было проводить еще, как минимум, в течение трех недель для достижения достаточного размера выборки.

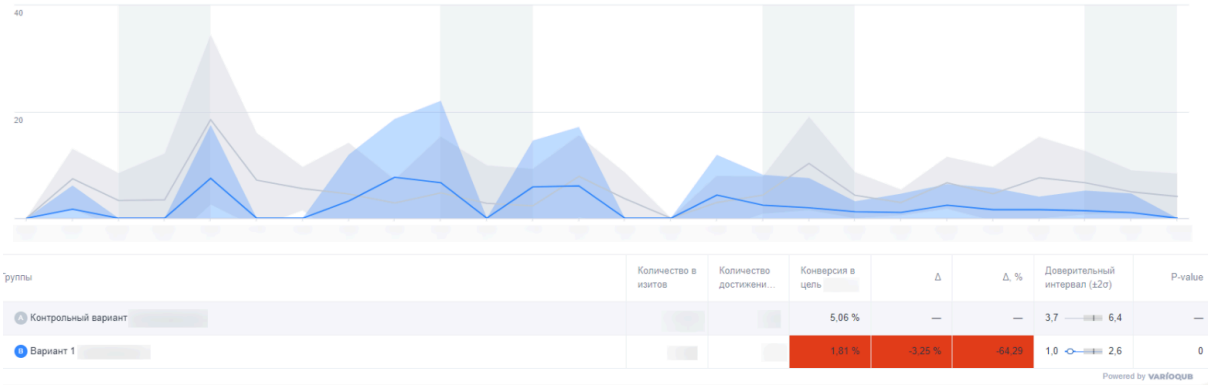
Завершение теста и интерпретация результатов

В эксперименте Яндекс Метрики тест останавливается автоматически по истечении срока, который был задан при создании.

Изначально мы задавали срок эксперимента — два месяца, но в ходе тестирования рассчитали, что понадобится месяц. В итоге по истечении четырех недель мы получили окрас всех целевых показателей. Еще раз с помощью калькулятора проверили, достаточная ли выборка, можно ли верить полученным показателям. Убедившись в корректности результатов, остановили тест вручную.

Как понимать полученные результаты?

Что мы получили по результатам эксперимента:



На графике линиями отображаются медианы целевых показателей, а закрашенные области выражают смоделированные диапазоны, в которые может попадать целевой показатель. Так можно заметить, что примерно до середины теста показатели колебались: показатель был выше то по контрольному варианту (A), то по сравниваемому (B). Во второй половине теста данные стабилизировались, и динамика показателя по контрольному варианту показывала значения стабильно выше, чем по сравниваемому. График наглядно показывает, что устойчивая разница была достигнута.

Чтобы сделать выводы по полученным значениям, обратите внимание на интенсивность окраса — столбец «P-value».

«P-value» — это вероятность того, что мы допустим ошибку, приняв результаты за статистически значимые. Хотя на самом деле они случайные, и целевой показатель в двух вариантах не изменился.

По умолчанию пороговое значение для «P-value» = 0.05. Т. е. если «P-value» < 0.05, допускается, что различия между вариантами статистически значимые. И, соответственно, чем меньше значение «P-value» от 0.05, тем ярче цвет окраса и тем больше вероятность, что полученные результаты не случайные.

Если «P-value» >= 0.05, то окраса не будет. И это означает, что тестируемые варианты значимо не отличаются друг от друга. Т. е. исследуемые изменения в варианте B не приведут к повышению или снижению целевого показателя.

Отсутствие окраса можно также наблюдать, если тест проводится мало времени и размера полученной выборки не хватает для определения статистически значимого различия. В таком случае не прерывайте тест и дождитесь окончания эксперимента, если правильно рассчитали длительность тестирования.

Результаты эксперимента показали: если вести исследуемый трафик на промостраницу из Варианта 1, то с вероятностью 99,99% ((1- P-value)*100%) конверсия по основному показателю снизится на 64%:

В целом по всем показателям мы получили следующую информацию:

Показатель	Контрольный вариант	Вариант 1	P-value	Вывод
Конверсия цели по отправкам форм	4,94 %	1,75 %	0,000014	Если будем вести трафик на страницу варианта 1, то с вероятностью 99% получим конверсию на 64% ниже, чем если будем вести трафик на страницу контрольного варианта
Конверсия цели по звонкам	1,35 %	0,36 %	0,004056	Если будем вести трафик на страницу варианта 1, то с вероятностью 99% получим конверсию на 73% ниже, чем если будем вести трафик на страницу контрольного варианта
Доля отказов	25,88 %	12 %	0	Если будем вести трафик на страницу варианта 1, то с вероятностью 100% получим показатель отказов на 53% ниже, чем если будем вести трафик на страницу контрольного варианта

На основе полученных данных мы выдвинули гипотезы, почему промостраница-лендинг показала результат по конверсии лучше, чем промостраница-сайт:

- **на лендинге простая форма заявки.** Для оформления заказа на сайте нужно авторизовываться и пройти несколько шагов оформления.
- **На лендинге пользователь видит одно предложение с простым описанием и фиксированной ценой.** На сайте сначала показывается одна цена, затем, когда открывается корзина, пользователь видит добавленную стоимость доставки. В результате может передумать оформлять заказ.
- **Те, у кого визит начинается с лендинга, могут также перейти на сайт** (если, например, пользователь уже является клиентом) и оформить заказ на нем. Плюс после отправки заявки с лендинга пользователь перенаправляется на главную страницу сайта, где также может ознакомиться с другими предложениями и, возможно, купить что-то еще.

Также мы предположили, над чем стоит поработать в варианте, где показатели хуже.

Таким образом, проведение эксперимента дало понимание, какое решение принять по ведению трафика на данном этапе и в каком направлении двигаться дальше, чтобы улучшать показатели.

Как итог

A/B-тестирования помогают принимать решения по улучшению продукта на основе данных о предпочтениях посетителей. Таким образом, вы можете быть уверены, что разрабатываете продукт, эффективно работающий на вашей целевой аудитории.

Для успешной реализации теста важно хорошо к нему подготовиться, учесть нюансы продукта и выбрать хороший сервис для проведения.

Эксперимент в Яндекс Метрике показал, что сервис Varioqub достаточно прост в настройке и понимании полученных результатов. Так что — тестируйте. Только так вы сможете подтвердить или опровергнуть ваши гипотезы.