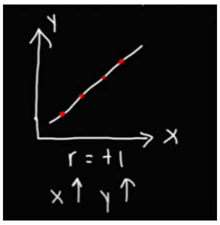
Korelácia, koeficient korelácie

Korelácia je vzťah medzi dvomi veličinami. Korelačný koeficient je číslo, ktoré vyjadruje silu tohto vzťahu. Skúma sa sila korelácie a tvar krivky, ktorá reprezentuje závislosť a hovorí o tom či je kladná alebo záporná.

Koeficient korelácie sa definuje na dvoch štatistických súboroch a preto sa dobre zobrazuje na grafe kedy jeden súbor je vynesený na x-ovú os a druhý na y-ovú os a preložená priamka vyjadruje koreláciu.

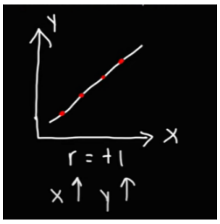
<u>Najprv sa ale naučme čo korelačný koeficient hovorí.</u>

https://www.youtube.com/watch?v=11c9cs6WpJU&t=2s

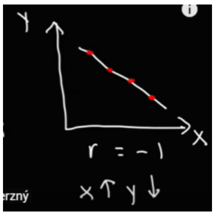


Najprv sa ale naučme čo korelačný koeficient hovorí.

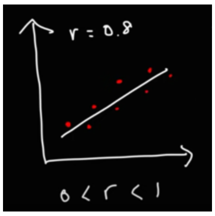
Povedzme že v súradnicovom systéme x,y máme takúto priamku a na priamke máme takto usporiadané body Koeficient korelácie **r** bude v tomto prípade kladný, lebo všetky body ležia na priamke a priamka má kladný sklon. Ide o **kladnú koreláciu**, lebo keď sa zvyšuje x, tak sa zvyšuje aj y. Existuje priamy vzťah (závislosť) medzi x a y.



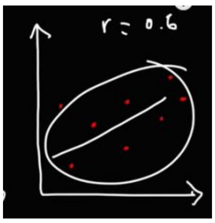
Pri ďalšom prípade máme body opäť na priamke, ale priamka klesá. Korelačný koeficient sa preto bude rovnať zápornej hodnote. Keď x rastie, y klesá. V tomto prípade máme nepriamy (inverzný) vzťah a **zápornú koreláciu**.



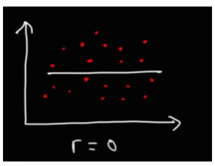
V treťom prípade nebu ležať body na priemke, ale **budú rozptýlené blízko priamky** a tým majú dosť lineárny vzťah. Nie však 1 ako to bolo v prvom prípade, ale **r** sa bude nachádzať niekde medzi **0** a **1.** Je to kladné lebo priamka stúpa a má kladný sklon. Ak by sme vypočítali korelačný koefičient mohol by byť napr. r = 0,8



Ak by sme mali podobnú priamku, ale body by boli od nej viac vzdialené (rozptýlené) hodnota r bude nižšia napr. 0,7 alebo 0,6. Hodnota r by sa teda opäť nachádzala medzi 0 a 1 ale oproti predchádzajúcemu prípadu by bola kvôli väčšiemu rozptilu nižšia.



Teraz nebude existovať žiadna korelácia lebo máme iba všade náhodné body. V takejto situácii je r blízko 0 čo vyjadruje to že neexistuje žiadna zjavná korelácia. Takto nám koeficient hovorí o sile linearneho (priameho) vzťahu medzi dvoma premennými



Ak premenné nebudú mať priamy vzťah r bude blízko 0. Ak existuje silný priamy vzťah r bude blízko 1 a podľa sklonu priamky bude mať vzťah kladnú alebo zápornú hodnotu.

V druhej časti uskutočníme výpočet korelačného koeficientu.

Majme dva štatistické súbory **X = [1, 2, 3, 4, 5, 6] a Y = [2, 4, 7, 9, 12, 14]** ktorých prvky sú reprezentované premennými malé x a y. Úlohou je vypočítať koeficient korelácie medzi týmito dvomi premennými:

1. Pripravíme si hodnoty ktoré budeme pri výpočte korelačného koeficientu potrebovať

	x	**y**	**xy**	**x^2**	**y^2**
	1	2	2	1	4
	2	4	8	4	16
	3	7	21	9	49
	4	9	36	16	81
	5	12	60	25	144
	6	14	84	36	196
$\sum$	21	48	211	91	490
	$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$

2. Použijeme vzorec pre výpočet korelačného koeficientu ktorý je :

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

Túto rovnicu sa musíš jednoducho naučiť a nájsť si spôsob na to ako si to zapamätať. Dole pod odmocninou sú tie zátvorky ktoré sa násobia rovnaké a odlišujú sa iba tým že najprv je to pre x a potom pre y. V zátvorke je to suma 2-hých mocnín x od ktorej je odpočítaná 2-há mocnina iba sumy x. To prvé číslo je ešte násobené počtom prvkov.

Hore to začína obdobne t.j. nasobkom n za ktorým nasleduje suma nasobku x a y. Od tohto je odpočítaný násobok oboch
súm x a y.

3. Dosadíme pripravené údaje z tabuľky do vzorca

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} = \frac{6x211 - 21x48}{\sqrt{\left[6x91 - (21)^2\right]\left[6x490 - (48)^2\right]}} = \frac{1\,266 - 1\,008}{\sqrt{\left[546 - 441\right]\left[2\,940 - 2\,304\right]}} =$$

$$= \frac{258}{\sqrt{[\,105\,][\,636\,]}} = \frac{258}{\sqrt{66\,780}} = \frac{258}{258{,}42} = 0{,}998$$

Takže hodnota korelačného koeficintu je veľmi vysoká čo naznačuje že existuje veľmi silný lineárny (priamy) vzťah medzi premennými x a y a skutočnosť že je kladná hovorí že sklon je pozitívny **x↑↑y**