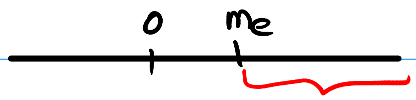
Када можемо користити тест знакова?

- посматрајмо 2 зависна узорка $(X_1, X_2, \dots, X_n)$ и $(Y_1, \dots, Y_n)$ т.д. су $(X_1, Y_1), \dots, (X_n, Y_n)$ независни.
- овакву поставку смо видели код упареног t теста где смо уз претпоставку да $D_i = X_i - Y_i \sim \mathcal{N}(m, \sigma^2)$ испитивали да ли постоји статистички значајна разлика између обележја X и Y, тестирајући хипотезе о вредности параметра m.

- Шта ако не знамо да ли $D_i \sim \mathcal{N}(m, \sigma^2)$?
  - извршићемо тестирање о медијани од D_i

- нека је m_e медијана од D_i ($D_1, \dots, D_n$ је и даље ПСУ)
- шта нам говори $m_e > 0$?



$P\{X_i - Y_i > 0\} > \frac{1}{2}$

вероватноћа да D_i узме вредност већу од m_e је 0,5

$\iff P\{X_i > Y_i\} > \frac{1}{2}$

$\Rightarrow$ вероватноћа да D_i узме вредност већу од 0 је **већа** од 0,5

- дакле, вредности m_e веће од 0 упућују на то да обележје X узима значајно веће вредности од обележја Y

48. A - дужина дејства 1. лека
B - дужина дејства 2. лека
$D = B - A$ - разлика дужине дејства два лека
m_e - медијана од D

- не постоји статистички значајна разлика у дужини дејства ових лекова $\Leftrightarrow$ $\boxed{me=0}$ $\frac{1}{2}=P\{D>0\}=P\{B>A\}$

- постоји статистички значајна разлика $\Leftrightarrow$ $\boxed{me\neq 0}$

$$H_0: me=0 \quad vs \quad H_1: me\neq 0$$

$$T=\sum_{i=1}^{n} I\{D_i>0\} \text{ - тест стат.}$$

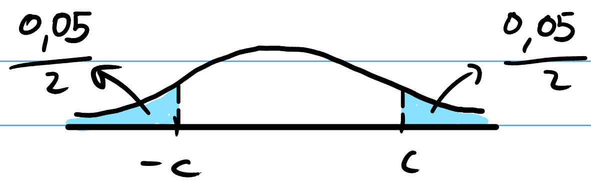
$n=12>10 \Rightarrow$ користићемо стандардизовану тест статистику:

$$T^*=\frac{T-\frac{n}{2}}{\sqrt{\frac{n}{4}}}\underset{\text{при } H_0}{\sim} N(0,1)$$

$W=\{|T^*|>c\}$ - критична област

тражимо c:

c је т.д. $P\{T^*\in W \mid H_0 \text{ тачна}\}=0,05$



$$\Rightarrow \Phi(c)=1-\frac{0,05}{2} \Rightarrow c=\Phi^{-1}(0,975)$$
$$\Rightarrow c=1,96$$

- реализована вредност од T^* је $\dfrac{9-\frac{12}{2}}{\sqrt{\frac{12}{4}}}=\sqrt{3}=1,732051$

$\Rightarrow$ T^* није упало у критичну област, па се не одбацује H_0. Дакле, не можемо рећи да постоји статистички значајна разлика у дужини дејства ових лекова

- Како би гласила H_1 ако желимо да испитамо да ли 2. лек има значајно дуже дејство од 1. лека?

$$H_1: \ me > 0$$

Вилкоксонов тест (означених рангова)

(Wilcoxon signed rank test)

- тестирамо хипотезе о вредности медијане симетрично и апсолутно непрекидно расподељеног обележја X

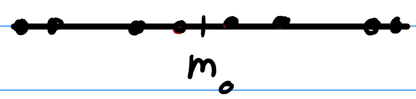
- $H_0: me = m_0$ vs $H_1: me \overset{>}{\underset{<}{\neq}} m_0$

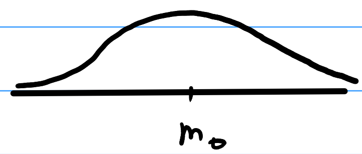
- Како је Вилкоксонову тест. стат. мало теже замислити, покушајмо са слике да стекнемо интуицију:

  - ако је H_0 тачна, очекујемо да тачке из узорка буду равномерно размештене око m_0

    ↓ јер је расподела обележја симетрична

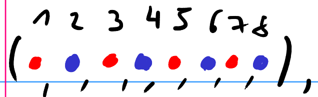
густина обележја при H_0





- приближно исти број тачака лево и десно од m_0
- када бисмо сортирали тачке по апсолутној вредности њихових растојања од m_0, то би отприлике изгледало овако:

