МГТУ им. Н.Э. Баумана
ТЕХНОПАРК
Mail.Ru Group

ТЕХНОАТОМ

Проверка гипотез. A/B тесты

Лекция 10

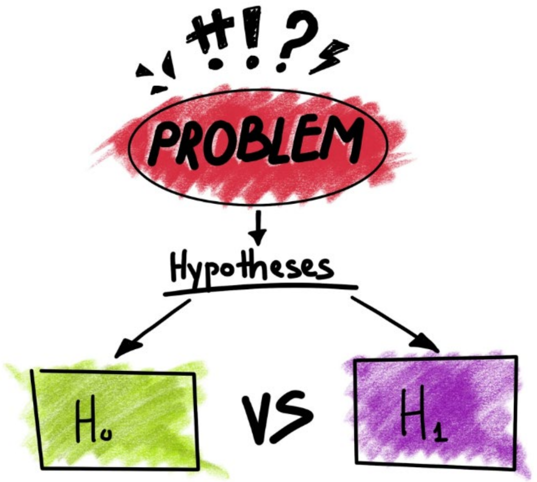
Иван Горбань

План занятия:

1. Проверка гипотез
2. Рандомизированный эксперимент
3. Перед AB тестом
4. Проведение теста
5. Оценка результатов
   - Простые проверки
   - CUPED
   - Difference in differences
   - ANCOVA
6. Отсутствие рандомизации
   - Propensity score
   - Regression Discontinuity
7. Другие темы
   - Uplift моделирование
   - Современные методы
   - Полезные библиотеки

#02

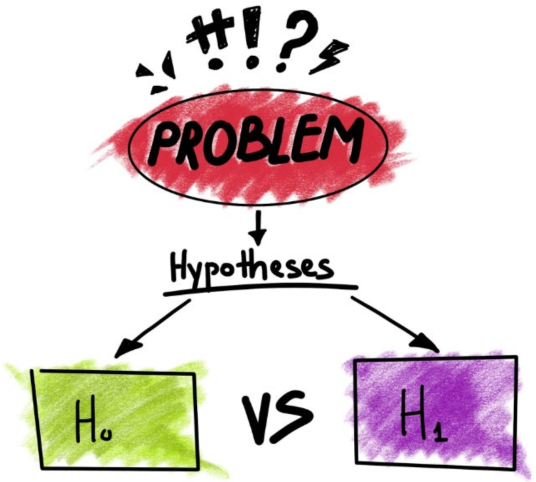
Проверка гипотез



- Первое что необходимо – сама гипотеза.
  Будем проверять нулевую гипотезу H_0 против альтернативы H_1.

#03

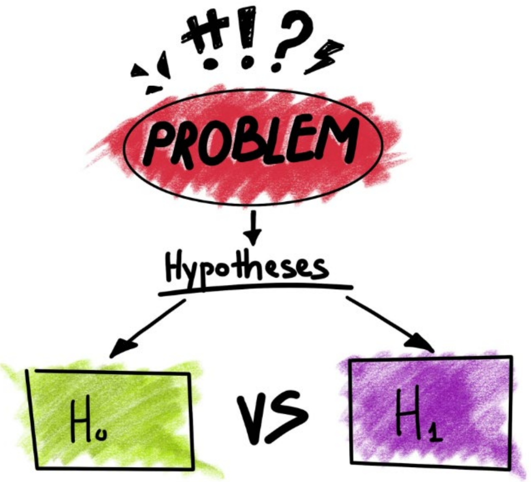
Проверка гипотез



- Первое что необходимо – сама гипотеза.
  Будем проверять нулевую гипотезу H_0 против альтернативы H_1.
- Второе – выборка реализаций случайной величины $\{x_1, ..., x_n\}$.

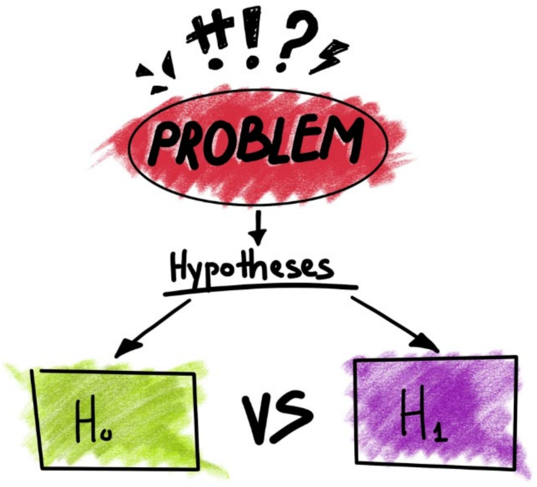
#04

Проверка гипотез



- Первое что необходимо – сама гипотеза.
  Будем проверять нулевую гипотезу H_0 против альтернативы H_1.
- Второе – выборка реализаций случайной величины $\{x_1, ..., x_n\}$.
- Третье, что нам понадобится – статистика критерия.
  Для различных гипотез используются различные статистики. В общем случае статистика – это некоторое значение, вычисляемое на основе выборки.

Проверка гипотез



- Первое что необходимо – сама гипотеза.
  Будем проверять нулевую гипотезу H_0 против альтернативы H_1.
- Второе – выборка реализаций случайной величины $\{x_1, ..., x_n\}$.
- Третье, что нам понадобится – статистика критерия.
  Для различных гипотез используются различные статистики. В общем случае статистика – это некоторое значение, вычисляемое на основе выборки.

Проверка выполняется таким образом, что при достижении некоторых значений статистики мы можем отвергнуть нулевую гипотезу в пользу альтернативной.

Проверка гипотез

Зададим некоторую α – вероятность, с которой мы можем отвергнуть верную нулевую гипотезу ($\alpha = P(\overline{H_0}|H_0)$), то есть ошибиться.
Обычно задается как 0.05 или 0.1.

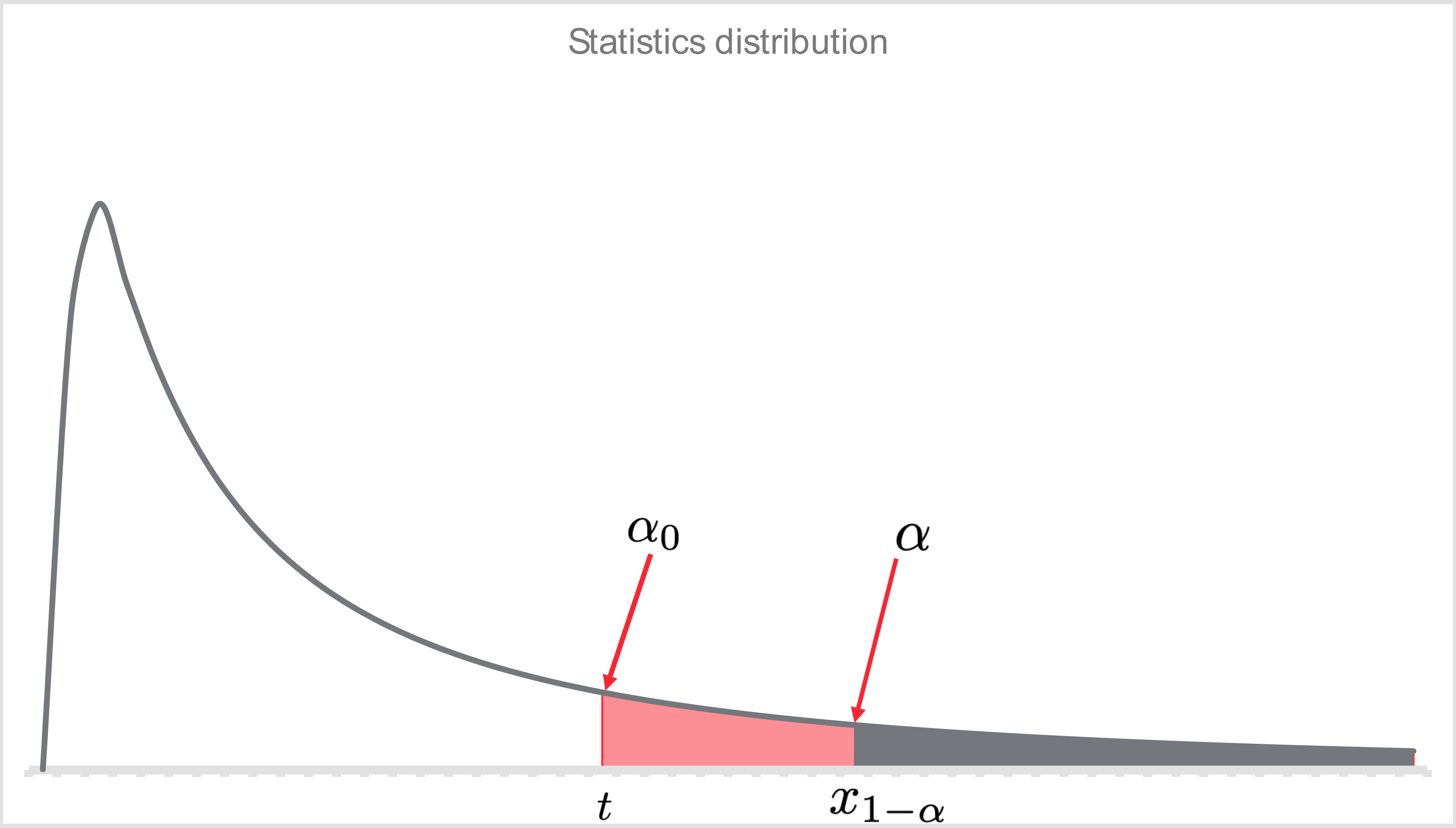
α называют уровнем значимости и определяют так:

$$P(T(x_1, ..., x_n) \geq x_{1-\alpha}) \leq \alpha,$$

где $x_{1-\alpha}$,как видно из определения, – квантиль уровня $1-\alpha$ для случайной величины T – выборочной статистики.

