Statistika za jezikoslovno istraživanje

Damir Ćavar

ožujak 2010.

Plan

- Silabus
- Uvod
- Pitanja

Statistika

- Tipovi statistike
  - Deskriptivna statistika
    - brojevi, niz, sredine, raspršenost
  - Vizualizacija statističkih osobina
  - Induktivna statistika
    - procesi, brojevi i/ili slike

Deskriptivna statistika

- Mjere za centralne tendencije podataka
  - Aritmetička sredina
  - Srednja vrijednost (medijan)
  - Dominantna vrijednost (mod)
- Mjere za varijaciju/varijancu
  - Raspršanost podataka
- Načini mjera

Deskriptivna statistika

- Aritmetička sredina
  - podaci:

datoteka	broj riječi
Flo031201.txt	10346
Flo031202a.txt	5031
Flo031202b.txt	11876
Flo031203.txt	12175
Flo031204.txt	10943

Deskriptivna statistika

- Aritmetička sredina

$$aritmetička\ sredina = \frac{suma\ svih\ rezultata}{broj\ rezultata}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- primjer:

$$\frac{10346 + 5031 + 11876 + 12175 + 10943}{5} = 10074.2$$

Deskriptivna statistika

- Centralna vrijednost (Medijan)
  - Sredina niza rezultata

tekst	broj riječi
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876
Flo031203.txt	12175

Deskriptivna statistika

- Srednja vrijednost (Medijan)
  - Smanjuje značajnost ekstremnih rezultata (“outlier”):

tekst	broj riječi
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876
Flo031203.txt	12175

Deskriptivna statistika

- Srednja vrijednost (medijan) s ravnim brojem rezultata:

tekst	broj riječi
Flo031202a.txt	5031
Flo031201.txt	10346
Flo031204.txt	10943
Flo031202b.txt	11876

- Aritmetička sredina dvaju srednjih vrijednosti:

$$\frac{10346 + 10943}{2} = 10644.5$$

Deskriptivna statistika

- Rezultati moraju biti poredani po veličini (smijer nije bitan)
- Formalno: Srednja vrijednost (medijan)

$$sv = \begin{cases} x_{k+1} & \text{ako je } n \text{ neparan, } n = 2k+1 \\ \frac{x_k + x_{k+1}}{2} & \text{ako je } n \text{ paran, } n = 2k \end{cases}$$

Deskriptivna statistika

- Aritmetička sredina: 10074.2
- Srednja vrijednost (medijan): 10943
- Aritmetička sredina je manja radi ekstremnih rezultata (“outlier”):
  - Flo031202a.txt 5031
- Srednja vrijednost može biti bolji indikator za osnovnu tendenciju ako imamo ekstremne rezultate.

Deskriptivna statistika

- Dominantna vrijednost (mod): 10943
  - Vrijednost koja se najčešće pojavljuje:

tekst	broj riječi
Flo031202a.txt	5031
Flo031201.txt	10943
Flo031204.txt	10943
Flo031202b.txt	6329
Flo031203.txt	12175

Deskriptivna statistika

- Aproksimacija
  - Dominantna vrijednost (mod)
    - *sredina* – 3 (*sredina* – *medijan*)
  - Srednja vrijednost (medijan)
    - (2 *sredina* + *mod*) / 3
  - Aritmetička sredina
    - (3 *medijan* – *mod*) / 2

Deskriptivna statistika

- Notacija
- Sredina (x-bar): $\overline{x}$
- Sredina populacije: μ
- Sumacija vrijednosti: $\sum$

Deskriptivna statistika

- Aritmetička sredina za skupine podataka:

tekstovi	broj riječi
35%	0-4999
30%	5000-9999
25%	10000-14999
10%	15000-19999

- S 100 primjera (tekstova) koja je aritmetička sredina?