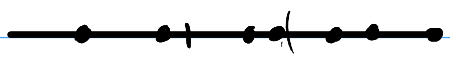
$$W = 2+4+6+8 = 20$$



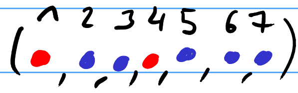
Где су плаве тачке оне које су веће од m_0, а црвене мање од m_0

- Ако је $me > m_0$, каква слика иде у прилог томе

$$W = 2+3+5+6+7 = 23$$



m_0



Више плавих тачака при десном крају низа

Дакле, у прилог $me > m_0$ иду веће редни бројеви плавих тачака у овако сортираном низу

- Сада можемо боље разумети тест статистику коју је Вилкоксон предложио:

$$\boxed{W = \sum_{i=1}^{n} R_i \, I\{X_i > m_0\}}$$ – Вилкоксонова тест статистика

R_i - "ранг" елемента $|X_i - m_0|$, тј. његов редни број у растуће сортираном низу $|X_1 - m_0|, \dots, |X_n - m_0|$

Т при H_0: $EW = \frac{n(n+1)}{4}$, $DW = \frac{n(n+1)(2n+1)}{24}$

За $n > 12$:

$$\boxed{T = \frac{W - EW}{\sqrt{DW}} \underset{\text{при } H_0}{\sim} N(0,1)}$$

Као што смо видели, у прилог $H_1: me > m_0$ иду веће вредности W, па је тад критична област: $\{W > c\}$

Ако је $H_1: m_e < m_0 \to$ кр. област: $\{W \le c\}$.

За $n > 12$ користићемо стандардизовану тест стат T и тада критичне области изгледају овако:

$$H_1: m_e > m_0 \to \{T \ge c\}$$
$$H_1: m_e < m_0 \to \{T \le c\}$$
$$H_1: m_e \ne m_0 \to \{|T| \ge c\}$$

49. Како нас занима да ли се променио просечан проценат неписмених, требало би извршити тестирање о очекивању расподеле. Пошто је расподела симетрична, очекивање и медијана су исти.

Израчунајмо тест стат W.

1. корак: рачунамо $X_1 - 0{,}7,\ X_2 - 0{,}7, \dots, X_n - 0{,}7$

-0.10 -0.20 -0.08 1.00 0.05 0.30 -0.01 0.10 0.10 -0.13 0.20 0.80 0.25 -0.17 0.40 0.50 1.30 -0.05 0.09 -0.09

2. корак: рачунамо $|X_1 - 0{,}7|, |X_2 - 0{,}7|, \dots, |X_n - 0{,}7|$ и памтимо знак

0.10	0.20	0.08	1.00	0.05	0.30	0.01	0.10	0.10	0.13	0.20	0.80	0.25	0.17	0.40	0.50	1.30	0.05	0.09	0.09
−	−	−	+	+	+	−	+	+	−	+	+	+	−	+	+	+	−	+	−

3. корак: сортирамо $|X_1 - 0{,}7|, \dots, |X_n - 0{,}7|$ и рангирамо их

0.01	0.05	0.05	0.08	0.09	0.09	0.10	0.10	0.10	0.13	0.17	0.20	0.20	0.25	0.30	0.40	0.50	0.80	1.00	1.3
1	2,5	2,5	4	5,5	5,5	8	8	8	10	11	12,5	12,5	14	15	16	17	18	19	20
	+				+		+	+			+		+	+	+	+	+	+	+

4. корак: сабирамо рангове изнад "+", односно сабирамо рангове од $|X_i - 0{,}7|$ ако је $X_i - 0{,}7 > 0$

$$W = 2{,}5 + 5{,}5 + 8 + 8 + 12{,}5 + 14 + 15 + 16 + 17 + 18 + 19 + 20 = 155{,}5$$

Како се врши рангирање?

- најпре се свим елементима сортираног низа $|X_i - 0{,}7|$ упишу редом бројеви од 1 до n. Ако у сортираном низу $|X_i - 0{,}7|$ нема понављања, готови смо
- ако има понављања, ти елементи 'деле' ранг, тј. њихов ранг је аритметичка средина претходно уписаних бројева
- тако 0,05 и 0,05 немају рангове 2 и 3, већ $\frac{2+3}{2}$ и сл.

Вилкоксонов тест за 2 независна узорка

(Wilcoxon rank-sum test)

- Сада ћемо испитивати да ли постоји статистички значајна разлика између вредности 2 независна обележја X и Y

- Нека су дата 2 независна узорка $X=(X_1,\dots,X_n)$ и $Y=(Y_1,\dots,Y_m)$ из истих расподела као X и Y, редом

- Претпостављаћемо да су расподеле X и Y апсолутно непрекидне и да важи: $X \overset{d}{=} Y+c$

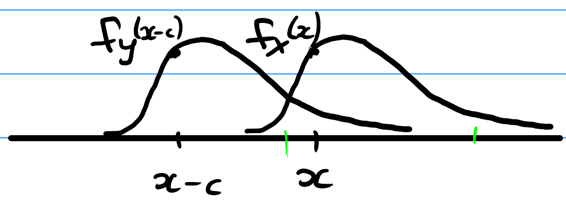
ознака $X \overset{d}{=} Y+c$ значи да X и $Y+c$ имају исту расподелу, али не нужно да су X и $Y+c$ једнаке као функције!

