KORELACIJA

STATISTIČKI PRAKTIKUM 2

1. VJEŽBE

Problem

Zanima nas razina *statističke povezanosti* dvije pojave/obilježja - možemo li, i koliko dobro, predvidjeti ponašanje jednog obilježja pomoću drugog.
U skladu s tom idejom, postavljaju se sljedeća pitanja:

- jesu li obilježja statistički povezana?
- ako jesu, koliko je ta veza snažna?
- 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?
- što ako imamo više od dva obilježja čiju međusobnu povezanost želimo ispitati?

Da bismo ispitali statističku povezanost, moramo imati neku funkciju koja će tu povezanost *mjeriti*, tj. moramo povezanost *kvantitativno* opisati. Dakle, rezultat mjerenja će biti opisan brojevno i veličina tog broja će dati odgovor na naša pitanja.

- funkcija ne može biti ista za sve varijable, odabir funkcije ovisit će 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 neki 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?

X \ Y	kvalitativna	kvantitativna
kvalitativna	χ^2-test; Cramerov V koeficijent	ANOVA; Kruskal-Wallisov test; logistička regresija
kvantitativna	ANOVA; Kruskal-Wallisov test; logistička regresija	Pearsonov koeficijent korelacije; Spearmanov koeficijent korelacije; linearna regresija

Cramerov V koeficijent

Opisuje jačinu veze između dvije kvalitativne varijable, a poprima vrijednost u intervalu $[0, 1]$ (1 prikazuje savršenu povezanost).

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

- $q =$ manji od broja redova i broja stupaca kontingencijske tablice

Primjenjiv na kontingencijskim tablicama raznih dimenzija pa se stoga može koristiti za uspoređivanje raznih χ^2 statistika. Također, na rezultat ne utječe veličina uzorka (koristan je u situaciji kada sumnjamo da bi povezanost mogla biti posljedica toga što imamo veliki uzorak, a ne stvarne statističke povezanosti).

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

Pearson vs. Spearman

Pearsonov koeficijent korelacije

- korelacija - linearna povezanost: $Y = aX$
- povećanje jedne varijable dovodi do linearnog povećanja (ako je $a > 0$) ili smanjenja (ako je $a < 0$) druge varijable
- koeficijent korelacije: $\rho \in [-1, 1]$ daje stupanj linearne povezanosti
- MLE procjenitelj za ρ je Pearsonov koeficijent korelacije $\hat{\rho}$

Pearson vs. Spearman

Spearmanov koeficijent monotone povezanosti

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

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?