

Прикладная статистика в машинном обучении

Семинар 7

A/B тестирование

И. К. Козлов
(Мехмат МГУ)

2022

A/A-тестирование

Q: Зачем нужно A/A-тестирование?

Пользователям в двух бакетах показывают тот же продукт.



A/A Testing

Рис.: Если нет разницы ...

A/A-тестирование

А: A/A-тестирование позволяет проверить, что сэмплирование правильно работает.

Заодно это может дать бейзлайн — насколько могут варьироваться метрики.

Shoot yourself in the foot

Обсудим теперь пару способов “выстрелить себе в ногу”.



Парадокс Беркsona

1 Парадокс Беркsona

2 Парадокс Симпсона

- Парадокс Симпсона в АБ

3 Тест Стьюдента

4 Критерий Манна-Уитни

Курение спасает жизнь?

Следующий пример взят из

Why most studies into COVID19 risk factors may be producing flawed conclusions - and how to fix the problem

В больницах (медперсонал и больных) тестировали на COVID-19.

Было установлено:

Курение уменьшает вероятность заболеть COVID-19!

Где же разгадка?

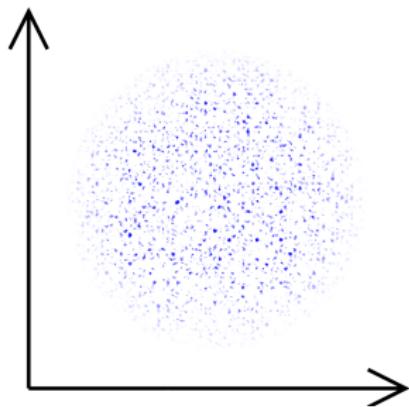
Q: В чём ошибка? Где подвох?



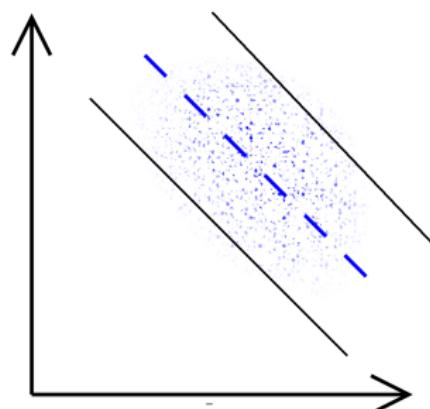
Рис.: Дело одной трубки

Где же разгадка?

A: Выборка не случайная — ничего не знаем про тех, кто не попадал в больницу.



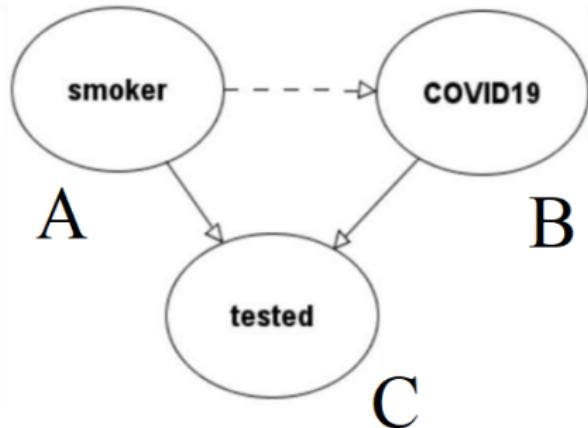
На всей выборке
зависимости НЕТ



При ограничениях
зависимость есть

Рис.: Парадокс Беркsonа

Убославливание на коллайдер



Вообще 2 независимых события A и B могут становиться зависимыми, если произошло некоторое третье событие C .

Событие C , зависящее и от A , и от B называют [коллайдером](#).

Наглядный пример



- События “друга похитили инопланетяне” и “мои часы спешат” явно независимы.
- Друг опаздывает.
Смотрим — часы спешат \Rightarrow вероятность похищения инопланетянами меньше.
События стали зависимы.

Мораль

Мораль:

Выборка должна быть случайной.

