Veze između promjenljivih

Dijagrami raspršenosti i korelacija

Bojan Baškot | EFBL | 2026.

Plan predavanja

Dio 1: Dijagram raspršenosti

1. Šta je dijagram raspršenosti?
2. Motivacijski primjer: ishrana u SAD
3. Zdravlje i bogatstvo nacija
4. Interpretacija: pravac, oblik, snaga
5. Pozitivna i negativna asocijativnost
6. Outlier-i i višestruke veze

Dio 2: Korelacija

1. Šta je korelacija r?
2. Primjer: Arheopteriks (fosili)
3. Izračun korelacije korak po korak
4. Pearsonov koeficijent — formula
5. Važna svojstva korelacije
6. Zamke i ograničenja

Ključno pitanje

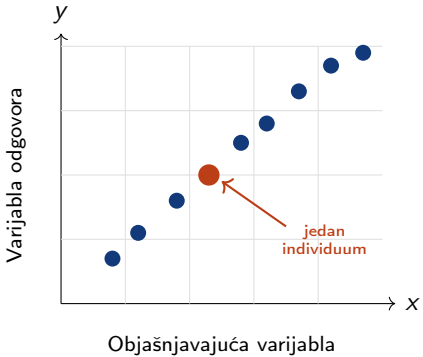
Kako opisujemo i mjerimo vezu između dvije kvantitativne varijable?

Dijagram raspršenosti (Scatter plot) — šta je?

Definicija

Dijagram raspršenosti pokazuje odnos između **dvije kvantitativne varijable** mjerenih na *istim* osobama/jedinicama.

- **Horizontalna os (x):** objašnjavajuća varijabla
- **Vertikalna os (y):** varijabla odgovora
- **Svaki pojedinac** = jedna tačka na grafikonu



Napomena

Ako nema jasne objašnjavajuće varijable, može se koristiti bilo koja os za bilo koju varijablu.

Motivacijski primjer — Ishrana u SAD

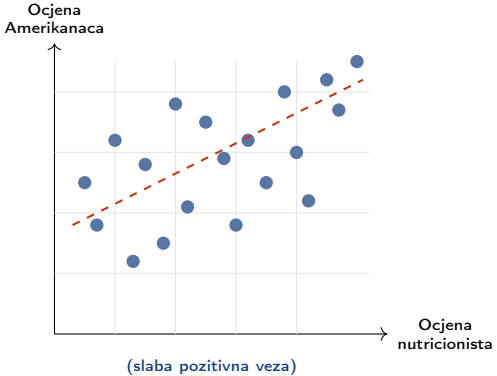
Kontekst istraživanja

Svake 5 godina, Ministarstvo zdravlja (HHS) i Ministarstvo poljoprivrede (USDA) objavljuju *Smernice o ishrani za Amerikance*.

Pitanje: Da li Amerikanci prepoznaju zdravu hranu isto kao stručnjaci za ishranu?

Metodologija

- Lista od **52 namirnice**
- **2 000** učesnika — obični Amerikanci
- **672 stručnjaka** — Američko društvo za ishranu
- Svaki ocjenjuje koliko je namirnica "zdrava"



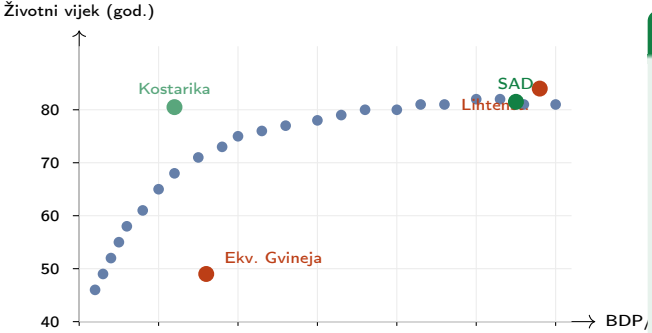
Zaključak

Veza postoji, ali nije savršena — stručnjaci i javnost se **ne slažu uvijek**.

Primjer: Zdravlje i bogatstvo nacija

Podaci Svjetske banke (2016.)

- **Individuum:** svaka nacija za koju postoje podaci
- x — **objašnjavajuća:** BDP po glavi stanovnika (USD)
- y — **varijabla odgovora:** očekivani životni vijek pri rođenju



Šta vidimo?

