Statistika

11. predavanje

Barbara Boldin

Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Univerza na Primorskem

Korelacija

Naj bosta X in Y številski slučajni spremenljivki na isti statistični populaciji, npr.:

- višina sistoličnega in diastoličnega krvnega tlaka pacientov izbranega zdravnika,
- vsebnost palmitinske in oleinske maščobne kisline v izbranih olivnih oljih,
- globina in magnituda potresov na izbranem območju,
- itd.

Spoznali bomo dve meri korelacije (oz. soodvisnosti) slučajnih spremenljivk:

- **Pearsonov koeficient** in
- **Spearmanov koeficient** korelacije.

Pearsonov koeficient korelacije

Par slučajnih spremenljivk (X, Y), ki je porazdeljen po **dvorazsežni normalni porazdelitvi**

$$N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

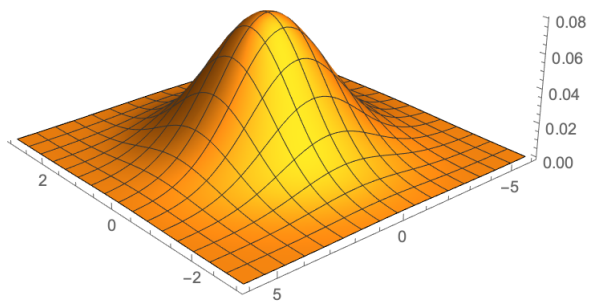
ima funkcijo gostote

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-\mu_X)}{\sigma_X}\frac{(y-\mu_Y)}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\Bigg)}$$

kjer $x, y \in \mathbb{R}$.

Število ρ je **koeficient korelacije**. Velja:

- $-1 \leq \rho \leq 1$
- $\rho_{(X,Y)} = \rho_{(Y,X)}$
- $\rho_{(X,X)} = 1$.



Korelacijo slučajnih spremenljivk X in Y bomo ugotavljali na podlagi vzorca. Denimo, da je v vzorcu n enot, za vsako enoto imamo podatek o vrednosti X in Y, torej

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Če razsevni grafikon prikaže približno linearno zvezo med X in Y, npr.:

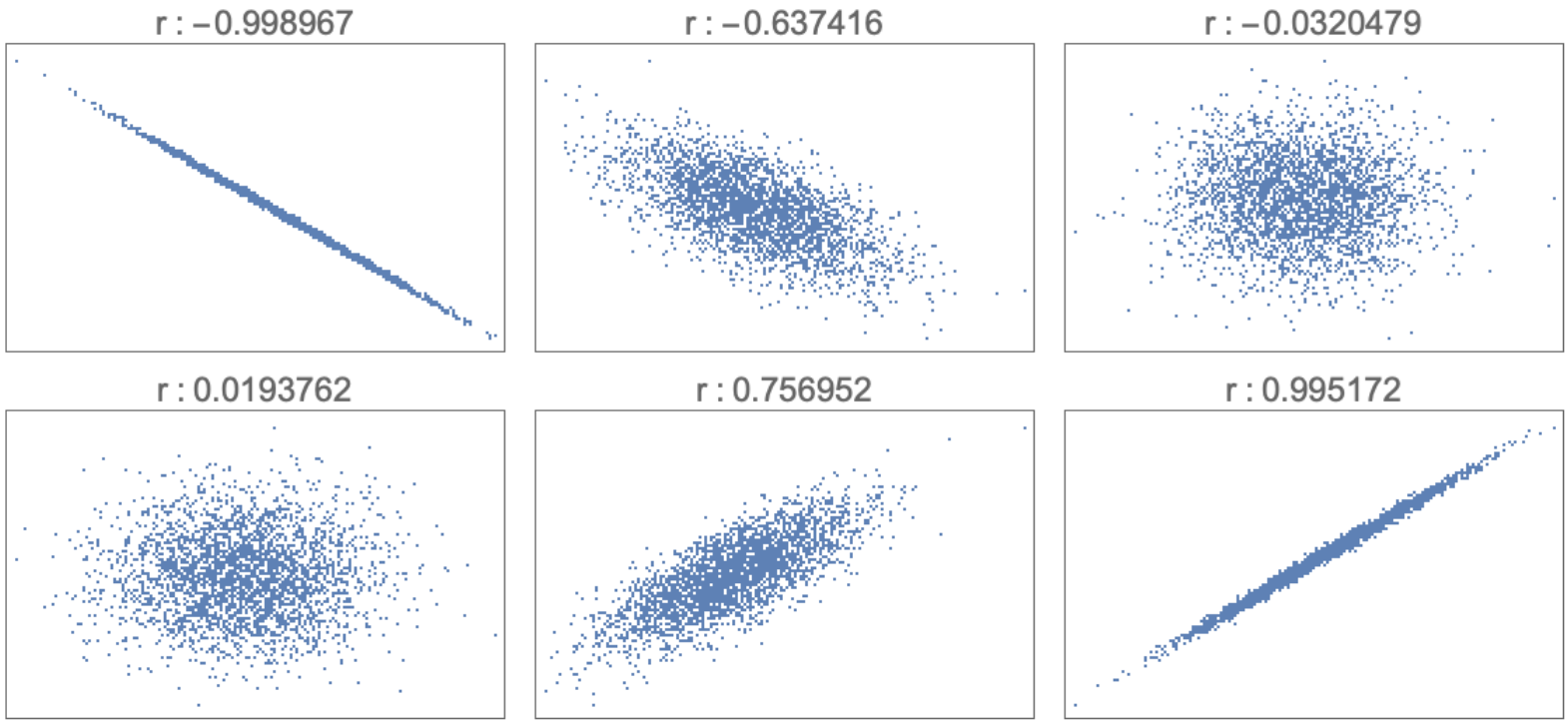
lahko privzamemo, da ima (X, Y) dvorazsežno normalno porazdelitev. Na podlagi vzorca izračunamo **oceno za koeficient korelacije**,

