



Проверка гипотез. A/B тесты

Лекция 10

Иван Горбань

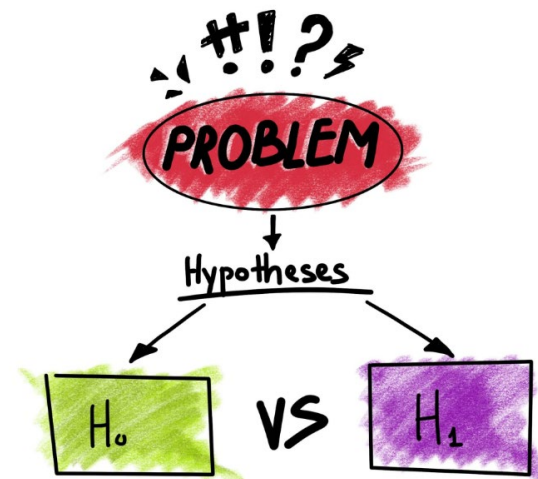


План занятия:

1. Проверка гипотез
2. Рандомизированный эксперимент
3. Перед АВ 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

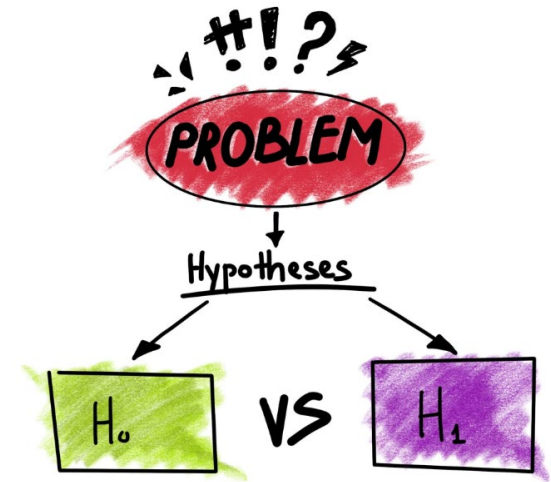
Проверка гипотез

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



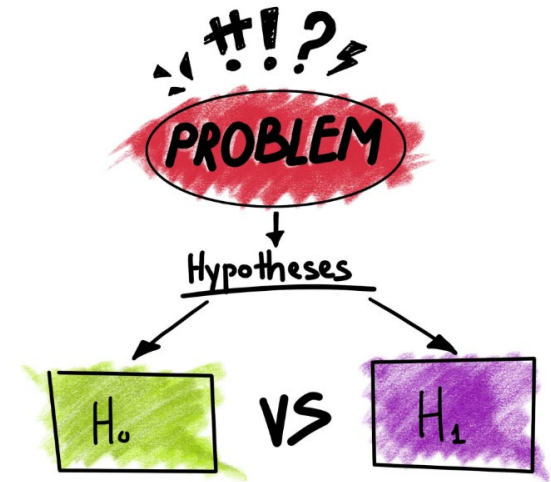
Проверка гипотез

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



Проверка гипотез

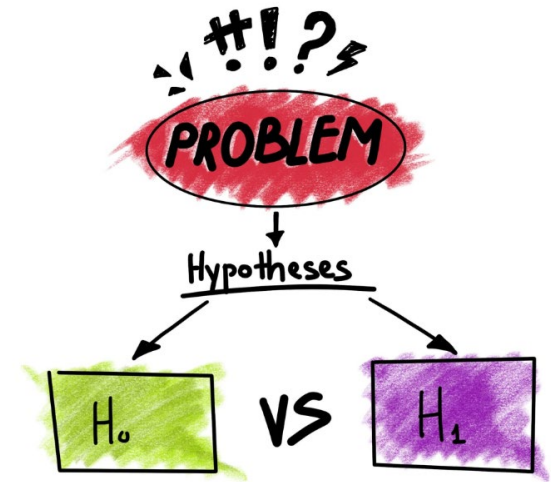
- Первое что необходимо – **сама гипотеза**.
Будем проверять **нулевую гипотезу** H_0 против **альтернативы** H_1 .
- Второе – **выборка** реализаций случайной величины $\{x_1, \dots, x_n\}$.
- Третье, что нам понадобится – **статистика критерия**.
Для различных гипотез используются различные статистики. 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 $\{x_1, \dots, x_n\}$.
- 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:

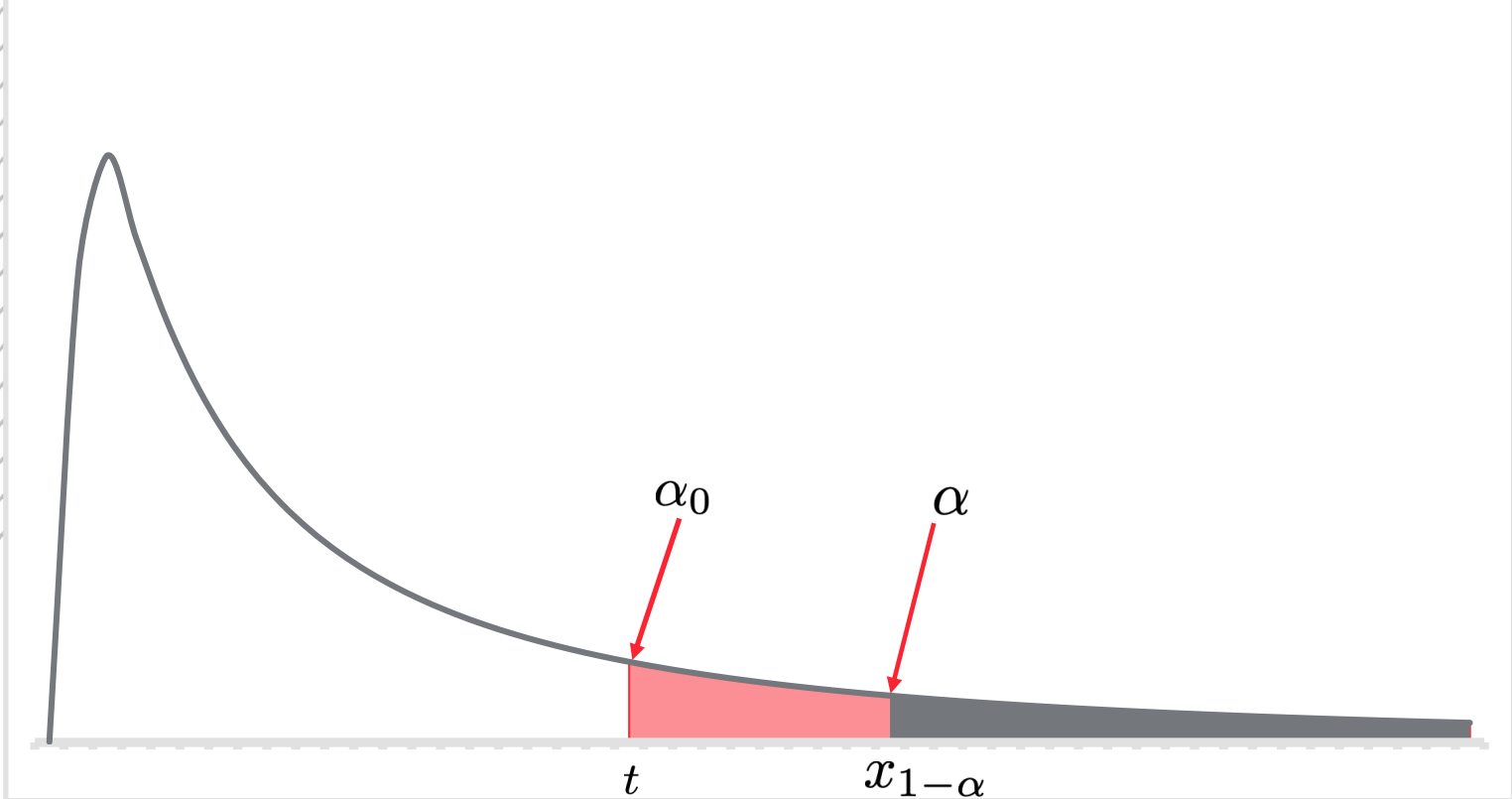
$$P(T(x_1, \dots, x_n) \geq x_{1-\alpha}) \leq \alpha,$$

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

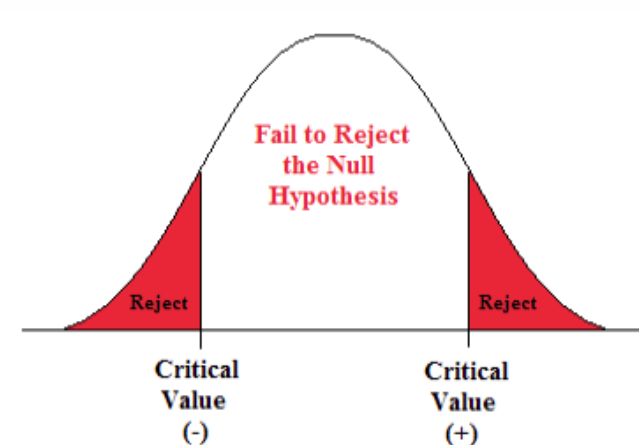
Statistics distribution



Пусть на выборке мы получили некоторое значение t для статистики.