Где же разгадка?

Этот парадокс в жизни повсюду — в каждой “ошибке выжившего”.



Парадокс Симпсона

1 Парадокс Берксона

2 Парадокс Симпсона

- Парадокс Симпсона в АБ

3 Тест Стьюдента

4 Критерий Манна-Уитни

Парадокс Симпсона

Парадокс Симпсона

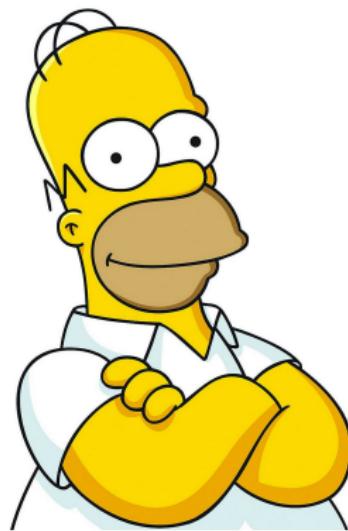


Рис.: Не тот Симпсон

Парadox Симпсона

Данные о поступлении на факультеты одного американского ВУЗа.
Скандал! Дискриминация!

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%
Total	8442	4321		

Q: Можно ли сказать, что мужчин явно дискриминируют?

Парадокс Симпсона

По всему ВУЗу целиком — динамика обратная! Парадокс!

Female

35% ADMITTED

4321 APPLICANTS

Male

44% ADMITTED

8442 APPLICANTS

Парадокс Симпсона

За парадоксом скрывается **простая арифметика**.

Может быть так, что

$$\frac{a_1}{b_1} < \frac{a_2}{b_2}, \quad \frac{c_1}{d_1} < \frac{c_2}{d_2},$$

но при этом

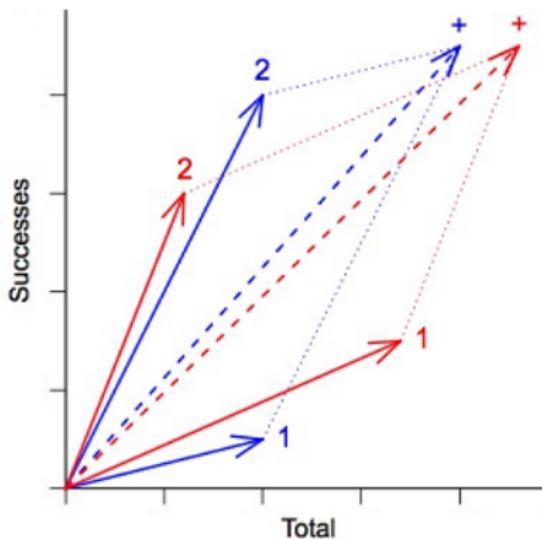
$$\frac{a_1 + c_1}{b_1 + d_1} > \frac{a_2 + c_2}{b_2 + d_2}.$$

Наглядно — на следующем слайде.

Парадокс Симпсона

Геометрическое объяснение.

Отношение $\frac{p}{q}$ — это коэффициент наклона вектора (q, p) .

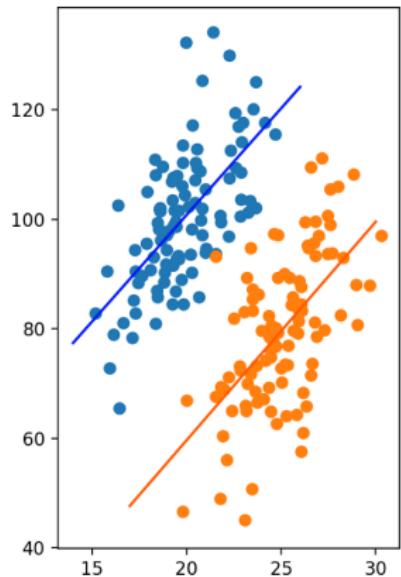


Красные векторы 1,2 менее наклонены, чем синие.