- **Pozitivan trend** — bogatije nacije žive duže
- Rast je **zakrivljen** — brzo raste pa izravna
- **Ekvatorijalna Gvineja:** visok BDP (nafta), ali nizak LE — prihod ide manjem sloju

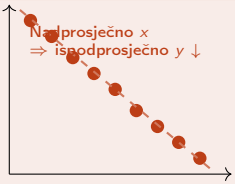
Interpretacija dijagrama raspršenosti

Tri dimenzije opisa

PRAVAC (pozitivno / negativno) → **OBLIK** (linearan / zakrivljen) → **SNAGA** (jak / umjeren / slab)

Pozitivna asocijacija

Negativna asocijacija



Primjeri iz stvarnog života

Pozitivna asocijacija

- Visina oca $\leftrightarrow$ visina sina
- Temperatura $\leftrightarrow$ prodaja sladoleda
- Obrazovanje $\leftrightarrow$ prihod
- Veličina stana $\leftrightarrow$ cijena stana

Više $x \rightarrow$ više y (u prosjeku)

Negativna asocijacija

- Cijena $\leftrightarrow$ potražnja
- Tjelesna težina $\leftrightarrow$ brzina trčanja
- Pušenje $\leftrightarrow$ životni vijek
- Broj grešaka $\leftrightarrow$ ocjena na testu

Više $x \rightarrow$ manje y (u prosjeku)

Koji pravac vidimo na grafikonu?

Scatter plot se **naginje prema gore** (pozitivno) ili **pada prema dolje** (negativno) kako se krećemo slijeva nadesno. Vizuelni pregled je uvijek prvi korak analize!

Šta još tražimo u dijagramu raspršenosti?

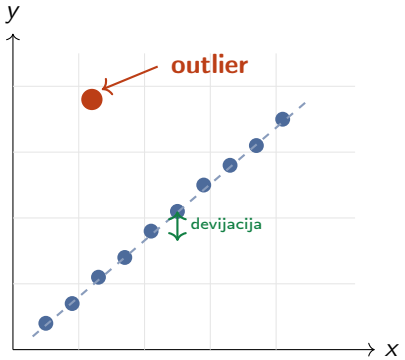
Tri dimenzije opisa

1. **Pravac:** pozitivno ili negativno?
2. **Oblik:** linearna (ravna linija) ili zakrivljena veza?
3. **Snaga:** koliko blizu su tačke linije?

Outlier (izdvojena vrijednost)

Individualna tačka koja **ne prati opšti obrazac**.

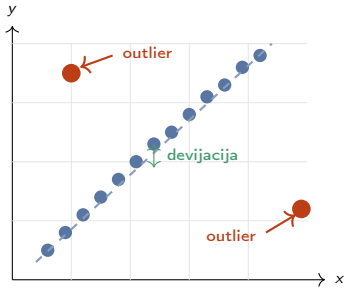
Uvijek je posebno istražiti! Može biti greška u podacima ili genuino zanimljiv slučaj.



Ispitivanje dijagrama raspršenosti

Generalno pravilo

U bilo kom grafikonu podataka potražite **opšti obrazac** i **upadljiva odstupanja** od tog obrasca.



Tri koraka analize

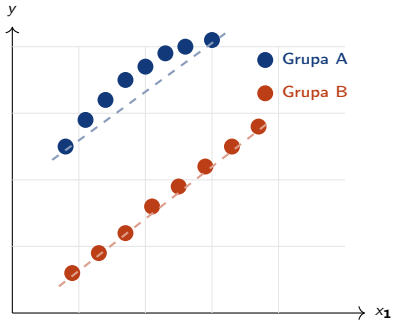
1. **Pravac**
   Raste li y s x (pozitivno) ili pada (negativno)?
2. **Oblik**
   Prate li tačke ravnu liniju ili krivulju?
3. **Snaga**
   Koliko su tačke blizu zamišljene linije?

Outlier

Povezanost više varijabli

Kada jedna varijabla nije dovoljna

Odnosi između varijabli mogu biti komplikovani. Jedan odgovor može se objasniti **kombinacijom** više varijabli. Scatter plot može prikazati i **treću varijablu** — bojom ili simbolom tačaka.



Primjer: vježba i zdravlje

- x_1: sati vježbe sedmično
- y: indeks tjelesne mase (BMI)
- Treća varijabla (boja): **pol**

Muškarci (plavo) i žene (crveno) mogu imati **isti trend** ali **različite nivoe**.

Zaključak

Ignorisanje treće varijable može dati

Korelacija

Mjerenje snage i pravca linearne veze

$r \in [-1,\ 1]$

Korelacija — šta je i zašto je koristimo?