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

Также принято название **p-value**.



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.

		PREDICTIVE VALUES	
		POSITIVE (CAT)	NEGATIVE (DOG)
ACTUAL VALUES	POSITIVE (CAT)	<div>TRUE POSITIVE</div> <div>3</div> <div>YOU ARE A CAT</div>	<div>FALSE NEGATIVE</div> <div>1</div> <div>YOU ARE A DOG</div> <div>TYPE II ERROR</div>
	NEGATIVE (DOG)	<div>FALSE POSITIVE</div> <div>2</div> <div>YOU ARE A CAT</div> <div>TYPE I ERROR</div>	<div>TRUE NEGATIVE</div> <div>4</div> <div>YOU ARE NOT A CAT</div>

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:

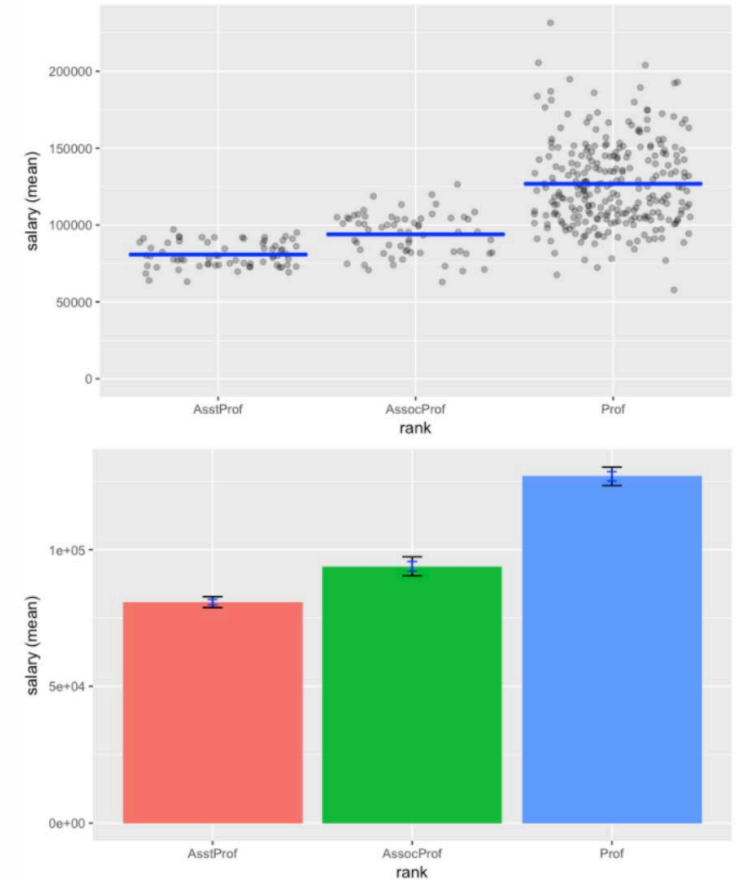
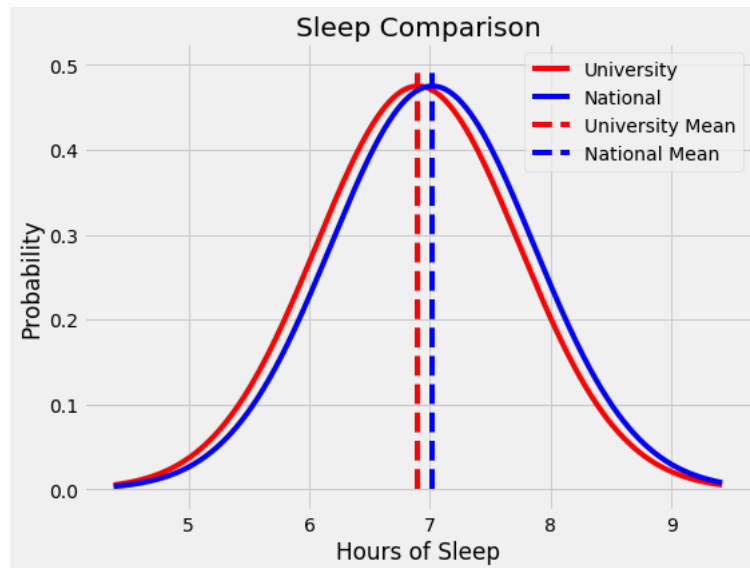
$$Power = P(\overline{H_0}|H_1) = 1 - P(H_0|H_1)$$

Если $Power \rightarrow 1$ при $n \rightarrow \infty$ 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://radiant-rstats.github.io/docs/basics/compare_means.html

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

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:

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

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:

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

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

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:

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

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

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

Проверка гипотез

Проверка на нормальность:

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

Гипотезы: $H_0 : X \sim N(\mu, \sigma^2)$, $H_1 : \bar{H}_0$.