Deskriptivna statistika

- Aritmetička sredina za skupine podataka:

$$\bar{x} = \frac{\sum fx}{n}$$

  - f = frequency
  - x = midpoint

Deskriptivna statistika

- Aritmetička sredina za razrede podataka:

datoteka	**sredina**	***fx***	**broj riječi**
35	2500	87500	0-4999
30	7500	225000	5000-9999
25	12500	312500	10000-14999
10	17500	175000	15000-19999

$$\bar{x} = \frac{\sum fx}{n} = \frac{87500 + 225000 + 312500 + 175000}{100} = \frac{800000}{100} = 8000$$

Deskriptivna statistika

- Srednja vrijednost za rezultate u razredima:

$$medijan = L + \frac{w}{f_{med}}(.5n - \sum f_b)$$

- L = donja granica medijalnog razreda
- n = zbroj frekvencija
- w = veličina medijalnog razreda
- f_{med} = frekvencija medijalnog razreda
- Σf_b = sumacija do medijalnog razreda

Deskriptivna statistika

- Srednja vrijednost za rezultate u razredima:

tekstovi	broj riječi
35	0-4999
30	5000-9999
25	10000-14999
10	15000-19999

$$sv = 5000 + \frac{4999}{30}(50 - 35) = 7499.5$$

Deskriptivna statistika

- Geometrijska sredina
  - Korijen umnožka rezultata (pozitivni i veći od 0)

$$gs = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

- Mjera za npr. brzinu neke promjene.