Эту квантиль называют критическим значением статистики.

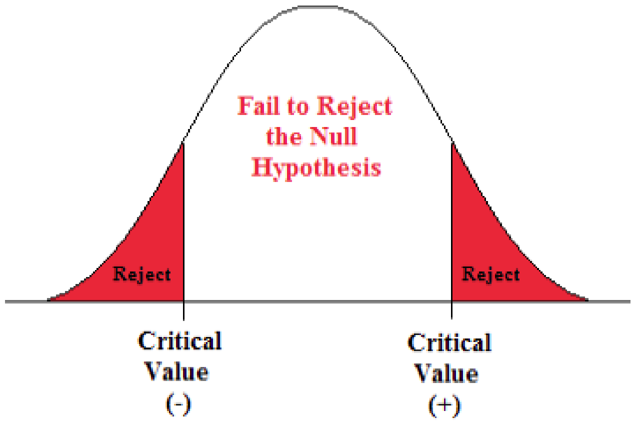
Проверка гипотез



Пусть на выборке мы получили некоторое значение t для статистики.

Тогда $P(T(x_1, ..., x_n) \geq t)$ называют фактическим уровнем значимости и обозначают α_0 .

Также принято название p-value.



https://towardsdatascience.com/demystifying-hypothesis-testing-with-simple-python-examples-4997ad3c5294

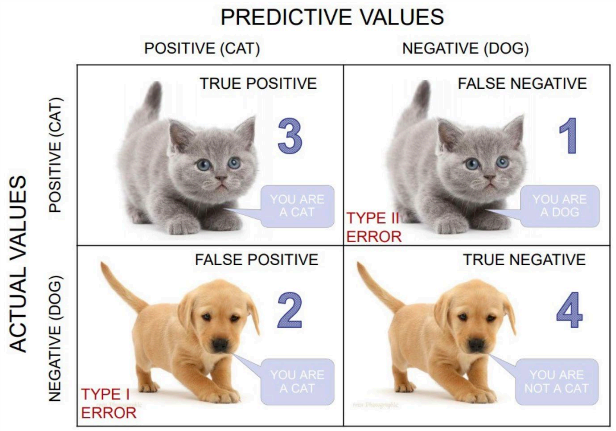
Проверка гипотез

Как мы поняли, проверки статистических гипотез ошибаются и с этим ничего не поделаешь. Единственное, что можно сделать – контролировать уровень ошибок.

Ошибки разделяют:

	H_0 верная	H_1 верная
H_0 принимается	Верно	Ошибка II рода
H_0 отвергается	Ошибка I рода	Верно

Как видно из таблицы, уровень значимости – вероятность ошибки первого рода.
(1-вероятность ошибки второго рода) называют мощностью критерия.



https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb

Проверка гипотез

Формально записать ошибки I и II родов можно следующим образом:

Ошибка I рода:

$$P(\overline{H_0}|H_0) = P(p_{value} \leq \alpha|H_0) \leq \alpha$$

Мощность:

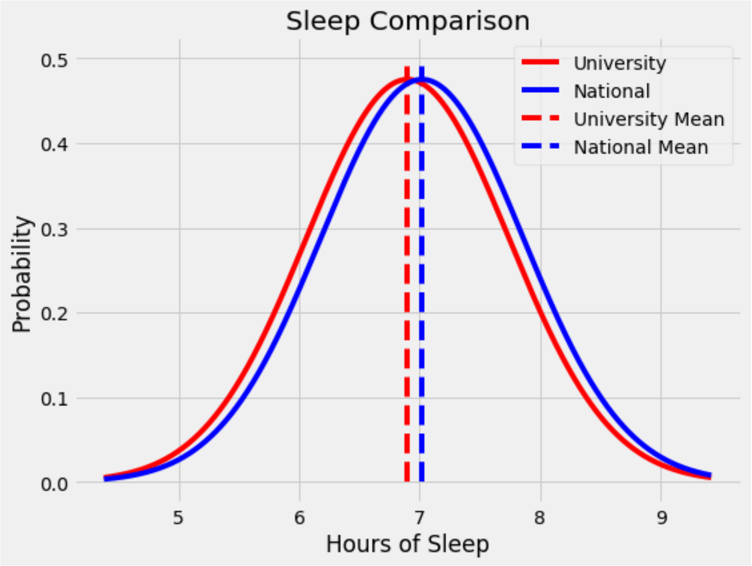
$$Power = P(\overline{H_0}|H_1) = 1 - P(H_0|H_1)$$

Если $Power \to 1$ при $n \to \infty$ критерий называют состоятельным.

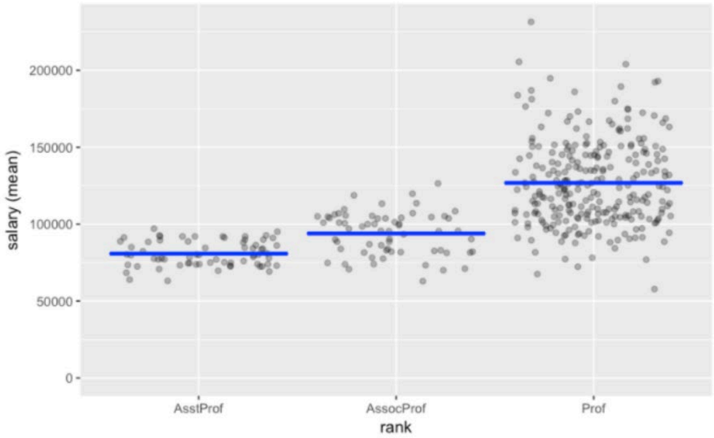
Проверка гипотез

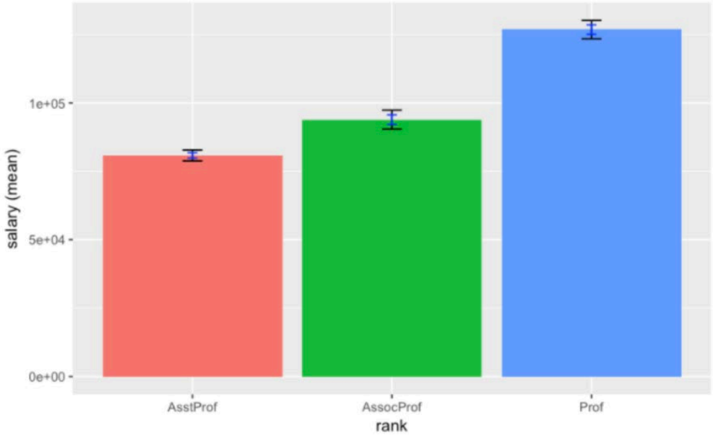
Гипотеза о равенстве средних двух выборок:

Часто проверяется после проведения A/B теста.
Есть две выборки, имеем их среднее каждой.
Какой процесс эффективнее?



https://towardsdatascience.com/statistical-significance-hypothesis-testing-the-normal-curve-and-p-values-93274fa32687





https://radiant-rstats.github.io/docs/basics/compare_means.html

Проверка гипотез

Будем рассматривать пример двух выборок, как это часто бывает в A/B тестах. Обозначим выборки как $X_1^{n_1}$ и $X_2^{n_2}$.

Перечислим наиболее популярные из тестов:

Z-критерий:

Предположения: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, известные σ_1, σ_2.

Гипотезы: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$

Статистика теста:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

При верной нулевой гипотезе $Z \sim N(0, 1)$

Проверка гипотез

t-критерий:

Предположения: $X_1 \sim N(\mu_1, \sigma_1^2),\ X_2 \sim N(\mu_2, \sigma_2^2)$, неизвестные σ_1, σ_2.

Гипотезы: $H_0 : \mu_1 = \mu_2,\ H_1 : \mu_1 \neq \mu_2$.

Статистика теста:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

При верной нулевой гипотезе $T \sim St(\nu)$, где $\nu = \dfrac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$

Проверка гипотез

t-критерий для связанных выборок:

Предположения: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, неизвестные σ_1, σ_2.

Гипотезы: $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$.

Статистика теста:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\frac{S}{\sqrt{n}}},$$

где $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \overline{D})^2, \;\; D_i = X_{1i} - X_{2i}$.

При верной нулевой гипотезе $T \sim St(n-1)$.

Проверка гипотез

Проверка на нормальность:

Имеем выборку $X^n = (x_1, ..., x_n)$.

Гипотезы: $H_0 : X \sim N(\mu, \sigma^2)$, $H_1 : \overline{H_0}$.

Разбиваем отсортированную выборку на Q частей, где каждая часть определяется границами $[q_i, q_{i+1}]$.

