

Statistika

Vanja Wagner, Ivan Biočić, Hrvoje Planinić
na temelju materijala B. Basraka i M. Marušića

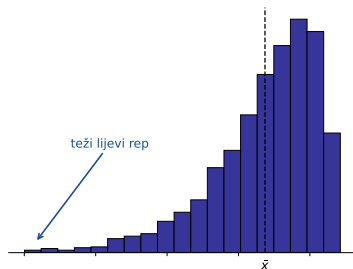
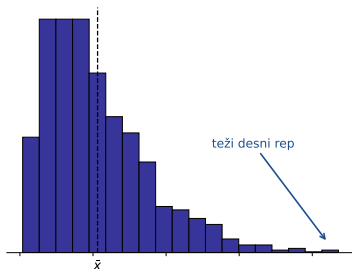
2. Deskriptivna statistika

Repovi razdiobe

Za podatke koji su značajno udaljeni od aritmetičke sredine $\bar{x}$ kažemo da pripadaju **repovima** razdiobe podataka.

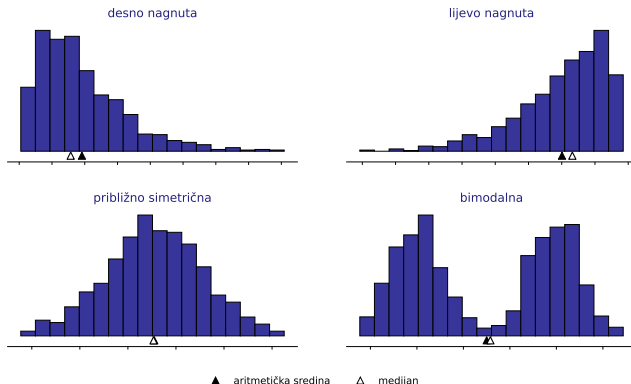
Grubo govoreći, jedan je rep **teži** ako su podaci na toj strani aritmetičke sredine značajnije udaljeni od nje nego podaci na drugoj strani. Drugi rep tada nazivamo **lakšim**.

Repove možemo usporediti **gledajući histogram**.



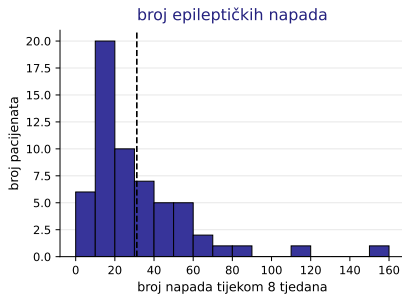
Oblik razdiobe uzorka

Histogrami numeričkih podataka mogu imati vrlo različite oblike.



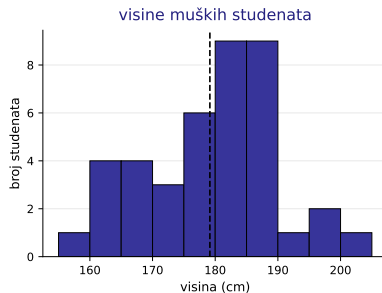
- jedan lokalni maksimum $\rightsquigarrow$ **"unimodalna"** razdioba (sve osim razdiobe dolje desno);
- teži desni rep $\rightsquigarrow$ **"desno nagnuta"**;
- kod približno simetričnih razdioba (donje dvije) medijan i prosjek se često ne razlikuju previše.

Primjeri repova razdiobe



Broj epileptičkih napada:

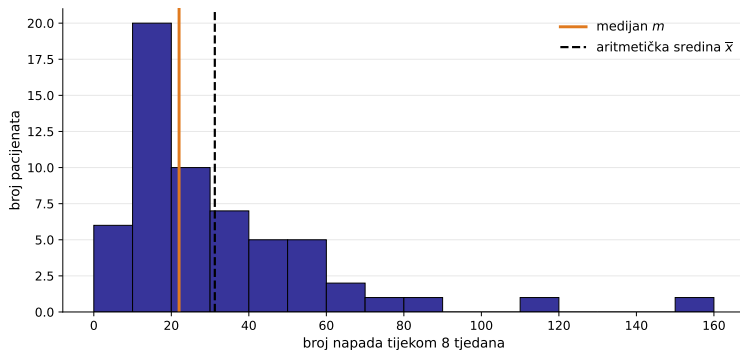
desni rep je teži; nekoliko pacijenata imalo je izrazito velik broj napada.



