

Deskriptivna statistika

Vježbe III

Deskriptivna statistika

- ❑ Mjere centralne tendencije
- ❑ Mjere varijacije

Centralna tendencija

- ❑ Centralna tendencija – mjere čije izračunavanje služi određivanju numeričke vrijednosti oko koje se rezultati grupišu
 - ❑ Aritmetička sredina – “prosjek”
 - ❑ Medijana – “središnja vrijednost”
 - ❑ Modus – “najfrekventnija vrijednost”

Mjere centralne tendencije

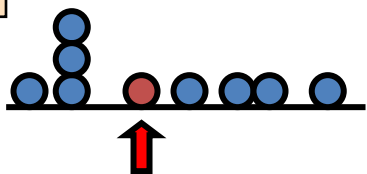
Pregled

Centralna tendencija

Aritmetička sredina

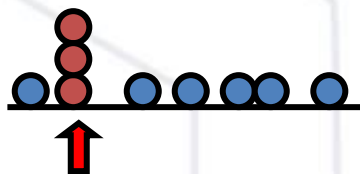
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

Medijana



Središnja
vrijednost
opservacije
(centralna
opservacija)

Modus



Opservirana
vrijednost sa
najvećom
frekvencijom

Geometrijska sredina

$$\bar{X}_G = (X_1 \times X_2 \times \dots \times X_n)^{1/n}$$

Centralne tendencije

- ☐ **Aritmetička sredina** – suma vrijednosti podijeljena brojem vrijednosti.
 - ☐ Zavisi od ekstremnih vrijednosti
 - ☐ U slučaju agregatnih podataka koristit sredinu intervala i broj observacija
- ☐ **Medijana** – središnja opservacija nakon što ih poredamo po veličini
 - ☐ Ne zavisi od ekstremnih vrijednosti
 - ☐ Uzim se središnja vrijednost (neparan broj), ili prosjek dvije središnje vrijednosti (paran)
 - ☐ $(n + 1) / 2$ je pozicija u uređenom nizu, ne vrijednost medijane
- ☐ **Modus** – Vrijednost koja se najčešće pojavljuje
 - ☐ I za numeričke i atributivne (nominalne) podatke
 - ☐ Multi-modalna distribucija

Aritmetička sredina

- Aritmetička sredina (sredina) je najčešće korišćena mjera centralne tendencije

– Za uzorak veličine n aritmetička sredina je:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

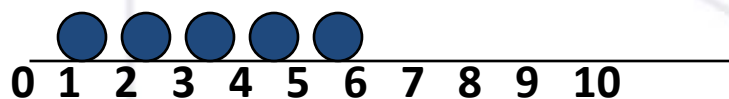
Veličina uzorka

Vrijednosti varijable

Aritmetička sredina

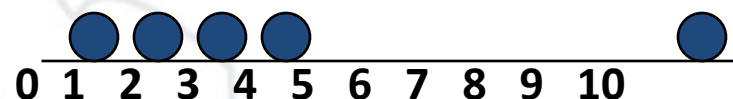
(nastavak)

- Sredina = suma vrijednosti podijeljena brojem tih vrijednosti (opservacija)
- Zavisi od ekstremnih vrijednosti



Sredina = 3

$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

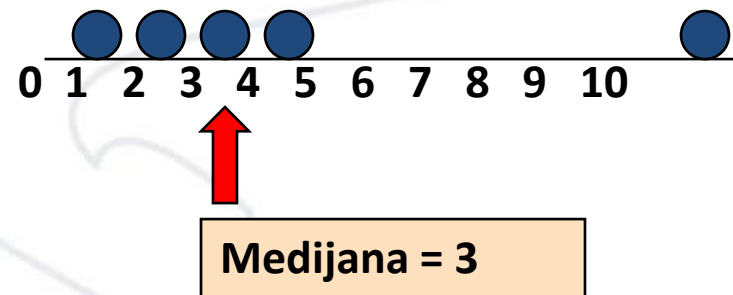
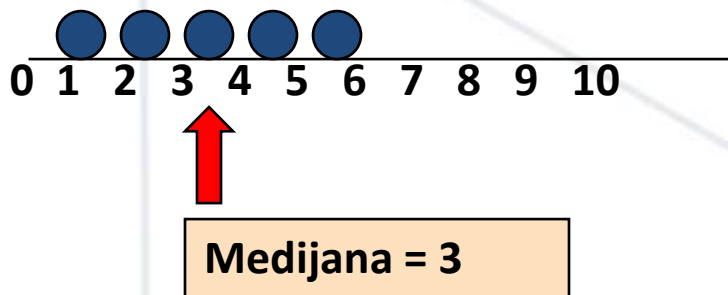