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

Статистика теста:

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

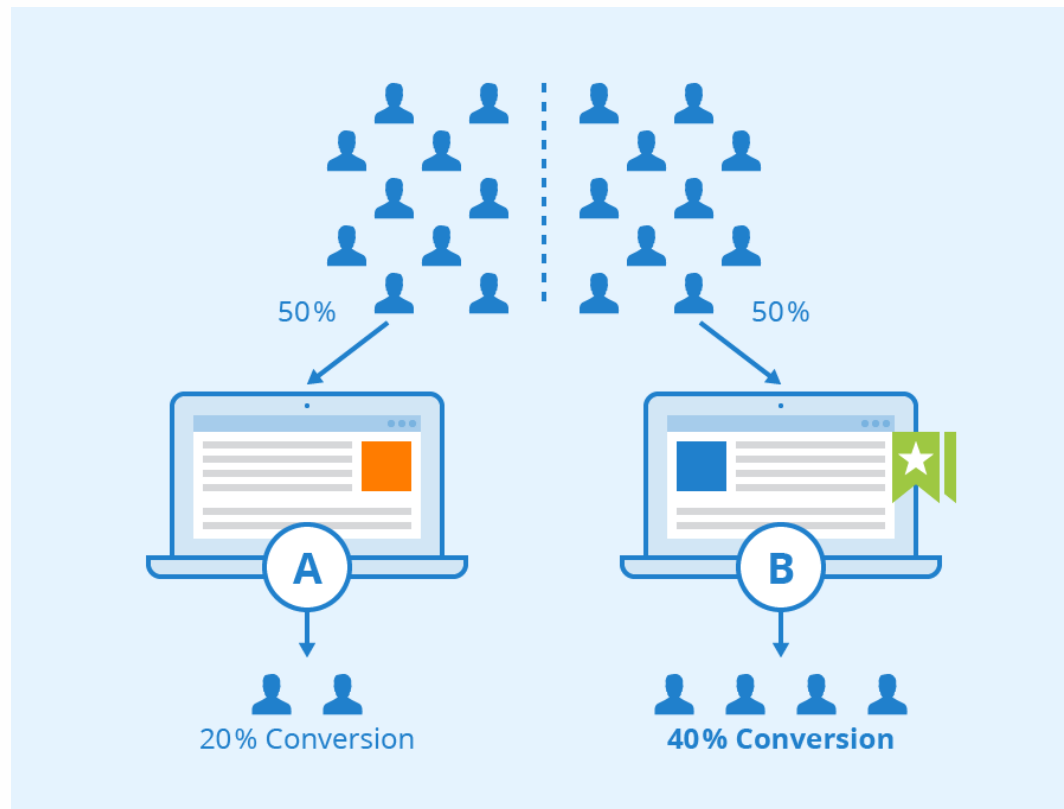
где n_i - число наблюдений в $[q_i, q_{i+1}]$. $p_i = F_{N(\mu, \sigma^2)}(q_{i+1}) - F_{N(\mu, \sigma^2)}(q_i)$

При верной нулевой гипотезе $\chi^2 \sim \chi_{K-1}^2$ или $\chi^2 \sim \chi_{K-3}^2$, 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:

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

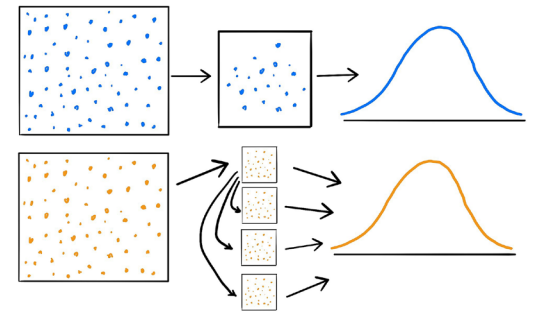
где f - элементы, которые равны 1 в первой выборке и 0 - во второй,
 g - наоборот.