Visine studenata:

lijevi i desni rep približno su jednaki.

Aritmetička sredina i medijan

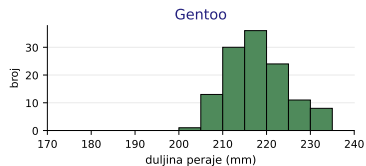
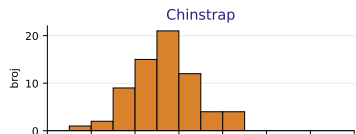
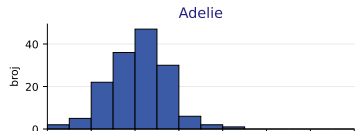
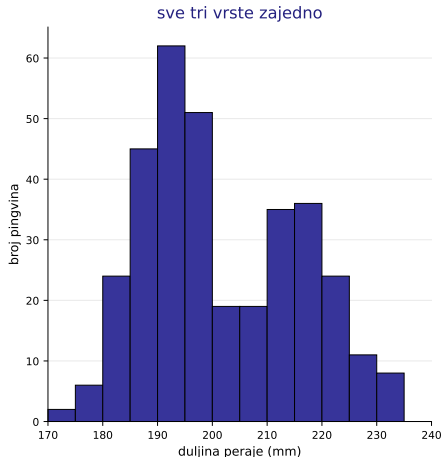


Kod desno nagnutih podataka često je $\bar{x}$ **desno** u odnosu na m zbog veće osjetljivosti na ekstremne vrijednosti.

Napomena: Površina histograma s lijeve i desne strane od medijana je **približno jednaka** (bilo da se radi o pravom histogramu ili histogramu frekvencija kao gore).

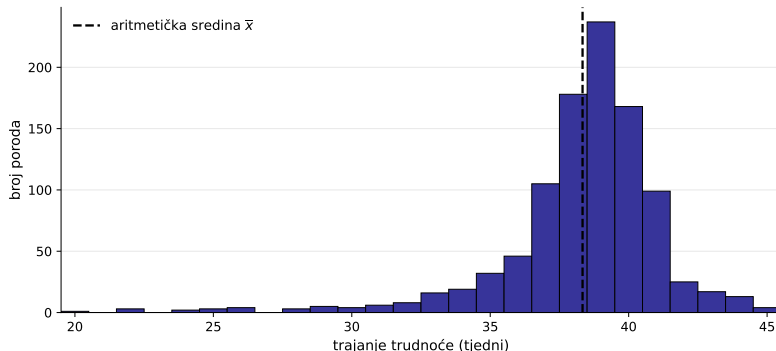
Primjer bimodalne razdiobe

Višemodalnost često upućuje na to da su u uzorku pomiješane **različite skupine**.



Zadatak: trajanje trudnoće

Histogram (apsolutnih frekvencija) prikazuje trajanje trudnoće za uzorak poroda u Sjevernoj Karolini, izraženo u tjednima.



- Je li razdioba približno simetrična, lijevo ili desno nagnuta?
- Kako biste objasnili uočeni oblik razdiobe?
- S koje strane aritmetičke sredine očekujete medijan? Očekujete li da je razlika velika?

Koeficijent asimetrije

- Pitanje je možemo li preciznije mjeriti "nagnutost" razdiobe po uzoru na medijan/prosjek (sredina) i raspršenje (kvantili/varijanca)?
- jedan prijedlog je **koeficijent asimetrije** α_3
- prije definicije, kratko ćemo objasniti kako tzv. **linearne transformacije** mijenjaju razdiobu podataka.