Sredina = 4

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Medijana

- U uređenom nizu, medijana je “srednja” opservacija (50% iznad, 50% ispod)



- Ne zavisi od ekstremnih vrijednosti

Određivanje medijane

- Lokacija medijane se određuje na sledeći način:

$$\text{Medijana} = \frac{n+1}{2} \text{ pozicija u uredjenom nizu}$$

- Ako je broj opservacija neparan, medijana je centralna opservacija
- Ako je broj opservacija paran, medijana je prosjek dvije središnje opservacije

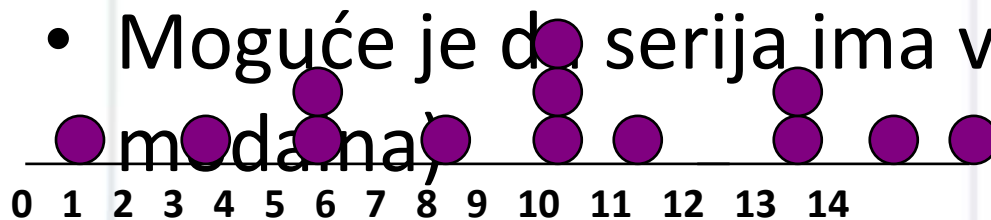
- Napomena nije *vrijednost* medijane, već samo *pozicija*

medijane u uredjenom nizu podataka

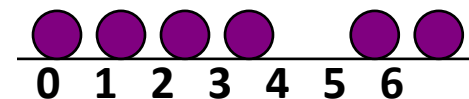
$$\frac{n+1}{2}$$

Modus

- Jedna od mjera centralne tendencije
- Vrijednost koja se najčešće pojavljuje
- Ne zavisi od ekstremnih vrijednosti
- Određuje se i za numeričke i za atributivne podatke
- Moguće je da serija nema modus
- Moguće je da serija ima više modusa (multi-



Modus = 9



Nema modusa

Primjer:

Cijene kuća:

\$2,000,000
500,000
300,000
100,000
100,000

Suma \$3,000,000

- **Sredina:** $(\$3,000,000/5)$
(prosječna cijena) = **\$600,000**
- **Medijana:** centralna vrijednost (cijena) u nizu
= **\$300,000**
- **Modus:** najčešća vrijednost (cijena)
= **\$100,000**

<div>Mjere</div> <div>Skale</div>	Prosjek	Medijana	Modus
Racio	+	+	(+)
Intervalna	+	+	(+)
Ordinalna	(-)	+	+
Nominalna	-	-	+

Varijacija

- ❑ **Mjere varijacije** – mjere koje opisuju „rasprostranjenost“ (disperziju) podataka.
 - ❑ Raspon
 - ❑ Interkvartalni raspon
 - ❑ Varijansa
 - ❑ Standardna devijacija

Mjere varijacije

☐ Raspon - $X_{max} - X_{min}$

- ☐ Zavisi od ekstremnih vrijednosti; distribucija nevažna; problem velikog uzorka