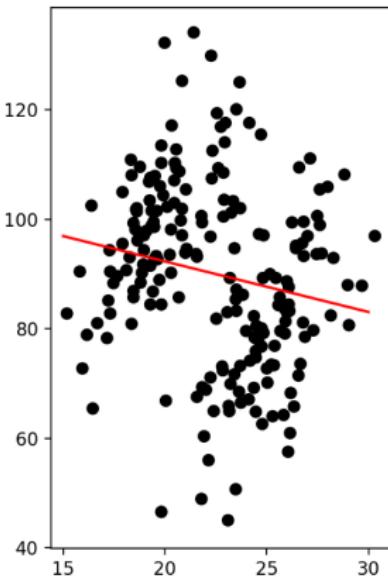
А сумма красных "+" — более пологая, чем сумма синих.

Парадокс Симпсона

Парадокс Симпсона часто изображают подобными картинками:



Позитивные тренды
для обоих групп



Негативный тренд
для всей популяции

Парadox Симпсона

На практике. Сэмплирование в группах может быть перекошено.

Разные размеры тестовой и контрольной группы — зло!

Treated



$$\Pr[A = 1 | X = \text{adult, male}] = 1/5$$

$$w_i = 5$$

Untreated



$$\Pr[A = 0 | X = \text{adult, male}] = 4/5$$

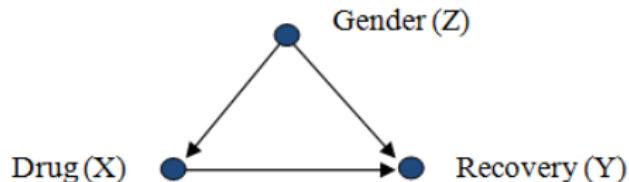
$$w_i = 5/4$$

Пример: мужчинам меньше выписывают лекарство.

Тогда наблюдаемый эффект может больше быть с связан с “бытием мужчиной”.

Группы разного размера следует перезвешивать.

Спутывающие переменные



Если Z влияет на X и Y , то Z — спутывающая переменная (конфаундер).

Парадокс Симпсона возникает из-за наличия таких переменных.

Спутывающие переменные

Что делать, если мы хотим доказать влияние X на Y ?

Нужно избавляться от спутывающих переменных.

Можно делать выборки более однородными (по возрасту, полу, региону и т.д.)

Парадокс Симпсона в АБ

1 Парадокс Беркsona

2 Парадокс Симпсона

- Парадокс Симпсона в АБ

3 Тест Стьюдента

4 Критерий Манна-Уитни

Одинаковый размер

Чему научил нас парадокс Симпсона?

Пилот и контроль разного размера — плохая идея.

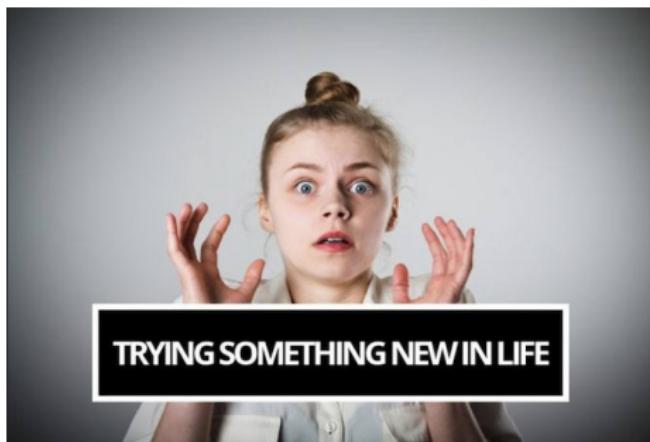
Постепенный запуск

Пилотная группа — это перемены сервиса.

Q: Цена эксперимента.

Мы не боимся испугать пользователей и потерять деньги?