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

Проверка гипотез

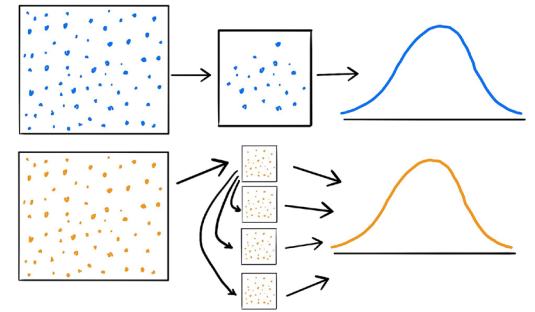
Непараметрические критерии – в Long Version

Bootstrap



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

Bootstrap

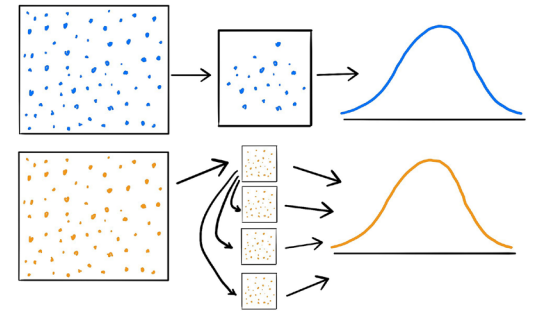


Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

Пусть у нас есть выборка из n элементов.

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

Bootstrap



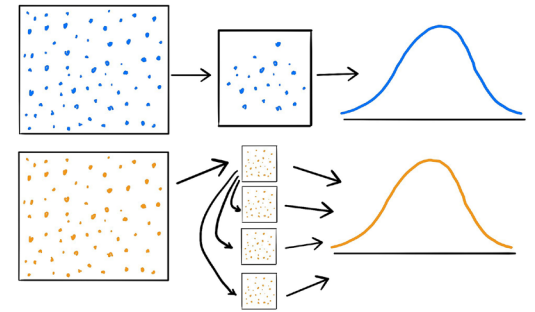
Бутстрапирование позволяет нам получить распределение произвольной статистики, которое невозможно или сложно рассчитать аналитически (например, квантили).

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.

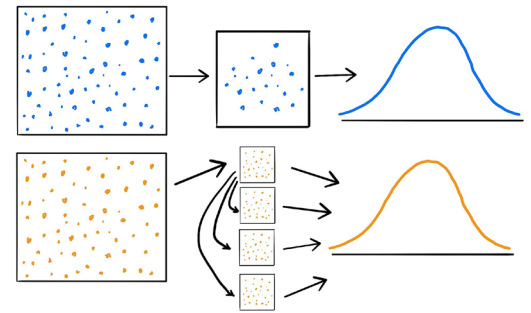
Генерируем B бутстрапированных выборок (сэмплируем с возвращением выборки размера n)

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

Подробнее – в long version





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:

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

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:

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

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

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:

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

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

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:

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

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:

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

В противном же случае:

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



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).

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



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:

- Требуется 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:

- Требуется 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 удастся, 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:

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

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

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

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

Оценка эффекта

Прежде чем рассматривать методы оценки ТЕ, сделаем следующие предположения:

Unconfoundedness:

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

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 ТЕ, сделаем следующие предположения:

Unconfoundedness:

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

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):

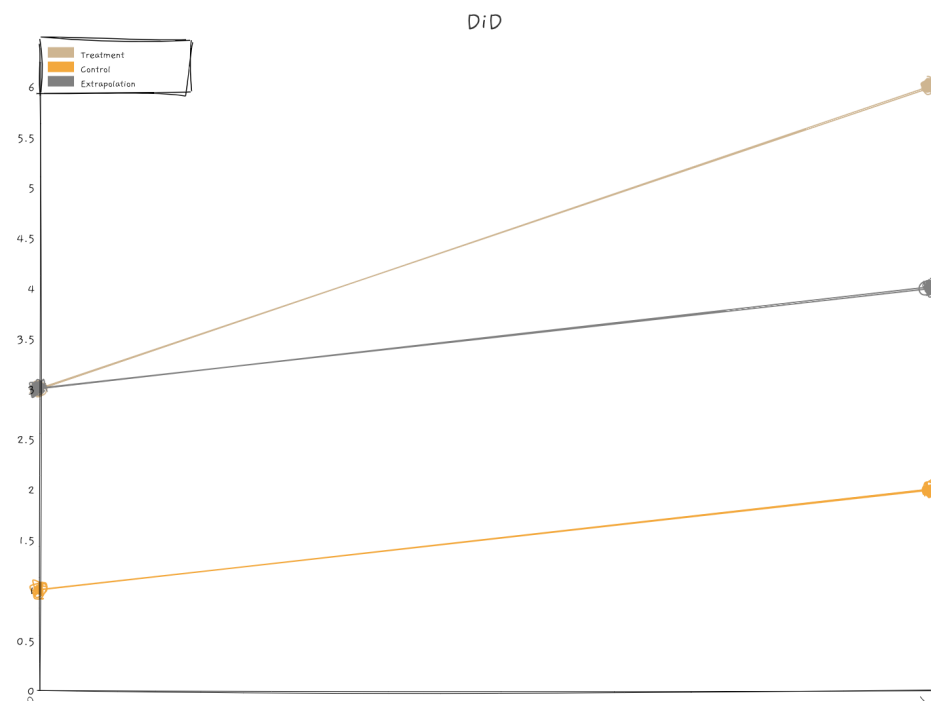
$$Y_i = \alpha + \beta X_i + \tau W_i + \epsilon_i$$