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}$$

ki jo imenujemo **Pearsonov koeficient korelacije.**

Pearsonov koeficent korelacije zavzame vrednosti med -1 in 1. Če je $r > 0$ govorimo o pozitivni povezanosti spremenljivk, če je $r < 0$ sta spremenljivki negativno povezani.

Nekaj primerov razsevnih grafikonov:



Običajno nas zanima, ali sta spremenljivki X in Y sploh povezani. S Pearsonovim koeficientom lahko preverjamo hipoteze o linearni povezanosti spremenljivk.

Preizkušamo ničelno domnevo

$H_0 : \rho = 0$ (spremenljivki nista linearno povezani)

ob alternativni domnevi

$H_1 : \rho \neq 0$ (spremenljivki sta linearno povezani)

Izkaže se, da je testna statistika

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

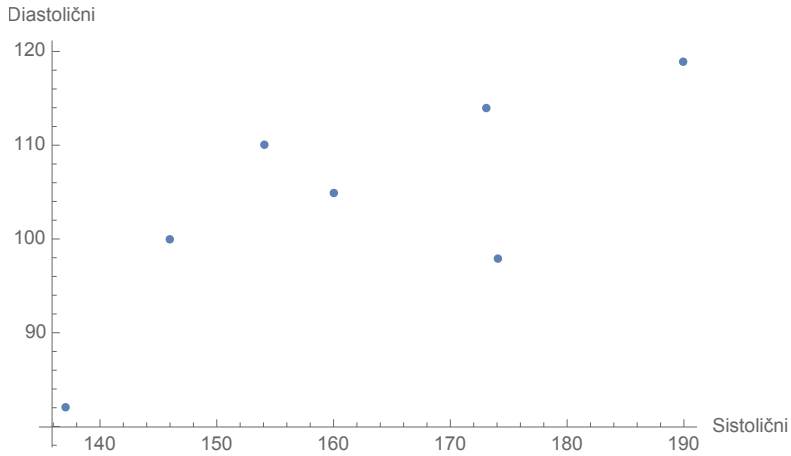
ničelna porazdelitev je Studentova s stopnjo prostosti $df = n - 2$.

Alternativen test: Izračunu testne statistike se lahko izognemo, če imamo na razpolago **tabelo kritičnih vrednosti Pearsonovih koeficientov** (glej e-učilnico).
Če je izračunan r po absolutni vrednosti večji od kritične vrednosti pri dani stopnji značilnosti α, H_0 zavrnemo v korist H_1.