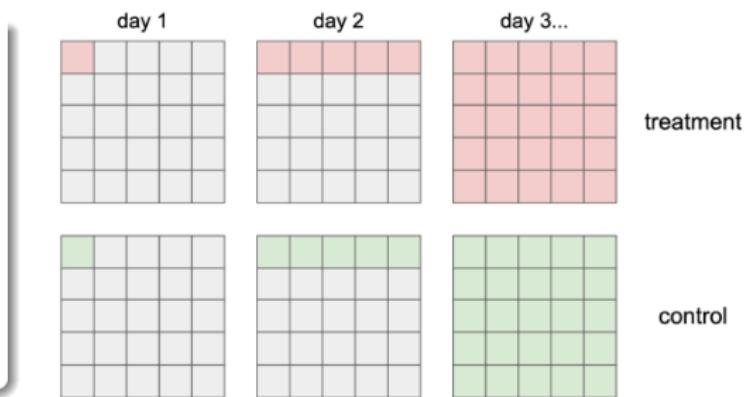
Стоит ли ставить эксперимент на 50% пользователей?



Постепенный запуск

Постепенное внедрение

- В первый день раскатываемся на 1% пользователей;
- Если всё в порядке, на следующий день раскатываемся на 10%;
- Если всё нормально, раскатываемся на всех пользователей.



Парadox Симпсона

		день 1	день 2	всего
A	Пользователи	1000	1000	2000
	Конверсии	400	100	500
	%	40%	10%	25%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

		день 1	день 2	всего
A	Пользователи	100	1000	1100
	Конверсии	40	100	140
	%	40%	10%	12.7%
B	Пользователи	100	1000	1100
	Конверсии	50	200	250
	%	50%	20%	23.7%

- Нужно внимательно работать со случаем, когда размеры групп меняются.
- Контрольная и экспериментальная группы должны быть одного размера!

Перерыв

Перерыв

Тест Стьюдента

1 Парадокс Беркsona

2 Парадокс Симпсона

- Парадокс Симпсона в АБ

3 Тест Стьюдента

4 Критерий Манна-Уитни

Тест Стъюдента

Тест Студентов)

Популярные у студентов напитки?

Тест Стъюдента



Рис.: Тест Стъюдента

Уильям Сили Госсет

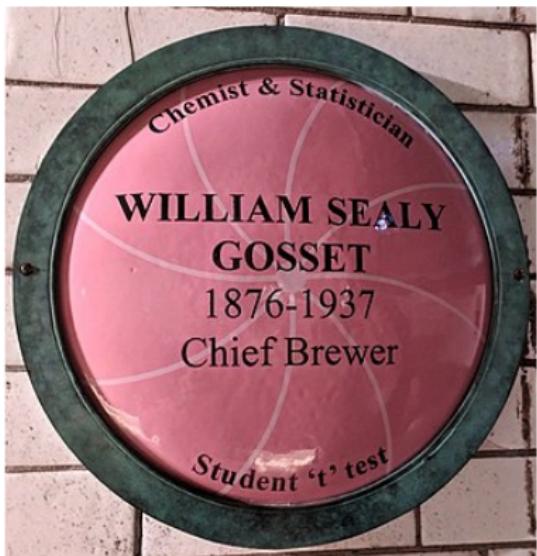


Рис.: Уильям Госсет. Был “под NDA”

Распределение Стьюдента

Ключевая идея. Как появляется распределение Стьюдента?

- Пусть X_1, \dots, X_n — i.i.d. из $\mathcal{N}(\mu, \sigma^2)$.
- Выборочные среднее и несмешённая дисперсия:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Распределение Стьюдента

Ключевая идея. Как появляется распределение Стьюдента?

- Пусть X_1, \dots, X_n — i.i.d. из $\mathcal{N}(\mu, \sigma^2)$.
- Выборочные среднее и несмешённая дисперсия:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Тогда $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ имеет стандартное нормальное распределение.

Распределение Стьюдента

Ключевая идея. Как появляется распределение Стьюдента?

- Пусть X_1, \dots, X_n — i.i.d. из $\mathcal{N}(\mu, \sigma^2)$.
- Выборочные среднее и несмешённая дисперсия:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Тогда $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ имеет стандартное нормальное распределение.
- А величина, которую можно посчитать на практике (зная μ)

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

имеет распределение Стьюдента с $n - 1$ степенью свободы.

