Непараметрические критерии

T_{obs} vs T_{cr} $\qquad \hat{\theta} - \theta_0$

Проверка гипотезы – смотрим на расстояния и их распределения

Другое

Параметрические	Непараметрические
Гипотезы были сформ. в терминах параметров (мат. ож.) + Предпосылки	Хочется ослабить предп. Хочется размышлять в терминах без парам. р.-ей
Z-тест	Тест Манна-Уитни

① Критерий знаков

Мэр: квартиры больше половины стоят дешевле 1 млн. ₽

Способ 1: $x_1, \dots, x_n$ $\qquad \bar{x} \overset{asy}{\underset{H_0}{\sim}} \mathcal{N}(\dots)$

Способ 2: ⊕, ⊖, ⊕ …

$H_0: p \overset{=}{\geq} 1/2$ честный мэр

$H_A: p < 1/2$ врёт

$$T = \sum_{i=1}^{n} X_i \underset{H_0}{\sim} Bin(1/2; n)$$

⊖ – дороже 1 млн.

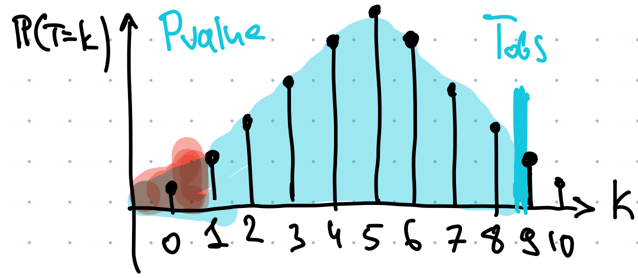
⊕ – дешевле

T_{cr}, α, T_{obs} => выбор

9 ⊕

1 ⊖

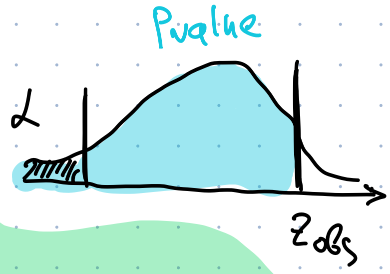
$T_{obs} = \underline{9}$



$\alpha = 0.01$

$Pvalue = 1\cdot\left(\frac{1}{2^{10}} + \frac{10}{2^{10}}\right)\cdot 2 \ (\approx 0{,}0107) > \alpha$

H_0 не отвергается



Мы теряем часть информации
Но зато можем ослабить предпосылки
(например, наплевать на выбросы)

$X_1, \ldots, X_n \sim iid$

$H_0: med(X) = m_0$

$H_A: med(X) \neq m_0$, $>$, $<$

Игнорируем величину изменений
Важно только направление

Тест знаков для одной выборки

$$T = \sum_{i=1}^{n} [X_i > m_0] \underset{H_0}{\sim} Bin\left(\tfrac{1}{2}; n\right)$$

$$\mathbb{E}(T) = \frac{n}{2}$$

$$Var(T) = \frac{n}{4}$$

$$\frac{T - \frac{n}{2}}{\sqrt{n/4}} \overset{asy}{\underset{H_0}{\sim}} N(0,1)$$

Приближение

Две зависимые выборки

$X_1, \dots, X_{n_x} \sim iid$

$Y_1, \dots, Y_{n_y} \sim iid$

выборки связанные

$Z_i = Y_i - X_i$

$H_0: med(Z_i) = 0$

$H_A: \ne 0$

$$T = \sum_{i=1}^{n} \underbrace{[Y_i > X_i]}_{Z_i > 0} \underset{H_0}{\sim} Bin\left(\frac{1}{2}; n\right)$$

А если выборки независимые?

$$U = \sum_{i,j} [Y_j > X_i]$$

Не будет никакого р.-я. Bin !!

n_Y

n_X

$n_Y \cdot n_X$ пар

② Ранговые критерии

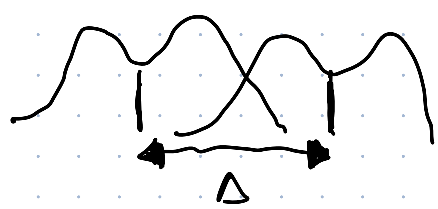
Тест Манна-Уитни

$X_1, \dots, X_{n_x} \sim iid$

$Y_1, \dots, Y_{n_y} \sim iid$

выборки нез.