☐ Varijansa – prosječno kvadratno odstupanje od prosjeka

☐ Standardna devijacija - kvadratni korijen iz varijanse

- ☐ Najčešće korišćena mjera

- ☐ Izražena u jedinicama kao i originalni podaci

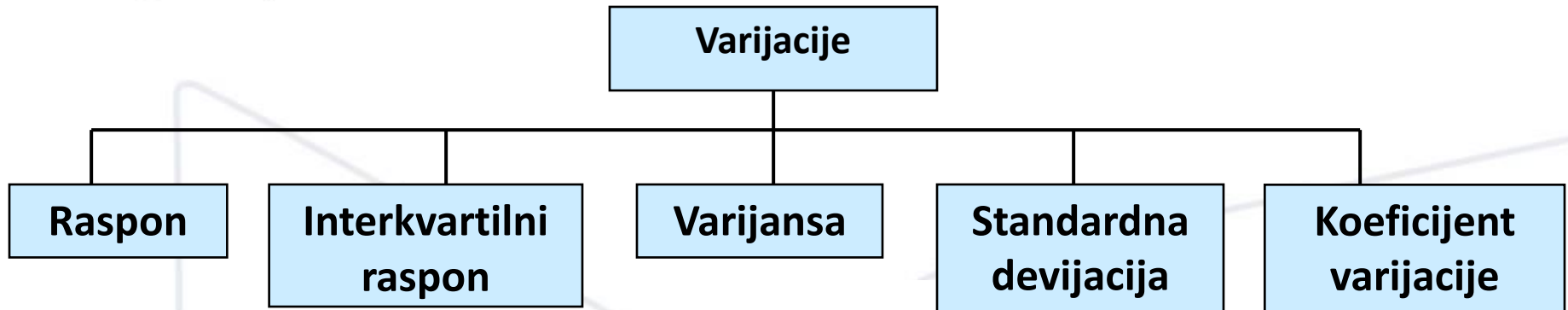
- ☐ Empirijsko pravilo: $\pm 1SD$ (68%)

$\pm 2SD$ (95%)

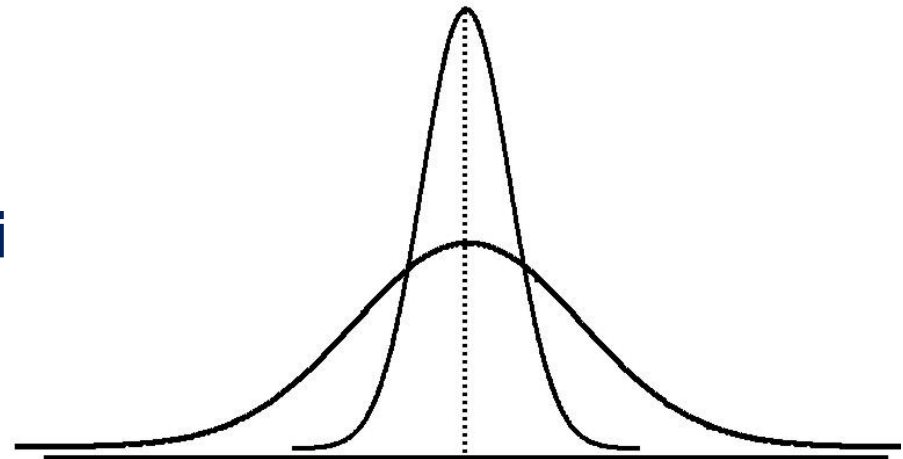
$\pm 3SD$ (99.7%)

- ☐ Z- skor (standardizovano odstupanje)

Mjere varijacije



- Mjere varijacije daju informaciju o **disperziji ili varijabilnosti** vrijednosti obilježja, odnosno podataka.



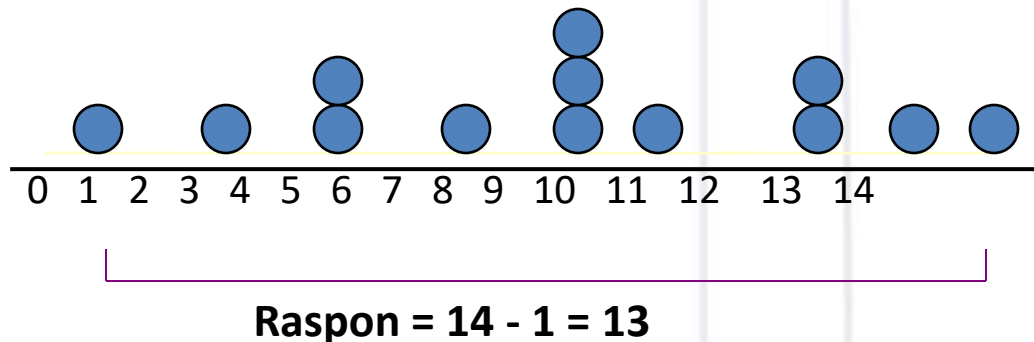
Isti centar,
različita varijacija

Raspon

- Najjednostavnija mjera varijacije
- Razlika između najveće i najmanje vrijednosti obilježja u statističkoj seriji:

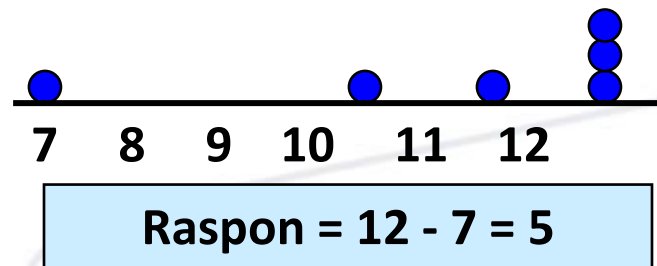
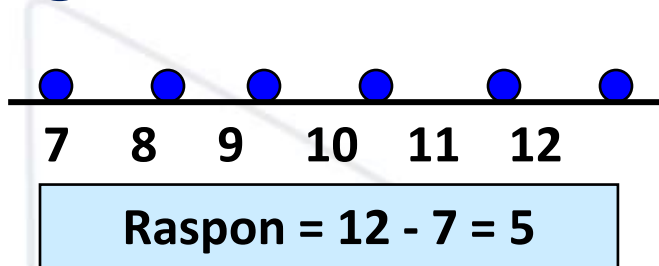
$$\text{Raspon} = X_{\max} - X_{\min}$$

Primjer:



Nedostaci raspona

- Ignoriše se distribucija podataka



- Osjetljiv na outlier-e

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

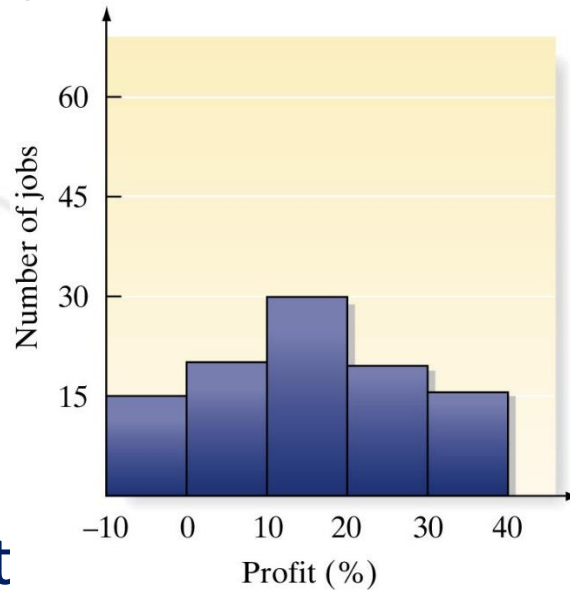
$$\text{Raspon} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

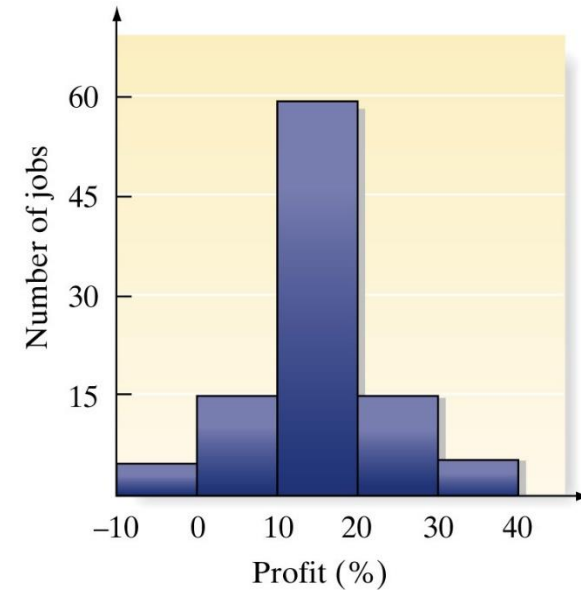
$$\text{Raspon} = 120 - 1 = 119$$

Nedostaci raspona

- Gubi smisao kod jako velikih uzoraka
- Ove 2 distribucije imaju isti raspon.
- Koliko vam raspon govori o varijabilnost podataka?



a. Cost estimator A



b. Cost estimator B

Varijansa

- Prosječno (približno) kvadratno odstupanje vrijednosti obilježja od aritmetičke sredine

– Uzoračka varijansa:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Đe je: \bar{X} = aritmetička sredina

n = veličina uzorka

X_i = i ta vrijednost varijable X

Standardna devijacija

- Najčešće korišćena mjera varijacije
- Pokazuje varijaciju oko aritmetičke sredine
- Izračunava se kao kvadratni koren iz varijanse
- Iskazuje se u **istim jedinicama kao i originalni podaci**

– Standardna devijacija
u uzorku:

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Interpretiranje standardne devijacije

- Koliko opservacija se nalazi u intervalu $\pm n s$ od aritmetičke sredine?