Распределение Стьюдента

Неожиданный факт! В формуле

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

числитель и знаменатель независимы!

Распределение Стьюдента

Неожиданный факт! В формуле

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

числитель и знаменатель независимы!

Доказательство.

- Числитель — функция от \bar{X} , а знаменатель — от $X_i - \bar{X}$.

Распределение Стьюдента

Неожиданный факт! В формуле

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

числитель и знаменатель независимы!

Доказательство.

- Числитель — функция от \bar{X} , а знаменатель — от $X_i - \bar{X}$.
- \bar{X} и $X_i - \bar{X}$ — линейные комбинации X_1, \dots, X_n , они нормально распределены.

Распределение Стьюдента

Неожиданный факт! В формуле

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

числитель и знаменатель независимы!

Доказательство.

- Числитель — функция от \bar{X} , а знаменатель — от $X_i - \bar{X}$.
- \bar{X} и $X_i - \bar{X}$ — линейные комбинации X_1, \dots, X_n , они нормально распределены.
- $\text{cov}(\bar{X}, X_i - \bar{X}) = 0$, значит \bar{X} и $X_i - \bar{X}$ (а следовательно и числитель с знаменателем) независимы.

Распределение Стьюдента

Чуть более общий вид случайно величины с **t-распределением Стьюдента с ν степенями свободы**:

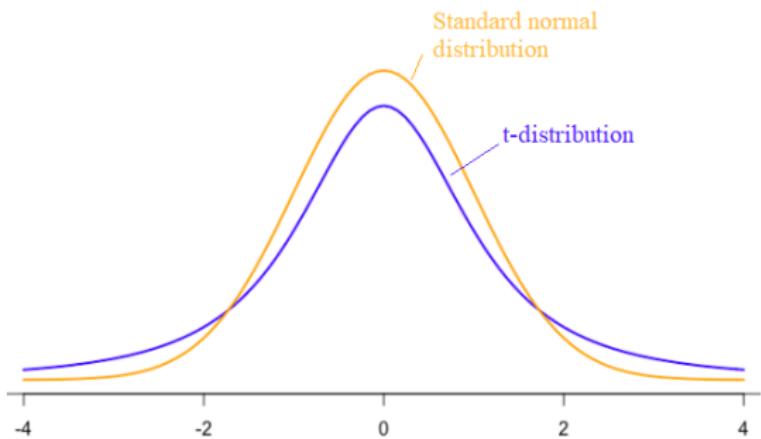
$$T = \frac{Z}{\sqrt{V/\nu}} = Z \sqrt{\frac{\nu}{V}},$$

- Z — это стандартное нормальное ($\mu = 0, \sigma = 1$),
- V имеет χ^2 -распределение с ν степенями свободы,
- Z и V независимы.

Распределение Стьюдента

Распределение Стьюдента по форме похоже на нормальное, но имеет более тяжёлые хвосты.

Распределение Стьюдента с 1 степенью свободы = распределение Коши.



Одновыборочный t-критерий

Есть ряд известных тестов Стьюдента про матожидания нормальных величин.

Одновыборочный t-критерий.

Даны X_1, \dots, X_n — i.i.d. из $\mathcal{N}(\mu, \sigma^2)$ (не знаем параметры).

Q: Как составить тест $H_0 : \mathbb{E}X = \mu_0$?

Одновыборочный t-критерий

Есть ряд известных тестов Стьюдента про матожидания нормальных величин.

Одновыборочный t-критерий.

Даны X_1, \dots, X_n — i.i.d. из $\mathcal{N}(\mu, \sigma^2)$ (не знаем параметры).

Q: Как составить тест $H_0 : \mathbb{E}X = \mu_0$?

