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.

A/B тест

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 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.

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

Рандомизация — средство борьбы с bias-ом

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 в последствии будет достаточно для устранения $BIAS$.

1 Применение Рандомизации в A/B-Тестировании

1.1 Зачем нужна рандомизация?

При проведении A/B-тестирования обычно используются случайные методы формирования групп, такие как рандомизация. Это делается для того, чтобы обеспечить однородное распределение переменных между группами, что в свою очередь приводит к созданию гомогенных групп.

Рандомизация помогает уменьшить влияние внешних факторов, которые могут исказить результаты тестирования. Например, если вы хотите проверить эффективность нового дизайна веб-страницы, вы можете

разделить пользователей на две группы: одна будет видеть старый дизайн, другая - новый. В то же время, если вы просто возьмете первых попавшихся пользователей и отправите их в разные группы, это может привести к тому, что в одной группе окажется больше людей одной категории. Например, преобладание пользователей одного пола или региона проживания может привести к искажению результатов тестирования.

При использовании рандомизации, в большинстве случаев каждая группа будет иметь примерно одинаковое соотношение пользователей со схожими характеристиками (пол, возраст, регион и т.д.), что делает группы более гомогенными.

Можно пояснить этот эффект иначе, при помощи ЦПТ из теории вероятностей. Действительно, при случайном выборе кандидатов и случайном распределении на группы $\mathcal{A}$ и $\mathcal{B}$, при достаточно большом размере групп средние значения потенциальных значений target ($Y_{(0)}$ и $Y_{(1)}$) будут близки к матожиданию этих величин. Следовательно, схожи между собой.

На самом деле в этих рассуждениях существует некоторое слабое место. А именно, для оценки точности расчета ATE необходимо знать, насколько схожи потенциальные значения target в группе. Мы могли бы использовать ЦПТ или другие теоремы, если бы знали дисперсии этих случайных величин. Но мы их не знаем.

В качестве прокси-значений $Y_(0)$ можно брать, например, прошлые значения целевой переменной (Y_{lag1}). И правда, в некоторых случаях логично предположить, что их распределения будут обладать некими общими свойствами, поскольку обе величины "замеряются" при условии отсутствия воздействия, хоть и в разные моменты времени. При этом для $Y_(1)$ дисперсия, предположительно, может отличаться от $Y_(0)$ и быть обусловлена величинами вспомогательных переменных X_1, X_2, X_3,.. Поэтому, за неимением потенциальных значений target, мы хотим получить группы, гомогенные по вспомогательным переменным, от которых target может быть зависим.

Данный трюк помогает предугадывать и контролировать интересующие нас значения дисперсий. Ввиду чего мы довольно уверенно сможем набрать достаточное количество объектов в экспериментальные группы для нивелирования bias.

Итак, для того, чтобы провести AB 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 AB 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.

2 Этапы проведения AB теста

2.1 формулирование гипотезы на естественном языке

2.2 формирование целевой метрики

- чэп - маржинальность - просмотры - клики - смертность

2.3 формулирование математической гипотезы

2.4 разница средних величин целевой метрики в группах A и B будет отлична от нуля

2.5 с учетом бизнес требований выбираем алгоритм разбиения, часто произвольно-случайный

2.6 контроль гомогенности вспомогательных переменных (AA test (плацебо тест))

- проводим 2000 разбиений - исследуем распределения p-value, тестируя набор гипотез, предполагающих отсутствие гомогенности по одной из вспомогательной переменной - прим.: если AA тест провалился, то имеет смысл изменить алгоритм разбиения (стратификация, изменение выборки, фильтрация) либо гипотезу, либо проверить данные на взаимозависимость

3 Пример

код примера

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.