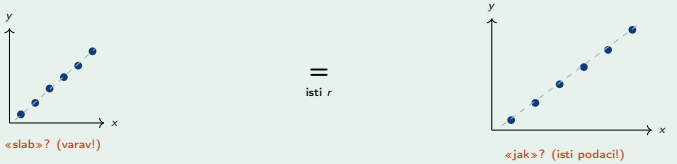
Definicija

Korelacija opisuje **pravac** i **jačinu linearnog odnosa** između dvije kvantitativne varijable.
Označava se sa r.

Zašto nam treba broj r?

- Scatter plot daje vizualnu sliku, ali **naše oči su loši suci** snage veze
- Dva grafa mogu izgledati potpuno različito samo zbog **razmjere osi**, iako prikazuju iste podatke i istu vezu
- r daje **jedinstven broj** koji mjeri snagu i pravac veze, nezavisno od mjernih jedinica i razmjere

Ilustracija: Isti podaci, različita razmjera osi



Razmjera osi vara naše oči — r je objektivan mjerač snage veze!

Primjer: Arheopteriks — fosilni ostaci

Kontekst

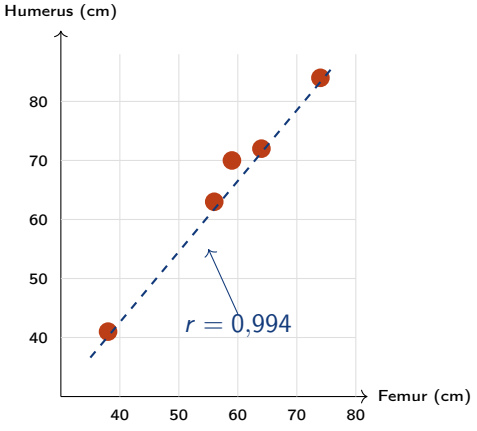
Arheopteriks je izumrla zvjer s perjem poput ptice, ali sa zubima i koštanim repom poput reptila. Poznato je samo **šest fosilnih primjeraka**.

Pitanje: Da li svi fossili pripadaju *istoj vrsti*?

Podaci (dužine u cm)

Femur	38	56	59	64	74
Humerus	41	63	70	72	84

Femur = butna kost (noga)
Humerus = nadlaktična kost (ruka/krilo)



Zaključak

Snažna, poz., pravolin. asocijacija. Fosili vjerovatno iste vrste - razlikuju se po starosti/veličini.

Interpretacija grafa — Arheopteriks

Šta nam govori scatter plot?

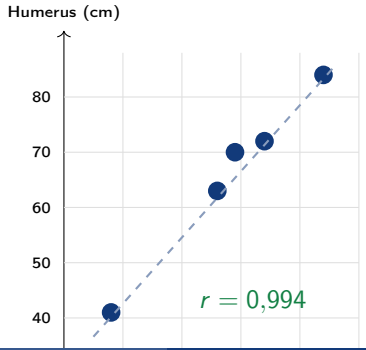
- **Oblik:** pravolinijski (linearan) — tačke prate ravnu liniju
- **Snaga:** jaka — tačke leže blizu linije
- **Pravac:** pozitivna — duži femur $\Rightarrow$ duži humerus

Biološka interpretacija

- Sva 5 fosila prate **isti razmjer** kostiju
- Razlike su u **veličini**, ne u vrsti
- Mlađe životinje = manje kosti (proporcijalno)
- Eventualni 6. primjerak koji *ne prati*

Ključni zaključak

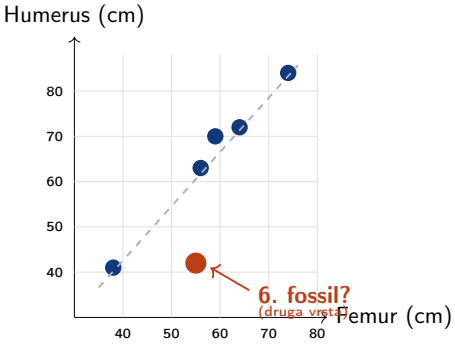
Jaka linearna asocijacija između dvije kosti sugerira da svi primjerci pripadaju **istoj vrsti** i da se razlikuju po starosti, ne po vrsti.



Šta bi se desilo s outlier-om? — Arheopteriks

Hipoteza

Ako bi 6. fosilni primjerak imao **neproporcionalne kosti** u odnosu na ostale, to bi ukazivalo da pripada **drugoj vrsti**.



