STATISTIKA

prof.dr.sc. Miljenko Marušić

Kontakt: miljenko.marusic@math.hr

WWW: http://www.math.hr/~rus/statistika

Vježbe:

Petra Daković

Literatura

Osnovna literatura:

Michael C. Whitlock, Dolph Schluter, The Analysis of Biological Data, Macmillan Learning (2020)

Dopunska literatura:

1. B. Petz, Osnovne statističke metode za nematematičare, 3. dopunjeno izdanje, Naklada Slap, Jastrebarsko, 1997.
2. Pivac, B. Šego, Statistika, Alka Script
3. Andy Hector, The New Statistics with R - An Introduction for Biologists, Oxford University Press (2021)
4. W.W. Daniel, C.L. Cross, Biostatistics: A Foundation for Analysis in the Health Sciences, 10th edition, Willey (2013)
5. A. Hector, The New Statistics with R: an Introduction for Biologists, 2nb edition, Oxford University Press (2021)
6. S. Siegel i N.J. Castellan, Jr., Nonparametric Statistics for the behavioral sciences, 2nd edition,McGraw-Hill, 1988.

Zadaci statistike:

- prikupljanje podataka,
- uređivanje i sažimanje,
- analiza podataka
- zaključivanje na osnovu prikupljenih podataka

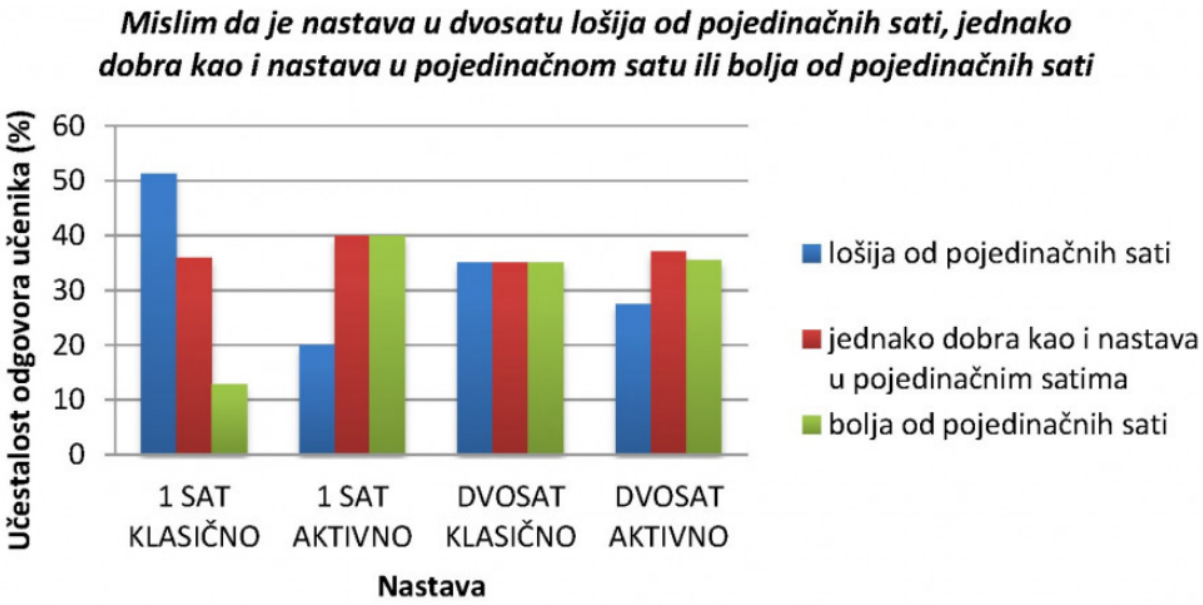
Ciljevi kolegija

- Sposobnost čitanja i razumijevanja raznih statističkih studija. Za razumijevanje ovakvih izvještaja osoba mora biti upoznata s riječnikom, simbolima, konceptima i postupcima korištenim u izvještaju.

- Sposobnost sažimanja podataka, donošenje generalnih zaključaka na osnovu rezultata istraživanja.

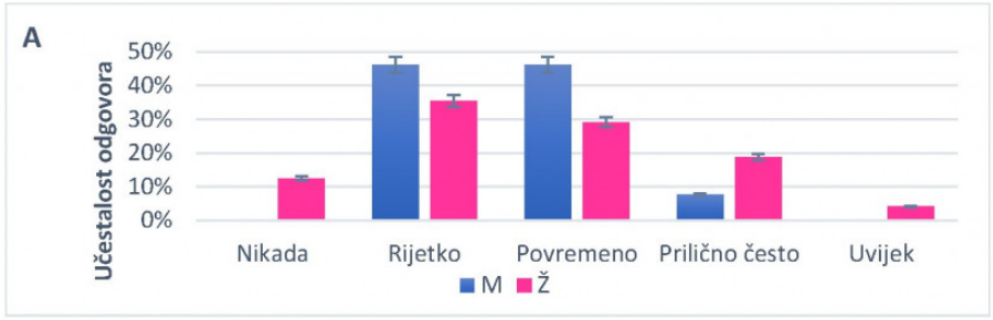
Primjeri

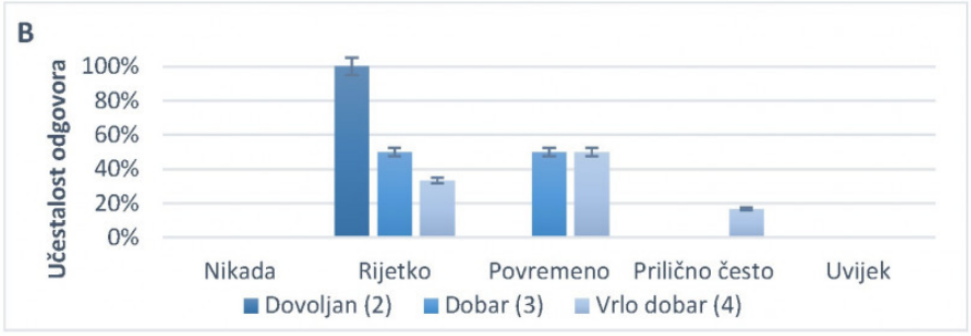
I. Labak, M. Heffer, I. Radanović, 'Stavovi učenika o nastavi prirode i biologije organiziranoj u dvosatu', Educatio biologiae : časopis edukacije biologije, 1 (2014)