A: Правильно, использовать статистику

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Двухвыборочный t-критерий

Также есть двухвыборочный t-критерий равенства матожиданий двух выборок

$$H_0 : \mathbb{E}X_1 = \mathbb{E}X_2, \quad H_1 : \mathbb{E}X_1 \neq \mathbb{E}X_2.$$

Формально, тест Стьюдента работает, если

- Выборки одного размера $n_1 = n_2$.
- Дисперсии равны $\sigma_1 = \sigma_2$.
- Выборки независимы.

Двувыборочный t -критерий

На практике эти условия на тест Стьюдента не выполняются.

Точного решения нет (Behrens–Fisher problem).

В качестве приближения используют t -тест Уэлча — небольшое обобщение теста Стьюдента.

Warning!

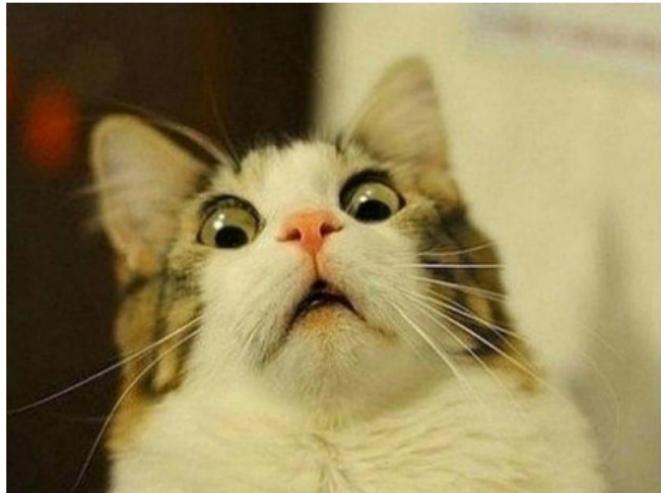


Рис.: Осторожно! Грядут страшные формулы!

Welch's t-test

t-тест Стьюдента Уэлча

- t-статистика:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{se_1^2 + se_2^2}}$$

Пока не страшно)

Welch's t-test

t-тест Стьюдента Уэлча

- t-статистика:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{se_1^2 + se_2^2}}$$

(Пока не страшно)

- Количество степеней свободы

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Информация собрана воеди на следующем слайде.

Welch's t-test

	one-tailed test		two-tailed test
hypothesis	$H_0: \mu_1 \geq \mu_2$ $H_1: \mu_1 < \mu_2$	$H_0: \mu_1 \leq \mu_2$ $H_1: \mu_1 > \mu_2$	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$
test statistic (t distribution)	$t = (\bar{x}_1 - \bar{x}_2) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$		
deg. of freedom	$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\frac{[s_1^2]}{n_1}^2 + \frac{[s_2^2]}{n_2}^2}$		round df down to the nearest integer number
rejection	reject H_0 if $t < -t_\alpha$	reject H_0 if $t > t_\alpha$	reject H_0 if $ t > t_{\alpha/2}$

Критерий Манна-Уитни

1 Парадокс Беркsonа

2 Парадокс Симпсона

- Парадокс Симпсона в АБ

3 Тест Стьюдента

4 Критерий Манна-Уитни

Критерий Манна-Уитни

U-критерий Манна-Уитни.

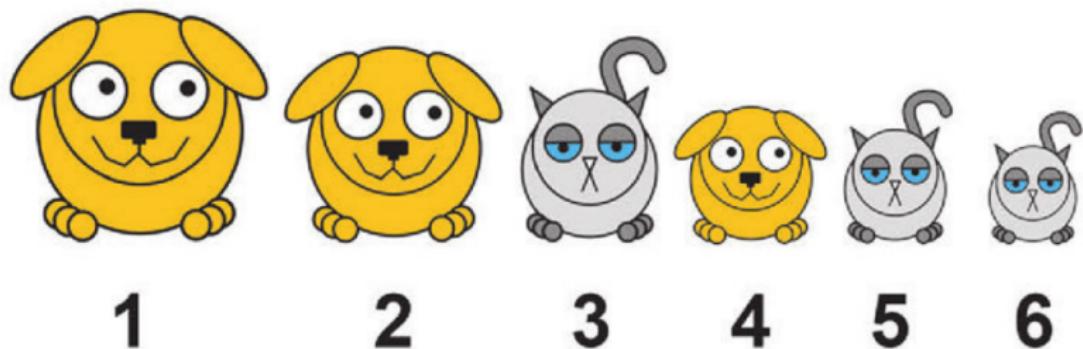
Объясняем на котиках)

Равны ли у двух выборок средние значения?