РАСПРЕДЕЛЕНИЕ одинаковое по форме



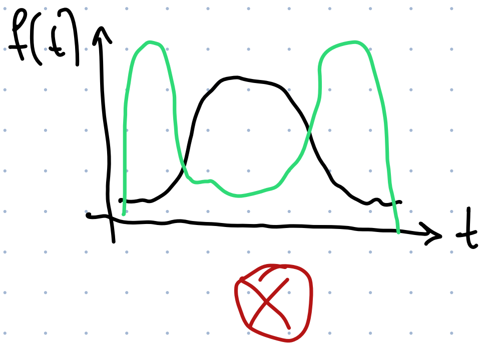
$H_0: f_X(t) = f_Y(t)$

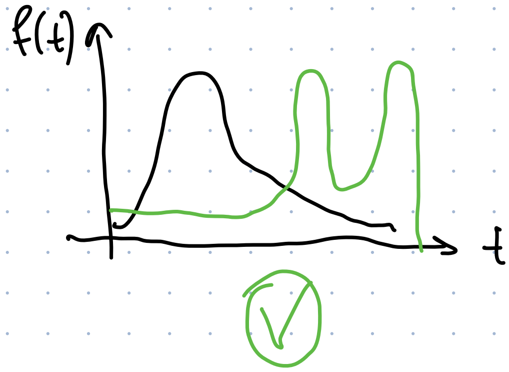
$H_A: f_X(t) = f_Y(t + \Delta)$

$\Delta \ne 0$

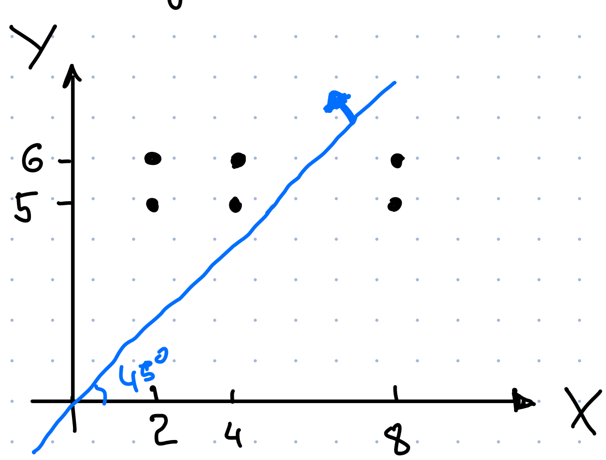
миф: Тест MW про медианы

Если наше воздействие поляризовало людей,
Тест MW его не поймает





$$U = \sum_{i,j} [Y_j > X_i]$$



6 пар

$n_x = 3$ $n_y = 2$

$U_{obs} = 4$

U_{cr} - ?

$$\underbrace{\overbrace{XXX}^{3!}\ \overbrace{YY}^{2!}}_{5!}$$

$$\mathbb{P}(U=6) = \frac{3!\,2!}{5!}$$

XXYXY

$$\mathbb{P}(U=5) = \frac{3!\,2!}{5!}$$

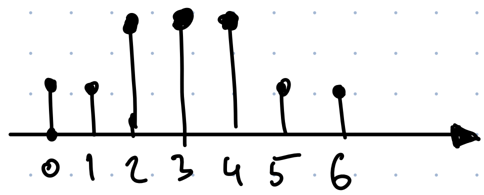
XXYYX

XYXXY

$$\mathbb{P}(U=4) = 2\cdot\frac{3!\,2!}{5!}$$

....

Решаем кучу однотипных комбинаторных задач и получаем р.-е.

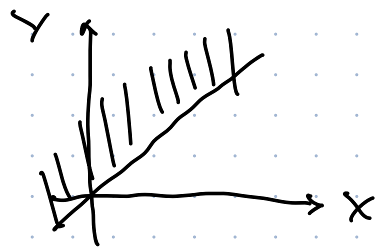


Табличное р.-е.