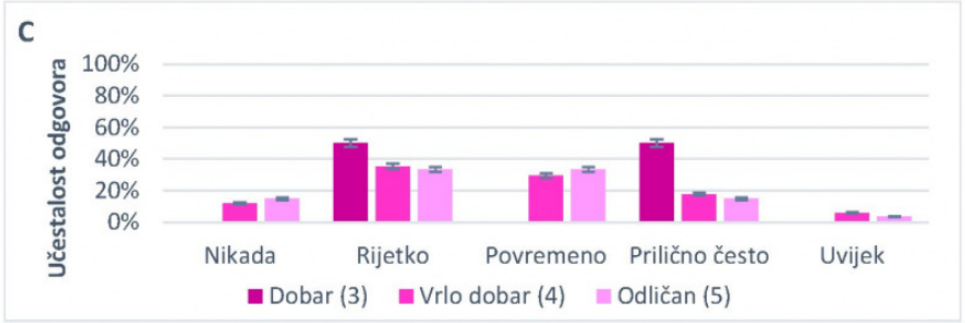


Slika 1 Usporedba odgovora svih učenika koji nastavu prate klasično/aktivno kao i kroz pojedinačan sat/dvosat na pitanje „Mislim da je nastava u dvosatu lošija od pojedinačnih sati, jednako dobra kao i nastava u pojedinačnom satu ili bolja od pojedinačnih sati"

I. Labak, I. Kligl, 'Navike učenika u samovrednovanju postignuća učenja', Educatio biologiae : časopis edukacije biologije, 5 (2019)







Slika 3 Usporedba odgovora učenika i učenica na tvrdnju „Tijekom nastavnog sata postavljam si različita pitanja kojima preispitujem vlastito razumijevanje sadržaja kojeg učimo." A usporedba odgovora s obzirom na spol; B usporedba odgovora s obzirom na zaključnu ocjenu iz biologije kod učenika; C usporedba odgovora s obzirom na zaključnu ocjenu iz biologije kod učenica

A. Bahat, Ž. Lukša, ’Primjena strategija aktivnog učenja i poučavanja u nastavi prirode i društva’, Educatio biologiae : časopis edukacije biologije, 5 (2019)

Tablica 1 Samoprocjena učitelja o učestalost primjene pojedinih strategija učenja i poučavanja (N=116)

Vrsta poučavanja	N	Min	Max	M	SD
Problemsko učenje	116	1	5	3,50	0,450
Istraživačko učenje	116	1	5	3,14	0,551
Suradničko učenje	116	1	5	3,70	0,488
Aktivnosti usmjerene na razvoj komunikacijskih kompetencija učenika i njihove kompetencije učenja	116	1	5	3,50	0,456
Aktivno učenje uz primjenu suvremene informacijsko-komunikacijske tehnologije	116	1	5	2,90	0,631
Direktno poučavanje	116	1	5	3,93	0,414
Ukupno za aktivno učenje	**116**	**1**	**5**	**3,36**	**0,374**

A. Bahat, Ž. Lukša, ’Primjena strategija aktivnog učenja i poučavanja u nastavi prirode i društva’, Educatio biologiae : časopis edukacije biologije, 5 (2019)

T-test za zavisne uzorke pokazao je da postoji statistički značajna razlika između primjene direktnog poučavanja i aktivnih strategija učenja i poučavanja ($t = 69,689$; $p < 0,01$). Učitelji češće koriste direktno poučavanje ($M = 3,93$; $SD = 0,414$) nego aktivne oblike učenja i poučavanja ($M = 3,36$; $SD = 0,374$). Analiza rezultata također pokazuje da učitelji smatraju kako aktivne oblike primjenjuju povremeno (1x mjesečno), dok direktno poučavanje koriste često (1x tjedno).

Rezultati pokazuju da ne postoji statistički značajna razlika ovisno o stupnju obrazovanja učitelja ni za jednu subskalu strategija učenja i poučavanja (tablica 2).

Zhengtong Lv, Shubin Lv, 'Clinical characteristics and analysis of risk factors for disease progression of COVID-19: A retrospective Cohort Study', Int. J. Biol. Sci., 17 (2021), pp. 1-7.

Table 2. Clinical characteristics of severe and non-severe COVID-19 patients

Items	All patients (n = 409)	Severe patients (n = 48)	Non-severe patients (n = 361)
Gender			
Male	188, 46.0%	24, 50.0%	164, 45.4%
Female	221, 54.0%	24, 50.0%	197, 54.6%
Age	50.47±12.43	62.98±6.80	48.80±12.05
Disease type			
Mild	46, 11.2%	0, 0%	46, 12.7%
Common	315, 77.0%	0, 0%	315, 87.3%
Severe	48, 11.8%	48, 100%	0, 0%
Fever			
Yes	391, 95.6%	46, 95.8%	345, 95.6%
No	18, 4.4%	2, 4.2%	16, 4.4%

Zhengtong Lv, Shubin Lv, 'Clinical characteristics and analysis of risk factors for disease progression of COVID-19: A retrospective Cohort Study', Int. J. Biol. Sci., 17 (2021), pp. 1-7.

Statistical analysis

Frequency and percentage are used to describe the categorical variables, and the continuous variables were described by mean and standard deviation. Categorical variables were analyzed using Chi-squared test. Continuous variables were analyzed using Student's t test. Variables showing a P value < 0.05 in univariable analysis were considered as candidates for the multivariate Cox regression model, and a forward stepwise method was used to determine the final multivariable model. SPSS 19.0 was used for the statistical analysis.

Zhengtong Lv, Shubin Lv, 'Clinical characteristics and analysis of risk factors for disease progression of COVID-19: A retrospective Cohort Study', Int. J. Biol. Sci., 17 (2021), pp. 1-7.

Sample size calculation

In this study, univariate and multivariate Cox regression analysis were performed to determine the independent effect of risk factors for progression from non-severe to severe COVID-19 or death. For sample size estimation, at least 10 samples are generally required as events per variable [9]. There were 17 independent variables in this study, so each group needed at least 170 cases, and the total sample size needed at least 340 cases.

Statistical analysis

Li-Li Zhu et al., 'Deciphering the genomic and lncRNA landscapes of aerobic glycolysis identifies potential therapeutic targets in pancreatic cancer ', Int. J. Biol. Sci., 17 (2021), pp. 107-118.