Primer. Zdravniki merijo dve vrsti krvnega tlaka, sistoličnega in diastoličnega. Zanima nas soodvisnost le teh. V vzorec je vključenih sedem oseb, njihovi podatki so:

Pacient	Sistolični tlak (X)	Diastolični tlak (Y)
1	160	105
2	146	100
3	154	110
4	190	119
5	174	98
6	173	114
7	137	82

Narišimo razsevni grafikon in pri stopnji značilnosti $\alpha = 0.05$ preverimo domnevo, da sta sistolični in diastolični tlak (linearno) povezana.

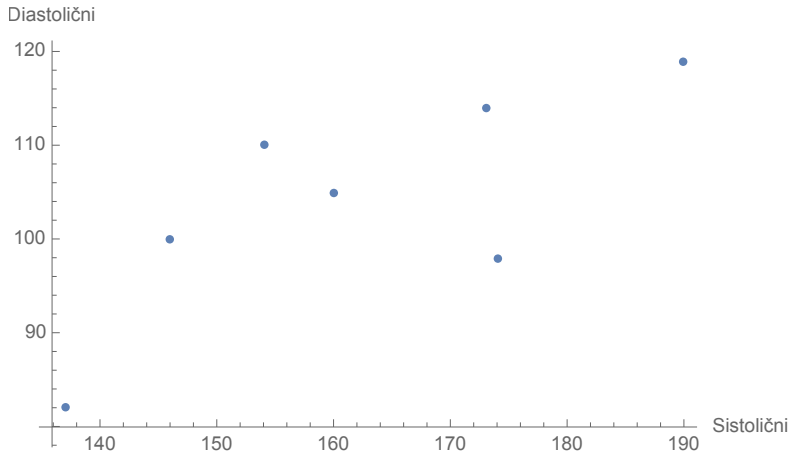


$$\bar{x} = 162, \bar{y} = 104$$

$$\sum_{i=1}^{7} x_i^2 = 185706, \sum_{i=1}^{7} y_i^2 = 76610, \sum_{i=1}^{7} x_i y_i = 118958$$

in torej

$$r = 0.76.$$



$$\bar{x} = 162, \bar{y} = 104$$

$$\sum_{i=1}^{7} x_i^2 = 185706, \sum_{i=1}^{7} y_i^2 = 76610, \sum_{i=1}^{7} x_i y_i = 118958$$

in torej

$$r = 0.76.$$

Preizkušamo ničelno domnevo

$H_0 : \rho = 0$ (sistolični in diastolični tlak nista linearno povezana)

ob alternativni domnevi

$H_1 : \rho \neq 0$.

