# KORELACIJA

STATISTIČKI PRAKTIKUM 2

2

# Problem

Zanima nas razina *statističke povezanosti* dvije pojave/obilježja - možemo li, i koliko dobro, predvidjeti ponašanje jednog obilježja pomoću drugog.
Pitanja:

- jesu li obilježja statistički povezana?
- ako jesu, koliko je ta veza snažna?
- ako jesu, u kojem smjeru se kreće ta veza?
- možemo li matematički modelirati tu vezu?
- postoji li zavisnost među tim obilježjima, u smislu da kretanje jednog obilježja određuje kretanje drugog, ili se kreću zajedno?
  - korelacija vs kauzalnost!
- što ako imamo više od dva obilježja čiju međusobnu povezanost želimo ispitati?

Statističku povezanost mjerimo *kvantitativno* funkcijom koja ovisi o *tipu varijabli* s kojima radimo.

Nakon procjene povezanosti slijedi analiza kvalitete modela (tj. procjene), eng. *goodness of fit*.

# Tipovi varijabli

- **kvalitativne varijable** (kategorijske)
  - poprimaju konačno mnogo vrijednosti (označene slovima ili brojevima) koje predstavljaju moguća stanja/kategorije koje varijabla može poprimiti
  - među kategorijama ne mora postojati uređaj
- **kvantitativne varijable** (numeričke, neprekidne)
  - vrijednosti su brojevi na nekom intervalu ili jako velikom skupu (npr. $[0, 1]$, $\mathbb{R}$, $\mathbb{N}$)
  - među vrijednostima postoji uređaj
  - skup vrijednosti može biti i konačan i beskonačan, ali u pravilu je velik (u suprotnom varijable u pravilu smatramo kategorijskima)

# Kako mjerimo povezanost?

| | kvalitativna | kvantitativna |
|---|---|---|
| kvalitativna | $\chi^2$-test;<br>Cramerov V koeficijent | ANOVA;<br>Kruskal-Wallisov test;<br>logistička regresija |
| kvantitativna | ANOVA;<br>Kruskal-Wallisov test;<br>logistička regresija | Pearsonov koeficijent;<br>Spearmanov koeficijent;<br>linearna regresija |

# Cramerov V koeficijent

- jačina veze između dvije kvalitativne varijable
- poprima vrijednost u intervalu $[0, 1]$ (0 - nema povezanosti, 1 - savršena povezanost)
- primjenjiv na kontingencijskim tablicama raznih dimenzija (može se koristiti za uspoređivanje raznih $\chi^2$ statistika)
- veličina uzorka ne utječe na rezultat (korisno kada sumnjamo da je povezanost posljedica velikog uzorka)

$$V = \sqrt{\frac{\chi^2}{n(q-1)}},$$

- $\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{j\cdot}}{n}\right)^2}{\frac{n_{i\cdot} n_{j\cdot}}{n}}$
- $q = \min\{r, c\}$ (manji od broja redova i broja stupaca kontingencijske tablice; tj. broja kategorija svake varijable)
- $n =$ ukupna veličina uzorka

# Cramerov V koeficijent

Interpretacija:

- $V \in [0, 0.1] \Rightarrow$ nema povezanosti
- $V \in \langle 0.1, 0.3] \Rightarrow$ slaba povezanost
- $V \in \langle 0.3, 0.5] \Rightarrow$ srednja razina povezanosti
- $V \in \langle 0.5, 1] \Rightarrow$ jaka povezanost

Simetrična mjera (nije bitan redoslijed varijabli).

```
> library(rcompanion)
> cramerV()
```

# Pearson vs. Spearman

**Pearsonov koeficijent korelacije**

- $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \in [-1, 1]$
- stupanj linearne povezanosti $Y = aX$
- povećanje jedne varijable dovodi do linearnog povećanja/smanjenja ($a > 0/a < 0$) druge varijable
- nezavisnost $\Longrightarrow$ nekoreliranost, ali obrat ne vrijedi!