Deskriptivna statistika

- Harmonična sredina: prosjek odnosa
  - ako su rezultati veći od 0

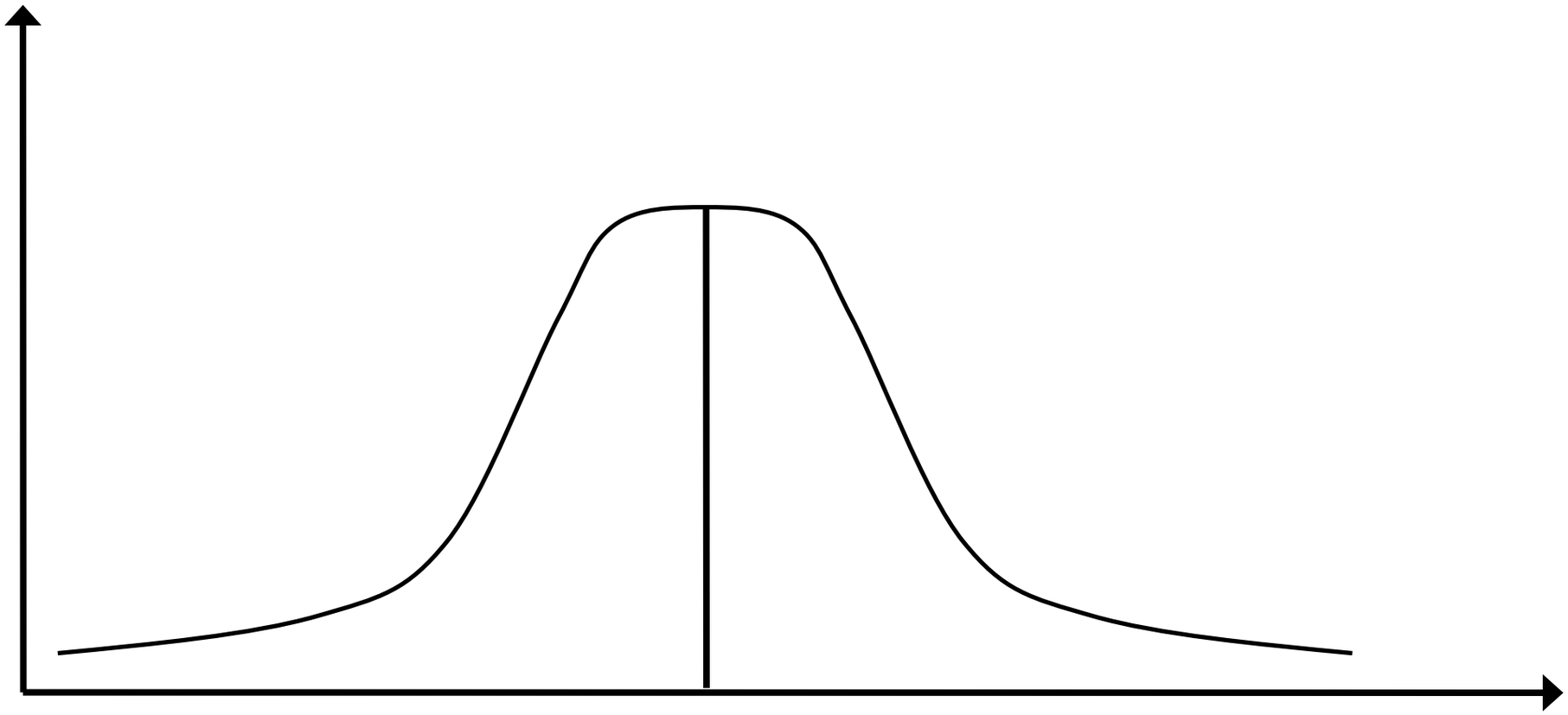
$$hs = \frac{N}{\sum \frac{1}{x}}$$

Deskriptivna statistika

- Distribucija
  - Simetrijska distribucija
  - Nesimetrijske distribucije

Deskriptivna statistika

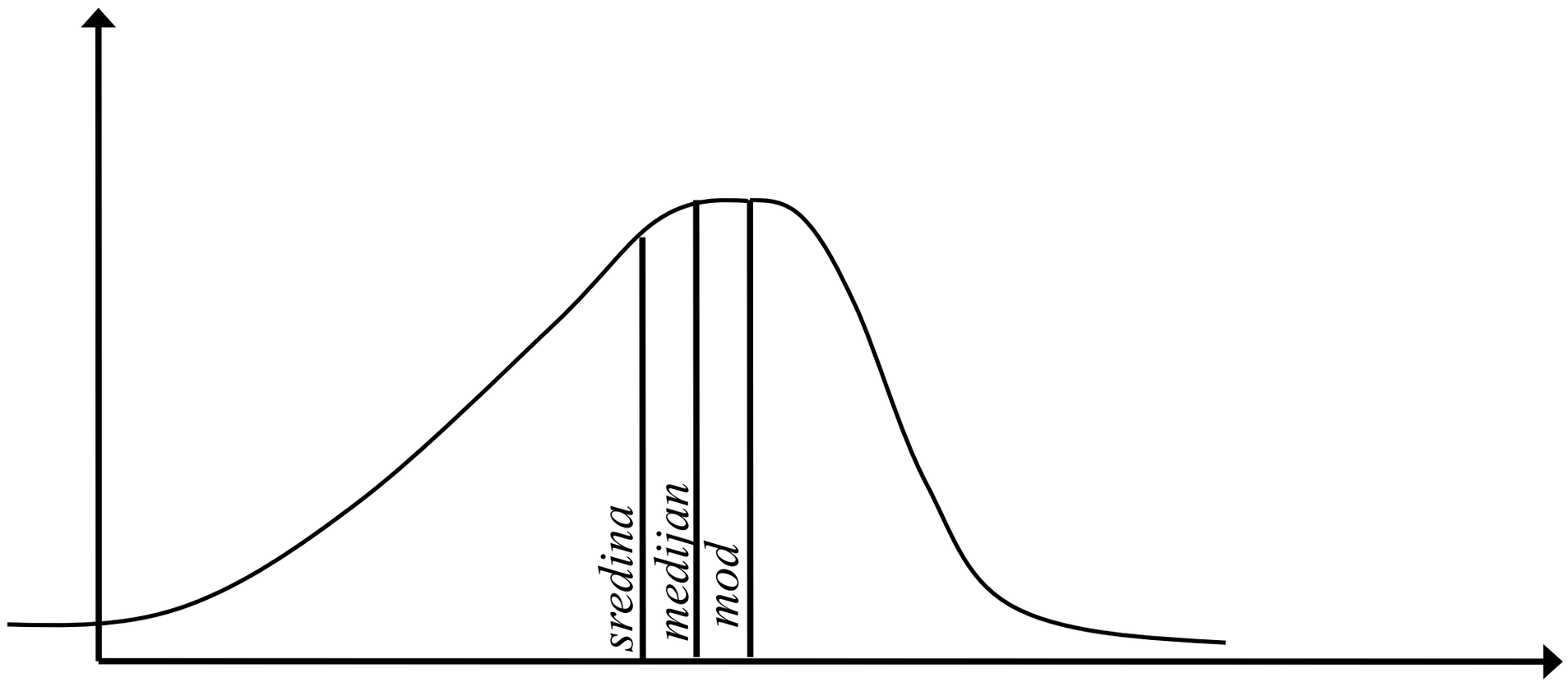
- Simetrijska distribucija: aritmetička sredina, srednja i dominantna vrijednost su isti