Logika zaključivanja

- Crvena tačka ima **sličan femur** kao ostali, ali **mnogo kraći humerus**
- Ne prati isti razmjer kostiju
- Takav fosilni primjerak bio bi **outlier** na grafikonu
- To bi bio statistički dokaz za **drugu vrstu**

Zaključak

Scatter plot nije samo alat za opis — može biti i alat za **naučne zaključke**!

Izračun korelacije — Koraci 1 i 2

Korak 1: Srednja vrijednost i standardna devijacija

Femur (x):

$$\bar{x} = \frac{38 + 56 + 59 + 64 + 74}{5} = 58{,}2 \text{ cm}$$

$s_x = 13{,}20$ cm

Humerus (y):

$$\bar{y} = \frac{41 + 63 + 70 + 72 + 84}{5} = 66{,}0 \text{ cm}$$

$s_y = 15{,}89$ cm

Korak 2: Standardni rezultati (z-score) za svaki par

Fossil	x	$z_x = (x - \bar{x})/s_x$	y	$z_y = (y - \bar{y})/s_y$	$z_x \cdot z_y$
1	38	$-1{,}53$	41	$-1{,}57$	$+2{,}40$
2	56	$-0{,}17$	63	$-0{,}19$	$+0{,}03$
3	59	$+0{,}06$	70	$+0{,}25$	$+0{,}02$
4	64	$+0{,}44$	72	$+0{,}38$	$+0{,}17$
5	74	$+1{,}20$	84	$+1{,}13$	$+1{,}36$
				$\sum z_x \cdot z_y$:	3,98

Izračun korelacije — Korak 3 i tumačenje

Korak 3: Konačna korelacija

$$r = \frac{\sum z_x \cdot z_y}{n-1} = \frac{3{,}98}{4} = 0{,}994$$

Tumačenje rezultata

- $r = 0{,}994$ — gotovo savršena pozitivna linearna veza
- Tačke leže **izuzetno blizu** trenda linije
- Svih 5 fosila prate isti razmjer femur/humerus
- Snažan statistički argument da su **iste vrste**

Zašto dijeli s $n-1$?

Dijeljenjem s $n-1$ (a ne n) dobijamo **nepristrasnu procjenu** korelacije u populaciji na osnovu uzorka.

Isti princip kao kod standardne devijacije uzorka s.

Zapamtite

Uvijek prvo **nacrtaj scatter plot**, pa tek onda računaj r!

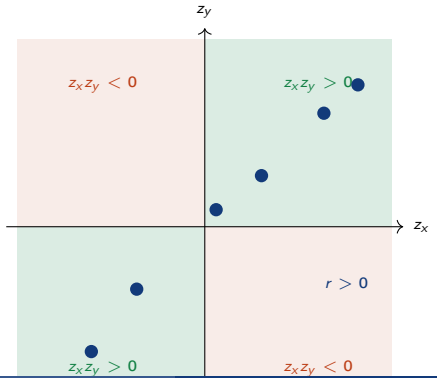
Pearsonov koeficijent korelacije — formula

Opšta formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Intuicija formule

- Standardizujemo svaki x i y (pretvorimo u z-score)
- Množimo z-score parove: ako oboje *idu zajedno* (oba visoka ili oba niska), produkt je **pozitivan**
- Ako idu u *suprotnim smjerovima*, produkt je **negativan**
- Prosječek tih produkata daje r



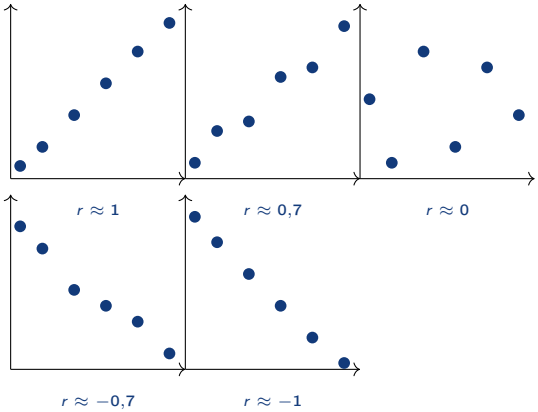
Važna svojstva korelacije r

Matematička svojstva

1. $-1 \leq r \leq 1$ **uvijek**
2. $r > 0 \rightarrow$ pozitivna asocijacija
3. $r < 0 \rightarrow$ negativna asocijacija
4. $r = 0 \rightarrow$ nema *linearne* veze
5. $|r| = 1 \rightarrow$ savršena linearna veza

Nezavisnost od mjernih jedinica

Ako femur mjerimo u inčima umjesto cm, r ostaje **isti** — jer radimo s z-score-ovima koji nemaju mjerne jedinice.



