Тестови сагласности са расподелом

- $(X_1, X_2, \dots, X_n)$ – ПСУ, F – ф-ја расподеле сл. вел. X_i

- тестираћемо следеће хипотезе:

$$H_0: F = F_0 \quad \text{vs} \quad H_1: F \neq F_0,$$

1. Колмогоров-Смирнов тест

- радимо само случај када F_0 не зависи од непознатих параметара
- претпоставка: F је апс. непрекидна

Када бисмо претпоставили да важи H_1? Поменули смо да $F(x)$ можемо оценити емпиријском функцијом расподеле:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \le x\}$$

и показали да ова оцена има неке лепе особине као што су непристрасност и постојаност (10. зад.)

Зато је идеја да упоредимо $F_n(x)$ и $F_0(x)$ $\forall x \in \mathbb{R}$. Ако F_n на неки начин 'значајно' одступа од F_0, тада ће расти наше уверење у H_1.

Како можемо измерити колико се F_n и F_0 разликују? Посматрајмо супремум норму:

$$\boxed{\|F_n - F_0\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|}$$

Зашто баш супремум норма?

Ⓣ (Гливенко – Кантели)

$(X_1, \dots, X_n)$ ПСУ F - ф-ја расподеле од X_i, F_n - емпиријска ф-ја расподеле.

Тада:
$$P\{\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0,\ n\to\infty\} = 1$$

тумачење: за n велико график ф-је F_n јако подсећа на график ф-је $F(x)$.

Дакле, велике вредности следеће статистике

$$\boxed{D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|}$$

упућују на то да $F \neq F_0$.

Ако је $H_0: F = F_0$ тачна и F апс. непр., може да се покаже да расподела од D_n не зависи од F_0.
Такође, ако важи H_0, позната је и тачна расподела од D_n, што нам говори да за тестирање хипотезе

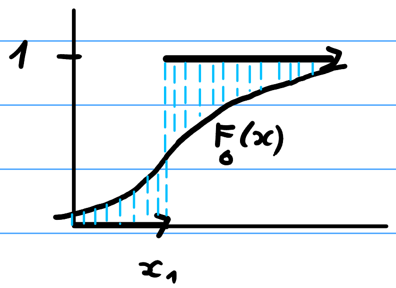
$$H_0: F = F_0 \quad vs \quad H_1: F \neq F_0$$

има смисла користити тест стат. D_n
- критична област: $\boxed{W = \{D_n \geq c\}}$
↳ читамо из таблица за К–С тест

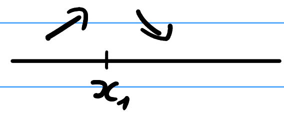
D_n - Колмогоров–Смирнова статистика

Како израчунати D_n?

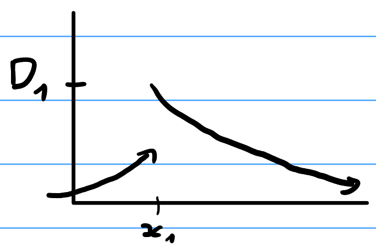
- нека је за почетак $n=1$:



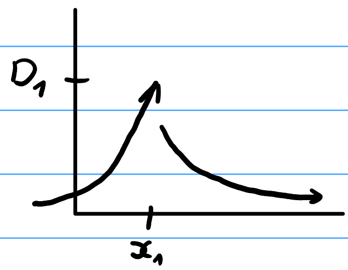
понашање $|F_n(x) - F_0(x)|$:



могуће слике графика $|F_n(x)-F_0(x)|$:



ако је $|1-F_0(x_1)|$ веће од $|F_0(x_1)-0|$

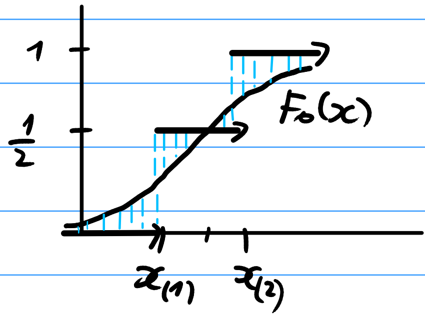


ако је $|F_0(x_1)-0|$ веће од $|1-F_0(x_1)|$

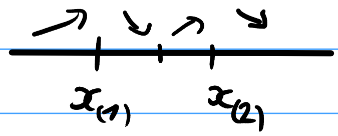
Дакле, D_1 је већа од вредности $|1-F_0(x_1)|$ и $|F_0(x_1)-0|$, односно, како је $1=F_1(x_1)$, а $0=F_1(x_1-)$ (леви лимес), можемо рећи да је D_1 већа од вредности $|F_1(x_1)-F_0(x_1)|$ и $|F_1(x_1-)-F_0(x_1)|$.