Статистика теста:

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i},$$

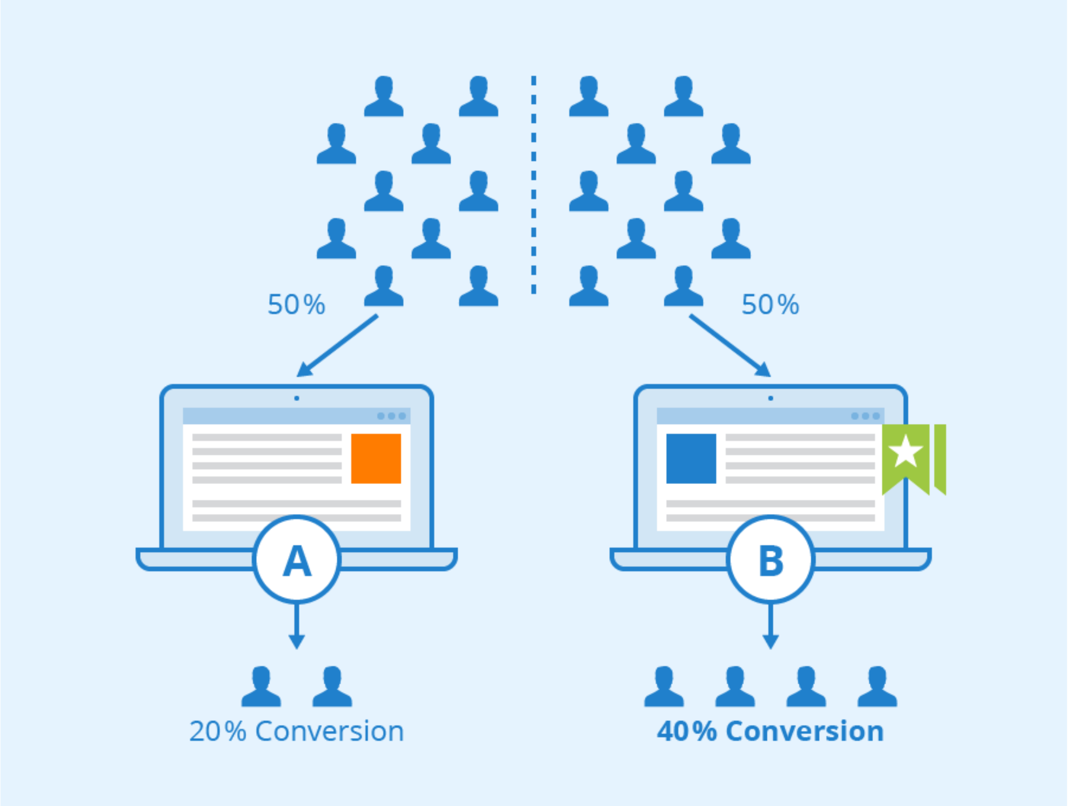
где n_i - число наблюдений в $[q_i, q_{i+1}]$. $p_i = F_{N(\mu,\sigma^2)}(q_{i+1}) - F_{N(\mu,\sigma^2)}(q_i)$

При верной нулевой гипотезе $\chi^2 \sim \chi^2_{K-1}$ или $\chi^2 \sim \chi^2_{K-3}$, в зависимости от известности μ, σ.

Проверка гипотез

Гипотеза о равенстве долей:

Применима, например, к конверсии по результатам A/B теста.



Проверка гипотез

z-критерий для доли:

Предположения: $X_1 \sim Ber(p_1)$, $X_2 \sim Ber(p_2)$.

Гипотезы: $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$.

Статистика теста:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} - \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

При верной нулевой гипотезе $Z \sim N(0,1)$.

Проверка гипотез

z-критерий для доли для связанных выборок:

Предположения: $X_1 \sim Ber(p_1)$, $X_2 \sim Ber(p_2)$.

Гипотезы: $H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$.

Статистика теста:

$$Z = \frac{f - g}{\sqrt{f + g - \frac{(f-g)^2}{n}}},$$

где f - элементы, которые равны 1в первой выборке и 0 – во второй,
g – наоборот.

При верной нулевой гипотезе $Z \sim N(0,1)$.

Проверка гипотез

Знаковый критерий для связанных выборок:

Гипотезы: $H_0 : P(X_1 > X_2) = \frac{1}{2}$, $H_1 : P(X_1 > X_2) \neq \frac{1}{2}$.

Статистика теста:

$$T = \sum_{i=1}^{n} I(X_{1i} > X_{2i})$$

При верной нулевой гипотезе $T \sim Bin(n, \frac{1}{2})$.

Проверка гипотез

Ранговый критерий Манна-Уитни:

Гипотезы: $H_0 : F_{X_1}(x) = F_{X_2}(x), \quad H_1 : F_{X_1}(x) = F_{X_2}(x+\Delta)$.

Перед расчетом статистики записываем объединенный вариационный ряд выборок из X_1, X_2.
Статистика теста:

$$U = min\left(\sum_{i=1}^{n_1+n_2} rank(X_{1i}) - \frac{n_1(n_1+1)}{2}, \sum_{i=1}^{n_1+n_2} rank(X_{2i}) - \frac{n_2(n_2+1)}{2}\right)$$

Распределение статистики табличное, но асимптотически сходится к:

$$U \sim N\left(\frac{n_1 n_2}{2}, \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right)$$

Проверка гипотез

Перестановочный критерий для независимых выборок:

Гипотезы: $H_0 : F_{X_1}(x) = F_{X_2}(x), \quad H_1 : F_{X_1}(x) = F_{X_2}(x + \Delta).$

Статистика теста:

$$T = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

При верной нулевой гипотезе распределение статистики порождается всеми значениями статистики размещений $C_{n_1+n_2}^{n_1}$ и $C_{n_1+n_2}^{n_2}$ объединенной выборки.

Проверка гипотез

Перестановочный критерий для связанных выборок:

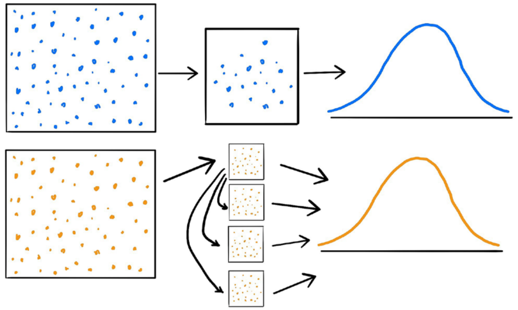
Гипотезы: $H_0 : \mathbb{E}(X_1 - X_2) = 0,\ H_1 : \mathbb{E}(X_1 - X_2) \neq 0\,.$

Статистика теста:

$$T = \sum_{i=1}^{n} (X_{1i} - X_{2i})$$

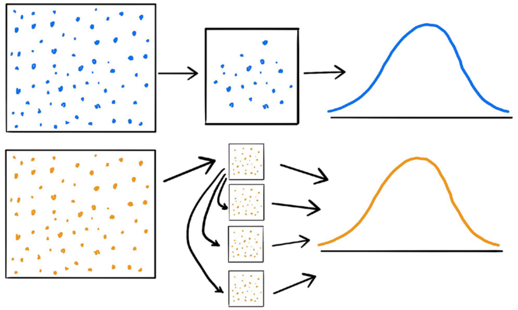
При нулевой гипотезе распределение статистики получаем путем перебора всех возможных знаков при суммируемых показателях.

Bootstrap



Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

Bootstrap

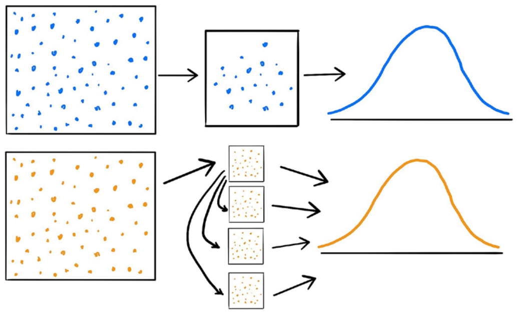


Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

Пусть у нас есть выборка из n элементов.

Генерируем В бутстрапированных выборок (сэмплируем с возвращением выборки размера n)

Bootstrap



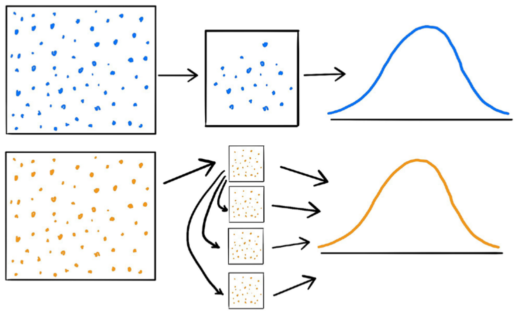
Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

Пусть у нас есть выборка из n элементов.

Генерируем B бутстрапированных выборок (сэмплируем с возвращением выборки размера n)

Для каждой выборки рассчитываем значение статистики (к примеру, некая T)

Bootstrap



Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

Пусть у нас есть выборка из n элементов.

Генерируем B бутстрапированных выборок (сэмплируем с возвращением выборки размера n)

Для каждой выборки рассчитываем значение статистики (к примеру, некая T)

В качестве квантилей $q_\alpha, q_{1-\alpha}$ берем $\hat{T}_{[B\alpha]}, \hat{T}_{[B(1-\alpha)+1]}$ в вариационном ряду для T.