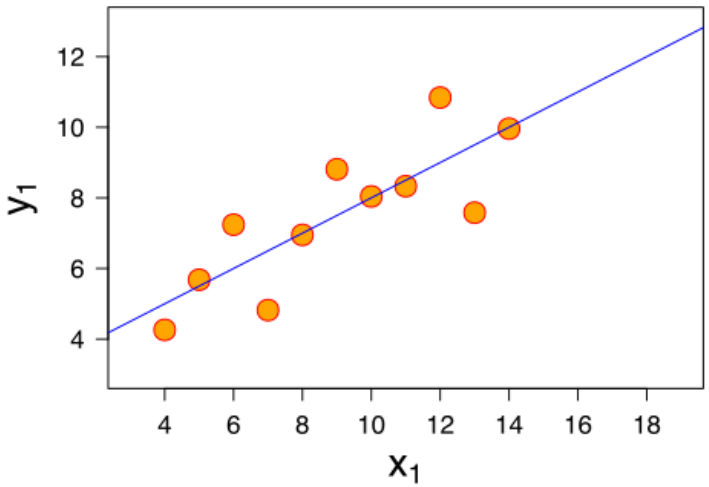
# Pearson vs. Spearman

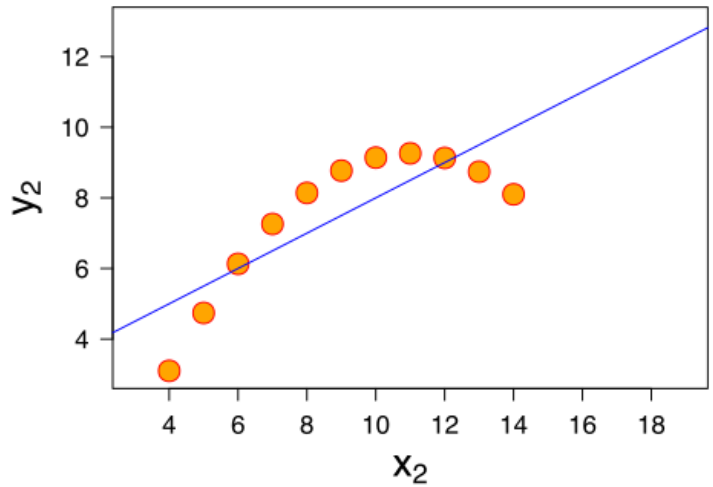
**Spearmanov koeficijent monotone povezanosti**

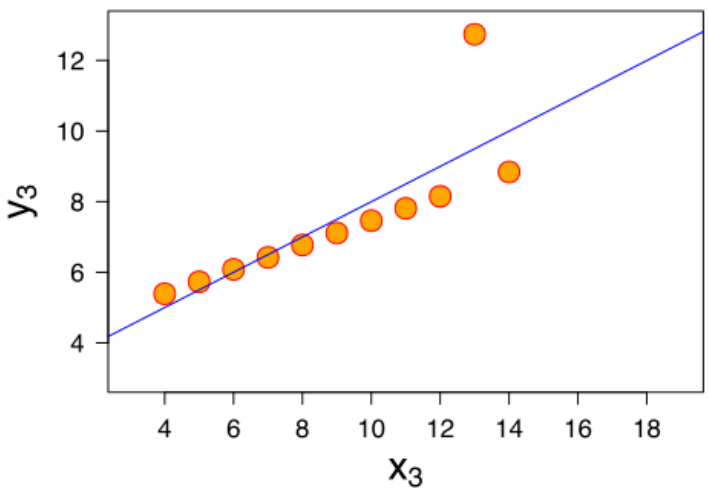
- stupanj *monotone povezanosti* dvije varijable (povećanje jedne varijable dovodi do povećanja ili smanjenja druge varijable, ali ne nužno proporcionalno)
- varijable mogu biti povezane i nekom drugom funkcijom (kretanje jedne varijable utječe na kretanje druge varijable)
- vrijednosti su u intervalu $[-1, 1]$
- Spearmanov koeficijent korelacije = Pearsonov koeficijent korelacije rangova tih varijabli
- ako u uzorcima nema ponavljajućih vrijednosti, vrijednosti $\pm 1$ predstavljaju situaciju kada je jedna varijabla jednaka savršenoj monotonoj funkciji druge varijable

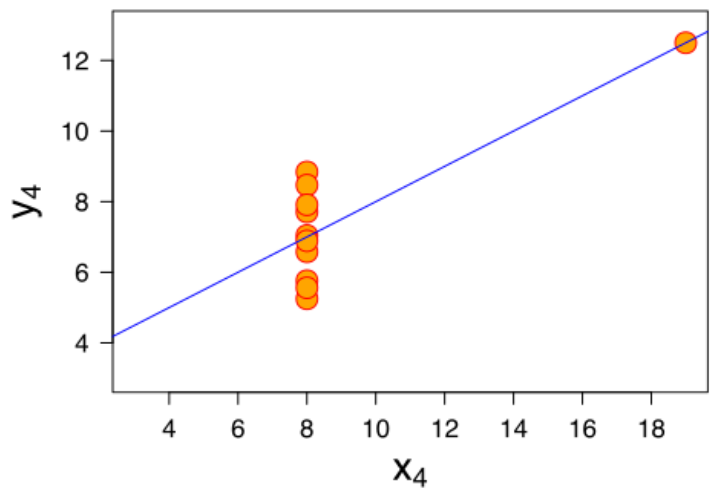
# Primjeri

- $x = (0, 10, 101, 102)$, $y = (1, 100, 500, 2000)$
- zanimljiv primjer s Wikipedije



# Zadatak

Učitajte podatke iz tablice `podaci1.csv`.

(a) Odredite tipove varijabli.

(b) Jesu li spol, stručna sprema i regija statistički povezane varijable? Koja od preostale dvije varijable više utječe na stručnu spremu osobe?

(c) Ovisi li plaća osobe o njenom spolu ili stručnoj spremi?

(d) Na razini značajnosti od 5%, jesu li dob i plaća osobe korelirane?