MDE

Подходы к определению длительности теста

Чаще всего мы сталкиваемся с упоминанием MDE в разрезе определения размера выборки/продолжительности теста.
Тут хочется коротко отметить, что есть несколько подходов к определению продолжительности теста, и первый из них — **Fixed time Horizon.**

Он предполагает заранее определить размер выборки и продолжительность теста запустить тест на этот срок принять решение об отклонении нулевой гипотезы. Этот подход мы как раз и используем на текущий момент. Он более практичен и понятен членам команды.

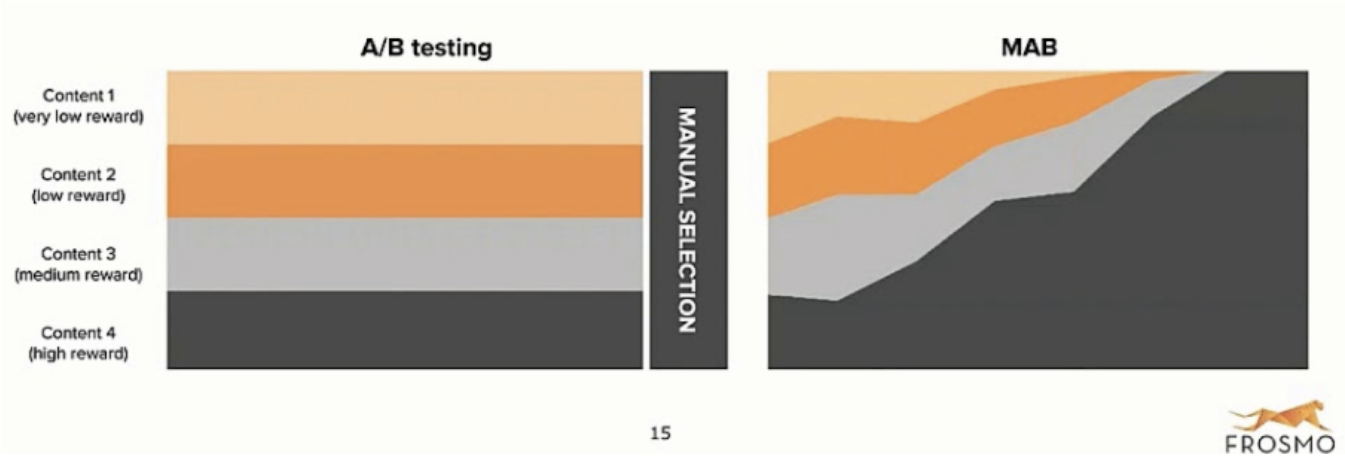
blocked URL

Вспомним заодно о нулевой и альтернативной гипотезах:

> Нулевая гипотеза: наши усилия не имели эффекта, средние значения в группах не отличаются.
>
> Альтернативная гипотеза: в результате изменений среднее значение сместилось, то есть эффект есть, но мы не знаем какой, но можем оценить его по выборке (рассматриваем двусторонний критерий).
>
> Задача состоит в том, чтобы опровергнуть нулевую гипотезу в пользу альтернативной, то есть доказать, что эффект есть и наблюдаемая разница средних между группами обусловлена не случайностью.
>
> Для этого мы, как правило, применяем критерий Стьюдента и считаем p-value (вероятность того, что мы будем наблюдать такое и более экстремальное различие между группы при верной нулевой гипотезе). А далее сравниваем p-value с выбранным уровнем значимости.

Второй подход к определению продолжительности теста — **Sequential testing.**

В Sequential testing используется семейство методов, при которых принято смотреть на границы остановки эксперимента в реальном времени и в связке с наблюдаемыми результатами. К данному подходу относятся и известные Байесовские бандиты, используемые в Google Optimize.



Чем отличаются lift и MDE? Зачем нам знать MDE?

Lift— это отличие метрики теста от метрики контроля. Считаем их разницу и делим на контроль. Умножаем на 100 и получаем %-ное изменение, относительный прирост.

Такой lift мы наблюдаем в дашборде, когда смотрим на метрики авторасчета (4 столбец). Он не связан с ошибками 1 и 2 рода, дисперсией и т.п.

AOV для товаров, добавленных из каталога	android	993,0219	1 310,6651	0,00	-24,24%	-317,64
Среднее кол-во товаров, добавленных из каталога	android	6,3676	8,8756	0,00	-28,26%	-2,51
Конверсия в добавление 1 товара из верхней категории (ratio)	android	0,0094	0,0085	0,00	10,60%	0,00
Время до добавление 1 товара из выдачи каталога	android	26,6439	24,9203	0,00	6,92%	1,72

MDE в свою очередь - это **минимальный эффект**, который можно обнаружить (при выбранных значениях уровня значимости и мощности). По сути дела, это тот минимальный эффект, который мы можем позволить себе задетектировать статзначимо на конкретном размере выборки, имеющемся в нашем "распоряжении".

