Causal Inference: прозрение и практика

Лекция 2. Рандомизированные контролируемые испытания

Юрашку Иван Вячеславович

July 1, 2024

Введение

Рандомизированные контролируемые испытания (РКИ) представляют собой наиболее объективную, прозрачную и эффективную методологию для проведения экспериментов. Они пользуются огромной популярностью и применяются в самых разных сферах, включая науку, медицину, маркетинг и технологии. С их помощью учёные и специалисты могут проверять эффективность новых методов лечения, лекарственных препаратов, продуктов или услуг, сравнивая результаты между двумя или более группами. РКИ встречаются гораздо чаще, чем может показаться на первый взгляд. Это невероятно популярный метод исследования причинно-следственных связей. Хотя они довольно просты в реализации, их точность значительно превосходит все другие методы аппроксимации ATE.

Существует несколько видов рандомизированных контролируемых исследований. Самые используемые из них:

- **Простая рандомизация** — каждому участнику испытания случайным образом назначается либо исследуемое вмешательство, либо контрольное.
- **Стратифицированная рандомизация** — участники сначала разделяются на страты на основе определённых характеристик, а затем внутри каждой страты происходит случайное распределение по группам исследования.
- **Кластерная рандомизация** — в этом случае рандомизация происходит по группам или «кластерам», а не по отдельным участникам.
- **Кроссоверное испытание** — сначала участники получают одно вмешательство, а после определённого периода времени – другое (и наоборот).
- **Факториальное испытание** — каждый участник случайным образом распределяется по группе, которая получает определённую комбинацию вмешательств, включая плацебо.

Анализ 616 РКИ, проиндексированных в PubMed в декабре 2006 года, показал, что 78% были исследованиями в формате простой рандомизации, 16% были кроссоверными, 2% были стратифицированными, 2% были кластерными и 2% были факториальными.

Смещение и предвзятость

В предыдущей статье мы столкнулись с ключевым препятствием для рассмотрения ATE как разницы средних значений между целевой и контрольной группами. Это препятствие называется bias, что в переводе с английского означает «смещение» или «предвзятость». Несмотря на их кажущуюся неродственность, вместе эти два термина хорошо описывают ситуацию.

Действительно, согласно выводу из предыдущей главы, если $BIAS(\mathcal{M})$ не равен нулю, то по крайней мере одно из значений $BIAS_0(\mathcal{M})$ или $BIAS_1(\mathcal{M})$ также не равно нулю. Предположим, к примеру, что $BIAS_0(\mathcal{M})$ сильно больше нуля.

$$BIAS(\mathcal{M}) = \frac{1}{|\mathcal{M}_{(1)}|} \sum_{i \in \mathcal{M}_{(1)}} Y^i_{(0)} - \frac{1}{|\mathcal{M}_{(0)}|} \sum_{i \in \mathcal{M}_{(0)}} Y^i_{(0)} \gg 0$$

Иначе говоря, это означает, что ещё до начала эксперимента эти две группы не были одинаковыми. Если бы наш эксперимент не проводился, то обе группы не подвергались бы воздействию, а их средние значения целевой переменной всё равно не были бы равны друг другу. Это и есть предвзятость. Она является атрибутом разбиения на группы и никак не связана с воздействием на объекты.

О чем мы мечтаем при проведении A/B теста

Чтобы преодолеть эту предвзятость, как один из способов, эксперты в области причинно-следственного вывода проводят рандомизированные контролируемые испытания.

В контексте решения поставленной задачи по оцениванию ATE при использовании двух групп, научимся применять РКИ первого типа. А именно РКИ, предполагающие наличие двух групп, контрольной и целевой, обозначаемых как $\mathcal{A}$ и $\mathcal{B}$, из-за чего также именуемые как A/B тесты.

В чем заключается суть. Из множества $\mathcal{U}$ набирается случайным образом достаточно большое подмножество объектов, после чего treatment распределяется также случайным образом между ними, образуя $\mathcal{A}$ и $\mathcal{B}$ группы. За счёт рандомизации распределения T группы оказываются похожими, а точнее гомогенными. Что такое гомогенность?

