

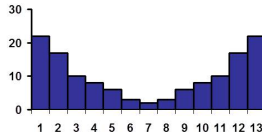
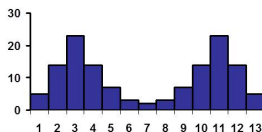
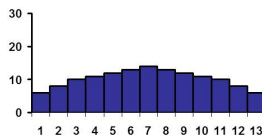
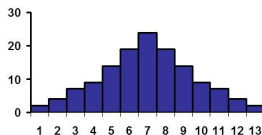
Statistika

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

2. Deskriptivna statistika

Mjere raspršenosti (disperzije)

- dva skupa podataka mogu imati slične mjere srednje vrijednosti (prosjeak i/ili medijan), ali biti jako različiti;
- dolje su histogrami (apsolutnih frekvencija) 4 skupa podataka koji svi imaju i medijan i prosječnu vrijednost jednaku 7. Ipak, promatrani podaci se očito razlikuju.
- U kojem od prikazana četiri uzorka su podaci najviše a gdje najmanje "raspršeni oko sredine"?
- u nastavku promatramo nekoliko različitih **mjera raspršenosti**.



Raspon

Najjednostavnija mjera raspršenja je **raspon**, odnosno razlika najveće i najmanje vrijednosti u uzorku:

$$d = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Uočite da uzorci

1, 1, 1, 2, 10

i

1, 4, 7, 8, 10

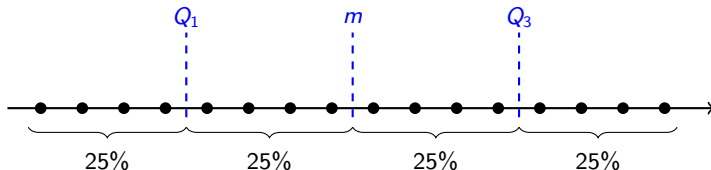
imaju isti raspon (jednak $d = 10 - 1 = 9$), premda očito nisu jednako disperzirani. Tražimo stoga bolju mjeru za disperziju.

Kvartili

Medijan dijeli uređeni uzorak na dva dijela: približno 50% podataka manje je, a 50% veće od njega.

Slično definiramo:

- **donji kvartil** Q_1 : približno 25% podataka manje je od Q_1 ;
- **gornji kvartil** Q_3 : približno 75% podataka manje je od Q_3 .



Ipak, slično kao kod medijana, često postoji **više kandidata** za Q_1 i Q_3 pa uvodimo formalnu (tehničku) definiciju.

Formalna definicija kvantila i kvartila

- Neka su podaci uređeni:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

- Ako je $s = k + r$, gdje je k cijeli dio broja s , a $0 \leq r < 1$, definiramo

$$x_{(s)} := x_{(k)} + r \cdot (x_{(k+1)} - x_{(k)}) = (1 - r)x_{(k)} + rx_{(k+1)}.$$

npr. za $s = 12.4$ imamo

$$k = 12, r = 0.4, \quad x_{(12.4)} = 0.6x_{(12)} + 0.4x_{(13)}$$

- Iako se pokaže da medijan m i za parne i za neparne n -ove možemo definirati kao

$$m := x_{((n+1)/2)}.$$

- općenito, za $\alpha \in (0, 1)$ definiramo α -**kvantil** kao

$$q_\alpha = x_{(\alpha(n+1))}.$$

- medijan je dakle upravo 0.5-kvantil, a definiramo **gornji/donji kvartil** kao

$$Q_1 = q_{0.25} = x_{(\frac{n+1}{4})}, \quad Q_3 = q_{0.75} = x_{(\frac{3(n+1)}{4})}.$$

Primjeri izračuna kvartila

Primjer 1

Promotrimo uređeni uzorak

$$2, 3, 5, 6, 8, 11, 13, \quad n = 7.$$

Budući da je

$$\frac{n+1}{4} = 2, \quad \frac{3(n+1)}{4} = 6,$$

slijedi

$$Q_1 = x_{(2)} = 3, \quad Q_3 = x_{(6)} = 11.$$

Primjer 2

Promotrimo uređeni uzorak

$$2, 4, 5, 8, 10, 13, \quad n = 6.$$

Budući da je

$$\frac{n+1}{4} = 1.75, \quad \frac{3(n+1)}{4} = 5.25,$$

slijedi

$$Q_1 = x_{(1.75)} = 0.25x_{(1)} + 0.75x_{(2)} = 0.25 \cdot 2 + 0.75 \cdot 4 = 3.5,$$

$$Q_3 = x_{(5.25)} = 0.75x_{(5)} + 0.25x_{(6)} = 0.75 \cdot 10 + 0.25 \cdot 13 = 10.75.$$

Interkvartilni raspon i centili

Interkvartilni raspon definiran je kao

$$d_Q = Q_3 - Q_1.$$

- Interval $[Q_1, Q_3]$ sadrži približno **središnjih 50% podataka**.
- d_Q mjeri raspršenost središnjeg dijela uzorka.
- Za razliku od običnog raspona, nije osjetljiv na nekoliko ekstremno malih ili velikih vrijednosti.

