

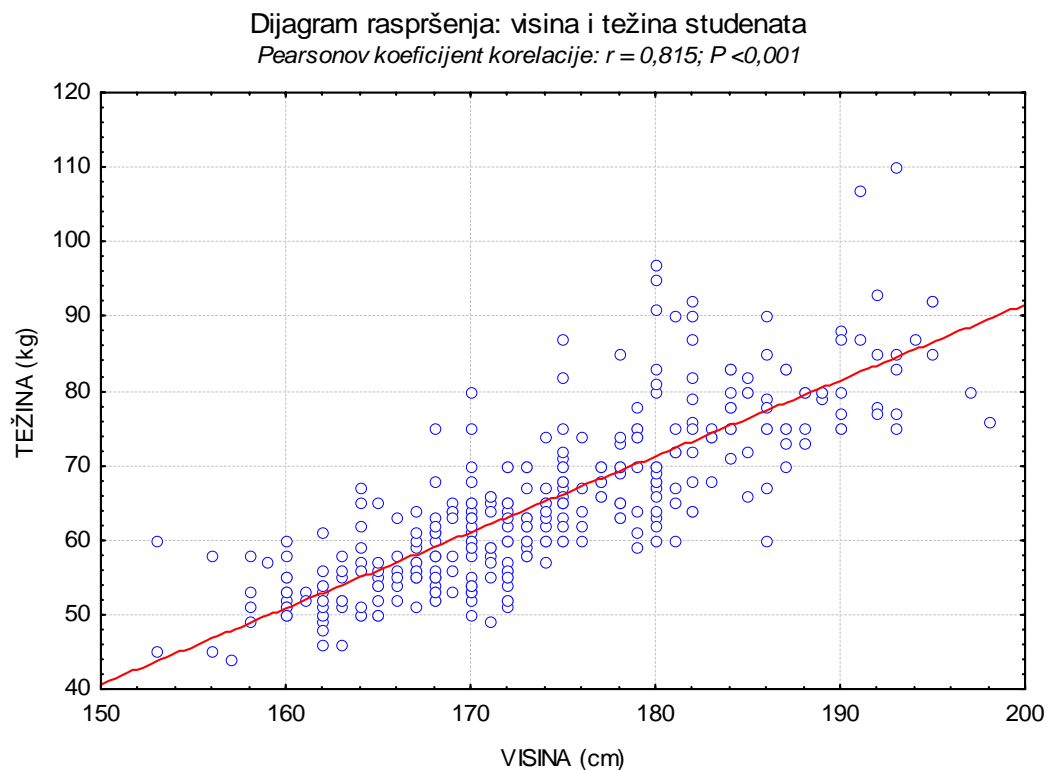
KORELACIJA (POVEZANOST)

Mirjana Kujundžić Tiljak i Davor Ivanković

Sukladnost u variranju vrijednosti dvaju ili više varijabli naziva se *korelacija*. Varijable su pri tom numeričke (kvantitativne), npr. visina i težina.

Potpuna korelacija ili funkcionalna veza postoji kada svakoj vrijednosti varijable x odgovara samo jedna vrijednost u drugoj varijabli y . Djelomična korelacija znači da određenoj vrijednosti varijable x odgovara više različitih vrijednosti varijable y . Što je korelacija manja, to je veća varijabilnost vrijednosti varijable y koje se pojavljuju uz neku određenu vrijednost varijable x . Između tjelesne visine i tjelesne težine u ljudi postoji korelacija. Viši ljudi su i teži, ali samo u prosjeku, jer svi ljudi iste visine nisu jednako teški.

Pretpostavimo da imamo par vrijednosti (x,y) , izmjerene na svakom od n ispitanika u našem uzorku. Možemo nacrtati pripadajuće točke svakog pojedinačnog para vrijednosti u dvodimenzionalnom *točkastom dijagramu* ili *dijagramu raspršenja* (engl., dotted /scatter diagram). Konvencionalno, na horizontalnoj osi je x , a na vertikalnoj y varijabla. Crtanjem točaka za n ispitanika, dobijemo njihov razmještaj koji može sugerirati povezanost dvije varijable.