$1+1s$ ili $1+1\sigma$

$1+2s$ ili $1+2\sigma$

$1+3s$ ili $1+3\sigma$

Empirijsko pravilo

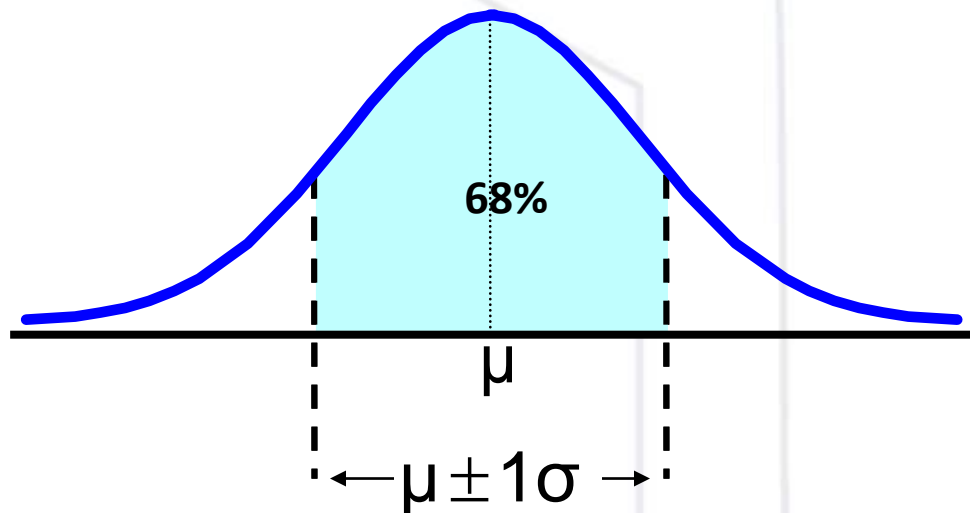
približno 68%

približno 95%

približno 99.7%

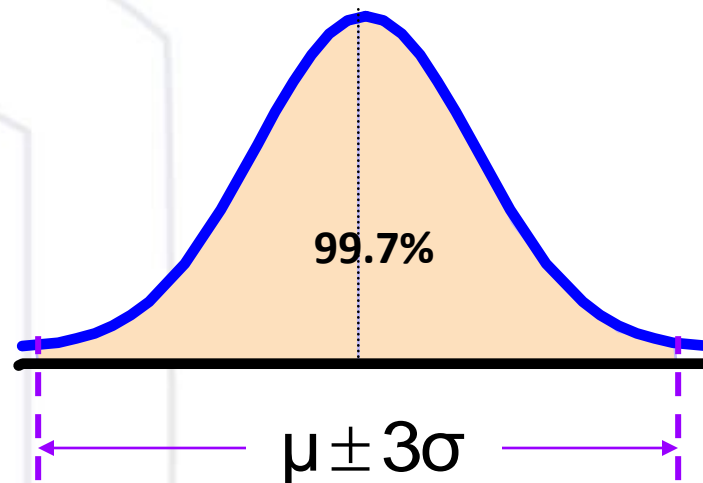
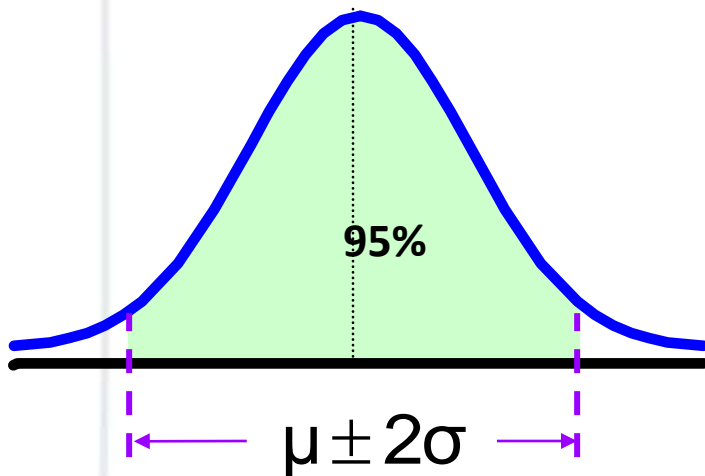
Empirijsko pravilo (pravilo 3σ)