Ako je $\alpha = k/100$, gdje je $k \in \{1, \dots, 99\}$, tada α -kvantil

$$q_{k/100}$$

nazivamo **k %-tnim centilom** (engl. *percentile*) uzorka.

Dakle, približno k % podataka manje je od k %-tnog centila, te su Q_1 , medijan i Q_3 redom 25%, 50% i 75% centil.

Zadatak

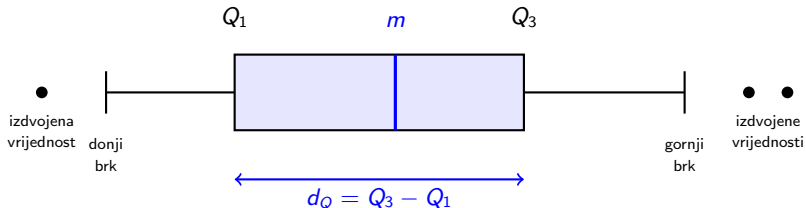
Odredite donji i gornji kvartil, te 5%-tni centil za donji uzorak (duljine $n = 20$):

1	2	2	2	2	2	3	3	3	3
3	3	4	5	5	5	6	6	6	6

rješenje: 2, 5, 1.05

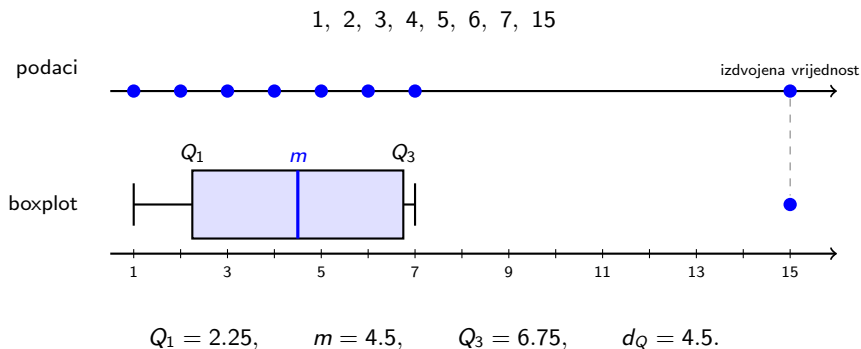
Pravokutni dijagram (*boxplot*)

Boxplot sažima **položaj** i **raspršenost** numeričkih podataka pomoću kvartila i medijana.



- Kutija od Q_1 do Q_3 sadrži približno **središnjih 50% podataka** ("tipične" vrijednosti).
- "Brkovi" završavaju na najudaljenijim podacima koji se nalaze najviše $1.5d_Q$ od pripadnog ruba kutije.
- Podatke izvan "brkova" posebno označavamo i nazivamo **outlieri** (ekstremne ili izdvojene vrijednosti).

Primjer konstrukcije boxplot-a



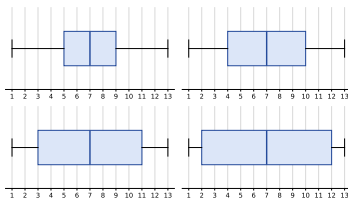
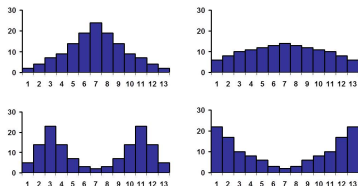
Granice za outliere su

$$Q_1 - 1.5d_Q = 2.25 - 1.5 \cdot 4.5 = -4.5, \quad Q_3 + 1.5d_Q = 6.75 + 1.5 \cdot 4.5 = 13.5.$$

Stoga brkovi završavaju u 1 i 7, dok vrijednost 15 označavamo kao outlier.

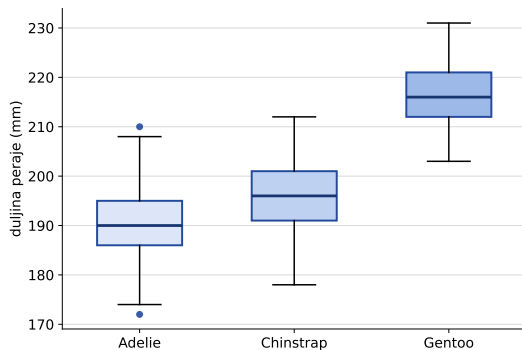
Usporedba raspršenosti pomoću boxplota

- promotrimo primjer s početka te pripadne boxplot-ove;



- Boxplot sažima **položaj i raspršenost** podataka – raspršenost se povećava od prvog prema četvrtom uzorku.
- Ipak, iz njih se **ne može** vidjeti da su oblici histograma vrlo različiti.
- Histogram i boxplot daju različite informacije o podacima te se **međusobno nadopunjuju**.

