Конспект по теме « A/B-тесты»

Что такое A/B-тест

A/B-тест — это инструмент, который позволяет делать надёжные выводы о влиянии изменения на продукт, за счёт использования статистических методов и параллельного сбора данных для сравниваемых групп.

Последовательность шагов при проведении A/B-теста:

1. Выбираем метрики и формулируем гипотезы.
2. Выбираем способ рандомизации и определяем параметры выборки.
3. Определяем необходимый размер выборки.
4. Запускаем эксперимент и собираем данные.
5. Проверяем валидность эксперимента.
6. Рассчитываем результаты и принимаем решение о раскатке фичи.

Разбиение на группы в рамках A/B-теста должно происходить параллельно и случайным образом. Это позволяет устранить влияние отличных от тестируемого изменения эффектов на метрику.

При планировании эксперимента важно учесть потенциальное наличие **сетевого эффекта**. Если группы влияют друг на друга, то вместо простой рандомизации по пользователям необходимо использовать более сложные способы.

Прежде чем проводить тест, нужно **рассчитать необходимый объём выборки**. Только после того, как он наберётся, можно анализировать результаты.

Важно не только проанализировать непосредственно результаты теста, но и дополнительно провести проверки валидности эксперимента. Такими проверками являются **A/A-тест** и проверка на **SRM**.

Количественные метрики

Все метрики делят на количественные, конверсионные и метрики-отношения.

Выручку можно считать в среднем на пользователя, платящего пользователя или заказ.

Связь между ARPU и ARPPU описывается формулой:

$$ARPU = ARPPU \cdot Paying\ share,$$

где Paying share — это доля пользователей, совершивших покупку.

Для расчёта размера выборки необходимо оценить дисперсию в тестовой и контрольной группах. Так как размер выборки рассчитывается до того, как сама выборка будет набрана, эти дисперсии оцениваются путём усреднения дисперсии за несколько прошедших периодов.

Для анализа изменений количественных метрик можно использовать Т-тест или бакетный тест.

Конверсии и метрики-отношения

Конверсией называют процент пользователей, совершивших целевое действие. Конверсию можно рассчитать по формуле:

$$CR_{X\ to\ Y} = \frac{K}{N} \cdot 100\%,$$

где K — количество пользователей, которые совершили целевое действие Y (дошли до шага Y),

N — количество пользователей, которые совершили действие X (дошли до шага X).

Дисперсия конверсионных метрик рассчитывается по формуле:

$$Var_{Bernoulli} = \bar{p} \cdot (1 - \bar{p}),$$

где $\bar{p}$ – рассчитанная конверсия.

Чтобы рассчитать необходимый размер выборки для конверсионных метрик, на этапе планирования эксперимента необходимо оценить дисперсии тестовой и контрольной групп. Для этого усредняют дисперсии конверсионной метрики за прошедшие периоды.

Метрики-отношения отличаются от количественных и конверсионных тем, что для них единица рандомизации, выбранная в рамках эксперимента, не

совпадает с единицей анализа. Для таких метрик нельзя в чистом виде применять T-тест и Z-тест для пропорций.

//конец кат-контейнера//

MDE и мощность

Вероятность ошибки первого рода, α — вероятность зафиксировать эффект там, где его на самом деле нет.

Вероятность ошибки второго рода, β — это вероятность не зафиксировать эффект там, где он на самом деле есть.

Мощность, $1 - \beta$ — это вероятность зафиксировать эффект там, где он на самом деле есть. Мощность также часто называют чувствительностью теста.

Мощность зависит от размера выборки, дисперсии в данных, уровня значимости и MDE:

$$z_{1-\beta} = \sqrt{\frac{n}{Var_{control} + Var_{test}}} \cdot MDE + z_{\alpha/2}.$$

Чтобы рассчитать MDE, необходимо оценить затраты на внедрение фичи и взять такой прирост метрики, который обеспечит их покрытие. Это значение является лишь желаемым значением MDE, то есть нашим предположением относительно того, каким может быть реальный эффект.

Так как реальное значение MDE может отличаться от желаемого, то и реальная мощность также может отличаться от той, которую мы использовали при расчёте размера выборки.

В случае если реальный эффект окажется меньше желаемого, мы всё равно сможем его зафиксировать, но с меньшей вероятностью, чем предполагали изначально.