Bootstrap

Предположим мы хотим бутстрапировать статистику $t = \frac{\hat{\theta}}{se(\hat{\theta})}$ для проверки $H_0 : \theta = 0$

Bootstrap

Предположим мы хотим бутстрапировать статистику $t = \frac{\hat{\theta}}{se(\hat{\theta})}$ для проверки $H_0 : \theta = 0$

Тогда бутстраповским аналогом нашей t-статистики будет:

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{se^*(\hat{\theta})}$$

Bootstrap

Предположим мы хотим бутстрапировать статистику $t = \frac{\hat{\theta}}{se(\hat{\theta})}$ для проверки $H_0 : \theta = 0$

Тогда бутстраповским аналогом нашей t-статистики будет:

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{se^*(\hat{\theta})}$$

Это дополнение к бутстрапу называется рецентрирование. Его используют для проверки гипотезы в статистике, содержащей расстояния между выборочными и популяционными объектами.

Bootstrap

Также бутстрап позволяет корректировать смещение статистики, вызванное конечностью выборки

$$\hat{\theta}_{BC} = 2\hat{\theta} - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

Bootstrap

Также бутстрап позволяет корректировать смещение статистики, вызванное конечностью выборки

$$\hat{\theta}_{BC} = 2\hat{\theta} - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

Для расчета доверительных интервалов для оценки параметра используют:

$$CI_{|\%t|} = \left[\hat{\theta} - se(\hat{\theta})q_{1-\alpha}^{*\%|t|}, \hat{\theta} + se(\hat{\theta})q_{1-\alpha}^{*\%|t|}\right],$$

где $q_{1-\alpha}^{*\%|t|}$ определяется как выборочная квантиль из бутстрап распределения $t^* = \frac{|\hat{\theta}^* - \hat{\theta}|}{se^*(\hat{\theta})}$

http://quantile.ru/03/03-SA.pdf

Рандомизированный эксперимент

Рандомизированный эксперимент (RCT, A/B test) проводится с целью выявить эффект воздействия.

Выполняется путем выделения контрольной (control) и целевой (treatment) групп, после чего к целевой группе применяется воздействие.

Рандомизированный эксперимент

Рандомизированный эксперимент (RCT, A/B test) проводится с целью выявить эффект воздействия.

Выполняется путем выделения контрольной (control) и целевой (treatment) групп, после чего к целевой группе применяется воздействие.

В чем смысл рандомизации?

В том, чтобы исключить возможность влияния неучтенных переменных на результат и исключить присутствие selection bias. Цель – выявить причинно-следственный эффект воздействия на таргет.

Рандомизированный эксперимент

Рандомизированный эксперимент (RCT, A/B test) проводится с целью выявить эффект воздействия.

Выполняется путем выделения контрольной (control) и целевой (treatment) групп, после чего к целевой группе применяется воздействие.

В чем смысл рандомизации?

В том, чтобы исключить возможность влияния неучтенных переменных на результат и исключить присутствие selection bias. Цель – выявить причинно-следственный эффект воздействия на таргет.

Примеры?

Рандомизированный эксперимент

Введем несколько формальных обозначений.

Пусть у нас есть два периода наблюдений – до эксперимента (T=0) и после (T=1).

Эксперимент описывается выборкой из троек: $S = \{(y_{iT}, W_{iT}, X_{iT})\}_{i=1, T\in\{0,1\}}^{N,T}$

Здесь y_{iT} - значение целевой переменной для i-го объекта в момент времени T.

W_{iT} - воздействие (treatment), равное 0 или 1.

X_{iT} - параметры объектов.

Рандомизированный эксперимент

Нас интересует эффект воздействия. И чаще всего, нас интересует не уровень y_{iT} , а его изменение $Y_i = y_{i1} - y_{i0}$. Также treatment $W_i = W_{i1} - W_{i0}$

Обозначим за $Y_i(W)$ потенциальное изменение таргета i-го объекта при воздействии или его отсутствии.

Рандомизированный эксперимент

Нас интересует эффект воздействия. И чаще всего, нас интересует не уровень y_{iT} , а его изменение $Y_i = y_{i1} - y_{i0}$. Также treatment $W_i = W_{i1} - W_{i0}$

Обозначим за $Y_i(W)$ потенциальное изменение таргета i-го объекта при воздействии или его отсутствии.

Различают несколько эффектов воздействия, приведем основные:

$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$ - Average Treatment Effect

Рандомизированный эксперимент

Нас интересует эффект воздействия. И чаще всего, нас интересует не уровень y_{iT} , а его изменение $Y_i = y_{i1} - y_{i0}$. Также treatment $W_i = W_{i1} - W_{i0}$

Обозначим за $Y_i(W)$ потенциальное изменение таргета i-го объекта при воздействии или его отсутствии.

Различают несколько эффектов воздействия, приведем основные:

$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$ - Average Treatment Effect

$\tau_{TOT} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]$ - Treatment On the Treated

Рандомизированный эксперимент

Нас интересует эффект воздействия. И чаще всего, нас интересует не уровень y_{iT}, а его изменение $Y_i = y_{i1} - y_{i0}$. Также treatment $W_i = W_{i1} - W_{i0}$

Обозначим за $Y_i(W)$ потенциальное изменение таргета i-го объекта при воздействии или его отсутствии.

Различают несколько эффектов воздействия, приведем основные:

$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$ - Average Treatment Effect

$\tau_{TOT} = \mathbb{E}[Y_i(1) - Y_i(0) | W_i = 1]$ - Treatment On the Treated

В рандомизированном эксперименте $W \perp\!\!\!\perp (Y(0), Y(1))$, из чего следует $\tau_{ATE} = \tau_{TOT}$

Рандомизированный эксперимент

Более того, при рандомизации, так как из $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$ следует $\mathbb{E}[Y_i(0)|W_i = 0] = \mathbb{E}[Y_i(0)|W_i = 1]$, то:

Рандомизированный эксперимент

Более того, при рандомизации, так как из $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$ следует $\mathbb{E}[Y_i(0)|W_i = 0] = \mathbb{E}[Y_i(0)|W_i = 1]$, то:

$$\tau_{naive} = \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] =$$

$$= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] + \mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] =$$

$$= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)] = \tau_{ATE}$$

Рандомизированный эксперимент

Более того, при рандомизации, так как из $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$ следует $\mathbb{E}[Y_i(0)|W_i = 0] = \mathbb{E}[Y_i(0)|W_i = 1]$, то:

$$\tau_{naive} = \mathbb{E}[Y_i(1)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] =$$
$$= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] + \mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] =$$
$$= \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)] = \tau_{ATE}$$

В противном же случае:

$$\tau_{naive} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1] + \mathbb{E}[Y_i(0)|W_i = 1] - \mathbb{E}[Y_i(0)|W_i = 0] =$$
$$= \tau_{ATT} + SelectionBias$$

Рандомизированный эксперимент

Разделение на группы может выполняться несколькими способами:

- Полностью рандомизированный эксперимент

Рандомизированный эксперимент

Разделение на группы может выполняться несколькими способами:

- Полностью рандомизированный эксперимент
- Стратифицированный рандомизированный эксперимент

Рандомизированный эксперимент

Разделение на группы может выполняться несколькими способами:

- Полностью рандомизированный эксперимент
- Стратифицированный рандомизированный эксперимент
- Парный рандомизированный эксперимент

Рандомизированный эксперимент

Разделение на группы может выполняться несколькими способами:

- Полностью рандомизированный эксперимент
- Стратифицированный рандомизированный эксперимент
- Парный рандомизированный эксперимент
- Кластеризованный рандомизированный эксперимент

Рандомизированный эксперимент

Зачем нужна стратификация?

Рандомизированный эксперимент

Зачем нужна стратификация?

- В большинстве случаев позволяет увеличить мощность теста, путем уменьшения дисперсии оценки.

Рандомизированный эксперимент

Зачем нужна стратификация?

- В большинстве случаев позволяет увеличить мощность теста, путем уменьшения дисперсии оценки.
- Позволяет изолировать на стратах влияние воздействия.

Рандомизированный эксперимент

Существуют требования к валидности эксперимента:

Рандомизированный эксперимент

Существуют требования к валидности эксперимента:

Внутренняя валидность:

- Субъекты могут выйти из эксперимента. Если это связано с таргетом, то для оценки это плохо – создает attrition bias

Рандомизированный эксперимент

Существуют требования к валидности эксперимента:

Внутренняя валидность:

- Субъекты могут выйти из эксперимента. Если это связано с таргетом, то для оценки это плохо – создает attrition bias
- Compliance. Нам необходимо понимать, что субъекты примут воздействие. Потому для корректной оценки ATE нам важно брать initial assignment для воздействия, а не факт.

Рандомизированный эксперимент

Внешняя валидность:

- Разница в популяции. В тесте участвует подвыборка, которая может значительно отличаться от генеральной совокупности. Эффект на всю популяцию может быть другим.

Рандомизированный эксперимент

Внешняя валидность:

- Разница в популяции. В тесте участвует подвыборка, которая может значительно отличаться от генеральной совокупности. Эффект на всю популяцию может быть другим.
- Эффект наблюдения. Если субъекты знают об участии в эксперименте – это может вносить свои корректировки в их поведение.

