MDE, sample size

Александр Сахнов
linkedin.com/in/amsakhnov

Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

Есть эффект?

Бывает, что после оценки пилота мы не видим статистически значимые отличия.
А могли ли заметить ожидаемый эффект на этих данных?

Эффект есть

Эффекта нет

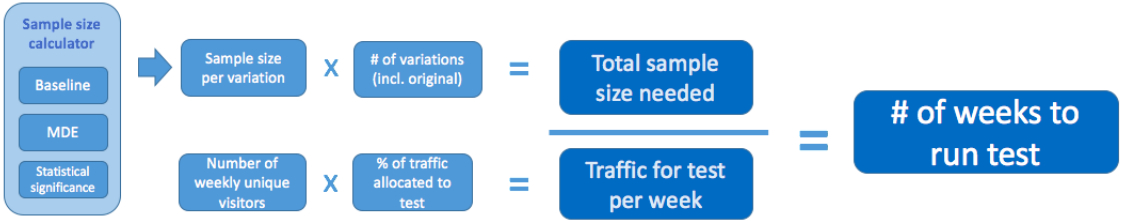
Что нужно посчитать до начала эксперимента

Устанавливаемые нами параметры:

- Минимальная величина эффекта
- Уровень статзначимости
- Долю пользователей в эксперименте

Вычисляемые параметры:

- Необходимый размер выборки
- Недельный трафик
- Продолжительность эксперимента



Разберемся как считать все эти параметры!

Выбор размера: статистика vs риски

Почему мы хотим большой размер эксперимента?

- Чем больше группы, тем репрезентативней они представляют генеральную совокупность.
- Мы получаем меньший разброс. Выше статзначимость.
- При том же уровне значимости можно быстрее получить результат.

С точки зрения статистики лучше всего поделить *всех* пользователей 50/50 между экспериментальной и контрольной группами.

Почему мы хотим маленький размер эксперимента?

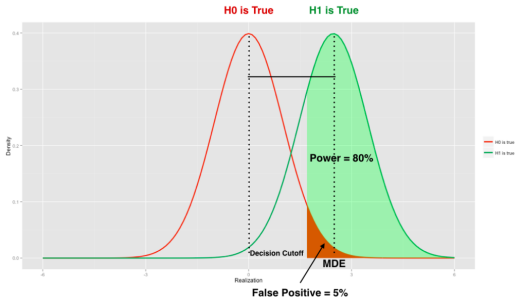
- Одновременно может идти несколько экспериментов. Мы не хотим, чтоб эксперименты влияли друг на друга.
- Проведение эксперимента может быть затратным.
- Любой эксперимент несет риски экономических потерь.

Необходимо искать правильный баланс.

Тестирование гипотез

Нулевая гипотеза

Мы предполагаем, что наши усилия не имели эффекта.
Нужно посчитать вероятность того, что статистически значимые различия появятся при выполнении нулевой гипотезы. Эта вероятность называется **уровнем значимости**.



Альтернативная гипотеза

В результате изменений среднее значение сместилось. Величину смещения мы не знаем, но можем оценить по выборке.
Вероятность верно принять альтернативную гипотезу называется **мощностью статистического критерия**.

$$\mathbb{P}\left(\frac{\bar{Y}-\bar{X}}{\sqrt{\sigma^2_{\bar{X}}+\sigma^2_{\bar{Y}}}} > C \,\middle|\, H_0\right) = FPR$$

$$\mathbb{P}\left(\frac{\bar{Y}-\bar{X}}{\sqrt{\sigma^2_{\bar{X}}+\sigma^2_{\bar{Y}}}} > C \,\middle|\, H_1\right) = Power$$

Односторонний и двусторонний тест

Какую гипотезу тестируем?

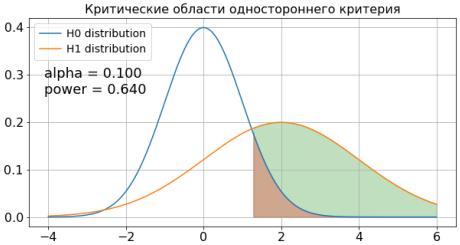
Мы можем выбирать разные постановки тестирования гипотезы.

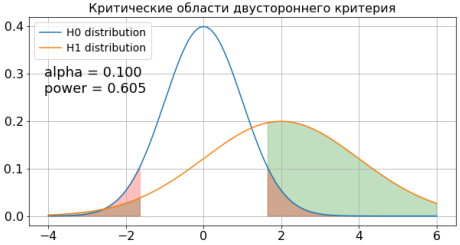
- Эффект больше пороговой величины.
- Эффект отличается от нуля больше, чем на пороговую величину.

Критические области

По выбранному уровню значимости мы можем определить критическую область и мощность критерия.

Обратите внимание, что мощность критерия у одностороннего теста выше.





Ошибки при принятии решений

Желание экспериментатора:

- Мы хотим уменьшить ошибку первого рода
- Мы хотим увеличить мощность критерия

Мы должны выбрать эти параметры. Это определяющие характеристики нашего эксперимента.

Ошибка первого рода

Большая величина ошибки первого рода означает, что мы часто будем находить эффект при его отсутствии. Потратим много денег на внедрения, которые ничего не дадут.
Типичное значение ошибки первого рода 5%.

Мощность критерия

Низкая мощность критерия означает, что мы будем часто пропускать позитивные изменения. У нас в руках идея, которая может заработать миллионы, а мы её отвергаем!
Можно выбрать мощность критерия в 80%. Тогда мы обнаружим четыре классные идеи из пяти.