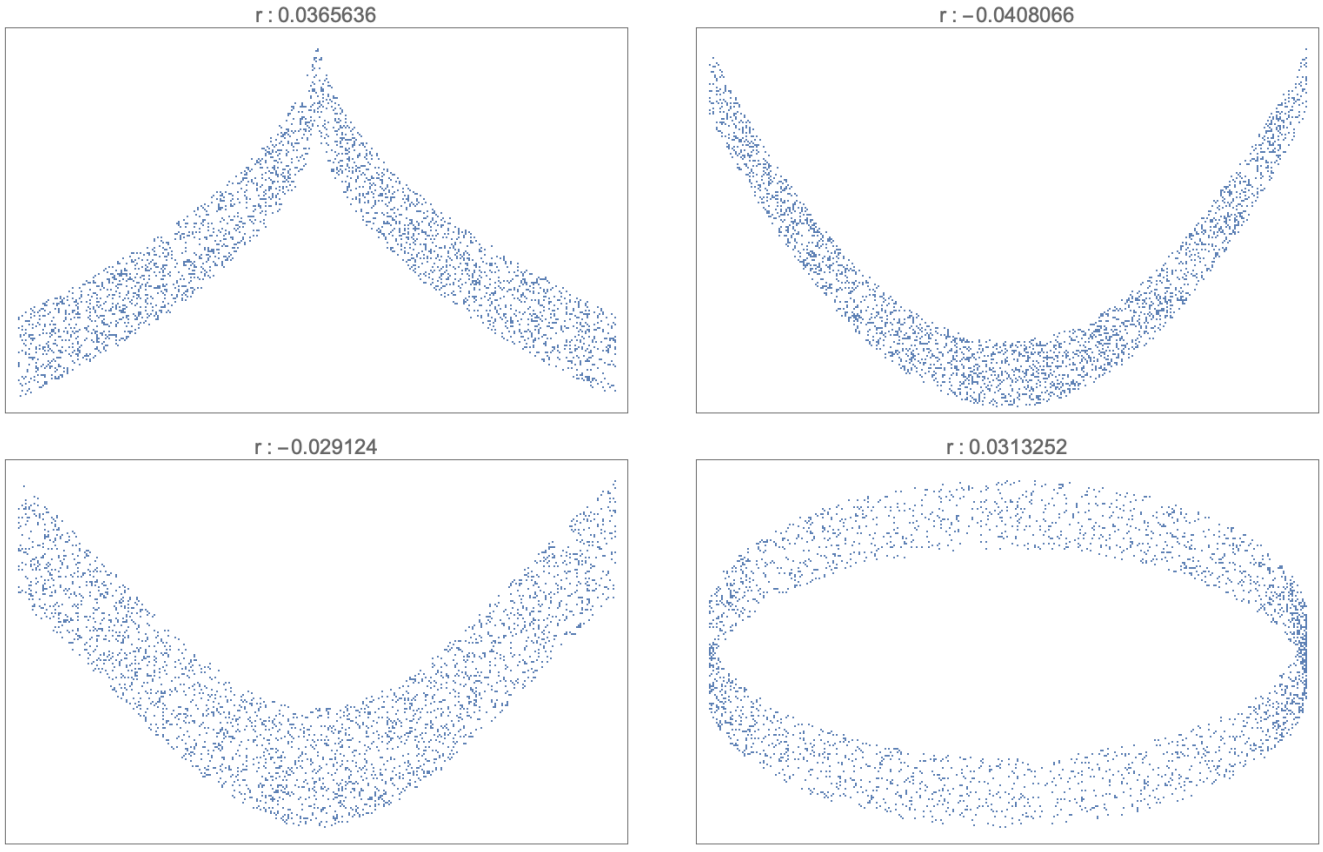
Iz tabele razberemo kritično vrednost $r_{\text{krit}} = 0.755$.

Ker $|r| > r_{\text{krit}}$, ničelno domnevo zavrnemo v korist H_1.

Pri stopnji značilnosti $\alpha = 0.05$ trdimo, da sta sistolični in diastolični tlak (linearno) povezana.

Kadar povezava med slučajnima spremenljivkama ni linearna, potem uporaba koeficienta korelacije kot mere povezanosti med spremenljivkama ni primerna.

Spodnja slika prikazuje primere, ko je Pearsonov korelacijski koeficient blizu 0, kar pa ne pomeni, da povezave ni! Povezava obstaja, vendar ni linearna.



Kakšna je povezava med linearno regresijo ter korelacijskim modelom?

- Pri regresiji je Y slučajna spremenljivka, ki ima pri dani vrednosti x določeno normalno porazdelitev.
- Pri korelaciji pa imamo par slučajnih spremenljivk (X, Y), ki ima določeno dvorazsežno porazdelitev.

Velja pa, da je koeficient korelacije ustrezno predznačen koren koeficienta determinacije

$$r = \pm\sqrt{r^2}.$$

Spearmanov koeficient korelacije

Pearsonov koeficient korelacije je mera **linearne** povezanosti dveh intervalskih slučajnih spremenljivk. Predpostavka o linearnosti ni vedno utemeljena.

Spearmanov koeficient korelacije je še ena mera povezanosti dveh številskih spremenljivk, ki pa temelji le na šibkejški predpostavki o monotoni povezavi med spremenljivkama.

Ideja: podatke nadomestimo z njihovimi rangi in izračunamo Pearsonov koeficient korelacije na rangih:

- Za spremenljivko X enotam priredimo njihove range (če ima več enot enako vrednost, jim priredimo njihov povprečni rang)
- Postopek ponovimo za spremenljivko Y
- Na rangih izračunamo Pearsonov koeficient korelacije, ga označimo z r_S in imenujemo Spearmanov koeficient korelacije.

Spearmanov koeficient korelacije

Pearsonov koeficient korelacije je mera **linearne** povezanosti dveh intervalskih slučajnih spremenljivk. Predpostavka o linearnosti ni vedno utemeljena.

Spearmanov koeficient korelacije je še ena mera povezanosti dveh številskih spremenljivk, ki pa temelji le na šibkejški predpostavki o monotoni povezavi med spremenljivkama.