$\Rightarrow U_{кр.}$

$$\mathbb{E}(U) = \mathbb{E}\Big(\sum_{ij} [Y_j > X_i]\Big) = \sum_{ij} \mathbb{E}\big([Y_j > X_i]\big) =$$

$$= \sum_{ij} \mathbb{P}(Y_j > X_i) = n_x \cdot n_y \cdot \underbrace{1/2}$$



$f_X(x) \cdot f_Y(y) = f(x,y)$

$\int\limits_{////} f(x,y) = 1/2$

$$Var(U) = \frac{n_x \cdot n_y \cdot (n_x + n_y + 1)}{12}$$

Офигенное св-во:

$$\frac{U - \mathbb{E}(U)}{\sqrt{Var(U)}} \xrightarrow[\substack{n_x \to \infty \\ n_y \to \infty}]{d} \mathcal{N}(0,1)$$

ЦПТ

$\boxed{\text{Сходимость будет быстрее, чем для средних}}$

Big Data

$\boxed{\alpha, \beta, \Delta = MDE \Rightarrow n}$

Калькулятор АБ-теста

На самом деле в тесте MW мы превращаем наблюдения в ранги

5	-4	5	3	22	наблюдения
3	1	4	2	5	порядковые номера
3.5	1	3.5	2	5	ранги

U – число преимуществ Y над X

R_Y – сумма рангов набл. из Y в выборке (X, Y)

набл.	кто здесь?	ранг
2	X	1
4	X	2
5	Y	3
6	Y	4
8	X	5

$R_Y = 7$

$R_X = 8$

Ранг = 3
набл. Y обогнало два X

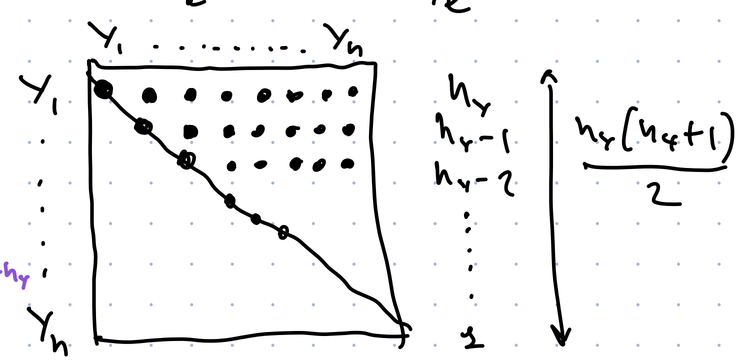
$$U = R_Y - \frac{n_Y \cdot (n_Y + 1)}{2}$$

Сколько раз в R_Y была заложена инф. о том что Y_k обогнал Y_e

Есть ещё третий способ

$$R_Y = \sum_{i=1}^{n_Y} rk(z_i)$$

$$\left.\begin{matrix} Y_1, \dots, Y_{n_Y} \\ X_1, \dots, X_{n_X} \end{matrix}\right\} \begin{matrix} Y_1\, X_1\, X_2\, Y_3, \dots \\ z_1\, z_2\, z_3\, z_4, \dots\, z_{n_X+n_Y} \end{matrix}$$



Найдём $\mathbb{E}(U)$ $Var(U)$

Но верна R_Y сумма n_Y наугад выбранных чисел из $1, 2, 3, 4, \ldots, n_X + n_Y$

$$R_Y = R_1 + R_2 + \ldots + R_n$$

$\boxed{\begin{array}{l} n \text{ чисел взял} \\ N \text{ чисел было} \end{array}}$ $n = n_Y$, $N = n_X + n_Y$

$$\mathbb{E}(R_Y) = n \cdot \mathbb{E}(R_i) =$$

$$= n \cdot \left(1 \cdot \frac{1}{N} + 2 \cdot \frac{1}{N} + \ldots + N \cdot \frac{1}{N}\right) = n \cdot \frac{1}{N} \cdot \frac{N(N+1)}{2}$$

$$\mathbb{E}(U) = n_Y \cdot \frac{n_X + n_Y + 1}{2} - \frac{n_Y(n_Y+1)}{2} = n_Y \cdot n_X \cdot \frac{1}{2}$$

$$Var(U) = Var(R_Y)$$

$$Var(R_Y) = n \cdot Var(R_1) \cdot \frac{N-n}{N-1}$$

$$Var(R_1) = \mathbb{E}(R_1^2) - \mathbb{E}^2(R_1)$$

$$\mathbb{E}(R_1^2) = \frac{1}{N}(1^2 + 2^2 + \ldots + N^2) = \frac{1}{N} \cdot \frac{N(N+1)(2N+1)}{6}$$