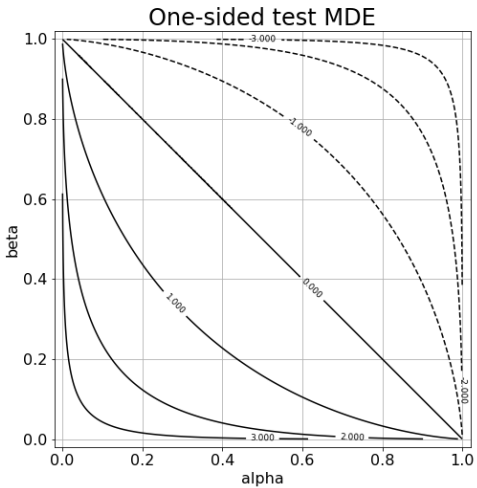
MDE - минимальный детектируемый эффект

Разнонаправленность ошибок

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

Минимальный детектируемый эффект

MDE — минимальный эффект, который можно обнаружить при выбранных значениях уровня значимости и мощности.



Математическое обоснование MDE

Статистическая гипотеза: $X_1, \ldots, X_n \sim N(a, \sigma_0^2)$
Нулевая гипотеза и альтернативная гипотезы: $H_0 : a = a_0, \quad H_1 : a = a_1, \quad a_0 < a_1$

Критерий отношения правдоподобия

$$T(X) = \frac{l_1}{l_0} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{(X_i - a_1)^2}{2\sigma_0}\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{(X_i - a_0)^2}{2\sigma_0}\right\}} = \exp\left\{\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n 2X_i(a_1 - a_0) + \sum_{i=1}^n (a_0^2 - a_1^2)\right)\right\}$$

Критерий имеет вид $T(X) \geq c^*(\alpha)$ или $\sum_{i=1}^n X_i \geq c(\alpha)$.

Ошибка первого рода $\mathbb{P}_{H_0}\left(\sum_{i=1}^n X_i \geq c\right) = \alpha$

$$\mathbb{P}\left(\frac{\sum_{i=1}^n X_i - na_0}{\sqrt{n}\sigma_0} \geq \frac{c - na_0}{\sqrt{n}\sigma_0}\right) = \alpha$$

Вывод MDE

Воспользуемся ЦПТ:

$$1 - \Phi\left(\frac{c - na_0}{\sqrt{n}\sigma_0}\right) = \alpha$$

Получаем выражение для границы критической области

$$c = \Phi^{-1}(1-\alpha)\sqrt{n}\sigma_0 + na_0$$

Заметим, что c не зависит от a_1 и верно $\forall a_1 : a_1 > a_0$.

Ошибка второго рода $\mathbb{P}_{H_1}\left(\sum_{i=1}^n X_i \geq c\right) \geq 1 - \beta$.

$$\mathbb{P}_{H_1}\left(\frac{\sum_{i=1}^n X_i - na_1}{\sqrt{n}\sigma_0} \geq \frac{c - na_1}{\sqrt{n}\sigma_0}\right) \geq 1 - \beta$$

Подставим выражение для c:

$$\mathbb{P}_{H_1}\left(\frac{\sum_{i=1}^n X_i - na_1}{\sqrt{n}\sigma_0} \geq \frac{\Phi^{-1}(1-\alpha)\sqrt{n}\sigma_0 + na_0 - na_1}{\sqrt{n}\sigma_0}\right) \geq 1 - \beta$$

Вывод MDE

Воспользуемся ЦПТ:

$$1 - \Phi\left(\Phi^{-1}(1-\alpha) + \frac{\sqrt{n}(a_0 - a_1)}{\sigma_0}\right) \geq 1 - \beta$$

$$\varepsilon = a_1 - a_0 \geq \frac{(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta))\sigma_0}{\sqrt{n}}$$

Покажем, что $\Phi^{-1}(\beta) = -\Phi^{-1}(1-\beta)$.

Пусть $\Phi(x) = \beta$.
Известно, что $\Phi(x) + \Phi(-x) = 1$, тогда

$$\Phi(-x) = 1 - \Phi(x) = 1 - \beta$$

$$-x = \Phi^{-1}(1-\beta)$$

С другой стороны $x = \Phi^{-1}(\beta)$.
Подставив, получим доказываемое равенство.

MDE

MDE (minimal detectable effect) - минимальный эффект который могли поймать.

ε - размер эффекта
α - допустимая ошибка первого рода
β - допустимая ошибка второго рода
σ_X^2, σ_Y^2 - дисперсии выборок
n - размеры выборок

$$\varepsilon^2 > \frac{\left[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right]^2 (\sigma_X^2 + \sigma_Y^2)}{n}$$

Sample Size

Оценим размер выборки, который необходим, чтобы обнаружить ожидаемый эффект при фиксированных ошибках первого и второго рода.

$$n > \frac{\left[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right]^2 (\sigma_X^2 + \sigma_Y^2)}{\varepsilon^2}$$

Variance reduction

$$n > \frac{\left[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right]^2 (\sigma_X^2 + \sigma_Y^2)}{\varepsilon^2}$$

Нужно много данных, что делать?

Снижать дисперсию

- повышать качество собираемых данных
- фильтровать выбросы
- CUPED
- ...

Заключение

1. Подготовка к эксперименту
2. Тестирование гипотез
3. MDE
4. Sample Size

Материалы

Материалы для самостоятельного изучения

1. Use minimum detectable effect to prioritize experiments
2. Power, minimal detectable effect, and bucket size estimation in A/B tests
3. A/B testing good practice – calculating the required sample size
4. The Minimum Detectable Difference (MDD) Concept for Establishing Trust in Nonsignificant Results: A Critical Review