sredina
medijan
mod

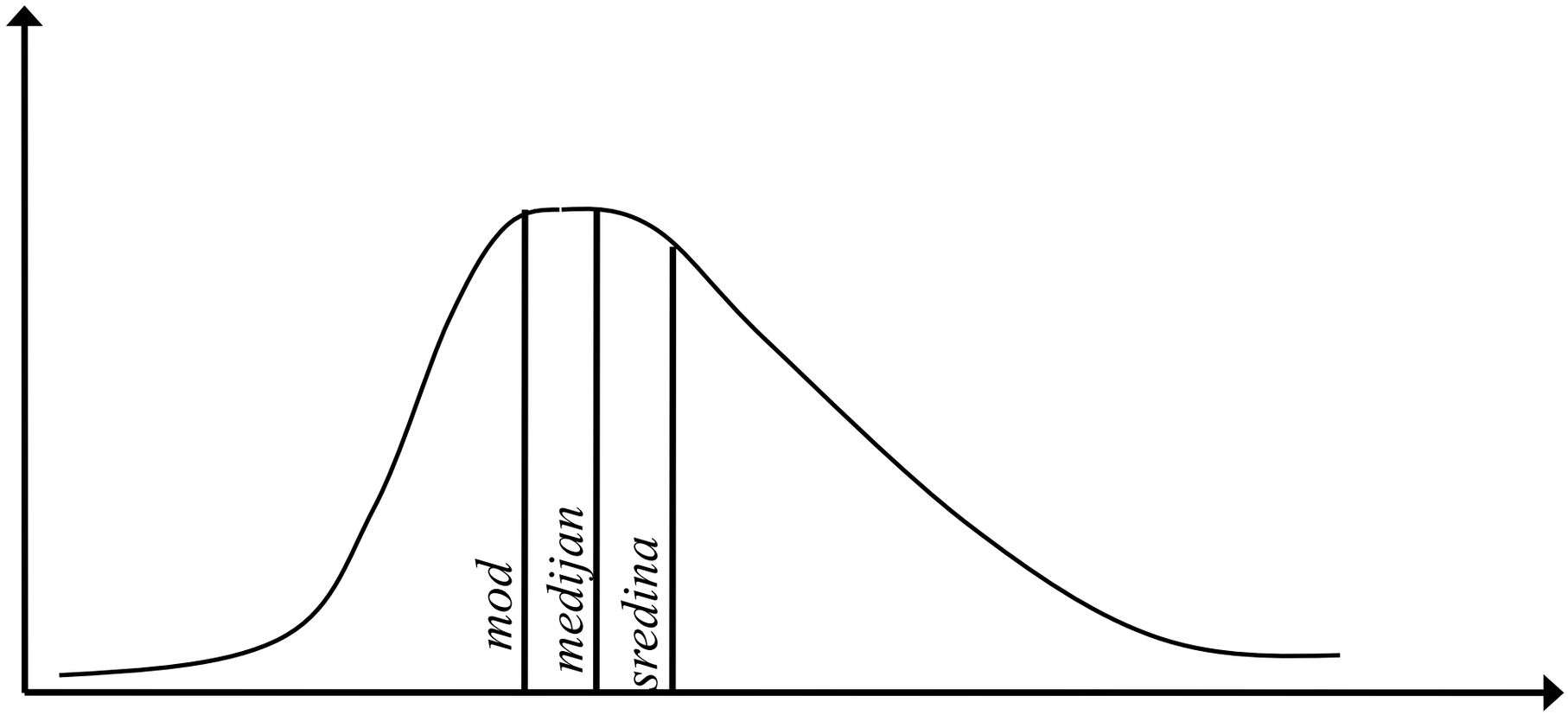
Deskriptivna statistika

- Asimetrična distribucija
  - Negativno asimetrična distribucija: sredina < medijan < mod



Deskriptivna statistika

- Asimetrična distribucija
  - Pozitivno asimetrična distribucija: mod < medijan < sredina



Deskriptivna statistika

- Varijacija i disperzija (ili raspršenost)

Eksperiment 1	Eksperiment 2
195	10
210	0
199	400
200	20
205	380
190	200
200	390
201	200

Deskriptivna statistika

- Varijacija i disperzija
  - Za oba eksperimenta dobijemo:
    - sredina: 200
    - mod: 200
    - medijan: 200
  - Eksperiment 2 ima veći raspon varijacije (promjenjivost) i disperzije (odstupanja prema srednjoj vrijednosti).
- Mjere varijacije: raspon varijacije, standardna devijacija, varijanca