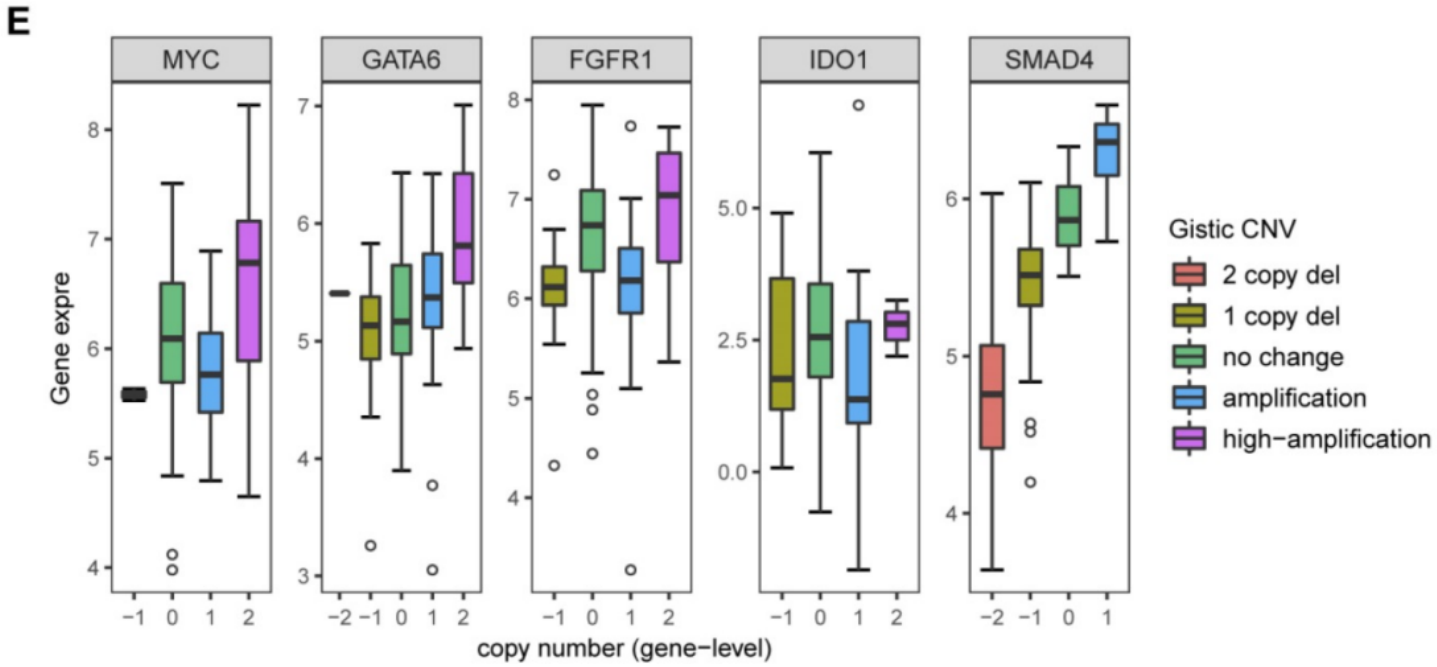


Figure 2. CNVs related to PDAC glycolysis. (A) Specific copy number profiles (gains in red and losses in blue) for glycolysis-high PDAC samples. **(B)** Significant CNVs from the glycolysis-high group. Each rectangle represents a PDAC subject. **(C)** Specific copy number profiles (gains in red and losses in blue) for glycolysis-low PDAC samples. **(D)** Significant CNVs from the glycolysis-low group. Each rectangle represents a PDAC subject. **(E)** Association of CNVs and gene expression in *MYC*, *GATA6*, *FGFR1*, *IDO1*, and *SMAD4*.

Sadržaj kolegija.

- Deskriptivna statistika
- Osnove vjerojatnosti
- Procjene
- Testiranje hipoteze
- Korelacija i jednostavna linearna regresija
- Neparametarski testovi

Osnovni pojmovi.

Varijabla ili **obilježje** je karakteristika (svojstvo, atribut) koja može poprimiti različite vrijednosti.

Primjeri varijable (obilježja):

- visina (osobe)
- težina (osobe)
- indeks tjelesne težine (BMI)
- visina stabla
- proteklo vrijeme
- broj sadnica u rasadniku
- ocjena na ispitu
- vrsta životinje

Podatak je izmjerena vrijednost varijable.

Skup svih podataka čini **skup podataka** (*engl.* data set)

Primjer:

- Marko je visok 172 cm

  Obilježje: visina
  Podatak: 172 cm

- Za drugog natjecatelja vrijednost obilježja može biti drugačije.

  Ivan je visok 175 cm.

  Obilježje: visina
  Podatak: 175 cm

- Andrija je težak 65 kg.

  Obilježje: Težina
  Podatak: 65 kg

Populacija je skup jedinki koje su predmet proučavanja.

Definicija populacije ovisi o istraživanju.

Primjer.
Zanima nas broj učenika neke škole koji se bave planinarenjem.

Rezultat: za svakog učenika provjeriti da li se bavi planinarenjem.
Prebrojiti sve takve učenike.

Populacija: Svi učenici dotične škole.

Obilježje: Bavljenje planinarenjem.
Podatak: da/ne

Uzorak je (bilo koji) podskup populacije.

Primjer.

Populacija: Svi učenici neke škole.

Uzorak:

- Svi učenici prvih razreda.
- Svi učenici sedmih razreda.
- Svi učenici 3.a razreda.

Deskriptivna statistika se sastoji od prikupljanja, organizacije, sažimanja i prikaza podataka.

Inferencijalna statistika (statističko zaključivanje) se sastoji od generalizacije s uzorka na populaciju, testiranja hipoteza, određivanja veza između varijabli i predviđanja.

Deskriptivna statistika - Zaključujemo o skupu o kojem imamo podatke.

Inferencijalna statistika - Na temelju podataka o manjem skupu (uzorku) donosimo zaključke o svojstvima šireg skupa (populaciji).

Deskriptivna statistika

Zaključujemo o skupu o kojem imamo podatke.

Primjer.

Od 920 učenika Osnovne škole X, njih 230 je uključeno u izvannastavne aktivnosti. Dakle, 25% učenika je je uključeno u izvannastavne aktivnosti.

- Populacija: svi učenici Osnovne škole X;
- Za svakog je učenika poznato je li uključen u izvannastavnu aktivnost ili nije.

Inferencijalna statistika

Na temelju podataka o manjem skupu (uzorku) donosimo zaključke o svojstvima šireg skupa (populaciji).

Primjer.

Od 30 učenika 2.a razreda Osnovne škole X, njih 12 je uključeno u izvannastavne aktivnosti. Dakle, 40% učenika Osnovne škole X je je uključeno u izvannastavne aktivnosti.

- Populacija: svi učenici Osnovne škole X;
- Uzorak: učenici 2.a razreda Osnovne škole X;
- Za svakog je učenika 2.a razreda poznato je li uključen u izvannastavnu aktivnost ili nije.
- Na osnovu podataka o jednom razredu zaključujemo o cijeloj školi.

Uzorak - poznati podaci

Populacija - nepoznati podaci ali donosimo zaključak o tom skupu

Klasifikacija varijabli (podataka)

Kvalitativne (kategorijske) varijable

- vrijednost varijable se nalazi u točno jednoj kategoriji.

Kvantitativne varijable

- vrijednost varijable je rezultat mjerenja na numeričkoj skali.

Primjeri kvalitativnih varijabli:

zemlja porijekla:
Austrija, Italija, Mađarska, Češka, Slovenija, Francuska, ...

spol
muški, ženski

porodica:
Talpa europaea, Rhinolophidae, Vespertilionide, Molossidae, Sciuridae, Myoxidae, Muridae, Delphinidae, Canidae, Mustelidae, Felidae, Phocidae, Accipitridae, Aegithalidae, Alaudidae, Alcedinidae, Anatidae, Apodidae, ...