- Ako podaci imaju distribuciju u obliku zvona, onda interval:
- $\mu \pm 1\sigma$ sadrži oko 68% vrijednosti obilježja u populaciji ili u uzorku



Empirijsko pravilo (pravilo 3σ)

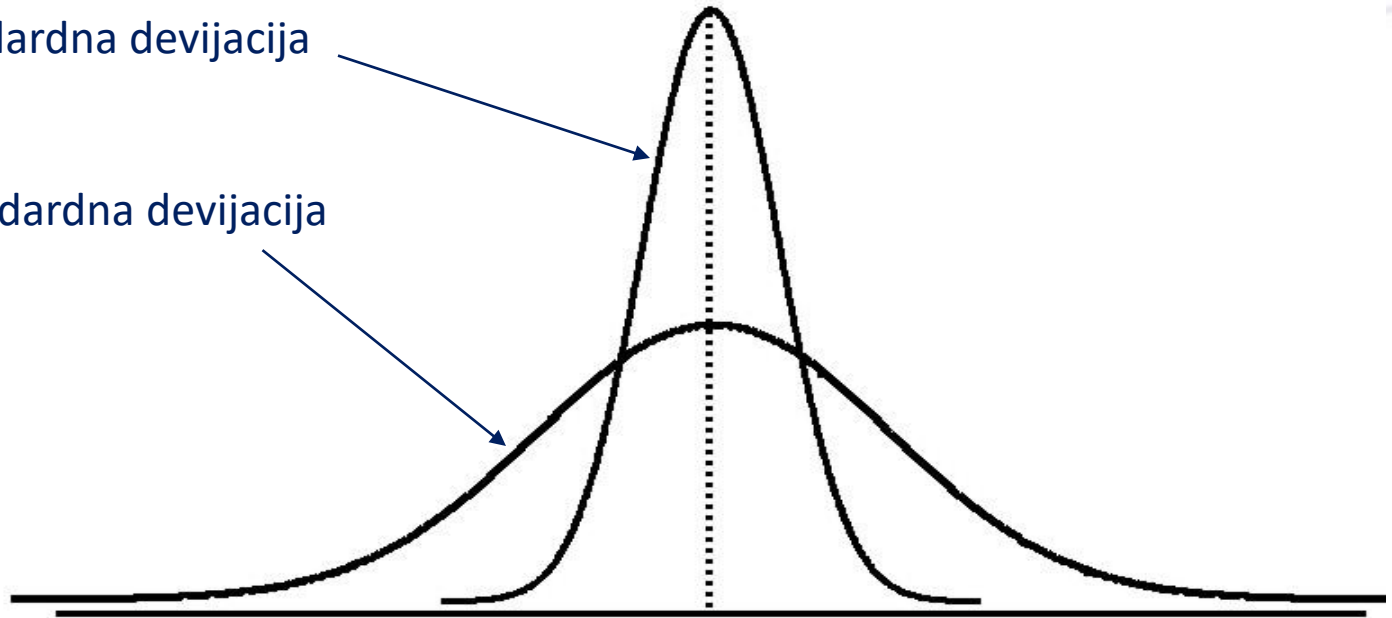
- $\mu \pm 2\sigma$ sadrži oko 95% vrijednosti obilježja u populaciji ili u uzorku
- $\mu \pm 3\sigma$ sadrži oko 99.7% vrijednosti obilježja u populaciji ili u uzorku