Что дает нам знание MDE на конкретном объеме трафика? Пример.

Представим, мы предполагаем, что наша фича прокрасит метрику Среднего чека на 0.5%. И, мы планируем запустить тест на 50% трафика, на классические две недели. Но если мы заглянем в MDE данной метрики, то увидим, что за две недели на 50% MDE = 1.49%.

То есть, эффект меньше 1.5% мы просто не сможем “рассмотреть” за эти две недели, а значит, нам либо нужно больше трафика для теста, либо другая метрика, либо нужны новые инструменты для его оценки, и т.д.

AOV

Android

Нажмите здесь для раскрытия...

platform	length	traffic_proportion	metric	alpha	mde_abs	mde_percent
android	2w	0.125	AOV	0.01	75.0404	2.96%
android	2w	0.125	AOV	0.05	61.52	2.43%
android	2w	0.25	AOV	0.01	57.1863	2.26%
android	2w	0.25	AOV	0.05	46.8804	1.85%
android	2w	0.5	AOV	0.01	38.1458	1.49%
android	2w	0.5	AOV	0.05	31.2668	1.22%

Про MDE подробнее и с большим погружением в статистику

Есть ли эффект?

Порой при подведении результатов АБ теста мы не видим статистически значимых отличий между группами. Значит, ли это что мы никак не повлияли на метрику?

На самом деле здесь может быть два варианта:

- либо различий между группами действительно нет,
- либо же эффект есть, но он мал, и наших данных не хватает, чтобы его обнаружить.

И тут возникает вопрос «а сколько данных нужно?».

Ведь было бы прекрасно заранее посчитать сколько нам необходимо данных для обнаружения сколь угодно малого эффекта и в каждом тесте именно столько их и набирать. Но, к сожалению, для этого потребовалось бы всего лишь бесконечное число пользователей и бесконечное число дней. «Многовато…» — наверное, подумали вы.

Так мы приходим к тому, что необходимо заранее определять:

1. какой эффект, мы хотим быть способными обнаружить в тесте
2. расчитывать размер выборки для этого эффекта
3. проводить тест столько времени, сколько требуется для сбора выборки такого размера.

Тогда, если метрика будет серой в тесте, то как минимум мы сможем в какой-то степени быть уверенными в том, что того эффекта, который мы хотели обнаружить, действительно нет.

И второй момент — это необходимость найти баланс между желанием обнаружить достаточно маленький эффект и рисками длительного проведения теста.

- Почему?

  Чем больше размер нашего эксперимента, тем репрезентативнее наши выборки представляют генеральную совокупность, тем меньше разброс в данных (дисперсия), и быстрее можно получить статзначимый эффект. Но при этом при большом (длительном / с большим кол-вом юзеров) эксперименте больше вероятность пересечься с другими тестами и получить влияние на метрики от другого теста, к тому же любой тест имеет риски экономический потерь.

Вспомним некоторые понятия из статистики

Уровень значимости (α): вероятность того, что статистически значимые различия появятся при выполнении нулевой гипотезы. Обычно берем 0.01 или 0.05

Мощность критерия (1 - β): вероятность верно принять альтернативную гипотезу. Обычно берем равным 0.8.

Ошибка первого рода (α): вероятность обнаружить стат значимый эффект, когда его на самом деле нет (то есть когда верна нулевая гипотеза), что и является уровнем значимости.

Ошибка второго рода (β): вероятность не обнаружить стат значимый эффект, когда он на самом деле есть (то есть когда верна альтернативная гипотеза).

При любой величине эффекта уменьшение ошибки первого рода ведет к увеличению ошибки второго рода. Одновременно их уменьшить нельзя.

Как рассчитывается MDE и размер выборки? Формулы и примеры.

MDE (minimum detectable effect) — минимальный эффект, который можно обнаружить при выбранных значениях уровня значимости и мощности (при условии, что мы зафиксировали размер эксперимента: трафик и продолжительность теста)

$$\mu^2 > \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\sigma_x^2 + \sigma_y^2\right)}{n}$$

μ - MDE

α - допустимая ошибка первого рода

β - допустимая ошибка второго рода

σ^2 - дисперсии выборок

n - размеры выборок

Проанализируем формулу:

Если мы уменьшаем ошибки первого/второго рода, то эффект, который можем поймать при заданных параметрах, увеличивается.

Если дисперсия наших данных увеличивается, то эффект, который можем поймать при заданных параметрах, увеличивается. То есть волатильность данных нам вредит.