Критерий Манна-Уитни

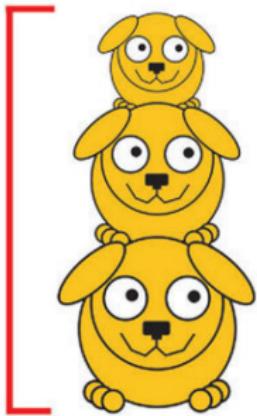
Шаг 1. Упорядочиваем выборки.



Критерий Манна-Уитни

Шаг 2. Считаем суммы рангов R_1 и R_2 .

Сумма рангов 1



Сумма рангов 2



$$1 + 2 + 4 = 7 \quad 3 + 5 + 6 = 14$$

Критерий Манна-Уитни

Шаг 3. Считаем:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}, \quad U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

Берём $U = \min(U_1, U_2)$.

Должно быть $U_1 + U_2 = n_1 n_2$. Если нет — ☺.

Критерий Манна-Уитни

Шаг 4. Сравниваем U с табличным значением.

Замечание. При достаточно больших сэмплах ($n_1, n_2 \geq 20$)
 U близко к нормальному распределению.

Критерий Манна-Уитни

Формально, статистику Манна-Уитни U можно определить как

$$U = \sum_{i=1}^n \sum_{j=1}^m S(X_i, Y_j),, \quad S(X, Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, . \\ 0, & \text{if } X < Y. \end{cases}$$

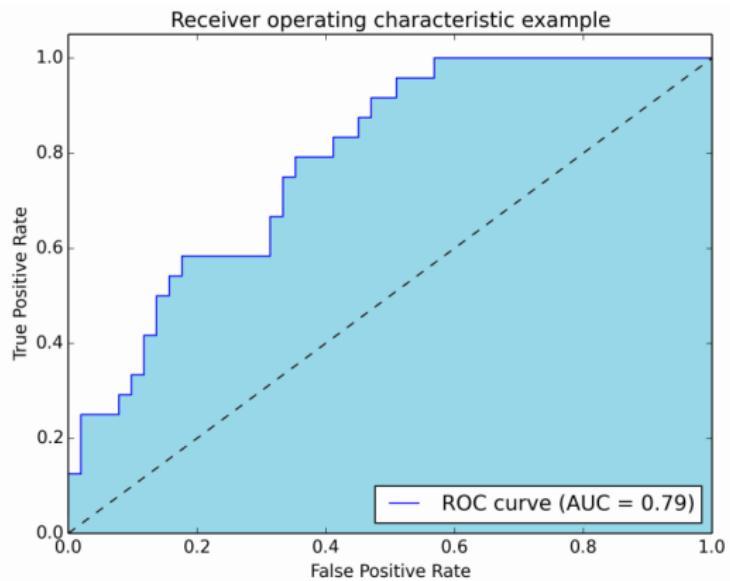
Критерий Манна-Уитни

Q: Где в ML мы сталкивались с долей перестановок для 2ух классовой классификации?

Критерий Манна-Уитни

Q: Где в ML мы сталкивались с долей перестановок для 2ух классовой классификации?

A: Площадь под ROC-кривой (ROC AUC).



ROC AUC

- Если быть точным:

$$\text{AUC} = \frac{U}{n_1 n_2},$$

где U — Манна-Уитни.

- Про ROC AUC можно прочитать в

<https://dyakonov.org/2017/07/28/аuc-roc-площадь-под-кривой-ошибок/comment-page-1/>