Slika 1. Dijagram raspršenja visine i težine studenata prve godine studija medicine (generacija 1998 godine; $N=347$; 141 student i 206 studentica)

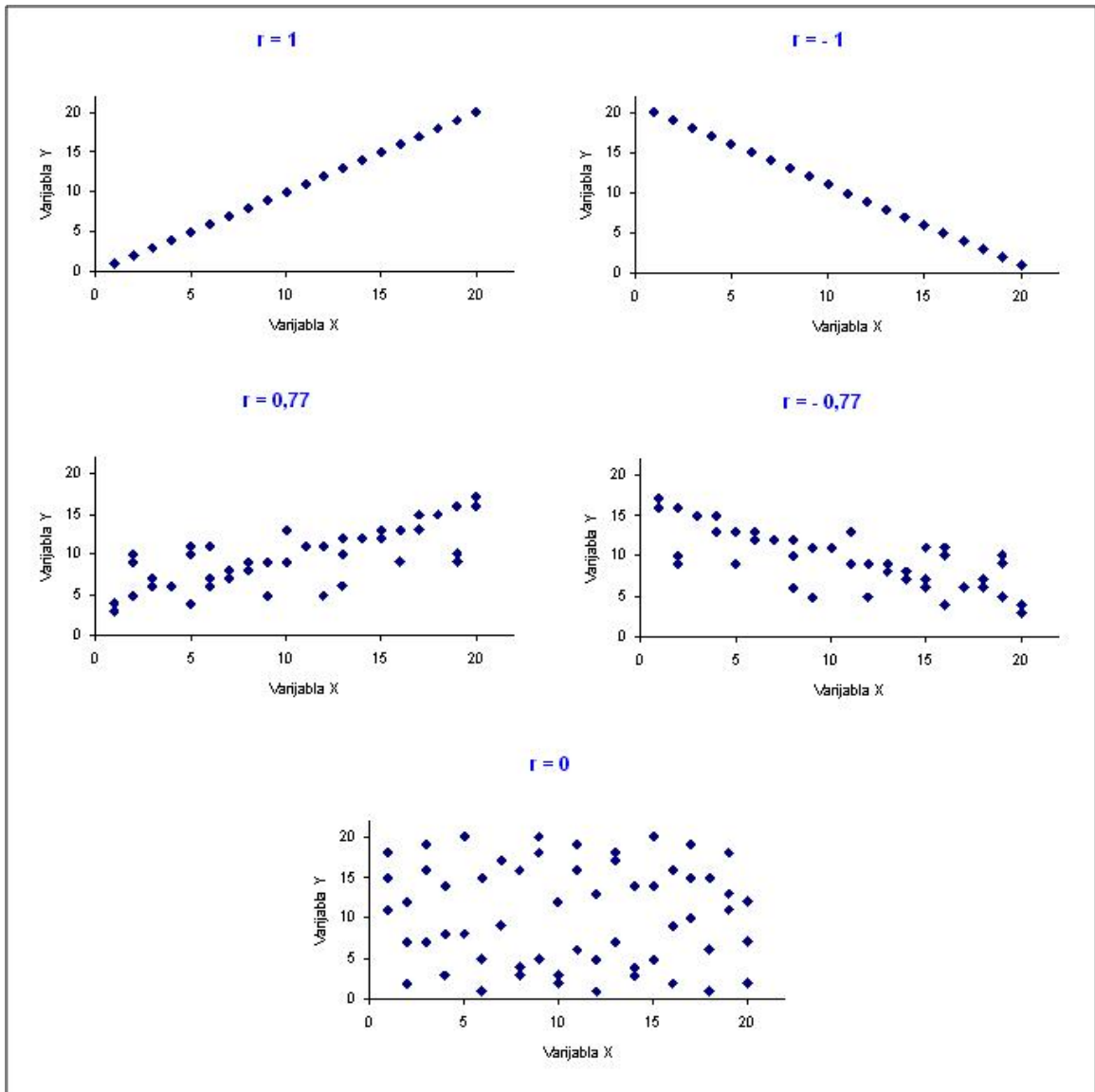
PEARSONOV KOEFICIJENT KORELACIJE (engl., Pearson correlation coefficient)