Boja kljuna:
žut, narančast, crven, ...

Primjeri kvalitativnih varijabli:

ocjena:

- nedovoljan (1)
- dovoljan (2)
- dobar (3)
- vrlo dobar (4)
- odličan (5)

stupanj zadovoljstva :

- jako nezadovoljan
- nezadovoljan
- zadovoljan
- jako zadovoljan

Primjeri kvantitativnih varijabli:

- broj stabala na parceli
- kapacitet rasadnika
- visina stabla
- duljina potoka

Podjela kvalitativnih varijabli:

nominalne varijable - ne postoji uređaj između kategorija

redoslijedne (ordinalne, uređajne) varijable - postoji uređaj između kategorija

Nominalne varijable: Nema bolje (gore) odnosno veće (manje) vrijednosti

Primjeri nominalnih varijabli:

zemlja porijekla:
Austrija, Italija, Mađarska, Češka, Slovenija, Francuska, ...

spol
muški, ženski

porodica:
Talpa europaea, Rhinolophidae, Vespertilionide, Molossidae, Sciuridae, Myoxidae, Muridae, Delphinidae, Canidae, Mustelidae, Felidae, Phocidae, Accipitridae, Aegithalidae, Alaudidae, Alcedinidae, Anatidae, Apodidae, ...

Boja kljuna:
žut, narančast, crven, ...

Redoslijedna varijabla:

- Za bilo koje dvije vrijednosti možemo reći koja je bolja (gora) odnosno veća (manja)
- Postoji uređaj

Primjer:

- stupanj zadovoljstva
- ocjena

Podjela kvantitativnih varijabli:

diskretne varijable - skup vrijednosti je konačan ili prebrojiv

neprekidne varijable - poprimaju sve vrijednosti unutar nekih granica

Diskretne varijable

- Konačan skup vrijednosti
  Broj pogodaka u gađanju puškom

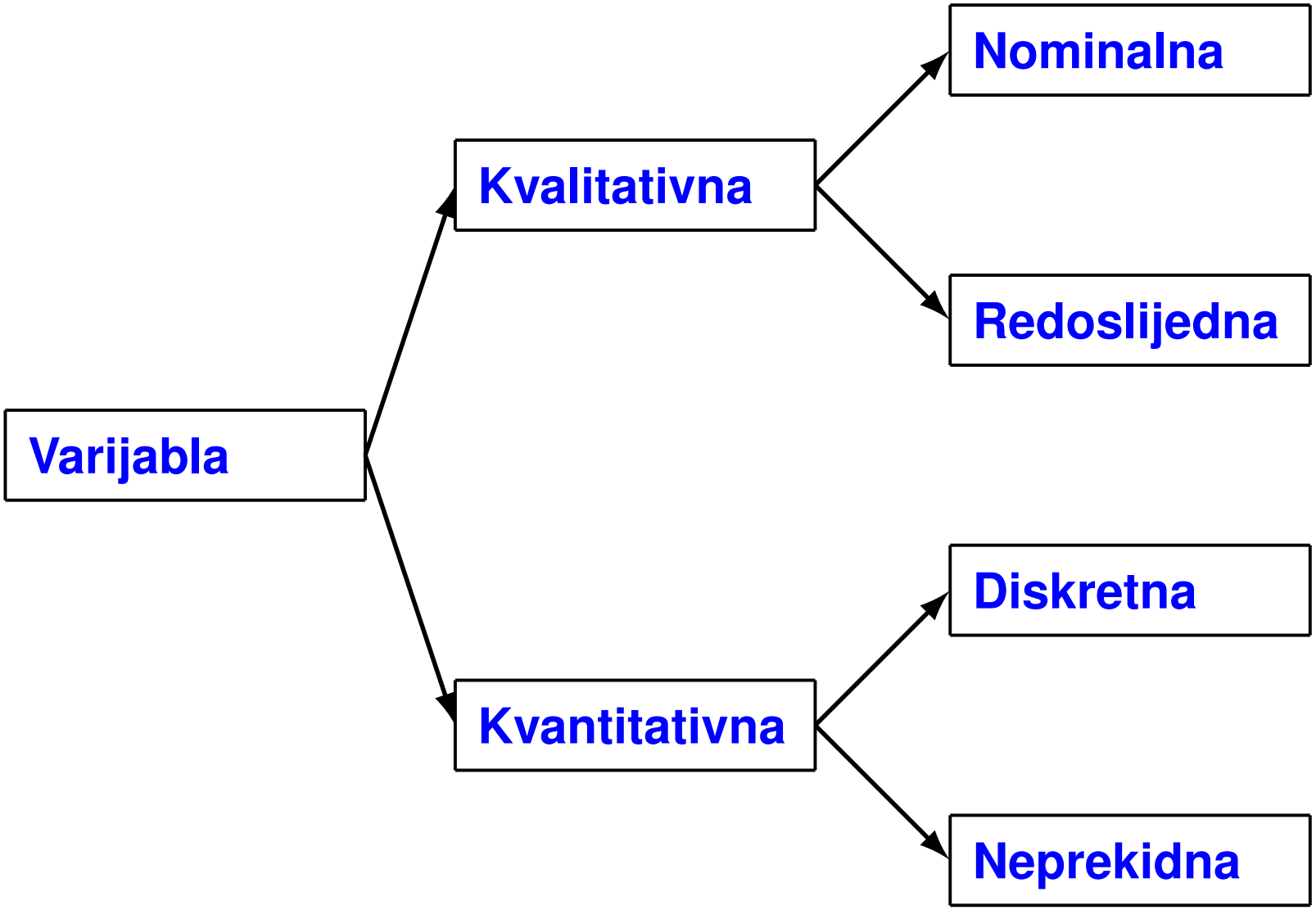
- Prebrojiv skup vrijednosti
  - npr. ukoliko vrijednost varijable može biti bilo koji cijeli broj
  - skup vrijednosti je beskonačan

- rezultat prebrajanja

Neprekidne varijable

- visina stabla, duljina potoka, vrijeme
- varijabla može poprimiti bilo koju vrijednost unutar nekih granica
- rezultat mjerenja

Klasifikacija varijabli.



Primjer.

Na stablu su 23 jabuke.

Varijabla: Broj jabuka na stablu.

Tip varijable: kvantitativna - diskretna

Primjer.

Ivan je odvozio slalom za 55 s.

Varijabla: Vrijeme vožnje.

Tip varijable: kvantitativna - neprekidna

Primjer.

Ivica je pretrčao maraton.

Varijabla: Pretrčan maraton.

Tip varijable: kvalitativna - nominalna

Primjer.

Marica je za svoj nastup ocjenjena s ocjenom 8.

Varijabla: Ocjena nastupa.

Tip varijable: kvalitativna - redoslijedna

Napomena. Varijable mogu biti i kombinacija navedenih varijabli. Npr. bodovi u skijaškim skokovima - kombinacija duljine skoka i ocjene sudaca.