Intermezzo: linearna transformacija podataka

- Često originalne numeričke podatke $x_1, \dots, x_n$ **transformiramo** na sljedeći način:

$$y_i = a + b \cdot x_i,$$

pri čemu je a bilo kakav realan broj, a $b > 0$.

- Primjerice, ako su x_i -evi temperature u stupnjevima Fahrenheita, transformirani podaci

$$y_i := \frac{x_i - 32}{1.8} = -\frac{160}{9} + \frac{5}{9}x_i$$

su temperature u stupnjevima Celzijusa.

- nije teško za pokazati da nakon transformacije vrijedi:

podaci x_i	podaci y_i
$\bar{x}$	$\bar{y} = a + b\bar{x}$
s_x	$s_y = bs_x$
$q_\alpha(x)$	$q_\alpha(y) = a + bq_\alpha(x)$

Primjer – dodavanje konstante

	x_i	$x_i + 10$	$x_i + 20$
	62.30	72.30	82.30
	62.28	72.28	82.28
	69.77	79.77	89.77
	68.95	78.95	88.95
	69.88	79.88	89.88
srednja vrijednost	66.64	76.64	86.64
medijan	68.95	78.95	88.95
gornji kvartil	69.83	79.83	89.83
varijanca uzorka	12.695	12.695	12.695

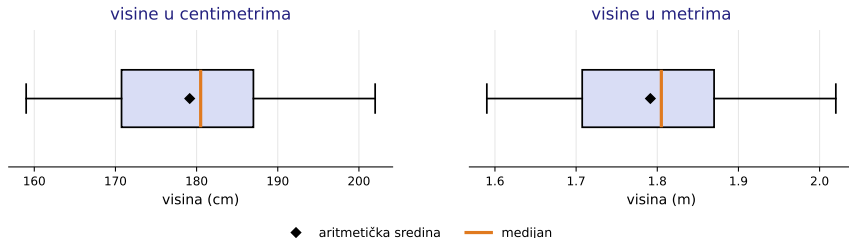
- ako je $y_i = a + x_i$, tj. $b = 1$, imamo $s_y = 1 \cdot s_x = s_x \rightsquigarrow$ standardna devijacija i varijanca **ostaju iste**;
- općenito, ako je $y_i = a + b \cdot x_i$ za $b > 0$, imamo $s_y^2 = b^2 \cdot s_x^2$.

Primjer – promjena skale

Ako su x_i -evi visine u centimetrima, visine izražene u metrima dobivamo transformacijom

$$y_i = \frac{x_i}{100}.$$

dakle, $b = 1/100 = 0.01$, $a = 0$.



- srednja vrijednost, medijan, kvartili i standardna devijacija dijele se sa 100;
- varijanca se dijeli sa 100^2 ;
- oblik boxplota, tj. relativni položaj podataka, **ostaje isti**.

Centriranje i standardizacija podataka

Podatke često "**centriramo**" oduzimanjem aritmetičke sredine:

$$y_i := x_i - \bar{x}.$$

- sada $y = 0$ predstavlja **prosječnu vrijednost**;
- vrijednost $y > 0$ kaže da je pripadna vrijednost x -a bila za y veća od $\bar{x}$;

Uz centriranje podatke često i **skaliramo** standardnom devijacijom:

$$z_i := \frac{x_i - \bar{x}}{s}.$$

- vrijednost $z > 0$ kaže da je pripadna vrijednost x -a bila za z **standardnih devijacija** veća od $\bar{x}$;
- $z_1, \dots, z_n$ zovemo **standardiziran uzorak**;
- standardiziranjem uklanjam mjerne jedinice i različite skupove podataka svodimo na **istu skalu**; preciznije, uvijek vrijedi (zašto?)

$$\bar{z} = 0, \quad s_z = 1.$$

Standardizacija ne ovisi o mjernoj jedinici

Neka su x_i visine izražene u centimetrima, a

$$y_i = \frac{x_i}{100}$$

iste visine izražene u metrima.