Svojstva korelacije — nastavak

r je simetričan

Korelacija ne pravi razliku između objašnjavajuće varijable i varijable odgovora.

Ako zamijenimo x i y mjesta, r **ostaje isti**.

$r(femur, humerus) = r(humerus, femur) = 0{,}994$

r mjeri samo **linearnu** vezu!

Korelacija ne opisuje zakrivljene odnose, ma koliko jaki bili.

Primjer: BDP i životni vijek — veza postoji, ali je zakrivljena $\Rightarrow$ r je podcijenjen!

Ilustracija: zakrivljena veza, nizak r

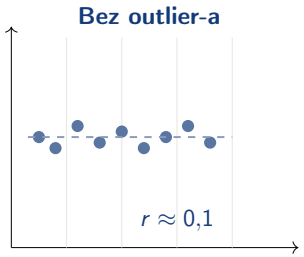
Poruka

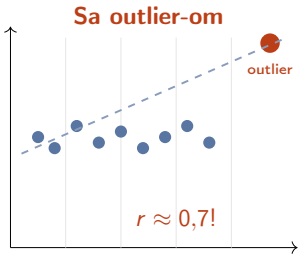
Uvijek pogledaj scatter plot — $r = 0$ ne znači automatski **nema veze**, znači samo nema **linearne** veze!

Uticaj outlier-a na korelaciju

Outlier-i mogu dramatično promijeniti r!

Jedna tačka koja je daleko od opšteg obrasca može **povećati ili smanjiti** korelaciju — ponekad do suprotnog predznaka.





Preporuka

Uvijek nacrtaj scatter plot prije nego računaš r. Nikad ne tumači korelaciju bez vizuelnog pregleda podataka!

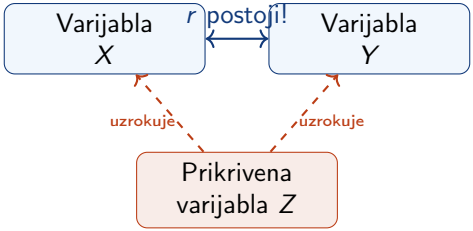
Korelacija $\neq$ uzročnost!

Najvažnija zamka u statistici

Visoka korelacija između x i y **ne znači** da x uzrokuje y.

Primjeri lažnih korelacija

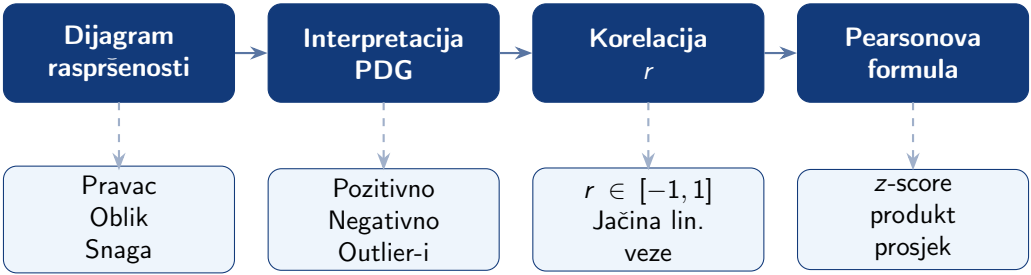
- Prodaja sladoleda $\leftrightarrow$ Utapanja
  (treća varijabla: ljeto/toplina)
- Broj vatrogasaca $\leftrightarrow$ Šteta od požara
  (treća varijabla: veličina požara)
- Cipele djeteta $\leftrightarrow$ Pismenost
  (treća varijabla: starost djeteta)
- Broj TV-a u kući $\leftrightarrow$ Dugovječnost
  (treća varijabla: životni standard)



Zaključak

Uzročnost se može tvrditi **samo** na osnovu dobro dizajniranog eksperimenta — ne na osnovu korelacije!

Sažetak — Veze između promjenljivih



Zapamtite

- Uvijek prvo **nacrtaj** scatter plot!
- r mjeri **samo linearnu** vezu
- Outlier-i mogu **iskriviti** r

Korelacija $\neq$ uzročnost

- Prikrivena varijabla može objasniti vezu
- Uzročnost traži **eksperiment**
- Scatter plot pomaže otkriti lažne veze

Hvala na pažnji!

Pitanja i diskusija

"Korelacija nije uzročnost —
ali je dobar početak za istraživanje."

— *osnovna ideja analize podataka*