A/B тесты

Сегодня

Эльбрус буткемп

- Что же такое A/B тесты
- Важные моменты
- достоинства и недостатки A/B тестов

Мощность в A/B тестировании

Эльбрус буткемп

> Давайте вспомним что такое мощность критерия?

Что влияет на мощность?

- Разные статистические критерии имеют разную мощность, нужно пробовать
- Величина эффекта — эффект, который мы хотим зарегестрировать. Чем больше эффект, тем больше мощность
- Количество наблюдений — чем больше наблюдений, тем легче зарегистрировать изменения в силу ЦПТ.
- Размер дисперсии - Можем ли мы как то влиять на него?

Мощность в A/B тестировании

Эльбрус буткемп

Вспоминая формулу доверительного интервала, можно увидеть влияние

$$(\bar{x} - \frac{s}{\sqrt{n}} t_\alpha(n-1)); (\bar{x} + \frac{s}{\sqrt{n}} t_\alpha(n-1))$$

Как можно уменьшить доверительный интервал?

CUPED

Эльбрус буткемп

CUPED (Controlled-experiment Using Pre-Experiment Data) — очень популярный в последнее время метод уменьшения вариации. Основная идея метода такова: давайте вычтем что-то из теста и из контроля так, чтобы математическое ожидание разницы новых величин осталось таким же, как было, а дисперсия уменьшилась.

$$T' = T - \theta \cdot A$$
$$C' = C - \theta \cdot B$$

A и B — некоторые случайные величины (ковариаты). Тогда утверждается, что если θ будет такой, как указано в формулах далее, то дисперсия будет минимально возможной для таких статистик.

CUPED

Эльбрус буткемп

$$\theta = \frac{cov(T, A) + cov(C, B)}{DA + DB}$$

Формула дисперсии примет вид:

$$D(T' - C') = (1 - \alpha^2) \cdot D(T - C),$$

где $\alpha = corr(T - C, A - B)$

A и B чаще всего берут значения той же метрики на предэкспериментальном периоде. Например, вы смотрите метрику выручки, тогда в роли ковариаты A и B можно взять выручку от пользователя за месяц до начала эксперимента. Чем хорош такой способ.

Бутстреп

Эльбрус буткемп

Бутстреп (bootstrap) - это метод для оценки стандартных отклонений и нахождения доверительных интервалов статистических функционалов.

В качестве оценки функции распределения будем использовать эмпирическую функцию распределения (ЭФР). ЭФР является несмещённой оценкой и сходится к истинной ФР при увеличении размера выборки.

Бутстреп

Эльбрус буткемп

- Генерируем пару подвыборок того же размера из исходных данных контрольной и экспериментальной групп;
- Считаем метрики (реализация оценки метрики) для каждой из групп;
- Вычисляем разность метрик, сохраняем полученное значение;
- Повторяем шаги 1-3 от 1000 до 10000 раз;
- Строим доверительный интервал с уровнем значимости \alpha;
- Если 0 не принадлежит ДИ, то отличия статистически значимы на уровне значимости \alpha, иначе нет.

Стратификация

Эльбрус буткемп

Логика стратифицирования очень простая, если у нас есть какой-то признак, который делит наших пользователей таким образом, что исследуемая метрика у этих страт различна, ты мы хотим сохранить пропорцию и при делении на контроль и тест