Mjerenje varijacije

Mala standardna devijacija

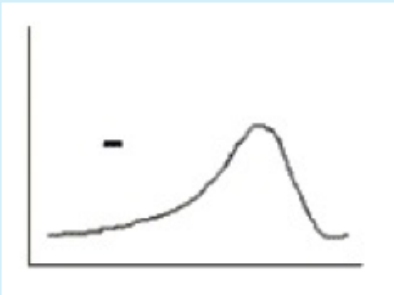
Velika standardna devijacija



Tipovi distribucije

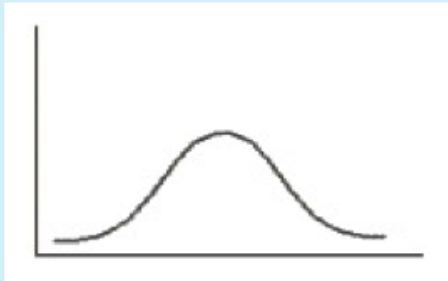
„Iskrivljena“ ulijevo

Prosjeak < Medijana



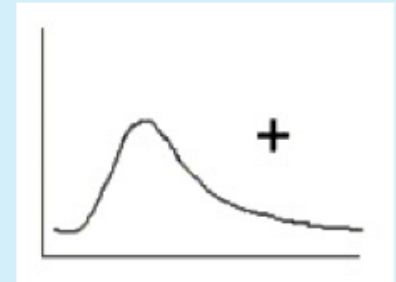
Normalna

Prosjeak = Medijana



„Iskrivljena“ udesno

Prosjeak > Medijana



Vježba I

Populacija za vrijednost aritmetičke sredine ima 30 i standardnu devijaciju 5.

Ukoliko dodamo vrijednost 5 svakom skoru u populaciji, koja bi bila nova vrijednost aritmetičke sredine i standardne devijacije?

Ukoliko svaki skor u populaciji pomnožimo sa 3, koja bi bila nova vrijednost aritmetičke sredine i standardne devijacije?

Vježba I

Populacija za vrijednost aritmetičke sredine ima 30 i standardnu devijaciju 5.

Ukoliko dodamo vrijednost 5 svakom skoru u populaciji, koja bi bila nova vrijednost aritmetičke sredine i standardne devijacije?

Arit. Sredina: 35 **Stand. devijacija:** 5

Ukoliko svaki skor u populaciji pomnožimo sa 3, koja bi bila nova vrijednost aritmetičke sredine i standardne devijacije?

Arit. Sredina: 90 **Stand. devijacija:** 15

Vježba II

Istraživač analizira donacije građana prema političkim partijama u zemlji X. Za populaciju od 6 partija izračunaj: aritmetičku sredinu, medijanu, modus, raspon i standardnu devijaciju.

***Napomena:** jedinica numeričke vrijednosti izražena je u „1.000 eura.“*

Vrijednost donacija: 11, 0, 2, 9, 9, 5

Vježba II

Arit. sredina:

Medijana:

Modus:

Vježba II

Arit. sredina:

$$(11+0+2+9+9+5)/6 = 36/6 = 6$$

U prosjeku, partije u zemlji X su primile 6.000 eura donacija od strane građana.

Medijana:

11, 9, 9, 5, 2, 0; niz ima paran broj vrijednosti

Dvije središnje vrijednosti su 9 i 5. Medijanu u ovom slučaju dobijamo iz aritmetičke sredine ova dva broja.

Vrijednost medijane je: 7

Modus:

Donacija:	11	0	2	9	5
Frekvenc.:	1	1	1	2	1

Vježba III

Vrijednost donacije: 11, 0, 2, 9, 9, 5

Raspon:

Varijansa: $(SS = \sum (X - \mu)^2) / N$

Standardna devijacija: korijenovana varijansa = 4

Vježba III

Vrijednost donacije: 11, 0, 2, 9, 9, 5

Raspon:

$$11 - 0 = 11$$

Varijansa: $(SS = \sum(X - \mu)^2)/N$

$$\begin{aligned} SS &= (11 - 6)^2 + (0 - 6)^2 + (2 - 6)^2 + (9 - 6)^2 + (9 - 6)^2 + (5 - 6)^2 \\ &= 25 + 36 + 16 + 9 + 9 + 1 = 96 \end{aligned}$$

$$\text{Var} = 96/6 = 16$$

Standardna devijacija: korijenovana varijansa = 4

Izračunavanje stand. dev.

- ❑ **Korak 1:** Pronađi **aritmetičku sredinu**.
- ❑ **Korak 2:** Od svake pojedinačne vrijednosti oduzmi aritmetičku sredinu (**devijaciju**).
- ❑ **Korak 3:** **Kvadriraj** svaku pojedinačnu devijaciju od aritmetičke sredine.
- ❑ **Korak 4:** **Saberi** sve kvadrirane devijacije.
- ❑ **Korak 5:** **Podijeli** sa brojem opservacija (veličinom populacije).
- ❑ **Korak 6:** Izračunaj **kvadratni korijen** iz dobijene vrijednosti.