Рандомизированный эксперимент

Внешняя валидность:

- Разница в популяции. В тесте участвует подвыборка, которая может значительно отличаться от генеральной совокупности. Эффект на всю популяцию может быть другим.
- Эффект наблюдения. Если субъекты знают об участии в эксперименте – это может вносить свои корректировки в их поведение.
- Ограниченное время. Мы можем учесть краткосрочный эффект, но часто не можем видеть долгосрочные последствия.

Рандомизированный эксперимент

Внешняя валидность:

- Разница в популяции. В тесте участвует подвыборка, которая может значительно отличаться от генеральной совокупности. Эффект на всю популяцию может быть другим.
- Эффект наблюдения. Если субъекты знают об участии в эксперименте – это может вносить свои корректировки в их поведение.
- Ограниченное время. Мы можем учесть краткосрочный эффект, но часто не можем видеть долгосрочные последствия.
- Побочные эффекты на популяцию. Когда воздействие внедрено на всю популяцию, то оно может привести к иным эффектам, за счёт нового равновесия в популяции.

Рандомизированный эксперимент

Внешняя валидность:

- Разница в популяции. В тесте участвует подвыборка, которая может значительно отличаться от генеральной совокупности. Эффект на всю популяцию может быть другим.
- Эффект наблюдения. Если субъекты знают об участии в эксперименте – это может вносить свои корректировки в их поведение.
- Ограниченное время. Мы можем учесть краткосрочный эффект, но часто не можем видеть долгосрочные последствия.
- Побочные эффекты на популяцию. Когда воздействие внедрено на всю популяцию, то оно может привести к иным эффектам, за счёт нового равновесия в популяции.
- Экстерналии! Необходимо учитывать, какие экстерналии может иметь наше воздействие на другие объекты.

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект
- Размеры выборки

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект
- Размеры выборки
- Долю целевой и контрольной групп

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект
- Размеры выборки
- Долю целевой и контрольной групп
- Тип рандомизации

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект
- Размеры выборки
- Долю целевой и контрольной групп
- Тип рандомизации
- Продолжительность теста

Перед A/B тестом

Итак, мы поняли, что хотим запустить A/B тест.

Что нам необходимо установить и проверить для запуска?

- Внутреннюю и внешнюю валидность теста
- Отсутствие spillover эффектов
- Уровень значимости и мощность
- Ожидаемый эффект
- Размеры выборки
- Долю целевой и контрольной групп
- Тип рандомизации
- Продолжительность теста
- Методологию оценки результатов (статистический критерий, гипотезы)

Перед A/B тестом

Часто нам необходимо рассчитать бюджет для эксперимента.

Для этого требуется учесть возможные издержки, а соответственно – ограничить длительность и выборку.

Как же выбрать группы и продолжительность теста?

Перед A/B тестом

Часто нам необходимо рассчитать бюджет для эксперимента.

Для этого требуется учесть возможные издержки, а соответственно – ограничить длительность и выборку.

Как же выбрать группы и продолжительность теста?

Здесь нам поможет Minimum Detectable Effect (MDE):

Он задается для конкретной мощности (k), уровня значимости, размера выборки (N), доли treatment группы (P).

$$MDE = (t_{(1-k)} + t_{\alpha})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

https://www.nber.org/system/files/working_papers/t0333/t0333.pdf

Перед A/B тестом

Часто перед A/B тестом проводят A/A тест.

Для целевой и контрольной групп производится идентичное воздействие или никакого.

Проводится тот же анализ, что и при A/B тесте.

Перед A/B тестом

Часто перед A/B тестом проводят A/A тест.

Для целевой и контрольной групп производится идентичное воздействие или никакого.

Проводится тот же анализ, что и при A/B тесте.

Плюсы:

- Позволяет выявить проблемы в дизайне эксперимента.

Перед A/B тестом

Часто перед A/B тестом проводят A/A тест.

Для целевой и контрольной групп производится идентичное воздействие или никакого.

Проводится тот же анализ, что и при A/B тесте.

Плюсы:

- Позволяет выявить проблемы в дизайне эксперимента.

Минусы:

- Требует времени, которое может быть потрачено на A/B тест.

Перед A/B тестом

Часто перед A/B тестом проводят A/A тест.

Для целевой и контрольной групп производится идентичное воздействие или никакого.

Проводится тот же анализ, что и при A/B тесте.

Плюсы:

- Позволяет выявить проблемы в дизайне эксперимента.

Минусы:

- Требует времени, которое может быть потрачено на A/B тест.

Возможная вариация:

A/A/B тест.

Перед A/B тестом

Качество рандомизации бывает полезно проверить, построив классификатор, используя в качестве таргета принадлежность к целевой или контрольной группе.

Если хороший классификатор построить не удается, то выборки разделены «хорошо».

Проведение теста

В процессе проведения теста нам необходимо регулярно мониторить целевые показатели и возможные проблемы, чтобы оперативно исправить дизайн, если будут выявлены проблемы.

Что не нужно делать – прекращать эксперимент в первый день, когда мы достигли значимого эффекта.

Оценка эффекта

По результатам пилота мы хотим оценить treatment effect от нашего воздействия.

Оценка эффекта

По результатам пилота мы хотим оценить treatment effect от нашего воздействия.

Формально:

$$\tau_{P,ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

$$\tau_{S,ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

Оценка эффекта

По результатам пилота мы хотим оценить treatment effect от нашего воздействия.

Формально:

$$\tau_{P,ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

$$\tau_{S,ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0))$$

$$\tau_{P,TOT} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i = 1]$$

$$\tau_{S,TOT} = \frac{1}{N_T}\sum_{i:W_i=1}(Y_i(1) - Y_i(0))$$

Оценка эффекта

Прежде чем рассматривать методы оценки TE, сделаем следующие предположения:

Unconfoudedness:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | X_i$$

Treatment assignment не зависит от результата эксперимента при условии контроля за X. Более слабое предположение, чем полная независимость

Оценка эффекта

Прежде чем рассматривать методы оценки TE, сделаем следующие предположения:

Unconfoudedness:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i | X_i$$

Treatment assignment не зависит от результата эксперимента при условии контроля за X. Более слабое предположение, чем полная независимость

Overlap:

$$0 < P(W_i = 1 | X_i) < 1$$

Нет таких характеристик объектов, для которых невозможен treatment.

Оценка эффекта

Использование регрессии в оценке эффекта:

В нашей выборке можем расписать: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$

Немного перегруппировав получим: $Y_i = Y_i(0) + W_i (Y_i(1) - Y_i(0))$

Оценка эффекта

Использование регрессии в оценке эффекта:

В нашей выборке можем расписать: $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$

Немного перегруппировав получим: $Y_i = Y_i(0) + W_i(Y_i(1) - Y_i(0))$

Видим, что последнее уравнение можно переписать следующим образом (предполагая линейность):

$$\boxed{Y_i = \alpha + \beta X_i + \tau W_i + \epsilon_i}$$

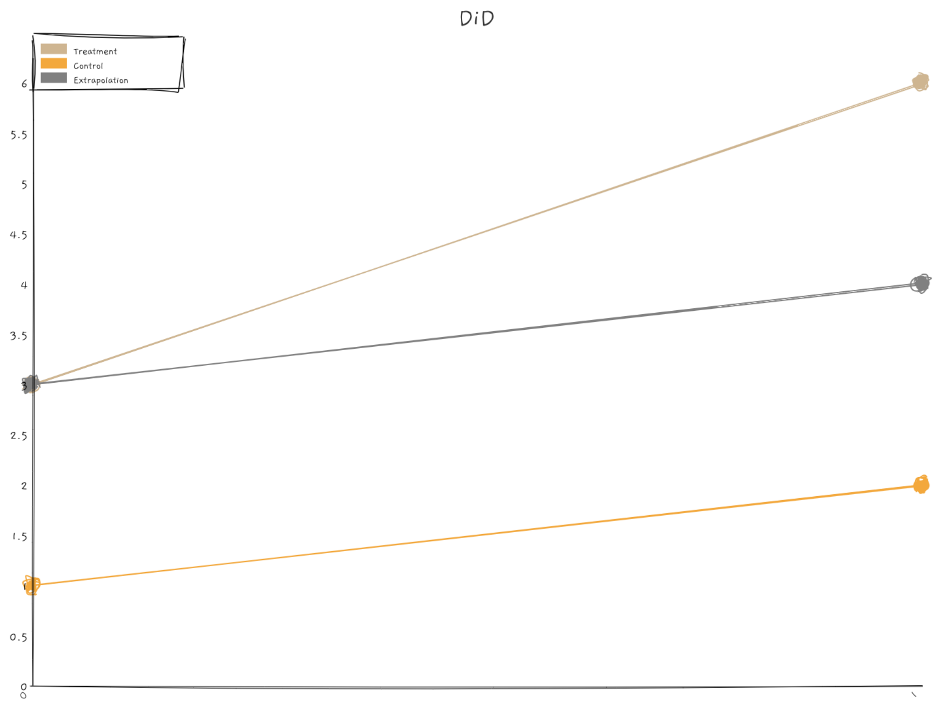
А это, как мы видим регрессионная модель, в результате чего оценкой ATE будет являться оценка $\hat{\tau}$.

http://www.nber.org/WNE/WNEnotes.pdf

Difference in Differences

Предположим мы выбрали группы, которые на старте отличаются. Запускаем эксперимент, получаем таргет.

И контрольная и целевая группы сдвинулись.