Если размер выборок увеличивается, то эффект, который можем поймать при заданных параметрах, уменьшается. То есть чем больше данных, тем меньший эффект можем поймать.

Пример

Пусть мы зафиксировали:

- ошибку первого рода α = 0.05
- ошибку второго рода β = 0.2
- дисперсия выборок для простоты σ^2 = 1
- размер выборки n = 1000

Тогда минимальный эффект, который мы сможем поймать, это

$$MDE > \frac{(1,96 + 0,84)(1+1)}{\sqrt{1000}} \approx 0.17$$

Размер выборки

Из формулы MDE можно получить формулу для размера выборки n:

$$n > 2 \cdot \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\mu^2}$$

Эта формула показывает какой размер выборки необходим, чтобы обнаружить ожидаемый эффект μ при заданных ошибках первого и второго рода.

Проанализируем формулу:

Если мы уменьшаем ошибки первого/второго рода, то размер выборки, необходимый для обнаружения заданного эффекта, увеличивается.

Если мы уменьшаем дисперсию данных, то размер выборки, необходимый для обнаружения заданного эффекта, уменьшается.

Если мы уменьшаем ожидаемый эффект, то размер выборки, необходимый для обнаружения этого эффекта, увеличивается.

Пример

Пусть мы зафиксировали:

- ошибку первого рода α = 0.05
- ошибку второго рода β = 0.2
- дисперсия выборок для простоты σ^2= 1

Эта формула показывает какой размер выборки необходим, чтобы обнаружить ожидаемый эффект μ при заданных ошибках первого и второго рода. Тогда для эффекта μ = 0.06, нам нужно

$$n > 2 \cdot \frac{(1,96 + 0,84)^2 1^2}{0,06^2} \approx 4360$$

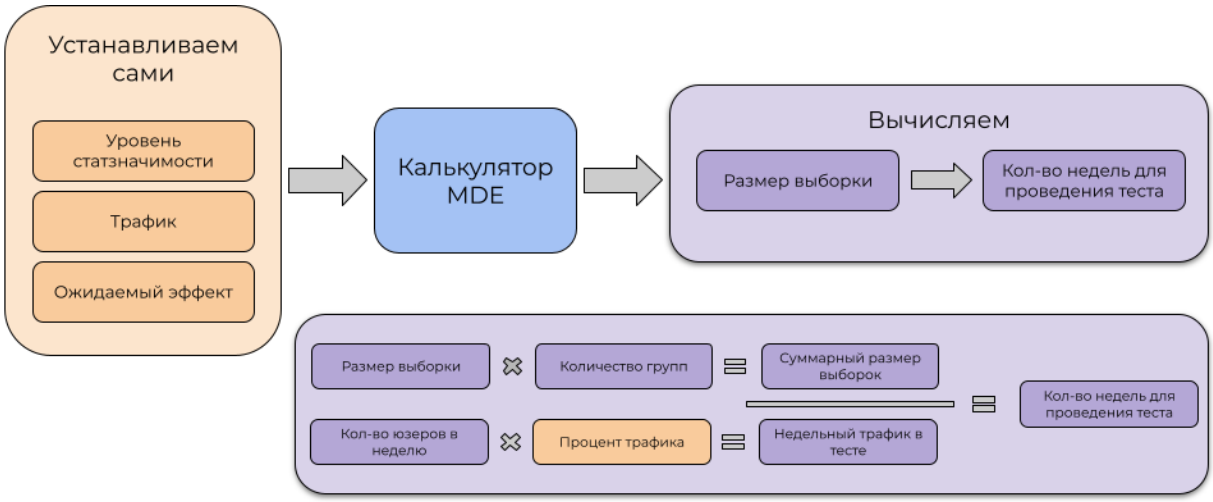
Но если мы захотим обнаружить эффект в два раза меньший, то нам потребуется

$$n > 2 \cdot \frac{(1,96 + 0,84)^2 1^2}{0,03^2} \approx 17442$$

Снижение дисперсии

Так как ошибки первого и второго рода и ожидаемый эффект обычно устанавливаются бизнесом, то есть на них мы как аналитики повлиять не можем, то единственным способом уменьшить размер выборки является снижение дисперсии. Можно применять современные методы: cuped, стратификация, в случае зависимых данных — разные виды линеаризации. А также повышать качество собираемых данных или фильтровать выбросы.

Итоги



Где искать MDE сейчас?

1. страница с MDE базовых метрик от аб платформы (конверсия, ср чек)
2. дашборд a2c по MDE
3. ваш аналитик (который может рассчитать mde, или сделать вам дашборд по mde, у Аделины для этого имеется ноутбук), либо можно использовать ноутбук АБ платформы
4. для аналитика – таблица sandbox.mde_metrics с mde базовых метрик. Таблица регулярно обновляется