Primjer.

Blue Storm je na cilj stigao šesti.

Varijabla: Plasman na utrci.

Tip varijable: kvalitativna - redoslijedna

Napomena. Iako je vrijednost numerička, radi se o kvalitativnoj varijabli.

Napomena. Ponekad se kvantitativna varijabla naziva i numerička (v. knjiga str. 76).
Mnogi to krivo interpretiraju da je vrijednost numerička.

Frekvencije

Na zahtjev vlasnika farme, radnik je bilježio spol zečeva. Sljedeći dan je rezultat proslijedio svome nadređenom:

M M Ž Ž MŽ M M Ž M Ž M Ž M Ž Ž M Ž Ž M Ž Ž M Ž M Ž M Ž M Ž M
Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M M M Ž M
Ž M Ž M Ž M Ž Ž M Ž M Ž M Ž Ž Ž M Ž M Ž Ž M Ž M Ž M Ž M Ž M Ž M
Ž Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M Ž M
Ž M Ž M Ž Ž M Ž M Ž M Ž Ž M Ž M Ž M Ž Ž Ž Ž M Ž M Ž M Ž M Ž M Ž M
Ž M Ž M Ž M Ž M Ž Ž M Ž Ž M Ž M Ž Ž M Ž M Ž M Ž Ž M Ž Ž M Ž Ž M
Ž M Ž M Ž M Ž Ž M Ž M Ž Ž Ž M Ž Ž M Ž Ž M Ž M Ž Ž Ž M Ž M Ž M Ž
M Ž Ž M Ž M Ž M Ž M Ž M Ž Ž M Ž Ž M Ž M Ž M Ž Ž M Ž M Ž M Ž M Ž
Ž M Ž Ž M Ž M Ž Ž M Ž M Ž M Ž Ž M Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž
Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž
Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž Ž M Ž M Ž Ž M Ž Ž M Ž Ž M Ž M Ž Ž
M Ž M Ž M Ž Ž M Ž M Ž M Ž M Ž M Ž M Ž Ž M Ž M Ž Ž M Ž Ž M Ž M Ž
Ž M Ž M Ž M Ž M Ž M Ž M Ž

Sirovi podaci

- originalni rezultat mjerenja

nepregledno!

Niz podataka

- podaci posloženi po veličini (redu)

Niz podataka

M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M M M M M M M M M M M M M M M M M M M M M M M M M
M M M M Ž
Ž Ž
Ž Ž
Ž Ž
Ž Ž
Ž Ž
Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž Ž

Niz podataka je još uvijek nepregledan.

Prebrojimo: U nizu se M pojavljuje 172 puta a Ž 225 puta.

Frekvencija je broj pojavljivanja određene vrijednosti varijable (u skupu podataka).

Frekvencija se ponekada naziva još i **apsolutna frekvencija**.

Oznaka: f_i - frekvencija i-te vrijednosti

U primjeru: 172 mužjaka i 225 ženki

$$f_1 = 172, \qquad f_2 = 225$$

veličina skupa = zbroj frekvencija ($N = \sum f_i$)

Distribucija frekvencija - niz vrijednosti obilježja i pripadajućih frekvencija.

Primjer. U prošlom primjeru bilo je 172 mužjaka i 225 ženki.

Distribucija frekvencija:

Spol	frekvencija
M	172
Ž	225

Relativna frekvencija je omjer frekvencije i veličine skupa:

$$r_i = \frac{f_i}{N}$$

Oznaka: r_i je relativna frekvencija i-te vrijednosti.

Kumulativna frekvencija je broj jedinki s vrijednošću obilježja manjim ili jednakim od pojedine vrijednosti obilježja.

postupno zbrajanje frekvencija: $c_i = f_1 + f_2 + \ldots + f_i$

frekvencije pozitivne $\Rightarrow$ kumulativne frekvencije su rastuće (nepadajuće)

Relativna kumulativna frekvencija je omjer kumulativne frekvencije i veličine skupa.

Relativna kumulativna frekvencija je **udio** jedinki s vrijednošću obilježja manjim ili jednakim od pojedine vrijednosti obilježja.

Primjer.

Dob	Frekvencija	Relativna frekvencija	Kumulativna frekvencija	Relativna kumulativna frekvencija
19	1	0.05	1	0.05
20	0	0.00	1	0.05
21	5	0.25	6	0.30
22	2	0.10	8	0.40
23	2	0.10	10	0.50
24	1	0.05	11	0.55
25	4	0.20	15	0.75
26	1	0.05	16	0.80
27	0	0.00	16	0.80
28	1	0.05	17	0.85
29	0	0.00	17	0.85
30	1	0.05	18	0.90
31	1	0.05	19	0.95
32	1	0.05	20	1.00

Prikazivanje podataka

- Tablično prikazivanje
- Grafičko prikazivanje

Primjer.

Podaci:

Spol	Dob
M	21
Ž	25
M	19
Ž	32
Ž	21
M	25
M	23
M	25
Ž	21
M	24

Spol	Dob
M	26
Ž	23
Ž	22
M	21
Ž	30
Ž	28
M	22
M	31
Ž	25
M	21

Tablica.