Budući da je

$$\bar{y} = \frac{\bar{x}}{100}, \quad s_y = \frac{s_x}{100},$$

vrijedi

$$\frac{y_i - \bar{y}}{s_y} = \frac{x_i/100 - \bar{x}/100}{s_x/100} = \frac{x_i - \bar{x}}{s_x}.$$

Dakle, standardizirane visine jednake su **bez obzira** na to jesu li izvorne visine izražene u metrima ili centimetrima.

Koeficijent asimetrije

Neka su

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \dots, n,$$

standardizirani podaci. **Koeficijent asimetrije** definiramo kao

$$\alpha_3 = \frac{1}{n-1} \sum_{i=1}^n z_i^3$$

Zašto podatke kubiramo?

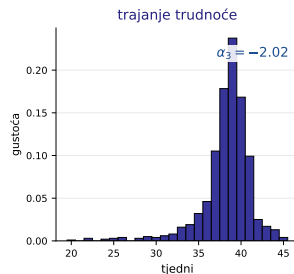
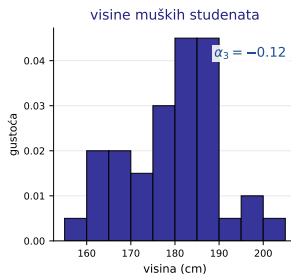
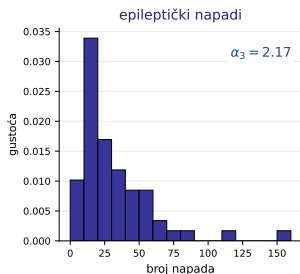
z_i	-2	-1	0	1	2
z_i^3	-8	-1	0	1	8

- kubiranjem se čuva informacija nalazi li se podatak **lijevo ili desno** od sredine;
- podaci udaljeniji od sredine dobivaju **znatno veći utjecaj**. Kod **unimodalnih** razdioba **tipično** vrijedi

$\alpha_3 < 0$	$\rightsquigarrow$	lijevo nagnuta razdioba,
$\alpha_3 \approx 0$	$\rightsquigarrow$	nema izražene asimetrije,
$\alpha_3 > 0$	$\rightsquigarrow$	desno nagnuta razdioba.

Primjeri koeficijenta asimetrije

Za prethodno promatrane podatke dobivamo:



Dvodimenzionalni numerički podaci

Kada na svakoj jedinki mjerimo **dvije numeričke varijable**, podatke zapisujemo kao parove

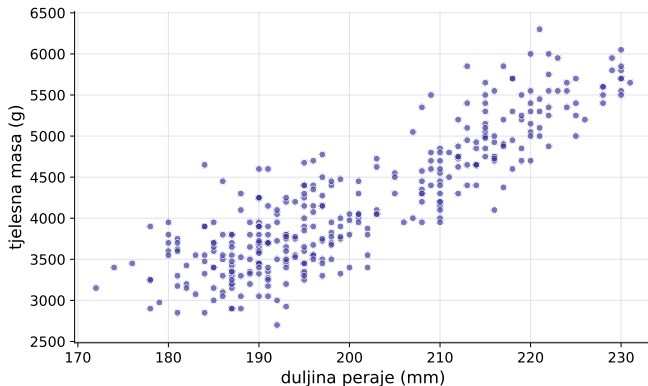
$$(x_1, y_1), \dots, (x_n, y_n).$$

Nakon što zasebno analiziramo nizove $x_1, \dots, x_n$ i $y_1, \dots, y_n$ (histogram, mjere srednje vrijednosti i raspršenja itd.) često nas zanima postoji li i kakva je **međuviznost** među varijablama x i y .