Ideja: podatke nadomestimo z njihovimi rangi in izračunamo Pearsonov koeficient korelacije na rangih:

- Za spremenljivko X enotam priredimo njihove range (če ima več enot enako vrednost, jim priredimo njihov povprečni rang)
- Postopek ponovimo za spremenljivko Y
- Na rangih izračunamo Pearsonov koeficient korelacije, ga označimo z r_S in imenujemo Spearmanov koeficient korelacije.

Zdaj lahko preizkusimo domnevo o monotoni povezanosti slučajnih spremenljivk.

Preizkušamo ničelno domnevo

$H_0 : \rho_S = 0$ (spremenljivki nista monotono povezani)

ob alternativni domnevi

$H_1 : \rho_S \neq 0$ (spremenljivki sta monotono povezani)

Kritične vrednosti Spearmanovega koeficienta za vrednosti stopnje značilnosti α najdemo v tabeli (glej e-učilnico).

Kadar je povezava med slučajnima spremenljivkama linearna, je razlika med Pearsonovim in Spearmanovim koeficientom majhna. V tem primeru uporabljamo Pearsonov koeficient korelacije, saj ima pripadajoči statistični test večjo moč.

Zdaj lahko preizkusimo domnevo o monotoni povezanosti slučajnih spremenljivk.

Preizkušamo ničelno domnevo

$H_0 : \rho_S = 0$ (spremenljivki nista monotono povezani)

ob alternativni domnevi

$H_1 : \rho_S \neq 0$ (spremenljivki sta monotono povezani)

Kritične vrednosti Spearmanovega koeficienta za vrednosti stopnje značilnosti α najdemo v tabeli (glej e-učilnico).

Kadar je povezava med slučajnima spremenljivkama linearna, je razlika med Pearsonovim in Spearmanovim koeficientom majhna. V tem primeru uporabljamo Pearsonov koeficient korelacije, saj ima pripadajoči statistični test večjo moč.

Primer. Profesorja zanima, ali obstaja povezava med uspehom študentov pri laboratorijskem delu in uspehom pri pisnem izpitu. V vzorec zajame deset študentov. Pri laboratorijskem delu so študenti rangirani (rang 1 ima najboljši študent, rang 10 najslabši). Pri pisnem izpitu je zabeleženo število točk (med 0 in 100). Rezultati so zbrani v tabeli.

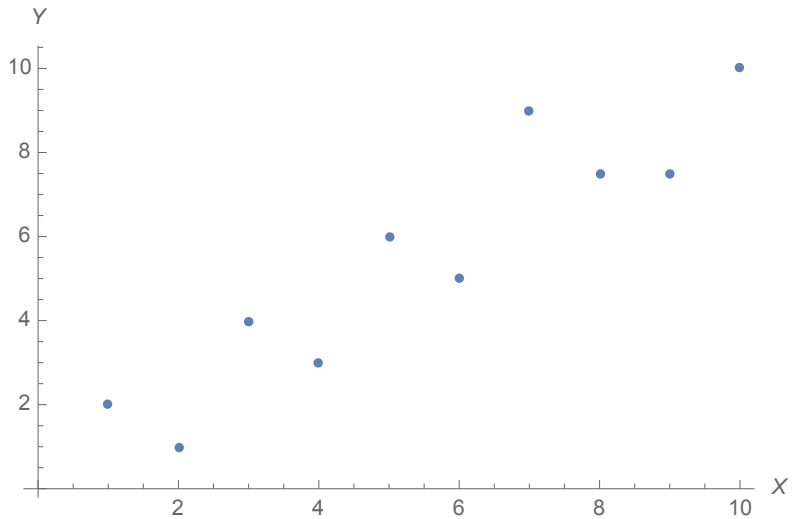
Laboratorij (rang) (X)	Pisni izpit (št. točk)	Pisni izpit (rang) (Y)
3	72	
8	35	
2	100	
9	35	
4	83	
1	89	
7	32	
5	44	
10	25	
6	45	

Pri stopnji značilnosti $\alpha = 0.05$ preverimo domnevo, da obstaja monotona povezava med uspehom pri laboratorijskem delu in uspehom na pisnem izpitu. Je ta povezava pozitivna ali negativna?