Spol	f
Muški	11
Ženski	9

ili

Spol	f
Muški	11
Ženski	9
Ukupno	20

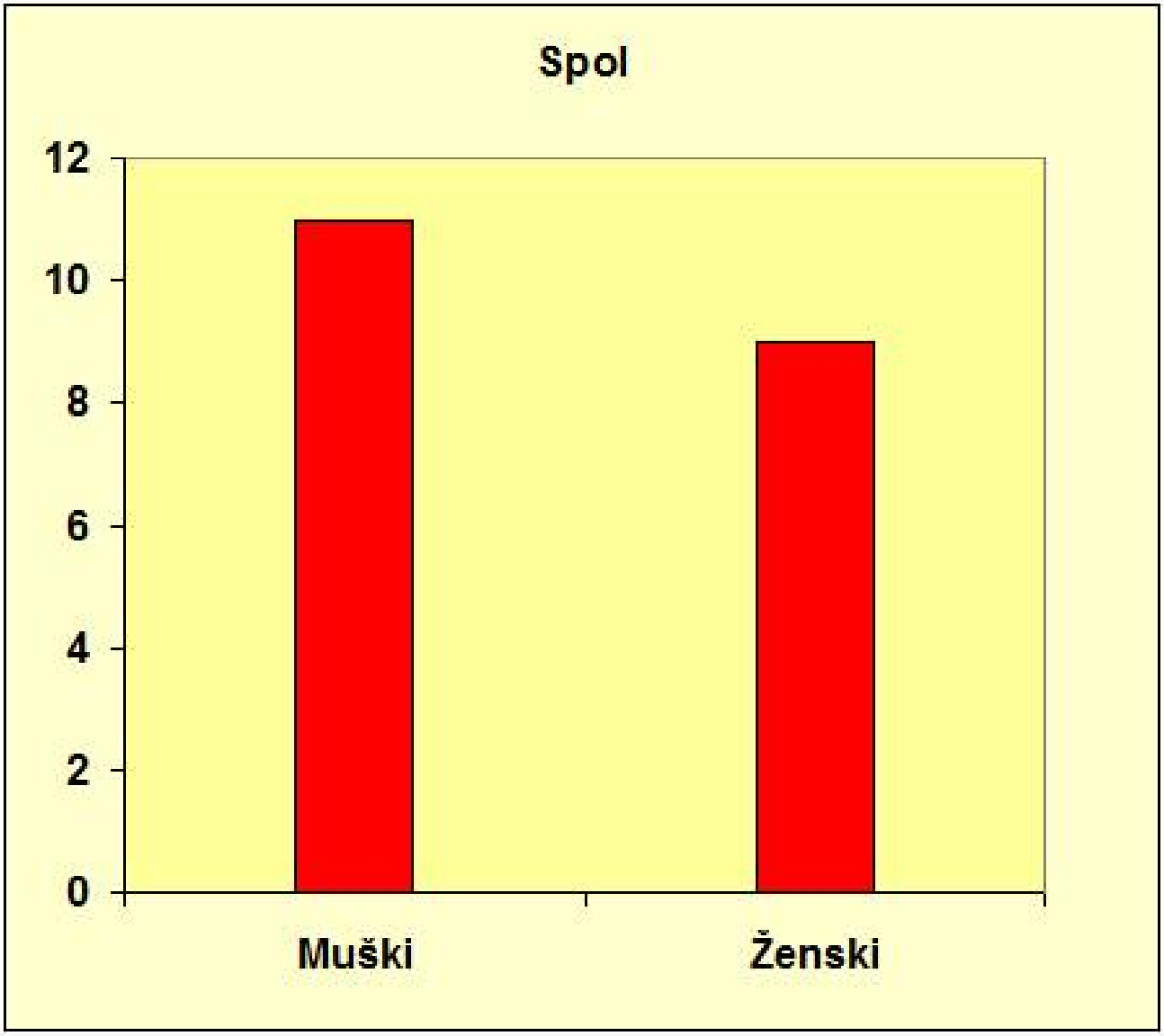
Spol	Frekvencija f	Relativna frekvencija r	Relativna frekvencija (%)
Muški	11	0.55	55
Ženski	9	0.45	45
Ukupno	20	1.00	100

Grafički prikaz podataka.

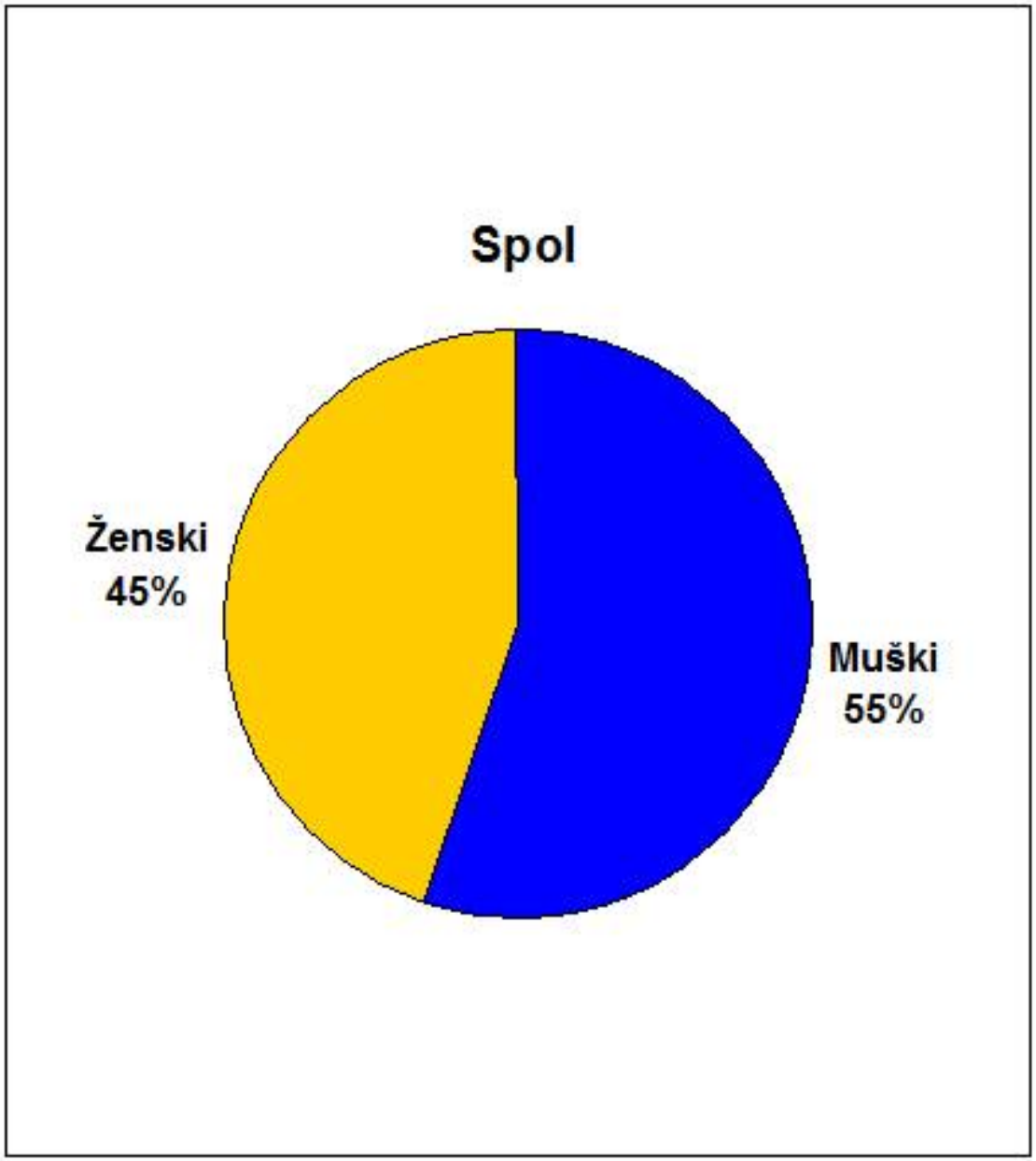
- jednostavnost
- preglednost
- ali ne i preciznost

- površinski grafikoni
  - grafikon stupaca
  - grafikon krugova i polukrugova
- linijski grafikoni
- kartogrami
- slikovni grafikoni (piktograf,..)