Možemo reći da postoji *linearna povezanost* između x i y ako je nacrtan pravac kroz točke koji omogućava najprikladniju procjenu opaženog odnosa. Mi mjerimo koliko su naša opažanja blizu pravcu koji najbolje opisuje njihovu linearnu povezanost računanjem *Pearsonovog koeficijenta korelacije umnožaka* (engl. Pearson product moment correlation coefficient), najčešće jednostavno zvan *koeficijent korelacije* (engl. Correlation coefficient). Njegova točna vrijednost u populaciji, ρ , procijenjuje se u uzorku s r , gdje je

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 (y - \bar{y})^2}}, \text{ što uobičajeno izračuna računalo.}$$

Obilježja Pearsonovog koeficijenta korelacije:

- *vrijednost r* je u rasponu *od -1 do +1*;
- *predznak r* ukazuje na smjer korelacije: *pozitivna korelacija (pozitivan r)* znači da kako vrijednost jedne varijable raste, tako raste i vrijednost druge varijable; *negativna korelacija (negativan r)* znači da kako vrijednost jedne varijable raste, tako pada vrijednost druge varijable;
- *veličina r* ukazuje na to koliko su točke blizu pravcu: ako je $r = +1$ ili $r = -1$, tada postoji potpuna povezanost sa svim točkama koje leže na pravcu (to je u praksi gotovo nemoguće); ako je $r = 0$, tada nema linearne povezanosti (premda može postojati ne-linearna povezanost); što je r bliži ekstremnim vrijednostima (+1, -1) to je i stupanj linearne povezanosti veći;
- r nema mjeru niti jedinicu mjerenja;
- korelacija između x i y nužno ne implicira „uzročno-posljedičnu“ vezu (engl. „cause and effect“ relationship);
- r^2 predstavlja udio varijabiliteta y koji se može pripisati njegovoj linearnoj povezanosti s x , a naziva se *koeficijent determinacije*.



Slika 2. Vrijednosti Pearsonovog koeficijenta korelacije

Pogrešno je računati Pearsonov koeficijent korelacije, r :

- kada postoji ne-linearna povezanost dvije varijable, kao npr. kvadratna povezanost (quadratic relationship)
- kada podaci uključuju više od jednog opažanja za svakog ispitanika,
- kada postoji jedan ili više „nepodobnih članova grupe“ (engl. outliers)
- kada se podaci sastoje od subgrupa pojedinaca za koje je prosječna vrijednost opažanja za barem jednu od varijabli različita.

TESTIRANJE HIPOTEZA ZA PEARSONOV KOEFICIJENT KORELACIJE

Želimo saznati da li postoji ikakva linearna povezanost (korelacija) između dvije numeričke varijable. Naš uzorak se sastoji od n nezavisnih parova vrijednosti x i y . Pretpostavljamo da barem jedna od dvije varijable slijedi normalnu distribuciju.

Postupak testiranja hipoteza za Pearsonov koeficijent korelacije:

1. Definirajte nul- i alternativnu hipotezu istraživanja
 $H_0: \rho=0$
 $H_1: \rho \neq 0$
2. Prikupite relevantne podatke na uzorku ispitanika
3. Izračunajte vrijednost statističkog testa specifičnog za H_0
Izračunajte r
 - Ako $n \leq 150$, r je vrijednost statističkog testa
 - Ako $n > 150$, izračunaj $T = \sqrt{\frac{(n-2)}{(1-r^2)}}$
koja slijedi t-distribuciju sa $n-2$ stupnjeva slobode
4. Usporedite vrijednost statističkog testa s vrijednostima poznate distribucije vjerojatnosti
 - Ako $n \leq 150$ upotrijebi Dodatak A10
 - Ako $n > 150$ upotrijebi dodatak A2
5. Interpretirajte P-vrijednost i rezultate
Izračunajte interval pouzdanosti za ρ . Obje date varijable otprilike slijede normalnu distribuciju, približno 95% interval pouzdanosti za ρ je

$$\left(\frac{e^{2z_1} - 1}{e^{2z_1} + 1}, \frac{e^{2z_2} - 1}{e^{2z_2} + 1} \right)$$

$$\text{gdje } z_1 = z - \frac{1,96}{\sqrt{n-3}}, \quad z_2 = z + \frac{1,96}{\sqrt{n-3}}$$

$$\text{i } z = 0,5 \log_e \left[\frac{(1+r)}{(1-r)} \right]$$

Imajte na umu da, ako je uzorak velik, H_0 se može odbaciti čak i kada je r vrlo blizu nuli. S druge strane, iako je r velik, H_0 se ne mora odbaciti ako je uzorak mali. Iz tog razloga, posebno je korisno izračunati r^2 , proporciju totalne varijance jedne varijable objašnjene s njezinom linearnom povezanošću s drugom varijablom. Na primjer, ako $r=0.40$ tada je $P < 0,05$ za uzorak veličine 25, ali povezanost jedino objašnjava 16% ($r^2=0,40^2 \times 100$) varijabilnosti jedne varijable.

SPEARMANOV KOEFICIJENT KORELACIJE RANGOVA (engl., Spearman 's rank correlation coefficient)

Spearmanov koeficijent korelacije rangova, neparametrijski ekvivalent Pearsonovom koeficijentu korelacije, računamo ako je točan jedan ili više od sljedećih navoda:

- barem jedna od varijabli, x ili y , mjerena je ordinalnom skalom;
- niti x niti y ne slijede normalnu distribuciju;
- uzorak je mali;
- trebamo mjeru povezanosti između dvije varijable kada ta povezanost nije linearna.

Za procjenu populacijske vrijednosti Spearmanovog koeficijenta korelacije rangova ρ_s , koristimo njegovu vrijednost izračunatu na uzorku, r_s :

1. Uredimo vrijednosti x u rastući niz počevši od najmanje vrijednosti, i dodijelimo im sukcesivne rangove (brojeve $1, 2, 3, \dots, n$). Jednake vrijednosti primaju prosjek rangova vrijednosti koje bi te vrijednosti poprimile kada ne bi bile identične.
2. Dodijelimo rangove vrijednostima varijable y na isti način;
3. r_s je Pearsonov koeficijent između rangova x i y .

Obilježja Spearmanovog koeficijenta korelacije:

- ima jednaka obilježja kao i Pearsonov koeficijent korelacije, označava se s r_s ;
- r_s omogućuje mjeru povezanosti, ne nužno linearne, između varijabli x i y ;
- ne računamo r_s^2 budući da on ne predstavlja proporciju ukupne varijacije u jednoj varijabli koja bi mogla biti pribrojena njezinoj linearnoj povezanosti s drugima.

Literatura:

1. *Ivanković D, i sur. Osnove statističke analize za medicinare. Zagreb: Medicinski fakultet Sveučilišta u Zagrebu, 1989.*
2. *Petrie A, Sabin C. Medical Statistics at a Glance (2nd Ed). Oxford: Blackwell Science Ltd, 2005.*
3. *Glantz. SA. Primer of Biostatistics (4th Ed). New York: McGraww-Hill: 1997.*
4. *Altman DG. Practical Statistics for Medical Research. London. Chapman & Hall, 1991.*
5. *Bland M. An Introduction to Medical Statistics (3rd Ed). Oxford: Oxford University Press, 2005.*
6. *Armitage P, Berry P. Statistical Methods in Medical Research. Oxford: Blackwell Science Ltd, 1994.*