Слично, за произвољно $n\in\mathbb{N}$ вредност D_n се добија као највећа од свих вредности $|F_n(x_i)-F_0(x_i)|$ и $|F_n(x_i-)-F_0(x_i)|$, где су $x_1,\dots,x_n$ реализоване вредности узорка.

слика и за $n=2$: (једна од могућих)



понашање $|F_2(x) - F_0(x)|$:



кандидати за $\sup\limits_{x\in\mathbb{R}} |F_2(x) - F_0(x)|$ су:

$|F_2(x_1-) - F_0(x_1)|$, $|F_2(x_1) - F_0(x_1)|$, $|F_2(x_2-) - F_0(x_2)|$ и $|F_2(x_2) - F_0(x_2)|$

53. рачунамо D_n:

- треба да одредимо $|F_n(x_i-) - F_0(x_i)|$ и $|F_n(x_i) - F_0(x_i)|$, $i \in \{1,2,\dots,n\}$, где су x_i вредности из узорка

- како F_n прави скокове у тачкама из узорка и неопадајућа је ф-ја, најлакше нам је да сортирамо узорак растуће и сукцесивно одредимо $F_n(x_{(1)}), F_n(x_{(2)}), \dots, F_n(x_{(n)})$

- одређивање $F_n(x_{(i)})$:
  - пратећи дефиницију F_n, видимо да она има прву ненулту вредност у $x_{(1)}$
  - та вредност је једнака: $\dfrac{\text{бр. тачака} \le x_{(1)}}{n}$,

што је у нашем примеру: $\frac{1}{n}$ (тј. $\frac{1}{16}$)

ВАЖНО: F_n прави скокове величине $\frac{1}{n}$ $\boxed{\text{ако и само ако}}$ су све тачке из узорка различите!
Зато приликом одређивања F_n треба водити рачуна о томе да ли се неке тачке у узорку понављају

- видимо у овом примеру да имамо 2 исте вредности 0,43, па $F_n(0,43)$ није за $\frac{1}{n}$ већа од претходне вредности, већ за $\frac{2}{n}$

- дакле, најбезбедније је F_n рачунати по дефиницији, а не само додавати $\frac{1}{n}$ на сваку претходну вредност
- у тачки $x_{(n)}$ вредност ф-је F_n је увек 1

- одређивање $F_n(x_{(i)}-)$:
  - вредност $F_n(x_{(i)}-)$ се може добити тако што се са $F_n(x_{(i)})$ 'спустимо за 1 степеник'
  - дакле, $F_n(x_{(1)}-)=0$, остале вредности $F_n(x_{(i)}-)$ се добијају 'шифтованем' колоне с вредностима $F_n(x_{(i)})$

- када смо одредили вредности $F_n(x_{(i)})$ и $F_n(x_{(i)}-)$, рачунамо $|F_n(x_{(i)})-F_0(x_{(i)})|$ и $|F_n(x_{(i)}-)-F_0(x_{(i)})|$

- вредност тест статистике D_n је највећа од свих вредности $|F_n(x_{(i)})-F_0(x_{(i)})|$ и $|F_n(x_{(i)}-)-F_0(x_{(i)})|$

- у нашем примеру F_0 је идентичко пресликавање на $[0,1)$, па је $F_0(x_{(i)})=x_{(i)}$, те није неопходно исписивати посебну колону за вредности $F_0(x_{(i)})$

$x_{(i)}$	$F_n(x_{(i)}^-)$	$F_n(x_{(i)})$	$\lvert F_n(x_{(i)}) - F_0(x_{(i)})\rvert$	$\lvert F_n(x_{(i)}^-) - F_0(x_{(i)})\rvert$
0,11	0	1/16	$\lvert 1/16 - 0{,}11\rvert = 0{,}0475$	$\lvert 0 - 0{,}11\rvert = 0{,}11$
0,18	1/16	2/16	$\lvert 2/16 - 0{,}18\rvert = 0{,}055$	$\lvert 1/16 - 0{,}18\rvert = 0{,}1175$
0,22	2/16	3/16	0,0325	0,095
0,31	3/16	4/16	0,06	0,1225
0,43×2	4/16	6/16	0,055	$\boxed{0{,}18}$
0,47	6/16	7/16	0,0325	0,095
0,56	7/16	8/16	0,06	0,1225
0,58	8/16	9/16	0,0175	0,08
0,59	9/16	10/16	0,035	0,0275
0,66	10/16	11/16	0,0275	0,035
0,72	11/16	12/16	0,03	0,0325
0,73	12/16	13/16	0,0825	0,02
0,78	13/16	14/16	0,095	0,0325
0,9	14/16	15/16	0,0375	0,025
0,96	15/16	16/16	0,04	0,0225

највећа вредност из ове две колоне је 0,18 и то је вредност тест стат. D_n

– критична област: $W = \{D_n \geq c\}$

читамо из таблице за К-С тест за $n=16$ и $\alpha = 0{,}05$