Deskriptivna statistika

- Raspon varijacije
  - Razlika između najveće i najmanje vrijednosti u nizu:
    - Eksperiment 1: 210 – 190 = 20
    - Eksperiment 2: 400 – 0 = 400
- Problem:
  - ekstremni rezultati mogu povećati raspon, bez značajnog utjecaja npr. na aritmetičku sredinu
  - raste s većim brojem rezultata

Deskriptivna statistika

- Srednja devijacija
  - Odstupanje rezultata od srednje vrijednosti:
    - Eksperiment 1: manje
    - Eksperiment 2: više

$$sd = \frac{1}{N} \sum_{i=1}^{N} |x_i - \bar{x}|$$

Deskriptivna statistika

- Varijanca
  - Sumacija kvadratnih odstupanja *n* mjera od aritmetičke sredine podjeljeno kroz (*n – 1*):
    - Eksperiment 1: ?
    - Eksperiment 2: ?

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad s_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Deskriptivna statistika

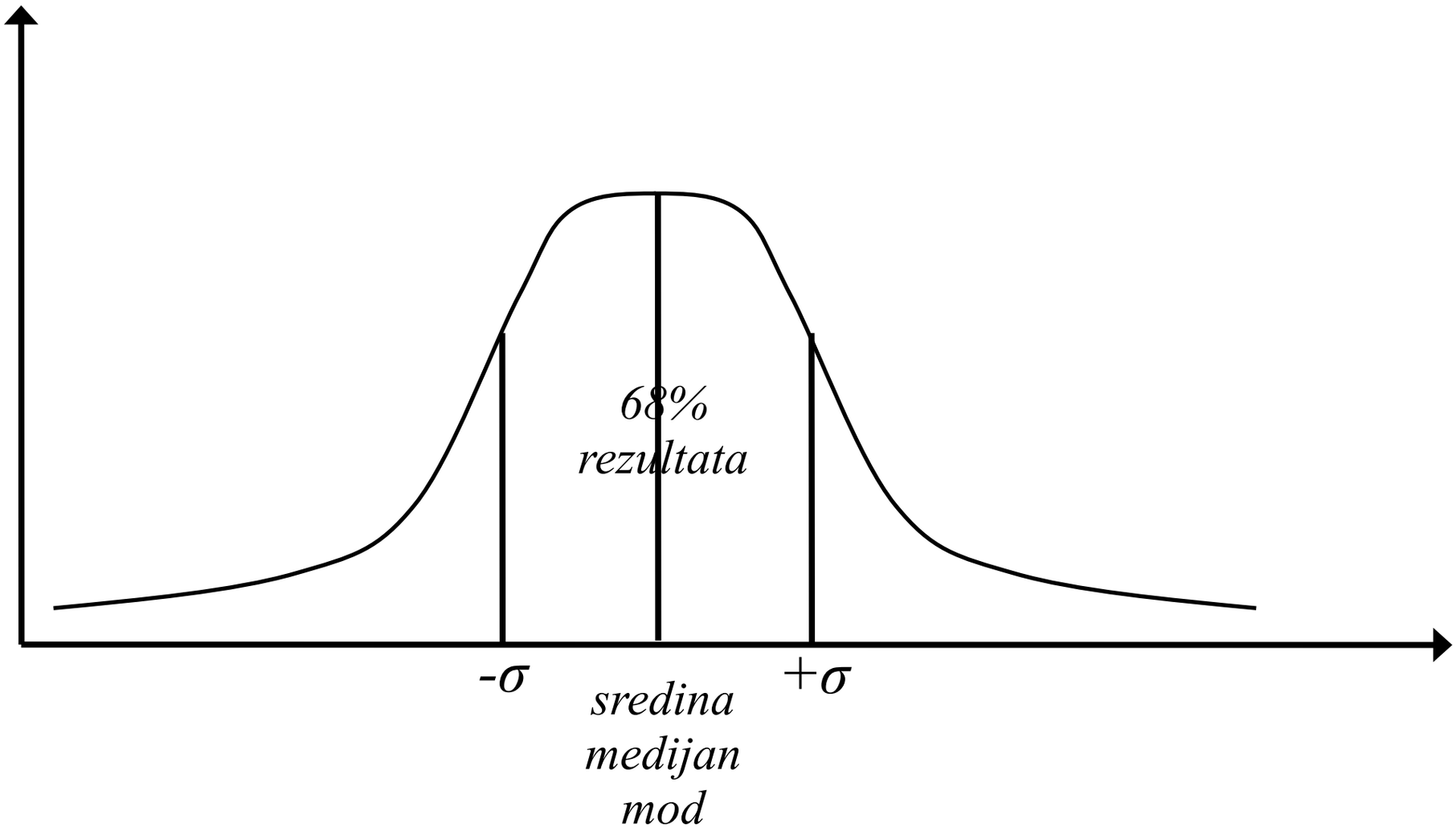
- Standardna devijacija
  - Positivni korijen iz varijance.
    - Eksperiment 1: ?
    - Eksperiment 2: ?