Объём групп и продолжительность теста

Обычно длительность эксперимента выбирают кратной периоду, в рамках которого может наблюдаться сезонность в поведении метрики.

Алгоритм расчёта длительности:

Шаг 0. Выбрать уровень значимости и мощность.

Шаг 1. Составить список из потенциальных длительностей эксперимента: 1 неделя, 2 недели и так далее.

Шаг 2. Для каждой такой длительности рассчитать на данных за прошлые периоды

- усреднённую дисперсию метрики,
- усреднённое среднее значение метрики,
- усреднённое количество пользователей, посетивших сайт или приложение.

Шаг 3. Для каждой длительности рассчитать значение MDE, которое можно обнаружить с заданной мощностью, при выбранном уровне значимости и оценённой на данных за прошедшие периоды дисперсии. Это можно сделать по формуле:

$$MDE = -(z_{\alpha/2} + z_{\beta}) \cdot \sqrt{\frac{4 \cdot Var_{hist}}{n_{hist}}},$$

Var_{hist} — дисперсия, оценённая на данных за прошлые периоды,

n_{hist} — количество пользователей, которые в среднем посещают сайт за период, равный выбранной длительности.

Шаг 4. Выбрать ту длительность, для которой рассчитанный MDE наиболее близок, но не превышает минимальный желаемый эффект. В таком случае мы будем уверены в том, что мощность теста для желаемого MDE будет не меньше, чем та, которую мы использовали на предыдущем шаге.

Чтобы преодолеть эффект накопления метрик, нужно смотреть не на абсолютное значение MDE, а на относительное. Его можно рассчитать, разделив MDE на усреднённое значение метрики.

Проверка валидности эксперимента

Прежде чем анализировать результаты A/B-теста, необходимо убедиться, что кроме тестируемой фичи нет факторов, которые могли бы повлиять на целевую метрику.

В этом помогают:

- A/A-тест на предпериоде, представляющий собой применение статистического критерия для сравнения значений метрики в тестовой и

контрольной группах на периоде, предшествующем периоду эксперимента.

- Проверка на SRM, представляющая собой проверку того, что наблюдаемое соотношение количества пользователей в тестовой и контрольной группах не отличается от ожидаемого.

Расчёт и интерпретация результатов

В ситуации, когда **A/B-тест оказался серым,** мы говорим о том, что не нашли доказательства того, что фича оказала отличный от 0 эффект. Однако это не означает, что эффект в точности равен 0 или точно меньше, чем желаемый эффект, который мы использовали в качестве MDE.

В ситуации, когда **A/B-тест оказался зелёным**, нам необходимо дополнительно проверить, что наблюдаемая разность средних позволяет сделать вывод о том, что истинный эффект не меньше желаемого. Для этого можно рассчитать t- или z-статистику для желаемого эффекта по формулам:

$$z = \frac{(\bar{p}_{test} - \bar{p}_{control}) - (p_{test} - p_{control})}{\sqrt{\frac{Var_{test}}{n_{test}} + \frac{Var_{control}}{n_{control}}}},$$

$$t = \frac{(\bar{x}_{test} - \bar{x}_{control}) - (\mu_{test} - \mu_{control})}{\sqrt{\frac{s^2_{test}}{n_{test}} + \frac{s^2_{control}}{n_{control}}}}.$$

A/B-тест не следует останавливать, как только была зафиксирована статистическая значимость, так как это ведёт к росту вероятности ошибки первого рода. Подводить итоги эксперимента следует только после того, как его длительность станет равна предрассчитанной.

Таблица принятия решений по результатам A/B-теста

Зафиксировано ли статистически значимое отклонение от 0?	Зафиксировано ли статистически значимое отклонение от MDE?	Что делать
Да, отрицательное	Да, отрицательное	Оставить всё как есть. Фичу не раскатывать.
Нет	Да, отрицательное	Оставить всё как есть. Фичу не раскатывать.

Зафиксировано ли статистически значимое отклонение от 0?	Зафиксировано ли статистически значимое отклонение от MDE?	Что делать
Нет	Нет	Если есть время на повторный A/B-тест, то провести его с большей выборкой. Если времени на повторный тест нет, отказаться от внедрения фичи.
Да, положительное	Нет	Желательно перезапустить A/B-тест с большей выборкой. Возможно, рискнуть и раскатить фичу.
Да, положительное	Да, положительное	Раскатить фичу.