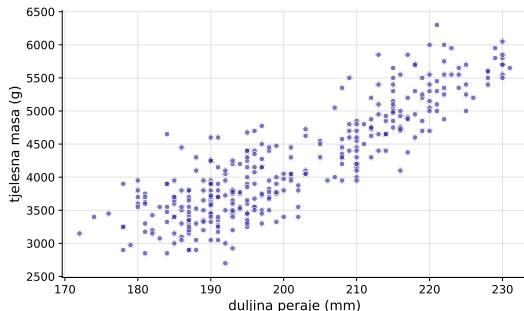
Dijagram raspršenosti

Primjer - **duljina peraje** (varijabla x) i **tjelesna masa** (varijabla y) 342 pingvina.

Dijagram raspršenosti (engl. *scatterplot*) dobijom crtanjem točaka $(x_1, y_1), \dots, (x_n, y_n)$ – točka (x_i, y_i) predstavlja duljinu peraje i masu i -tog pingvina.



Dijagram raspršenosti



- za veće duljine peraja i masa je **u pravilu** veća;
- točke **približno** leže na pravcu – približno "**linearna**" veza
- koliko je **veza jaka**, tj. koliko su točke raspršene oko zamišljenog pravca?

Koeficijent korelacije

Za podatke

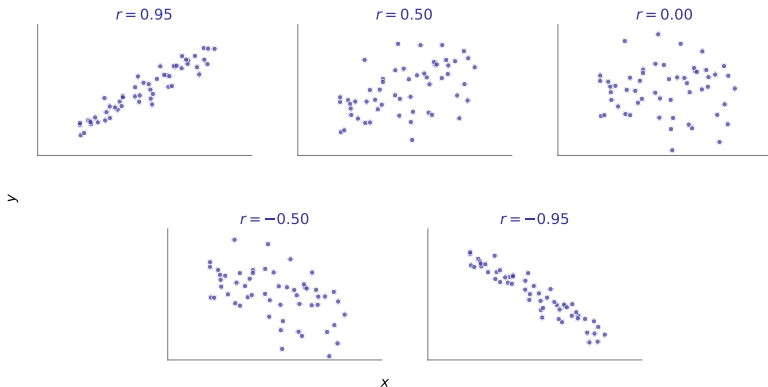
$$(x_1, y_1), \dots, (x_n, y_n)$$

definiramo **koeficijent korelacije**

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

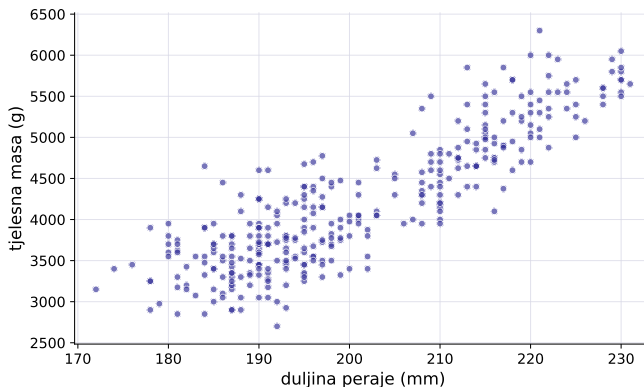
- ako je veza približno **linearna**, imat ćemo
 - $r > 0$ ako je nagib (zamišljenog) pravca **pozitivan**, te će r biti **bliže 1** što je veza **jača**;
 - $r < 0$ ako je nagib (zamišljenog) pravca **negativan**, te će r biti **bliže -1** što je veza **jača**;
- uvijek vrijedi $-1 \leq r \leq 1$;
- korelacija **ne ovisi** o mjernim jedinicama varijabli x i y (što je posljedica standardizacije);
- više o korelaciji kasnije u kolegiju.

Ilustracije



- predznak korelacije opisuje **nagib** (zamišljenog) pravca;
- što je $|r|$ bliže 1, to su točke **manje raspršene** oko pravca;
- r blizu nule upućuje na slabu ili nepostojeću **linearnu** povezanost (što god to značilo – primjer kasnije!)