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad s_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

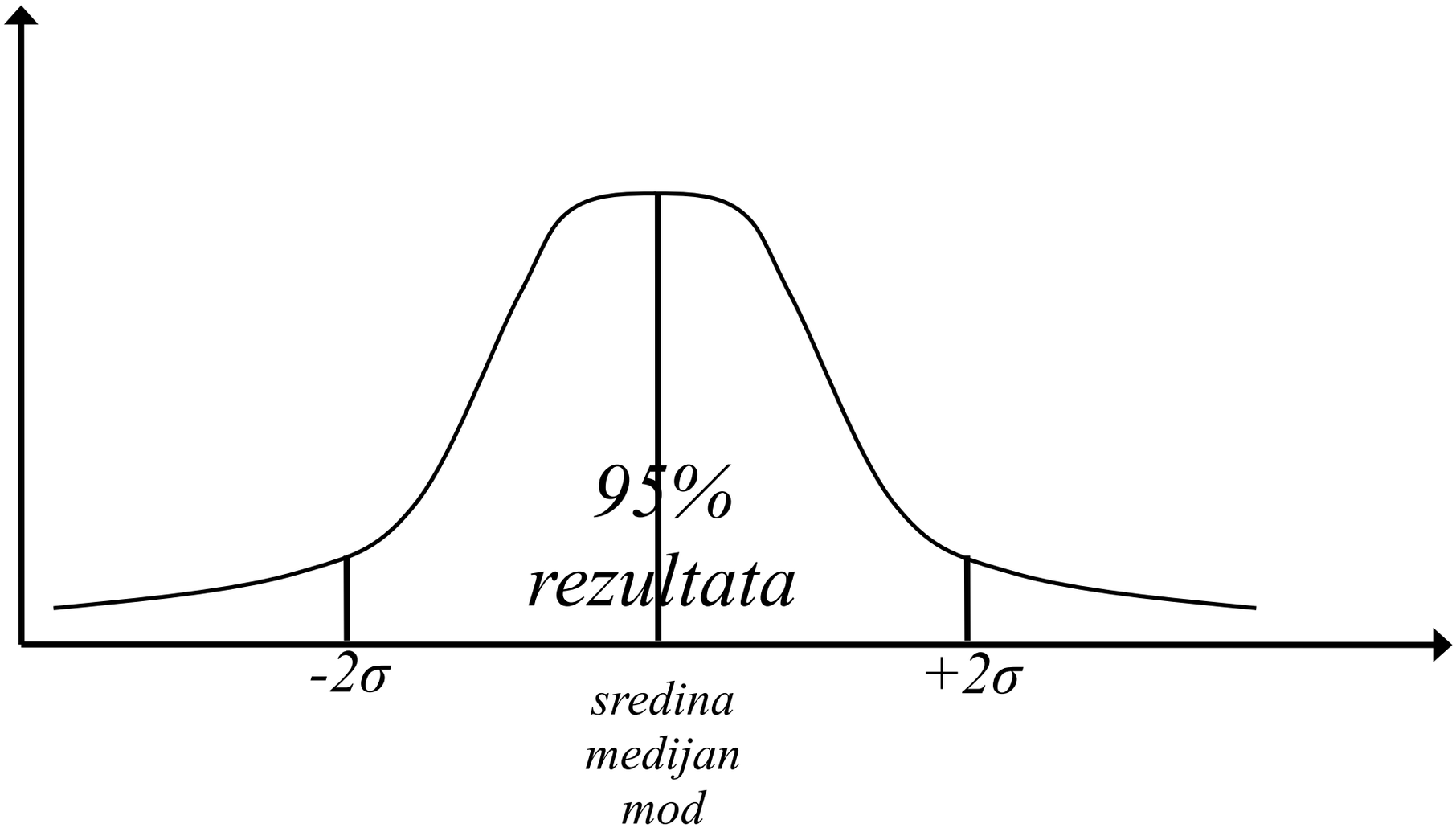
Deskriptivna statistika

- Notacija
  - s^2 = varijanca
  - σ^2 = varijanca populacije
  - s = standardna devijacija
  - σ = standardna devijacija populacije