Duljina peraje pingvina



Iz boxplota vidimo:

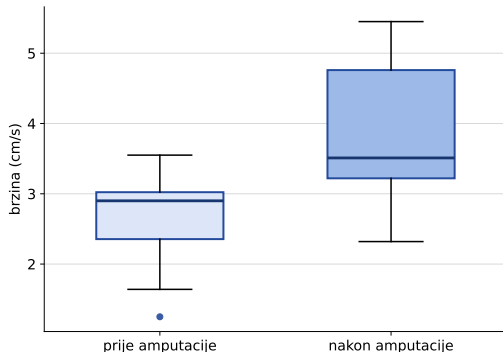
- Gentoo pingvini imaju znatno dulje peraje od ostalih vrsta.
- Raspršenost središnjih 50% ("tipičnih") duljina peraja približno je jednaka za sve tri vrste.
- Kod Adelie pingvina pojavljuju se dvije potencijalno izdvojene vrijednosti.

Brzina pauka prije i nakon amputacije

- Promotrimo brzinu 16 mužjaka pauka vrste *Tidarren sisyphoides* **prije i nakon** što su tijekom sazrijevanja odbacili jedan pedipalp.
- Zanima nas ima li to naizgled neobično ponašanje funkcionalnu prednost?

pauk	brzina prije	brzina poslije
1	1.25	2.40
2	2.94	3.50
3	2.38	4.49
4	3.09	3.17
5	3.41	5.26
6	3.00	3.22
7	2.31	2.32
8	2.93	3.31
9	2.98	3.70
10	3.55	4.70
11	2.84	4.94
12	1.64	5.06
13	3.22	3.22
14	2.87	3.52
15	2.37	5.45
16	1.91	3.40

Brzina pauka prije i nakon amputacije



Iz boxplota vidimo:

- Medijan brzine nakon amputacije veći je nego prije amputacije.
- Brzine nakon amputacije više su raspršene (za središnjih 50% podataka).
- Promijenio se i raspored podataka "unutar kutije": nakon amputacije podaci iznad medijana (dakle, njih približno 25%) znatno su raspršeni od podataka ispod medijana.

Varijanca uzorka

Za uzorak $x_1, \dots, x_n$ definiramo njegovu **variјancu** kao

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

tj. variјanca je **prosječno kvadratno odstupanje** podataka od aritmetičke sredine.

- postoji razlog zašto dijelimo s $n - 1$ umjesto s brojem podataka n , ali u praksi to ne igra preveliku ulogu osim ako n nije jako mali;
- uvijek je $s^2 \geq 0$;
- imamo $s^2 = 0$ (tj. nema raspršenja) samo u slučaju ako su **svi podaci jednaki** (čemu?);
- intuitivno, što su podaci raspršeniji oko $\bar{x}$, to je variјanca veća, tj. variјanca je još jedna mjera raspršenosti podataka oko (aritmetičke) sredine;
- često koristimo i oznaku σ^2 umjesto s^2 .

Standardna devijacija

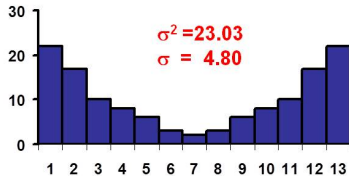
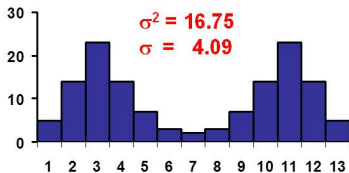
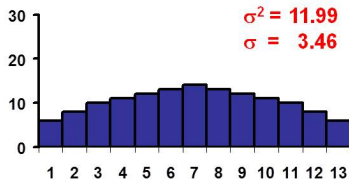
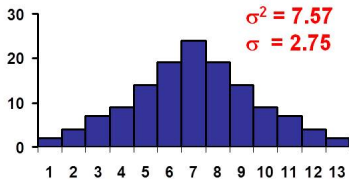
Varijanca je izražena u **kvadratu** izvornih mjernih jedinica.

Zato definiramo **standardnu devijaciju** uzorka:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{s^2}.$$

- Standardna devijacija izražena je u istim mjernim jedinicama kao i podaci.
- Intuitivno, možemo je shvatiti kao "tipično" odstupanje podataka od aritmetičke sredine.