Difference in Differences

Как это оценить?

https://www.nber.org/WNE/lect_10_diffindiffs.pdf

Difference in Differences

Как это оценить?

Метод, позволяющий оценить подобный эксперимент называется Difference in differences (DiD).

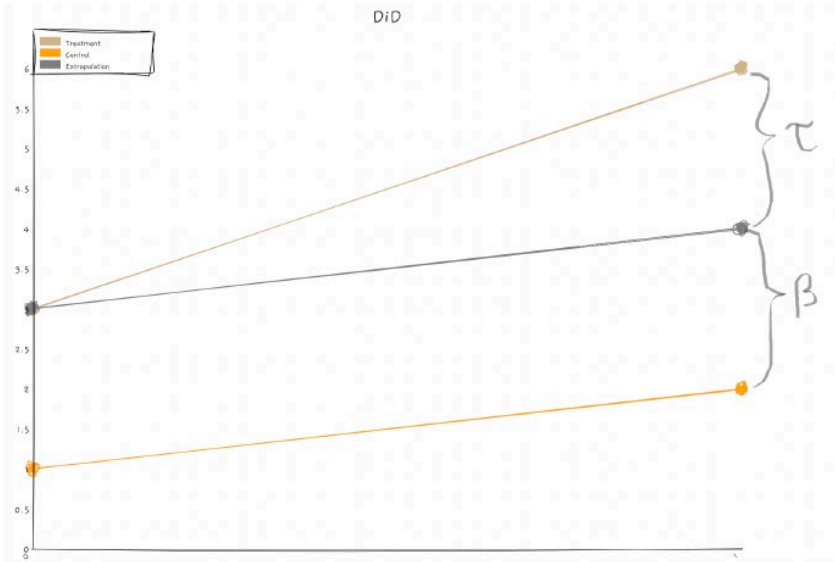
Difference in Differences

Как это оценить?

Метод, позволяющий оценить подобный эксперимент называется Difference in differences (DiD).

Оценка производится с помощью регрессии, модель выглядит следующим образом:

$$y_{iT} = \alpha + \beta I(T=1) + \gamma I(W_i = 1) + \tau I(T=1) I(W_i = 1) + \epsilon$$

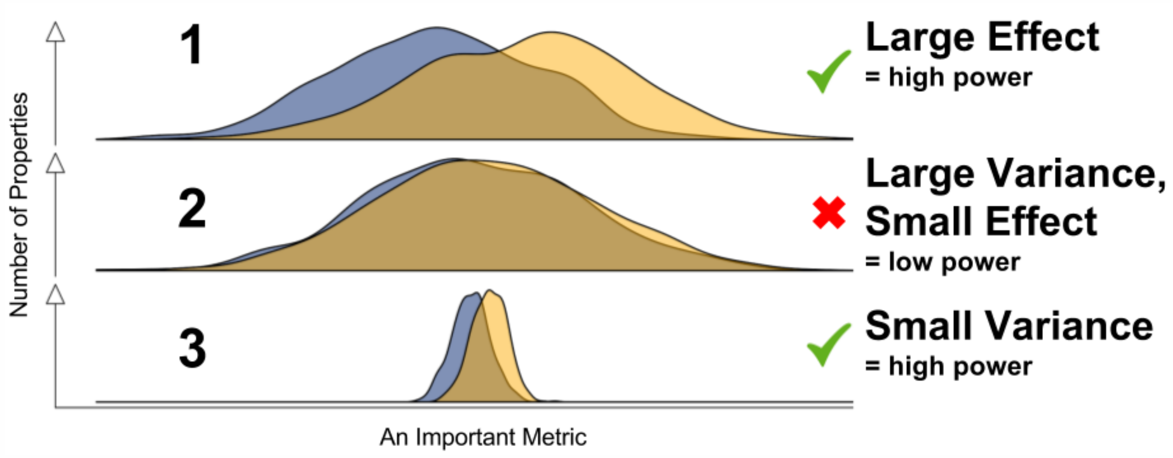


https://www.nber.org/WNE/lect_10_diffindiffs.pdf

CUPED

Следующий метод называется Controlled experiment Using Pre-Experiment Data (CUPED) и может рассматриваться как обобщение DiD.

Позволяет увеличить мощность оценки ATE.



CUPED

Как работает CUPED?

Во-первых, зададим параметр $\theta = \dfrac{cov(y_1, y_0)}{var(y_0)}$.

Далее, оценка CUPED может быть записана следующим образом:

$$\boxed{\tau_{CUPED} = y_1 - (y_0 - \overline{y}_0)\theta}$$

Так как CUPED увеличивает мощность теста, то в результате мы можем получить значимый результат раньше и с меньшей выборкой, что удешевляет эксперимент.

https://www.exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf

https://booking.ai/how-booking-com-increases-the-power-of-online-experiments-with-cuped-995d186fff1d

ANCOVA

Теперь рассмотрим подход под названием Analysis of Covariation.

Он также используется для оценки ATE в RCT.

Выглядит она следующим образом:

Пусть y_{iT} – у нас таргет как и прежде.

Тогда для простейшего вида ANCOVA предлагается оценить следующую регрессию:

$$\boxed{y_{i1} = \alpha + \beta y_{i0} + \tau I(W_i = 1) + \epsilon_{i1}}$$

Отсюда - $\hat{\tau}$ будет оценкой ANCOVA для ATE.

ANCOVA

Теперь рассмотрим подход под названием Analysis of Covariation.

Он также используется для оценки ATE в RCT.

Выглядит она следующим образом:

Пусть y_{iT} – у нас таргет как и прежде.

Тогда для простейшего вида ANCOVA предлагается оценить следующую регрессию:

$$\boxed{y_{i1} = \alpha + \beta y_{i0} + \tau I(W_i = 1) + \epsilon_{i1}}$$

Отсюда - $\hat{\tau}$ будет оценкой ANCOVA для ATE.

Вместо y_{i0} можем использовать произвольные бейзлайн переменные X_{i0} , которые позволяют с достаточной точностью предсказывать y_{i1} при сохранении независимости воздействия от них.

ANCOVA

Существуют модификации для повышения эффективности ANCOVA.

В частности, есть непараметрическая ANCOVA, которая позволяет использовать произвольный ML алгоритм для оценки ATE.

ANCOVA

Существуют модификации для повышения эффективности ANCOVA.

В частности, есть непараметрическая ANCOVA, которая позволяет использовать произвольный ML алгоритм для оценки ATE.

- Строим наивную оценку ATE по модели, например $y_{i1} = \alpha + \delta I(W_i = 1) + \epsilon_{i1}$, что является бейзлайном для оценки $\hat{y}_{i1}(W)$.

ANCOVA

Существуют модификации для повышения эффективности ANCOVA.

В частности, есть непараметрическая ANCOVA, которая позволяет использовать произвольный ML алгоритм для оценки ATE.

- Строим наивную оценку ATE по модели, например $y_{i1} = \alpha + \delta I(W_i = 1) + \epsilon_{i1}$, что является бейзлайном для оценки $\hat{y}_{i1}(W)$.
- Строим для каждого объекта $\delta = \mathbb{E}[y_{i1}(1)] - \mathbb{E}[y_{i1}(0)]$ по первой модели.

ANCOVA

Существуют модификации для повышения эффективности ANCOVA.

В частности, есть непараметрическая ANCOVA, которая позволяет использовать произвольный ML алгоритм для оценки ATE.

- Строим наивную оценку ATE по модели, например $y_{i1} = \alpha + \delta I(W_i = 1) + \epsilon_{i1}$, что является бейзлайном для оценки $\hat{y}_{i1}(W)$.
- Строим для каждого объекта $\delta = \mathbb{E}[y_{i1}(1)] - \mathbb{E}[y_{i1}(0)]$ по первой модели.
- Находим $\hat{m}(y_{i1}, W_i, \theta) = y_{i1} - \hat{\alpha} - \hat{\delta} I(W_i = 1)$ для каждого объекта.

ANCOVA

Существуют модификации для повышения эффективности ANCOVA.

В частности, есть непараметрическая ANCOVA, которая позволяет использовать произвольный ML алгоритм для оценки ATE.

- Строим наивную оценку ATE по модели, например $y_{i1} = \alpha + \delta I(W_i = 1) + \epsilon_{i1}$, что является бейзлайном для оценки $\hat{y}_{i1}(W)$.
- Строим для каждого объекта $\delta = \mathbb{E}[y_{i1}(1)] - \mathbb{E}[y_{i1}(0)]$ по первой модели.
- Находим $\hat{m}(y_{i1}, W_i, \theta) = y_{i1} - \hat{\alpha} - \hat{\delta} I(W_i = 1)$ для каждого объекта.
- Строим произвольные модели $\hat{m}(y_{i1}, W_i, \theta) \sim a_W(X_i)$, где в качестве X_i можем использовать любые бейзлайновые значения и характеристики объектов, включая pre-treatment уровни целевой переменной

ANCOVA

Решаем:

$$\hat{\tau} = arg\min_{\tau}\left\{\mathbb{E}\left[m(y_{i1}, W_i, \tau) - \sum_{j=0}^{|W|}(I(W_j = 1) - P_j)\hat{a}_j(X_i)\right]^2\right\},$$

где $m(y_{i1}, W_i, \tau) = y_{i1} - \tau_0 - \tau_1 I(W_i = 1)$ и P – доля наблюдений в группе.