Primer. Profesorja zanima, ali obstaja povezava med uspehom študentov pri laboratorijskem delu in uspehom pri pisnem izpitu. V vzorec zajame deset študentov. Pri laboratorijskem delu so študenti rangirani (rang 1 ima najboljši študent, rang 10 najslabši). Pri pisnem izpitu je zabeleženo število točk (med 0 in 100). Rezultati so zbrani v tabeli.

Laboratorij (rang) (X)	Pisni izpit (št. točk)	Pisni izpit (rang) (Y)
3	72	4
8	35	7.5
2	100	1
9	35	7.5
4	83	3
1	89	2
7	32	9
5	44	6
10	25	10
6	45	5

Pri stopnji značilnosti $\alpha = 0.05$ preverimo domnevo, da obstaja monotona povezava med uspehom pri laboratorijskem delu in uspehom na pisnem izpitu. Je ta povezava pozitivna ali negativna?

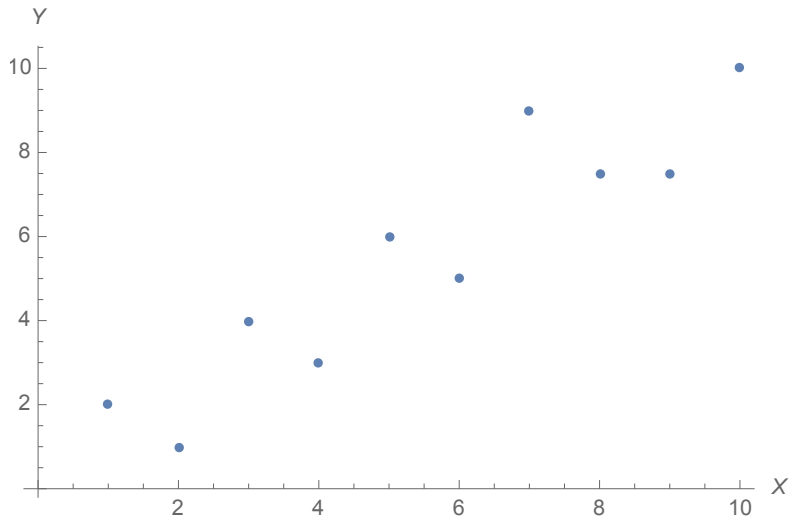


$$\bar{x} = 5.5, \bar{y} = 5.5$$

$$\sum_{i=1}^{10} x_i^2 = 385, \sum_{i=1}^{10} y_i^2 = 384.5, \sum_{i=1}^{10} x_i y_i = 378.5$$

in torej

$$r = 0.924.$$



$$\bar{x} = 5.5, \bar{y} = 5.5$$

$$\sum_{i=1}^{10} x_i^2 = 385, \sum_{i=1}^{10} y_i^2 = 384.5, \sum_{i=1}^{10} x_i y_i = 378.5$$

in torej

$$r = 0.924.$$

Preizkušamo ničelno domnevo

$H_0 : \rho_S = 0$ (rezultata v laboratoriju in na pisnem izpitu nista monotono povezani)

ob alternativni domnevi

$H_1 : \rho_S \neq 0$.

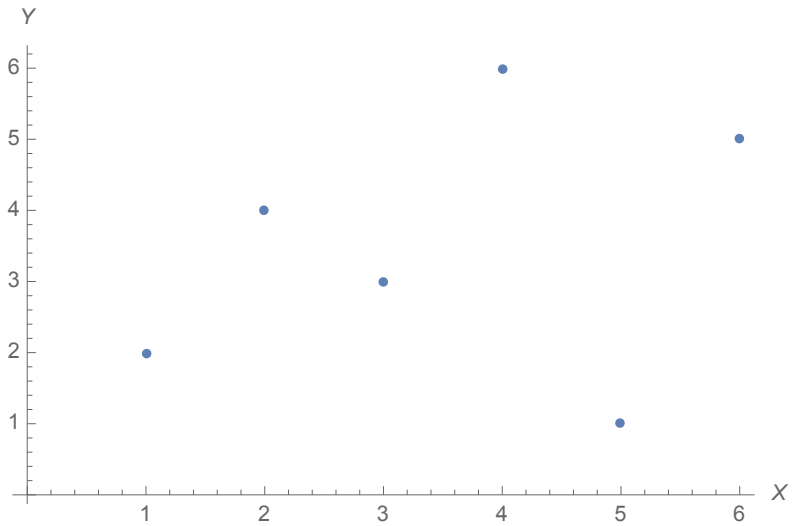
Iz tabele razberemo kritično vrednost $r_{\text{krit}} = 0.648$.