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

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?

Difference in Differences

Как это оценить?

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

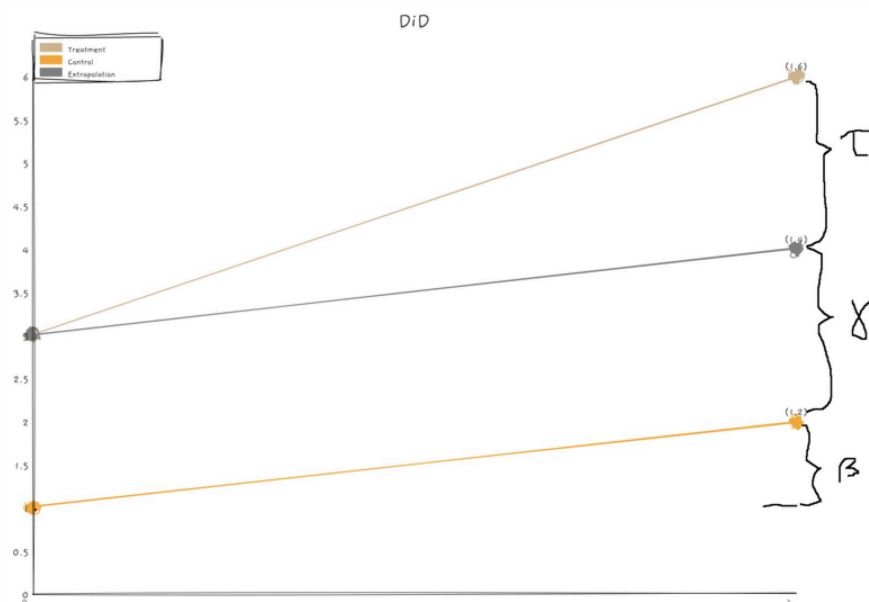
Difference in Differences

Как это оценить?

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

Оценка производится с помощью регрессии, модель выглядит следующим образом:

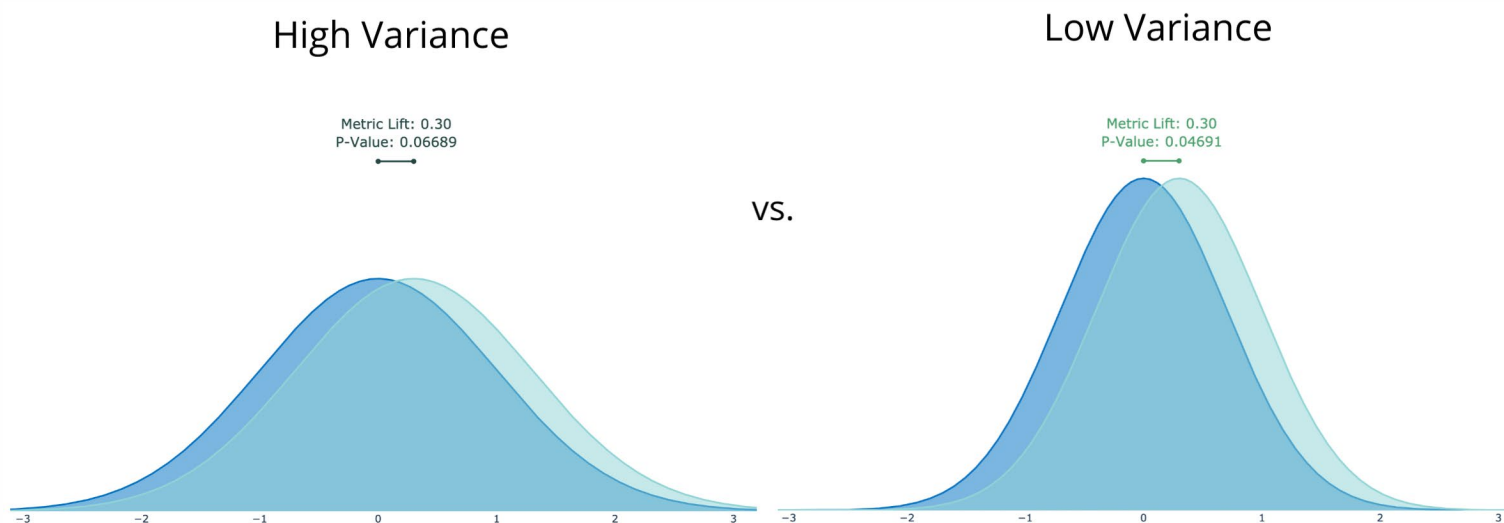
$$y_{iT} = \alpha + \beta I(T = 1) + \gamma I(W_i = 1) + \tau I(T = 1)I(W_i = 1) + \epsilon$$