ANCOVA

Решаем:

$$\hat{\tau} = arg\min_{\tau}\left\{\mathbb{E}\left[m(y_{i1}, W_i, \tau) - \sum_{j=0}^{|W|}(I(W_j = 1) - P_j)\hat{a}_j(X_i)\right]^2\right\},$$

где $m(y_{i1}, W_i, \tau) = y_{i1} - \tau_0 - \tau_1 I(W_i = 1)$ и P – доля наблюдений в группе.

Это можно представить как OLS:

$$\boxed{y_{i1} - \sum_{j=0}^{|W|}(I(W_j = 1) - P_j)\hat{a}_j(X_i) = \tau_0 + \tau_1 I(W_i = 1) + \epsilon_{i1}}$$

Propensity Score

Часто случается, что нам необходимо оценить эффект от воздействия, тогда как провести рандомизированный эксперимент невозможно. Независимости воздействия нет.

Предположим, проводилась кампания и по каким-то, неизвестным нам критериям, выделили целевую группу и на ней провели.

Пусть, как и прежде, мы заинтересованы в оценке ATE. Идеи?

Propensity Score

Часто случается, что нам необходимо оценить эффект от воздействия, тогда как провести рандомизированный эксперимент невозможно. Независимости воздействия нет.

Предположим, проводилась кампания и по каким-то, неизвестным нам критериям, выделили целевую группу и на ней провели.

Пусть, как и прежде, мы заинтересованы в оценке ATE. Идеи?

Вспомним, его оценка – это:

$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{Y}_i(1) - \hat{Y}_i(0) \right)$$

И, как мы знаем, вся проблема в том как мы будем оценивать игреки для counterfactual.

В случае простых статистических тестов и разницы средних оценка будет смещена.

Propensity Score

Для коррекции смещения используют Propensity Score (PS).

$$e(X) = P(W = 1|X) = E[W|X]$$

Propensity Score

Для коррекции смещения используют Propensity Score (PS).

$$e(X) = P(W = 1|X) = E[W|X]$$

Предположения, которые необходимо сделать для его использования:

- Unconfoudedness.
- Overlap

Тогда $W \perp\!\!\!\perp (Y(0), Y(1)) \,|e(X)$.

Propensity Score

Мы получили PS для каждого объекта, что теперь будем делать?

Propensity Score

Мы получили PS для каждого объекта, что теперь будем делать?

Существуют несколько подходов для оценки ATE с PS:

- Blocking: разбиваем выборку на блоки по PS, затем оцениваем naïve ATE внутри каждого из блоков, затем – взвешенно суммируем.

Propensity Score

Мы получили PS для каждого объекта, что теперь будем делать?

Существуют несколько подходов для оценки ATE с PS:

- Blocking: разбиваем выборку на блоки по PS, затем оцениваем naïve ATE внутри каждого из блоков, затем – взвешенно суммируем.
- Weighting: $E\left[\frac{(W-e(X))Y}{e(X)(1-e(X))}\right]$

Propensity Score

Мы получили PS для каждого объекта, что теперь будем делать?

Существуют несколько подходов для оценки ATE с PS:

- Blocking: разбиваем выборку на блоки по PS, затем оцениваем naïve ATE внутри каждого из блоков, затем – взвешенно суммируем.
- Weighting: $E\left[\frac{(W - e(X))Y}{e(X)(1 - e(X))}\right]$
- Также мы можем представить это все в виде регрессии $Y_i = \alpha + \tau I(W_i = 1) + \epsilon_i$, где каждому наблюдению из treatment будут сопоставлены веса $\frac{1}{e(X_i)}$, а из control $\frac{1}{(1 - e(X_i))}$.

Propensity Score

Мы получили PS для каждого объекта, что теперь будем делать?

Существуют несколько подходов для оценки ATE с PS:

- Blocking: разбиваем выборку на блоки по PS, затем оцениваем naïve ATE внутри каждого из блоков, затем – взвешенно суммируем.
- Weighting: $E\left[\frac{(W - e(X))Y}{e(X)(1 - e(X))}\right]$
- Также мы можем представить это все в виде регрессии $Y_i = \alpha + \tau I(W_i = 1) + \epsilon_i$, где каждому наблюдению из treatment будут сопоставлены веса $\frac{1}{e(X_i)}$, а из control $\frac{1}{(1 - e(X_i))}$.
- Для учета ковариатов, можем модифицировать регрессию из пред. пункта в

$$Y_i = \alpha + \beta X_i + \tau I(W_i = 1) + \epsilon_i$$

Помимо этого, можно сделать взвешенную регрессию для каждого из блоков в blocking, а затем оценки усреднить как в п.1

Regression Discontinuity

Также случается, что некоторое воздействие на объекты произведено в прошлом без разделения на контрольную и целевую группы, но мы хотим понять каков же ATE.

Здесь нам поможет Regression Discontinuity Design (RDD).

Regression Discontinuity

Также случается, что некоторое воздействие на объекты произведено в прошлом без разделения на контрольную и целевую группы, но мы хотим понять каков же ATE.

Здесь нам поможет Regression Discontinuity Design (RDD).

- Пусть мы имеем гетерогенную выборку, на часть из которой произведено воздействие. Мы понимаем, что воздействие произведено на основе одной из переменных (пример – в армию по ЕГЭ, оценка влияния армии).

Regression Discontinuity

Также случается, что некоторое воздействие на объекты произведено в прошлом без разделения на контрольную и целевую группы, но мы хотим понять каков же ATE.

Здесь нам поможет Regression Discontinuity Design (RDD).

- Пусть мы имеем гетерогенную выборку, на часть из которой произведено воздействие. Мы понимаем, что воздействие произведено на основе одной из переменных (пример – в армию по ЕГЭ, оценка влияния армии).
- Однако, есть множество неучтенных факторов, информации о которых у нас нет.

Regression Discontinuity

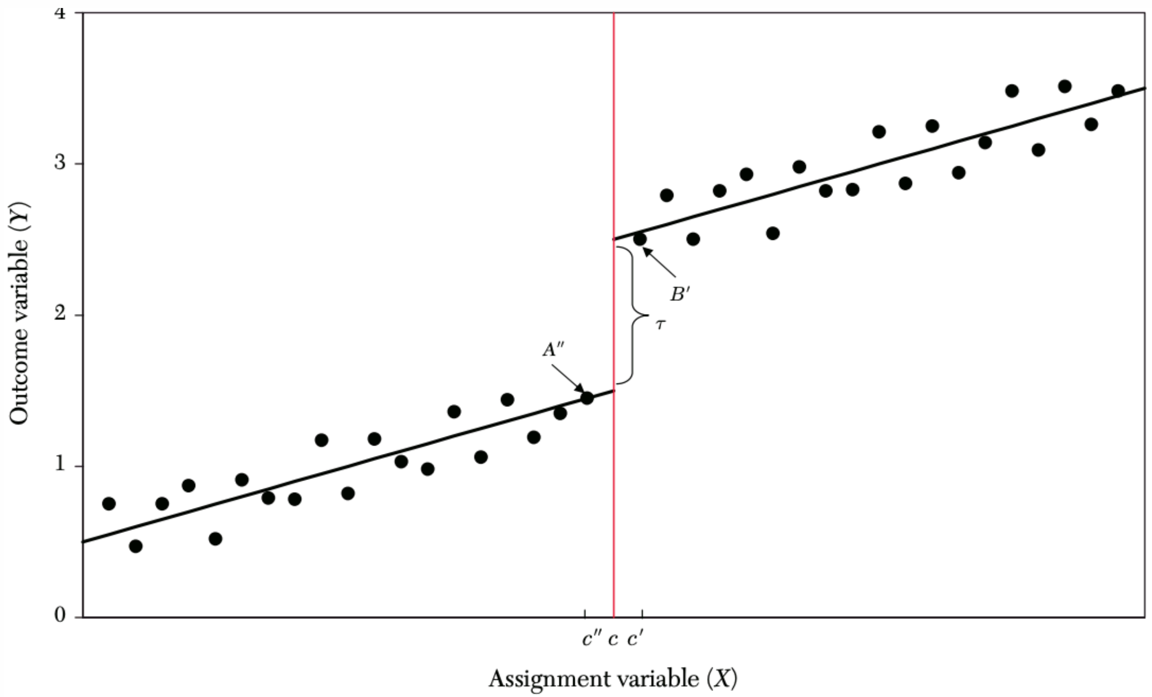
Также случается, что некоторое воздействие на объекты произведено в прошлом без разделения на контрольную и целевую группы, но мы хотим понять каков же ATE.

Здесь нам поможет Regression Discontinuity Design (RDD).

- Пусть мы имеем гетерогенную выборку, на часть из которой произведено воздействие. Мы понимаем, что воздействие произведено на основе одной из переменных (пример – в армию по ЕГЭ, оценка влияния армии).
- Однако, есть множество неучтенных факторов, информации о которых у нас нет.
- Берем людей в окрестности порога.

Regression Discontinuity

Простейший RD Design:



Требование - Y непрерывна по X.

https://www.princeton.edu/~davidlee/wp/RDDEconomics.pdf

Regression Discontinuity

Определим теперь RD модель строго:

$$y_i = \alpha + \beta X_i + \tau I(x_i > c) + \epsilon_i,$$

где c – порог.