Ker $|r| > r_{\text{krit}}$, ničelno domnevo zavrnemo v korist H_1. Pri stopnji značilnosti $\alpha = 0.05$ trdimo, da je uspeh v laboratoriju monotono povezan z uspehom na pisnem izpitu. Povezava je pozitivna.

Primer. Na vinskem sejmu sta izkušeni sommelier in naključno izbran obiskovalec sejma šest izbranih vin razvrstila od najboljšega (rang 1) do najslabšega (rang 6). Rezultati so v naslednji tabeli.

Vino	Sommelier (rang) (X)	Obiskovalec (rang) (Y)
1	2	4
2	5	1
3	3	3
4	6	5
5	1	2
6	4	6

Pri stopnji značilnosti $\alpha = 0.1$ preverite domnevo o povezanosti ocene vina izkušenega sommelierja in naključnega obiskovalca sejma.

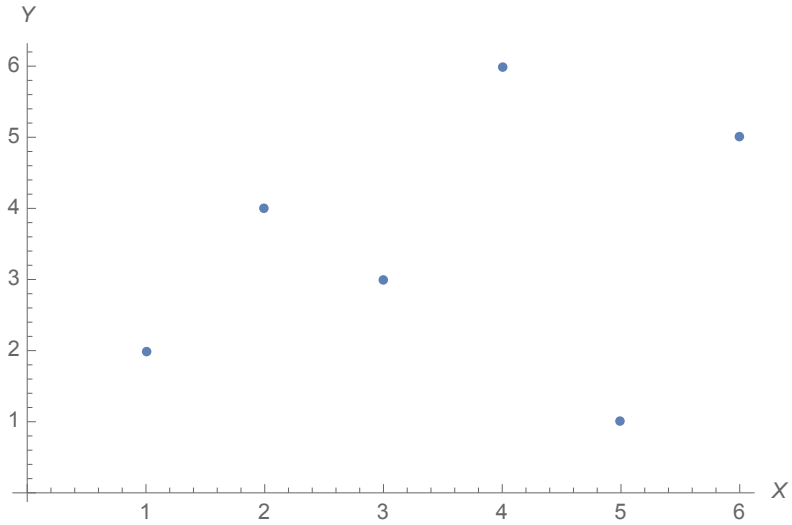


$$\bar{x} = 3.5, \bar{y} = 3.5$$

$$\sum_{i=1}^{10} x_i^2 = 91, \sum_{i=1}^{10} y_i^2 = 91, \sum_{i=1}^{10} x_i y_i = 78$$

in torej

$$r = 0.257.$$



$$\bar{x} = 3.5, \bar{y} = 3.5$$

$$\sum_{i=1}^{10} x_i^2 = 91, \sum_{i=1}^{10} y_i^2 = 91, \sum_{i=1}^{10} x_i y_i = 78$$

in torej

$$r = 0.257.$$

Preizkušamo ničelno domnevo

$H_0 : \rho_S = 0$ (oceni sommelierja in obiskovalca nista monotono povezani)

ob alternativni domnevi

$H_1 : \rho_S \neq 0$.

Iz tabele razberemo kritično vrednost $r_{\text{krit}} = 0.829$.

Ker $|r| < r_{\text{krit}}$, ničelno domnevo obdržimo. Pri stopnji značilnosti $\alpha = 0.1$ trdimo, da med oceno sommelierja in obiskovalca ni monotone povezave.

Kdaj torej lahko uporabljamo Pearsonov in Spearmanov test?

- Pearsonov koeficient korelacije uporabljamo za ugotavljanje linearne povezave med intervalskima slučajnima spremenljivkama.
- Spearmanov koeficient korelacije uporabljamo za ugotavljanje monotone povezave med slučajnima spremenljivkama, ki sta lahko intervalski ali urejenostni.

Primerjava kritičnih vrednosti Pearsonovega in Spearmanovega koeficienta za različne velikosti vzorca in $\alpha = 0.05$:

Velikost vzorca (n)	Pearson	Spearman
10	0.632	0.648
15	0.514	0.525
20	0.444	0.450
25	0.396	0.400
30	0.361	0.364