CUPED

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

Позволяет увеличить **МОЩНОСТЬ** оценки ATE.



<https://towardsdatascience.com/how-to-double-a-b-testing-speed-with-cuped-f80460825a90>

CUPED

Как работает CUPED?

- Определяем целевую переменную y_1 для измерения
- Выбираем пред-экспериментальный ковариат (например значение до эксперимента y_0)
- Задаем параметр: $\theta = \frac{cov(y_1, y_0)}{var(y_0)}$
- Вычисляем CUPED-скорректированную метрику: $\hat{y}_1 = \bar{y}_1 - \theta \cdot \bar{y}_0$
- Вычисляем стандартное отклонение этой метрики: $var(\hat{y}_1) = var(y_1) \cdot (1 - corr(y_0, y_1))$

Далее, оценка CUPED может быть записана следующим образом:

$$\tau_{CUPED} = \hat{y}_1^T - \hat{y}_1^C$$

CUPED

Для проверки значимости ATE $\tau_{CUPED} = \hat{y}_1^T - \hat{y}_1^C$ используют t-test для независимых выборок с неизвестной и различной дисперсией.

Есть небольшая хитрость. Если присмотреться, CUPED-корректированная целевая метрика:

$$\hat{y}_{1i} = y_{1i} - \theta y_{0i}$$

подозрительно похожа на результат преобразования OLS регрессии центрированных:

$$y_{1i}^{centered} \sim \theta y_{0i}^{centered}$$

А значит и дисперсию легко получить как: $var(\hat{y}_{1i}) = var(y_{1i} - \theta y_{0i})$

$$= var(y_{1i}) + \theta^2 var(y_{0i}) - 2\theta cov(y_{0i}, y_{1i}) = var(y_{1i}) + \left(\frac{cov(y_{0i}, y_{1i})}{var(y_{0i})} \right)^2 var(y_{0i}) - 2\theta cov(y_{0i}, y_{1i})$$

$$= var(y_{1i}) - \frac{cov(y_{0i}, y_{1i})^2}{var(y_{0i})} = \boxed{var(y_{1i})(1 - \rho^2)}$$

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:

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

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:

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

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.

Смотрите в Long Version.

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 **АТЕ**. Идеи?

Вспомним, его оценка – это:

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

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:

- Unconfoundedness.
- 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 деления 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