Regression Discontinuity

Определим теперь RD модель строго:

$$y_i = \alpha + \beta X_i + \tau I(x_i > c) + \epsilon_i,$$

где с – порог.

Можем ослабить требование к одинаковому наклону:

$$y_i = \alpha_l + \beta_l(x_i - c) + (\beta_r - \beta_l) I(x_i > c)(x_i - c) + \epsilon_i$$

Regression Discontinuity

Определим теперь RD модель строго:

$$y_i = \alpha + \beta X_i + \tau I(x_i > c) + \epsilon_i,$$

где с – порог.

Можем ослабить требование к одинаковому наклону:

$$y_i = \alpha_l + \beta_l(x_i - c) + (\beta_r - \beta_l)I(x_i > c)(x_i - c) + \epsilon_i$$

Если мы опасаемся, что при удалении от с зависимость Y от X может сильно меняться, то можно построить локальную регрессию, ограничившись наблюдениями, удовлетворяющими $x_i \in [c - h, c + h]$ для некоторого h.

Regression Discontinuity

Существует модификация RD, где в качестве целевого регрессора используется время. Regression Discontinuity in Time (RDiT).

Аналогично:

$$y_i = \alpha + \tau I(t > T) + \epsilon_i,$$

Regression Discontinuity

Существует модификация RD, где в качестве целевого регрессора используется время. Regression Discontinuity in Time (RDiT).

Аналогично:

$$y_i = \alpha + \tau I(t > T) + \epsilon_i,$$

Пример – проблема перетоков.

Regression Discontinuity

Существует модификация RD, где в качестве целевого регрессора используется время. Regression Discontinuity in Time (RDiT).

Аналогично:

$$y_i = \alpha + \tau I(t > T) + \epsilon_i,$$

Пример – проблема перетоков.

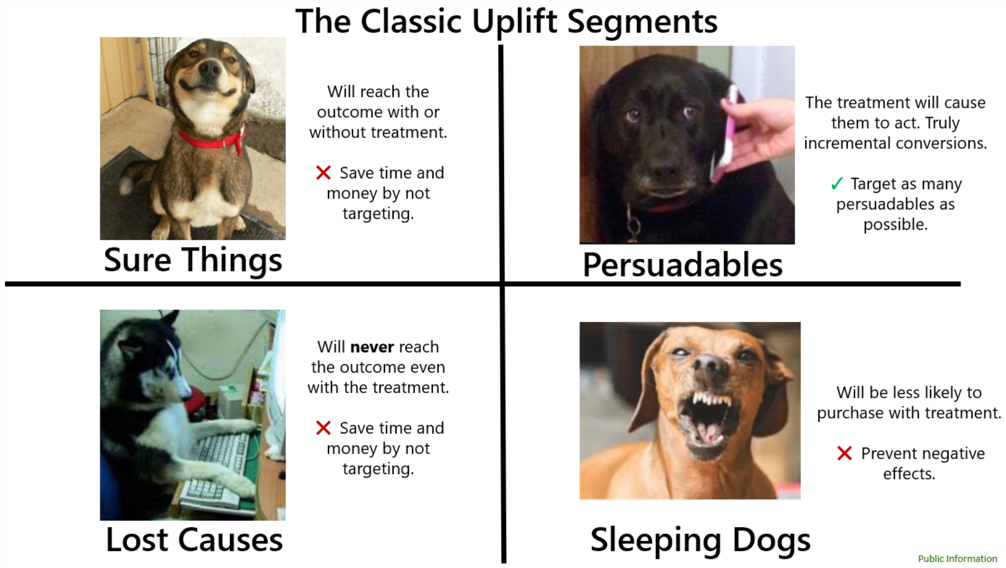
- Важно очистить ряд от сезонности и тренда и быть уверенными, что одновременно с воздействием не происходило иных структурных изсменений.
- Во всех регрессиях возможно добавление полиномиальной формы и иные игры со спецификациями.

Uplift моделирование

В данное время становится популярным термин uplift моделирование.

По сути, это другое название для Conditional ATE или Heterogenous ATE.

До этого мы сталкивались с оценкой ATE на всю выборку или популяцию, здесь же нас интересует эффект на индивидуальный объект, так как предполагается, что для разных объектов эффект разный.



https://towardsdatascience.com/a-quick-uplift-modeling-introduction-6e14de32bfe0

Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:
   - Оцениваем $\mu(x) = \mathbb{E}[Y|X = x, W]$
   - $\boxed{\hat{\tau}_{CATE}(x) = \hat{\mu}(x, W = 1) - \hat{\mu}(x, W = 0).}$

Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:
   - Оцениваем $\mu(x) = \mathbb{E}[Y|X = x, W]$
   - $\boxed{\hat{\tau}_{CATE}(x) = \hat{\mu}(x, W = 1) - \hat{\mu}(x, W = 0).}$
2. T-learner:
   - Оцениваем $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ и $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$
   - $\boxed{\hat{\tau}_{CATE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)}$

Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:
   - Оцениваем $\mu(x) = \mathbb{E}[Y|X = x, W]$
   - $\boxed{\hat{\tau}_{CATE}(x) = \hat{\mu}(x, W = 1) - \hat{\mu}(x, W = 0).}$
2. T-learner:
   - Оцениваем $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ и $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$
   - $\boxed{\hat{\tau}_{CATE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)}$
3. X-learner:
   - Оцениваем $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ и $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$
   - Вычисляем $D_i^1 = Y_i^1 - \hat{\mu}_0(x_i^1)$ и $D_i^0 = \hat{\mu}_1(x_i^0) - Y_i^0$, затем оцениваем $\tau_1(x) = \mathbb{E}[D^1|X = x]$ и $\tau_0(x) = \mathbb{E}[D^0|X = x]$
   - $\boxed{\tau_{CATE} = e(x)\tau_0(x) + (1 - e(x))\tau_1(x)}$

Современные методы

Основное развитие в данной сфере происходит в области непараметрических моделей и адаптации методик выше под ML, так как это позволяет увеличить эффективность оценок.

В качестве примера приведу подход для оценки CATE с применением Random Forest.

Он основывается на подходе Double ML, который часто применяется в новых методологиях оценки причинно-следственных связей.

В частности, этот подход широко применим для uplift моделирования.

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

1. Выберем случайно s наблюдений и разделим их на две равные части I и J.

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

1. Выберем случайно s наблюдений и разделим их на две равные части I и J.
2. Выберем minimum leaf size for group = k – минимальное число наблюдений из группы в листе.

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

1. Выберем случайно s наблюдений и разделим их на две равные части I и J.
2. Выберем minimum leaf size for group = k – минимальное число наблюдений из группы в листе.
3. Строим лес из «честных” деревьев:
   - Построим модель на J следующего вида: $Y_i \sim tree(X_i, W_i)$

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

1. Выберем случайно s наблюдений и разделим их на две равные части I и J.
2. Выберем minimum leaf size for group = k – минимальное число наблюдений из группы в листе.
3. Строим лес из «честных” деревьев:
   - Построим модель на J следующего вида: $Y_i \sim tree(X_i, W_i)$
   - Для I – предсказываем $\hat{Y}_i = t\hat{r}ee(X_i, W_i)$ по дереву, построенному на предыдущем шаге

Double Sample Trees

Пусть у нас есть n наблюдений вида (Y_i, X_i, W_i).

1. Выберем случайно s наблюдений и разделим их на две равные части I и J.
2. Выберем minimum leaf size for group = k – минимальное число наблюдений из группы в листе.
3. Строим лес из «честных” деревьев:
   - Построим модель на J следующего вида: $Y_i \sim tree(X_i, W_i)$
   - Для I – предсказываем $\hat{Y}_i = \hat{tree}(X_i, W_i)$ по дереву, построенному на предыдущем шаге
   - Оцениваем $\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i$ и

$$\hat{\tau}(x) = \frac{1}{|i : W_i = 1, X_i \in L|} \sum_{|i:W_i=1,X_i \in L|} Y_i - \frac{1}{|i : W_i = 0, X_i \in L|} \sum_{|i:W_i=0,X_i \in L|} Y_i$$

Double Sample Trees

Для всех деревьев в лесе усредняем:

$$\hat{RF}(x) \approx \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b(x)$$

Double Sample Trees

Можно показать, что данная оценка является несмещенной оценкой гетерогенного эффекта воздействия, а также произвести её асимптотическую оценку:

$$\sqrt{n}(\hat{RF}(x) - \tau_{CATE}(x)) \sim N(0, V_{IJ})$$

Double Sample Trees

Можно показать, что данная оценка является несмещенной оценкой гетерогенного эффекта воздействия, а также произвести её асимптотическую оценку:

$$\sqrt{n}(\hat{RF}(x) - \tau_{CATE}(x)) \sim N(0, V_{IJ})$$

Здесь

$$\hat{V}_{IJ}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} cov_*[\hat{\tau}_b^*(x), N_{ib}^*]^2$$

где $N_{ib}^* = 1$, если i-й объект появляется либо в I либо в J сэмпле.

Полезные библиотеки

1. CausalML – библиотека от Uber для оценки CATE и ITE.
2. EconML – библиотека от Microsoft для оценки CATE.
3. Dowhy – библиотека от Microsoft со множеством инструментов для causal analysis.
4. ExpAn – библиотека для простых статистических тестов.
5. Rdd – библиотека для RDD.