Дакле, $X \overset{d}{=} Y+c \Rightarrow F_X(x) = F_{Y+c}(x) = P\{Y+c \le x\} = F_Y(x-c) \,/'_x$

Одавде следи: $f_X(x) = f_Y(x-c)$

Када ћемо рећи да X узима знатно веће вредности од Y? Посматрајмо c.

Нека је нпр $c>0$. Како изгледају густине обележја X и Y?



Са слике видимо да се вредности обележја X и Y сконцентришу на 2 транслирана скупа A и $A-c$, редом.
$A-c=\{a-c,\ a\in A\}$.

- Већина елемената скупа A је веће од елемената скупа $A-c$, па очекујемо да, ако је c знатно позитивно, узорак $(X_1,\ldots,X_n)$ садржи више већих вредности него узорак $(Y_1,\ldots,Y_m)$

- Односно, ако плавим тачкама означимо вредности $(X_1,\ldots,X_n)$, а црвеним вредности $(Y_1,Y_2,\ldots,Y_m)$, очекујемо следећу ситуацију:

Дакле, испитивање да ли је X значајно веће од Y се своди на испитивање да ли је $c>0$

Вилкоксоновим тестом ћемо дакле тестирати:

$$H_0:\ c=0 \qquad \text{vs} \qquad H_1:\ c \overset{>}{\underset{<}{\neq}} 0$$

$\Updownarrow$

не постоји значајна разлика између обележја X и Y

Предложена тест стат. је $\boxed{T=\sum_{i=1}^{n} R_i}$, где је

R_i "ранг" елемента X_i у обједињеном узорку (X,Y).

Односно, R_i је редни број X_i у растуће сортираном низу $X_1,X_2,\ldots,X_n,Y_1,\ldots,Y_m$.

Видели смо да у прилог $H_1: c>0$ иду веће вредности R_i (плаве тачкице су углавном десно од црвених), па у том случају критична област изгледа: $W=\{T\geq q\}$

- ако је $H_1: c<0 \Rightarrow W=\{T\leq q\}$

(Т) Ако је H_0 тачна, $ET=\frac{n(n+m+1)}{2}$, $DT=\frac{nm(n+m+1)}{12}$

и тада за $n,m>10$ $\boxed{T^*=\frac{T-ET}{\sqrt{DT}}\sim \mathcal{N}(0,1)}$

Дакле, за велике m,n користићемо тест стат T^* и следеће критичне области:

$$H_1: c>0 \rightarrow W=\{T^*\geq q\}$$
$$H_1: c<0 \rightarrow W=\{T^*\leq q\}$$
$$H_1: c\neq 0 \rightarrow W=\{|T^*|\geq q\}$$

50. Нека X представља количину падавина у региону A, а Y количину падавина у региону B. Треба тестирати да ли је $EX=EY$. Како је претпоставка овог теста: $X\overset{d}{=}Y+c$, значи да је $\boxed{EX=EY+c}$, па је $EX=EY \Leftrightarrow c=0$

Зато хипотезе изгледају овако:

$$\underset{\text{или } c=0}{H_0: EX=EY} \quad \text{vs} \quad \underset{\text{или } c\neq 0}{H_1: EX\neq EY}$$

Да бисмо израчунали $T=\sum_{i=1}^{n} R_i$, потребно је да сортирамо обједињени узорак падавина у оба региона

1.3 2.5 2.6 2.7 2.8 2.9 3.0 3.3 3.4 3.4 3.5 4.0 4.1 4.8 4.9 4.9 5.0 5.1 5.2 6.0 7.0 10.3

затим је потребно да га рангирамо, што се чини исто као у претходном задатку:

1.3	2.5	2.6	2.7	2.8	2.9	3.0	3.3	3.4	3.4	3.5	4.0	4.1	4.8	4.9	4.9	5.0	5.1	5.2	6.0	7.0	10.3
1	2	3	4	5	6	7	8	9.5	9.5	11	12	13	14	15.5	15.5	17	18	19	20	21	22
A	A	A	A	A		A		A				A	A	A		A					

Следеће што нам преостаје јесте да саберемо рангове изнад 'A', тј. да саберемо редне бројеве елемената низа падавина у региону A у сортираном обједињеном узорку

$$T = 1+2+3+4+5+7+9{,}5+13+14+15.5+17 = 91$$