> ***Гомогенность*** (однородность) — свойство исследуемых значений иметь схожие качества или показатели. В контексте сравнения двух выборок гомогенность означает степень сходства или однородности этих выборок по определенным параметрам или характеристикам. Если выборки гомогенны, это означает, что они очень похожи друг на друга по исследуемым признакам. В противном случае, если выборки гетерогенны, они различаются по этим признакам.

В частности, как уже говорилось, нас интересует межгрупповое равенство средних как для $Y_{(0)}$, так и для $Y_{(1)}$:

$$\frac{1}{|\mathcal{M}_{(1)}|} \sum_{i \in \mathcal{M}_{(1)}} Y^i_{(0)} = \frac{1}{|\mathcal{M}_{(0)}|} \sum_{i \in \mathcal{M}_{(0)}} Y^i_{(0)}$$
$$\frac{1}{|\mathcal{M}_{(1)}|} \sum_{i \in \mathcal{M}_{(1)}} Y^i_{(1)} = \frac{1}{|\mathcal{M}_{(0)}|} \sum_{i \in \mathcal{M}_{(0)}} Y^i_{(1)}$$

Итак, все, чего мы хотим – получить разбиение на группы, которые будут гомогенны по $Y_{(0)}$ и $Y_{(1)}$, чего нам будет вполне достаточно для успешного устранения $BIAS$.

Что дает рандомизация

Гомогенность групп в A/B тесте объясняется Центральной предельной теоремой. Эта теорема говорит о том, что если мы имеем достаточно большое количество случайных независимых элементов (например, участников теста), то сумма их влияний на результаты становится похожей на нормальное распределение, независимо от их исходных распределений. В контексте A/B тестов это означает, что если выборка случайным образом разделена на контрольную и тестовую группы, и количество участников в каждой группе достаточно велико, то основные характеристики групп такие, как среднее, будут схожими.

Признаковые переменные

Естественно, если мы хотим контролировать эксперимент, узнать, с какой вероятностью тест «соврет», с какой погрешностью мы считаем ATE, какой объем выборки нам необходим и тому подобное - необходимо обладать некоей информацией о распределении случайных величин $Y_{(0)}$ и $Y_{(1)}$. Но мы их пока ничего о них не знаем. В том числе потому, что чаще всего значения этих величин до начала эксперимента еще даже не замеряны.

Что делать?

В качестве прокси-значений для $Y_{(0)}$ можно использовать, например, прошлые значения целевой переменной ($Y_{lag\,1}$). Логично предположить, что их распределения будут обладать общими свойствами, поскольку обе величины измеряются при отсутствии воздействия, хоть и в разные моменты времени. Для $Y_{(1)}$ дисперсия может отличаться и быть обусловлена вспомогательными переменными (X_1, X_2, X_3, и т.д.). Таким образом, для получения гомогенных групп мы ориентируемся на вспомогательные переменные, от которых может зависеть целевая переменная.

Этапы проведения A/B теста

Итак, для того, чтобы провести A/B тест «руками», нужно довольно хорошо разбираться в этой теме. Хоть алгоритм и прост, но содержит довольно большое количество подводных камней. На эту тему существует огромное количество книг и методичек. В данной статье мы не станем дублировать эту информацию. Для иллюстрации мы разберем самые классические этапы проведения A/B теста и рассмотрим пример на python с использованием готовых инструментов, помогающих формализовать и строго контролировать такие понятия, как гомогенность и её статзначимость по каждой переменной.

Этап 1: Формулирование гипотезы

Первым шагом является формулирование гипотезы. Гипотеза должна быть конкретной и проверяемой. Например:

> *Изменение цвета кнопки "Купить" на веб-странице с синего на зелёный увеличит количество кликов на 10%.*

Этап 2: Определение целевой метрики

Далее необходимо определить целевую метрику, которая будет использоваться для оценки эффективности изменений. В нашем примере целевая метрика — это количество кликов на кнопку "Купить".