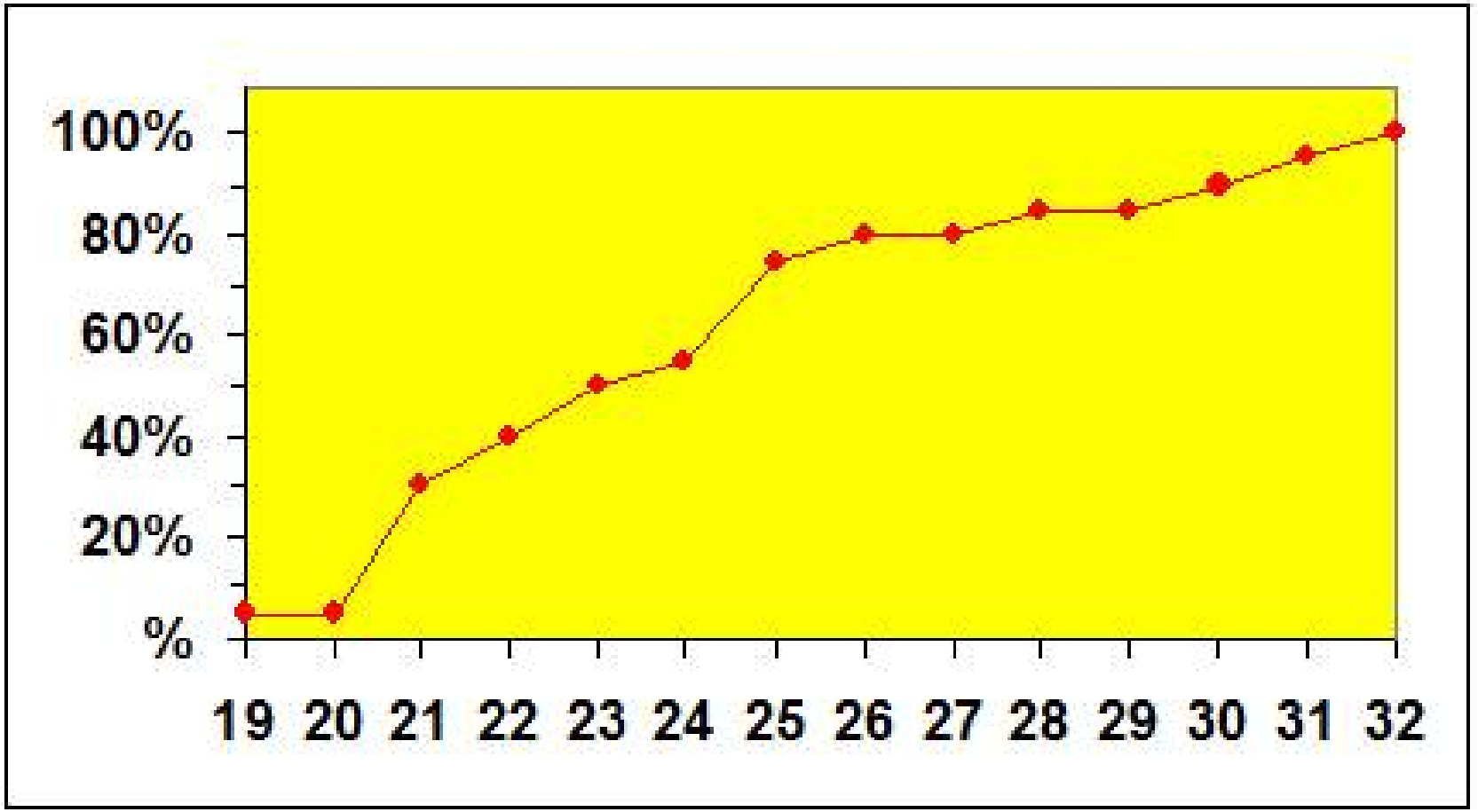
Grafikon stupaca.



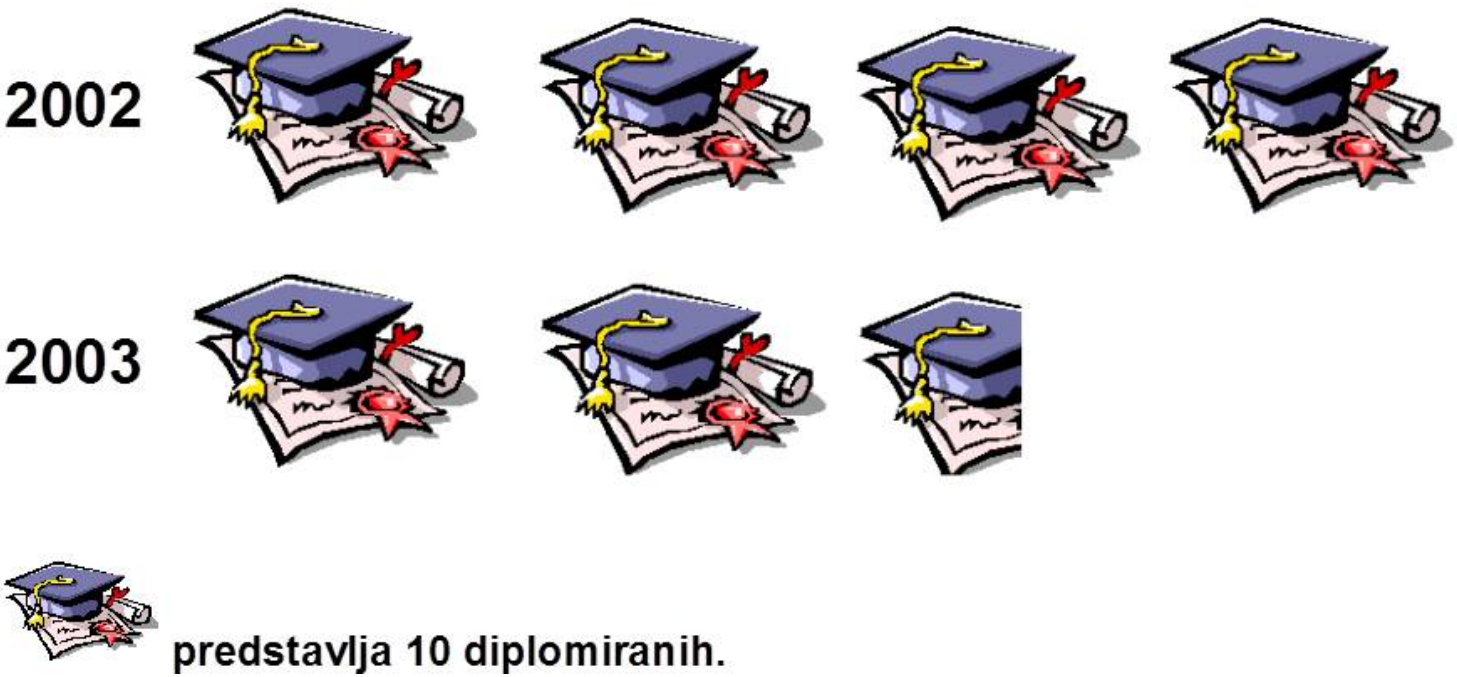
Grafikon krugova (pita).