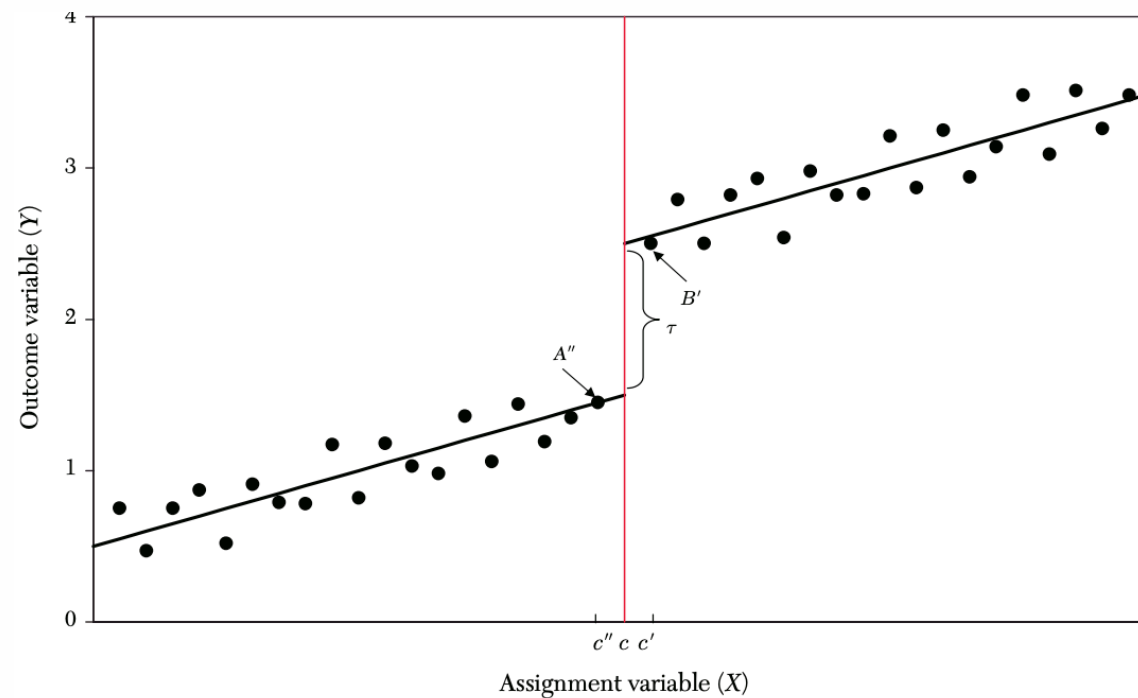
Также случается, что некоторое воздействие на объекты произведено в прошлом без деления на контрольную и целевую группы, но мы хотим понять каков же ATE.

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 c – 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 c – 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 c 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.

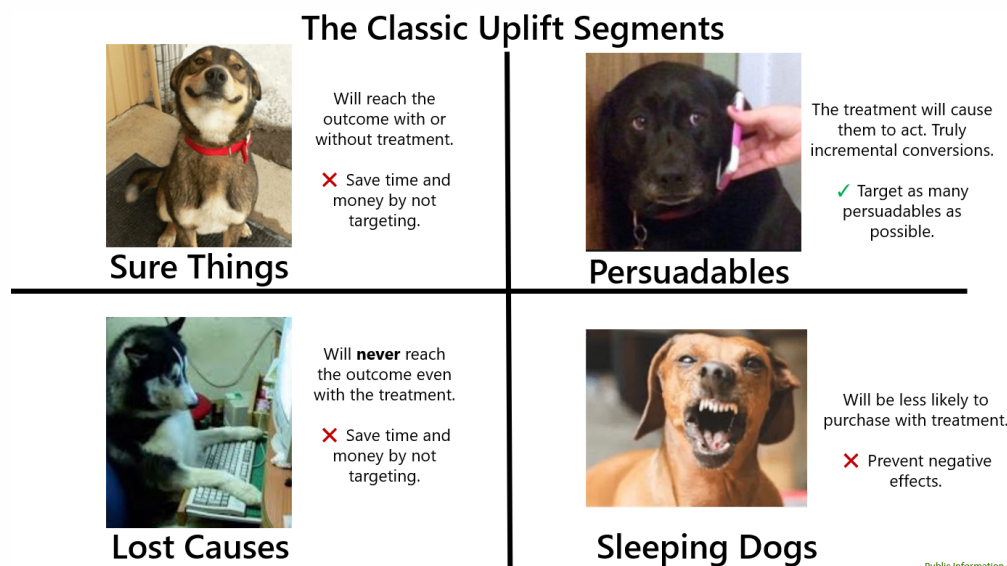
- Важно очистить ряд от сезонности и тренда и быть уверенными, что одновременно с воздействием не происходило иных структурных изменений.
- Во всех регрессиях возможно добавление полиномиальной формы и иные игры со спецификациями.

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.



Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:

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

Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:

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

2. T-learner:

- Оцениваем $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ и $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$
- $\hat{\tau}_{CATE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

Uplift моделирование

Существует множество оценок CATE. Начнем с простых:

1. S-learner:

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

2. T-learner:

- Оцениваем $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ и $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$
- $\hat{\tau}_{CATE}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

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 $\tau_0(x) = \mathbb{E}[D^0|X = x]$
- $\tau_{CATE} = e(x)\tau_0(x) + (1 - e(x))\tau_1(x)$



Современные методы

Приходите на ML-club и узнаете 😊

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.