$$Var(R_1) = \frac{n_X \cdot n_Y \cdot (n_X + n_Y + 1)}{12}$$

Тест Уилкоксона
Wilcoxon Signed Rank Test

$X_1, X_2, \ldots X_n \sim iid$

$F_X(x)$ – симм. отн. med

$H_0: Med(X) = m_0$
$H_A: Med(X) \neq m_0$

$d_i = X_i - m_0$

T^+ – сумма рангов положит. d_i

Связные выборки

$X_1, \dots, X_n \sim iid$
$Y_1, \dots, Y_n \sim iid$

$d_i = X_i - Y_i$

$H_0: Med(d_i) = 0$
$H_A: Med(d_i) \neq 0$

$T^+ = 1 \cdot I_1 + 2 \cdot I_2 + \dots + n \cdot I_n$

$I_j = 1$ если $d_j > 0$

Все d_j отсортированы

$\mathbb{P}(I_j = 1) = 1/2$

$\mathbb{E}(T^+) = \frac{1}{2} \cdot (1 + 2 + \dots + n) = \frac{n(n+1)}{4}$

$\mathbb{E}(T^{+2}) = \frac{1}{2}(1^2 + 2^2 + \dots + n^2) = \dots$

$Var(T^+) = \frac{n(n+1)(2n+1)}{24}$

Можно исп. нормальную аппрокс.

③ Как Roc-Auc связан с U?

$h_0 \; h_1 \quad p(x_i) = \mathbb{P}(y_i = 1 \mid x_i)$ классиф.

$h_0 \cdot h_1 \quad y_i = 1 \quad p(x_i) = 0.8$
$\qquad\qquad y_i = 0 \quad p(x_i) = 0.9$

Плохая пара

$X_1, \dots, X_{n_x}$ — $P(x_1)$ $P(x_2)$ … $P(x_{n_x})$

$y_1 = 1$ $y_2 = 1$ … $y_{n_x} = 1$

$Y_1, \dots, Y_{n_y}$ — $P(x_{n_x+1}), P(x_{n_x+2})$ … $P(x_{n_x+n_y})$

$y_{n_x+1} = 0$ $y_{n_x+2} = 0$ … $y_{n_x+n_y} = 0$

$$U = \sum_{i,j} [X_i > Y_j]$$

Кол-во пар, которые алгоритм верно отсортировал

$$\text{Roc-Auc} = \frac{U}{\underbrace{n_0 \cdot n_1}_{\text{всего пар}}}$$

Я могу проверить гипотезу:

$H_0: \text{Roc-Auc} = \frac{1}{2}$

$H_A: \text{Roc-Auc} \neq \frac{1}{2}$

4 Перестановочные критерии

Давайте делать всё то же самое, но в ранги наблюдения превращать не будем.

$X_1, \dots, X_{n_x} \sim iid$

$Y_1, \dots, Y_{n_y} \sim iid$

| нез. $f_X(t) = f_Y(t + \Delta)$

n_x n_y

$H_0: \Delta = 0$

$H_A: \Delta \neq 0$

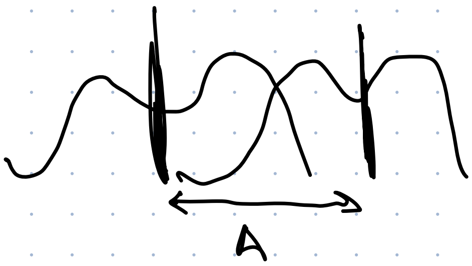
$Z_1, \dots, Z_{n_x} | Z_{n_x+1}, \dots, Z_{n_x+n_y}$

n_x n_y

$$T = \frac{1}{n_x}\sum_{i=1}^{n_x} Z_i - \frac{1}{n_y}\sum_{j=1}^{n_y} Z_j \quad \Big| \quad T_{obs} = \bar{X} - \bar{Y}$$

Табличное р.-е. для T

$C_{n_x+n_y}^{n_x}$



Фишер

Очень похожа на бутстрэп

На больших выборках делать перестановочный тест очень долго.

1. Метод максимального правд.-я
2. Критерий согласия

Про что ещё рассказать?

1. Бутстрап
2. Гипотезы в доля х
3. Методы ↓ дисперсии
   - Стратификация
   - CUPED, CUPAC

   ....
4. Специфические процедуры АБ
   - → DnD
   - → LATE
   - → Мэтчинг
   - → Uplift-моделирование

Линейная регрессия