Этап 3: Формулирование математической гипотезы

Затем формулируется математическая гипотеза. Например, нулевая гипотеза H_0 может утверждать, что нет разницы в количестве кликов между двумя версиями кнопки, в то время как альтернативная гипотеза H_1 будет утверждать обратное.

$$H_0 : \mu_A = \mu_B$$
$$H_1 : \mu_A \neq \mu_B$$

Этап 4: Разбиение на группы

На этом этапе пользователи случайным образом разделяются на две группы. Рандомизация, как мы помним, необходима для того, чтобы обеспечить равномерное распределение характеристик пользователей между группами.

Этап 5: Контроль гомогенности вспомогательных переменных (A/A тест)

Перед проведением самого A/B-теста часто проводят AA-тест для проверки гомогенности вспомогательных переменных. Этот тест заключается в том, что две группы формируются и анализируются без введения изменений. Цель — убедиться, что группы действительно являются случайными и схожими по важным характеристикам. Для этого нужно:

1. Повторение большого количества A/B тестов с тем же сценарием, что и исходный. Отличием будет являться фиктивность воздействия.
2. Сбор информации о целевых и вспомогательных показателях испытания. Контроль и, при необходимости, калибровка доверительных интервалов, распределений p-value и т.д.
3. Успешное завершение A/A теста поможет нам исправить возможные ошибки дизайна исходного эксперимента и удостовериться в работоспособности выстроенного механизма.
4. Случается, что A/A тест провалился. Часто в таком случае следует изменить алгоритм разбиения (например, использовать стратификацию), проверить гипотезу, либо исследовать данные на взаимозависимость.

Этап 6: Принятие решения

На основе анализа результатов принимается решение о том, была ли гипотеза подтверждена или опровергнута. Если p-value меньше заранее установленного уровня значимости (например, 0.05), то нулевая гипотеза отвергается, и считается, что изменение имеет значимый эффект.

Пример анализа данных на Python

В данном примере мы рассмотрим проведение A/B теста на синтетических данных с использованием библиотеки `hypex`.

0.1 Шаг 1: Установка и импорт необходимых библиотек

Для начала установим библиотеку `hypex`, если она еще не установлена, и импортируем необходимые модули.

```
!pip install hypex
import numpy as np
import pandas as pd
import hypex as hx
```

0.2 Шаг 2: Генерация синтетических данных

Создадим синтетические данные для двух групп: контрольной (A) и тестовой (B).

```
# Установка случайного зерна для воспроизводимости
np.random.seed(42)

# Размер выборок
n_A = 1000
n_B = 1000

# Генерация данных
```

```
data_A = np.random.normal(loc=50, scale=10, size=n_A)
data_B = np.random.normal(loc=55, scale=10, size=n_B)

# Создание DataFrame
df = pd.DataFrame({
    'group': ['A'] * n_A + ['B'] * n_B,
    'value': np.concatenate([data_A, data_B])
})
```

0.3 Шаг 3: Проведение A/B теста с использованием hypex

Используем библиотеку hypex для проведения A/B теста.

```
# Инициализация эксперимента
experiment = hx.Experiment(data=df, test_group='B', control_group='A', metric='value')

# Проведение теста
results = experiment.run()

# Вывод результатов
print(results)
```

0.4 Шаг 4: Интерпретация результатов

Результаты A/B теста включают в себя p-значение, доверительный интервал и другие статистические показатели, которые помогут принять решение о наличии значимых различий между группами.

```
# Пример вывода результатов
print(f"P-value: {results['p-value']}")
print(f"Mean difference: {results['mean_difference']}")
print(f"Confidence interval: {results['confidence_interval']}")
```

Заключение

Как и любой другой метод, рандомизированные исследования имеют свои ограничения и недостатки. К ним относятся некоторая дороговизна, длительность процесса и, иногда, практическая невозможность полного случайного распределения объектов по группам. Например, при проведении эксперимента по изучению вреда курения на беременных женщинах, невозможно принудительно назначить участника в ту или иную группу. В следующей статье мы рассмотрим метод, который поможет нам в подобных случаях определить ATE.