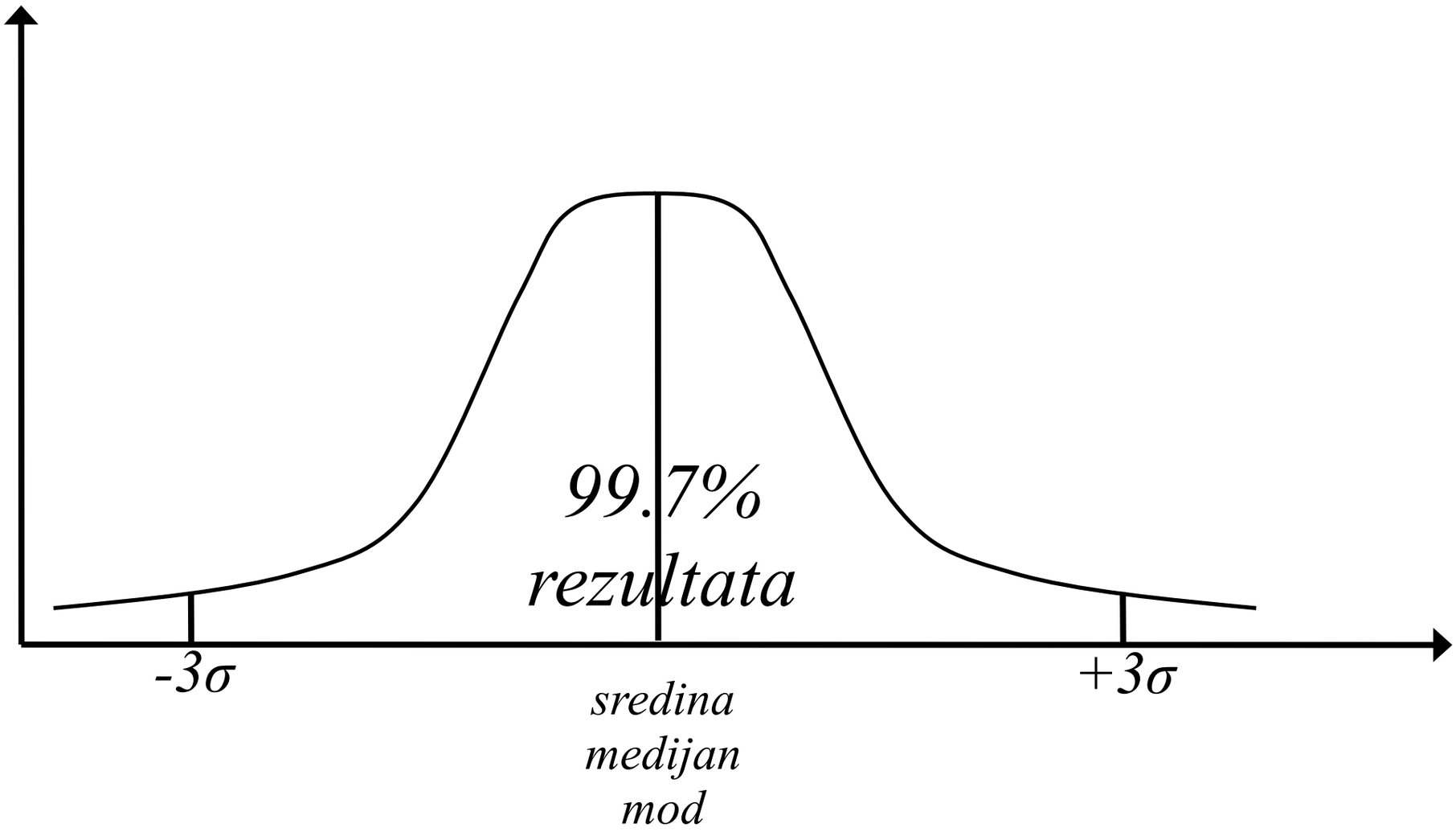
Deskriptivna statistika



Deskriptivna statistika



Deskriptivna statistika



Domaći

- Čitati
  - MS99: Poglavlje 2
  - Stat. Poglavlje 5