Duljina peraje i tjelesna masa

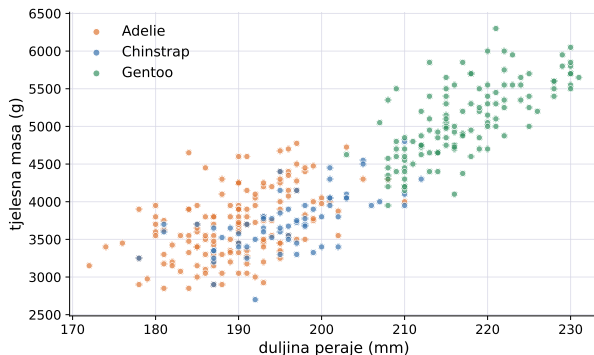


Za 342 promatrana pingvina dobivamo

$$r = 0.87.$$

što upućuje na **jaku pozitivnu** linearnu ovisnost.

Duljina peraje i masa prema vrsti



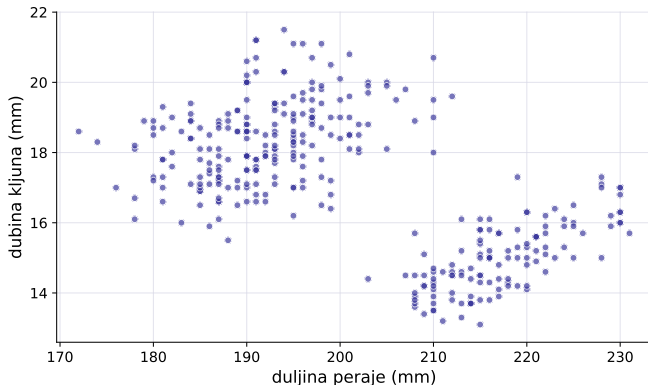
pingvini	<i>r</i>
svi zajedno	0.871
<i>Adelie</i>	0.468
<i>Chinstrap</i>	0.642
<i>Gentoo</i>	0.703

- i unutar svake vrste dulja peraja u pravilu dolazi uz veću tjelesnu masu;
- ipak, u sva tri slučaja korelacija (tj. zavisnost) je **manja** nego ukupna korelacija!
- Ukupna korelacija uz povezanost unutar skupina, odražava i **razlike među skupinama**.

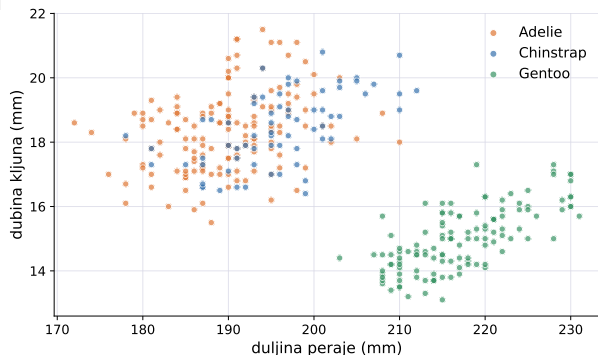
Duljina peraje i dubina kljuna

Promatramo **duljinu peraje** i **dubinu kljuna** istih 342 pingvina.

- Kakav predznak korelacije očekujete?
- Očekujete li da je korelacija bliže 1/ – 1 ili bliže 0?
- Može li se ovdje primijetiti više skupina pingvina? Utječe li to na zaključke o zavisnosti?



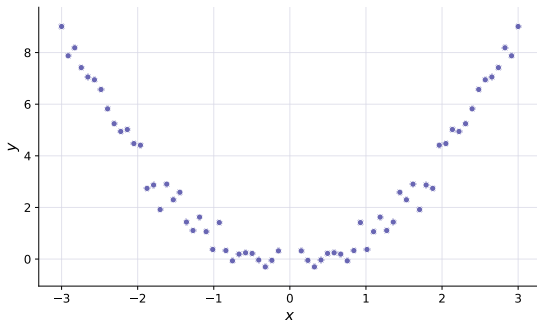
Duljina peraje i dubina kljuna prema vrsti



pingvini	<i>r</i>
svi zajedno	-0.584
<i>Adelie</i>	0.308
<i>Chinstrap</i>	0.580
<i>Gentoo</i>	0.707