Varijanca i standardna devijacija za podatke iz uvodnog primjera:



Računanje varijance

Za računanje varijance možemo koristiti i formulu

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}.$$

Primjer. Za 20 bacanja kocke dobivene su frekvencije

broj na kocki x_j	1	2	3	4	5	6
frekvencija f_j	1	5	6	1	3	4

Aritmetička sredina iznosi $\bar{x} = 3.6$. Zato je

$$\begin{aligned} s^2 &= \frac{1 \cdot 1^2 + 5 \cdot 2^2 + 6 \cdot 3^2 + 1 \cdot 4^2 + 3 \cdot 5^2 + 4 \cdot 6^2 - 20 \cdot 3.6^2}{19} \\ &= \frac{310 - 259.2}{19} \approx 2.674. \end{aligned}$$

Stoga je standardna devijacija

$$s = \sqrt{2.674} \approx 1.635.$$

Koeficijent varijacije

Standardna devijacija ovisi o mjernoj jedinici u kojoj su izraženi podaci. Primjerice, ne znači nam isto standardna devijacija 2.1 ako se podaci tipično reda veličine 10 ili 100000.

Za nenegativne podatke s $\bar{x} > 0$ definiramo **koeficijent varijacije**

$$v = \frac{s}{\bar{x}}.$$

- Mjeri raspršenost podataka **u odnosu na njihovu srednju vrijednost**.
- Omogućuje usporedbu varijabilnosti podataka izraženih u različitim mjernim jedinicama ili na različitim skalama.

Primjer: prinos i rizik ulaganja

Promatramo prosječne godišnje prinose i standardne devijacije prinosa triju dionica (kroz neki period godina):

	dionica A	dionica B	dionica C
prosječni prinos $\bar{x}$	6%	6%	7%
standardna devijacija s	4.5%	2.2%	5.0%
koeficijent varijacije v	0.75	0.37	0.71

- ako gledamo samo prinose, dionica C se čini najbolja;
- ipak, ako standardnu devijaciju prinosa dionice poistovjetimo s "rizičnošću" dionice, mogli bismo zaključiti da je B manje rizična nego A, te dodatno da je C prerizična za ulaganje;
- Ekonomisti promatraju i tzv. Sharpov omjer $\frac{\bar{x}}{s}$ (što je zapravo vrijednost obrnuto proporcionalna koeficijentu varijacije v) – u našem slučaju Sharpe-ov omjer iznosi

$$\frac{\bar{x}}{s} : \quad 1.33, \quad 2.73, \quad 1.40.$$

Prema tom omjeru dionica B ima najpovoljniji odnos prinosa i rizika.

Numerički podaci grupirani u razrede

Katkad ne poznajemo točne vrijednosti podataka, nego samo intervale kojima pripadaju i njihove frekvencije.

interval	frekvencija
$[a_1, b_1]$	f_1
$[a_2, b_2]$	f_2
$\vdots$	$\vdots$
$[a_k, b_k]$	f_k

Jedna opcija je tada **svaki podatak** iz intervala $[a_j, b_j]$ zamijeniti njegovom **sredinom**

$$c_j = \frac{a_j + b_j}{2}.$$

Tako dobivamo pomoćni uzorak u kojem se vrijednost c_j pojavljuje f_j puta.

Primjer: duljina sijamskih krokodila

Izmjerene su duljine trupa (u metrima) 14 primjeraka vrste *Crocodylus siamensis*:

interval	f_j	c_j
[1.0, 1.2]	1	1.10
[1.3, 1.4]	3	1.35
[1.5, 1.6]	6	1.55
[1.7, 1.8]	1	1.75
[1.9, 2.0]	3	1.95

Svaki podatak zamjenjujemo sredinom pripadnog razreda i dobivamo **pomoćni uzorak**

1.10, 1.35, 1.35, 1.35, 1.55, 1.55, 1.55,
1.55, 1.55, 1.55, 1.75, 1.95, 1.95, 1.95.

Za pomoćni uzorak dobivamo

$$\bar{x} = \frac{22.05}{14} = 1.575 \text{ m}, \quad m = 1.55 \text{ m.}$$

$$s^2 = \frac{35.5625 - 14 \cdot 1.575^2}{13} \approx 0.0641 \text{ m}^2, \quad s \approx 0.253 \text{ m.}$$

Zadatak

Zadani su grupirani numerički podaci:

interval	frekvencija
[2, 3)	3
[3, 4)	4
[4, 5)	3
[5, 6)	2

Koristeći sredine razreda, odredite približnu:

- aritmetičku sredinu podataka;
- varijancu i standardnu devijaciju podataka.

$$\text{Rj. } \bar{x} = \frac{23}{6} \approx 3.83, \quad s^2 = \frac{38}{33} \approx 1.15, \quad s = \sqrt{\frac{38}{33}} \approx 1.07$$