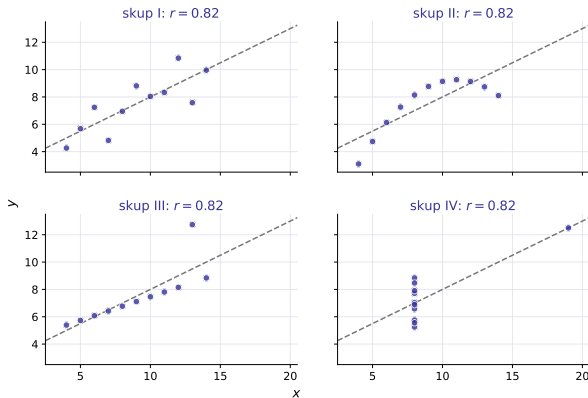
- kada promatramo sve pingvine zajedno, korelacija je **negativna**: dulje peraje povezane su s manjom dubinom kljuna;
- ipak, unutar svake pojedine vrste korelacija je **pozitivna**: pingvini s duljim perajama u pravilu imaju i dublje kljunove;
- ovo je primjer tzv. **Simpsonovog "paradoksa"**: povezanost promatrana unutar skupina može imati suprotan predznak od povezanosti dobivene spajanjem svih skupina.

Korelacija mjeri linearnu povezanost



- Za prikazane podatke vrijedi $r \approx 0$, iako između x i y postoji **jaka veza**: ako znamo x , možemo s velikom preciznosti odrediti y ;
- ipak, ta veza je približno **kvadratna** (dakle, **nelinearna**);
- korelacija mjeru jačinu **linearne** zavisnosti – intuitivno, korelacija je blizu nule jer je “najbolji” pravac koji opisuje podatke približno **vodoravan**, odnosno ima nagib blizu nule. (više o najboljem pravcu kroz podatke kasnije u kolegiju)

Što nam korelacija i ostale mjere ne otkrivaju

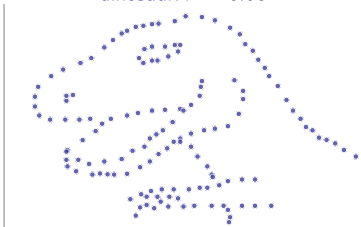
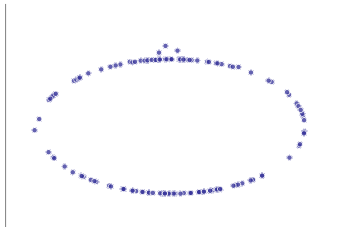
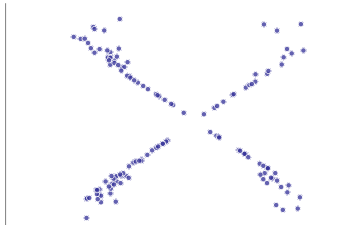


Četiri skupa imaju gotovo jednake osnovne numeričke karakteristike:

$$\bar{x} = 9, \quad \bar{y} \approx 7.5, \quad s_x \approx 3.32, \quad s_y \approx 2.03, \quad r \approx 0.82.$$

Poruka: uvijek grafički prikazati podatke!

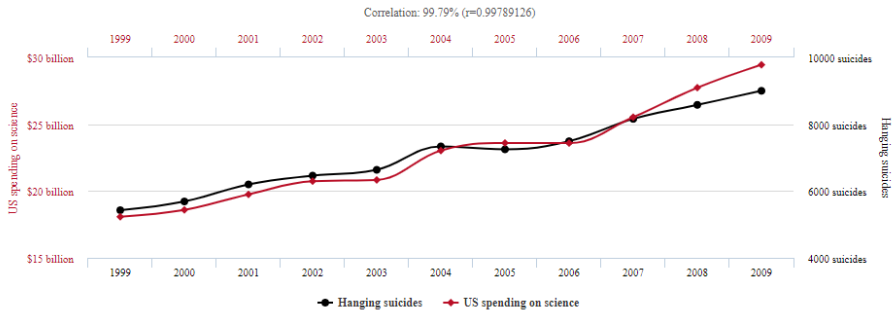
Još bolje...

dinosaur: $r = -0.06$ kružnica: $r = -0.07$ zvijezda: $r = -0.06$ slovo X: $r = -0.07$ 

Izvor: Matejka i Fitzmaurice (2017), *The Datasaurus Dozen*.

Korelacija nije kauzalnost

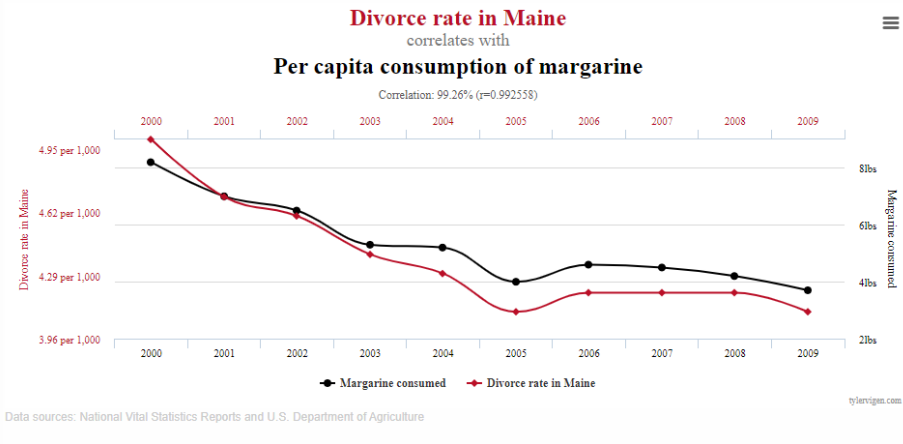
US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation



Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

tylervigen.com

Još jedan primjer prividne korelacije



Kauzalnost vs. koreliranost

$$A \longleftrightarrow B$$

- A uzrokuje B
- B uzrokuje A
- A i B se međusobno uzrokuju
- A i B su uzrokovani nekim trećim faktorom C
- B je prouzročen od C koji je koreliran s A
- A i B nisu povezani već je dobivena koreliranost rezultat slučajnosti
- ...

Koreliranost i zajednički uzrok

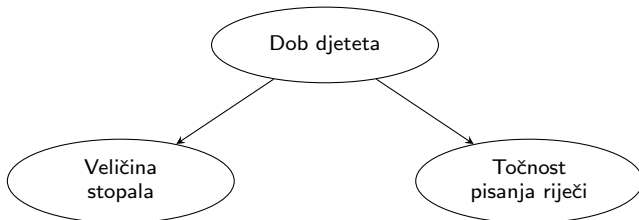
- „Istraživanje je pokazalo da djeca s većim stopalima točnije pišu riječi od djece s manjim stopalima.”

Izvor: *Aha! A two volume collection*, Mathematical Association of America, 2006, str. 177.

Koreliranost i zajednički uzrok

- „Istraživanje je pokazalo da djeca s većim stopalima točnije pišu riječi od djece s manjim stopalima.”

Izvor: *Aha! A two volume collection*, Mathematical Association of America, 2006, str. 177.



- „Pokazalo se da je broj pojedenih sladoleda visoko koreliran s brojem ljudi koji se utopio na plaži!”

Koreliranost i zajednički uzrok: još primjera

- „Korisnici Office 365 paketa računalnih programa koji vide više grešaka vjerniji su korisnici.”

Izvor: *Thrustworthy Online Controlled Experiments*, Cambridge University Press, 2020, str. 145.



- „Statistike pokazuju da se vrlo malo nesreća događa autima koji voze brzinom većom od 150 kilometara na sat.”