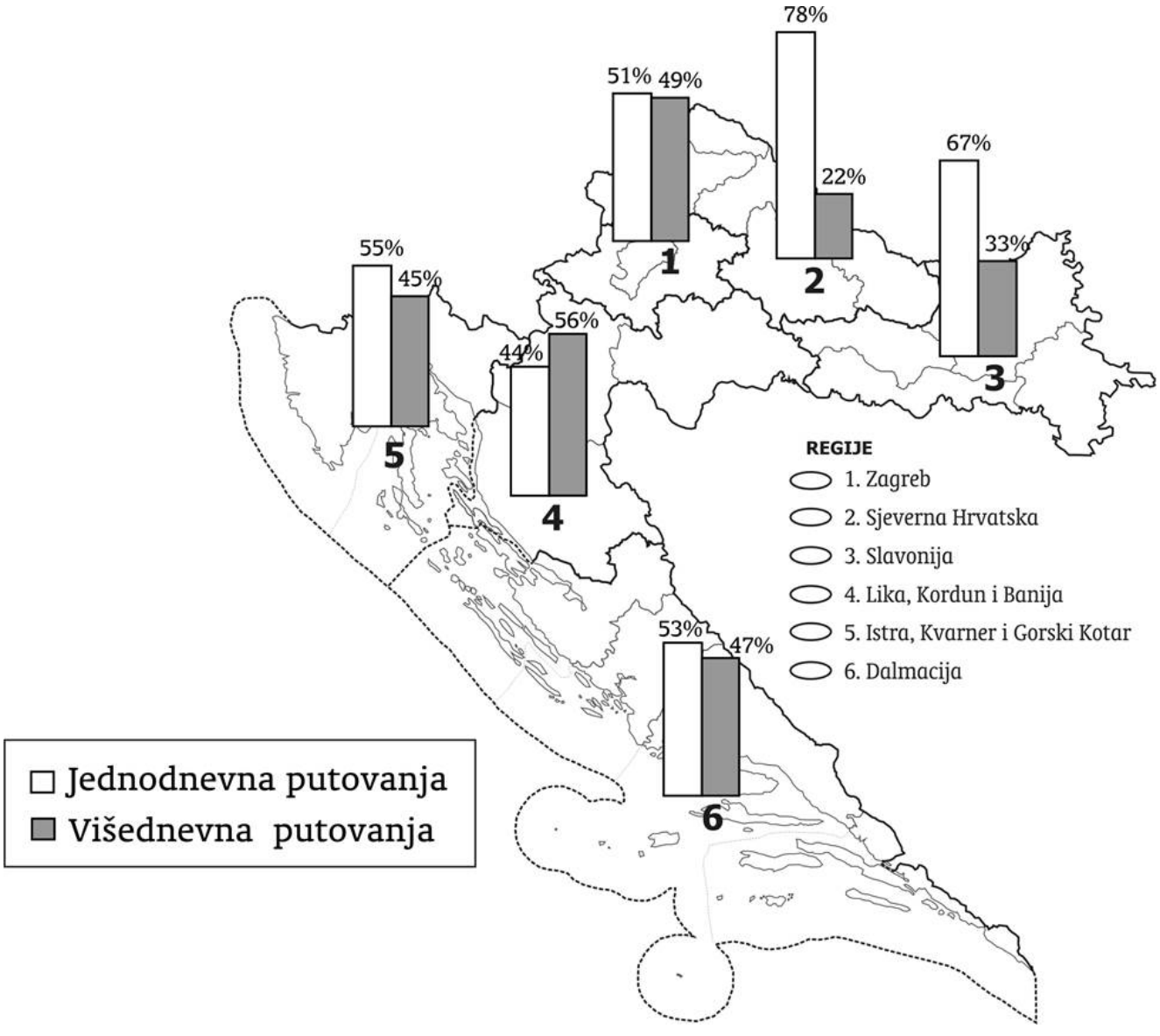
Linijski grafikon.



Piktograf.



Kartogram.

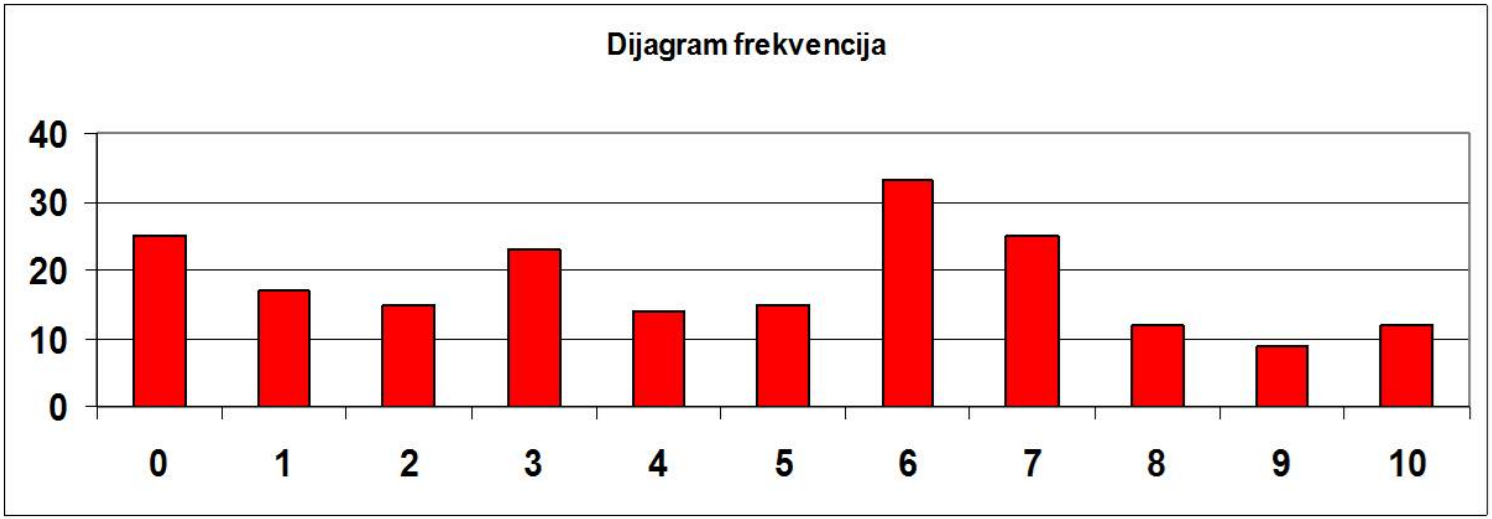

78%
51% 49%
67%
22%
55%
45%
33%
56%
44%
1
2
3
4
5
6
53%
47%
REGIJE
1. Zagreb
2. Sjeverna Hrvatska
3. Slavonija
4. Lika, Kordun i Banija
5. Istra, Kvarner i Gorski Kotar
6. Dalmacija
Jednodnevna putovanja
Višednevna putovanja

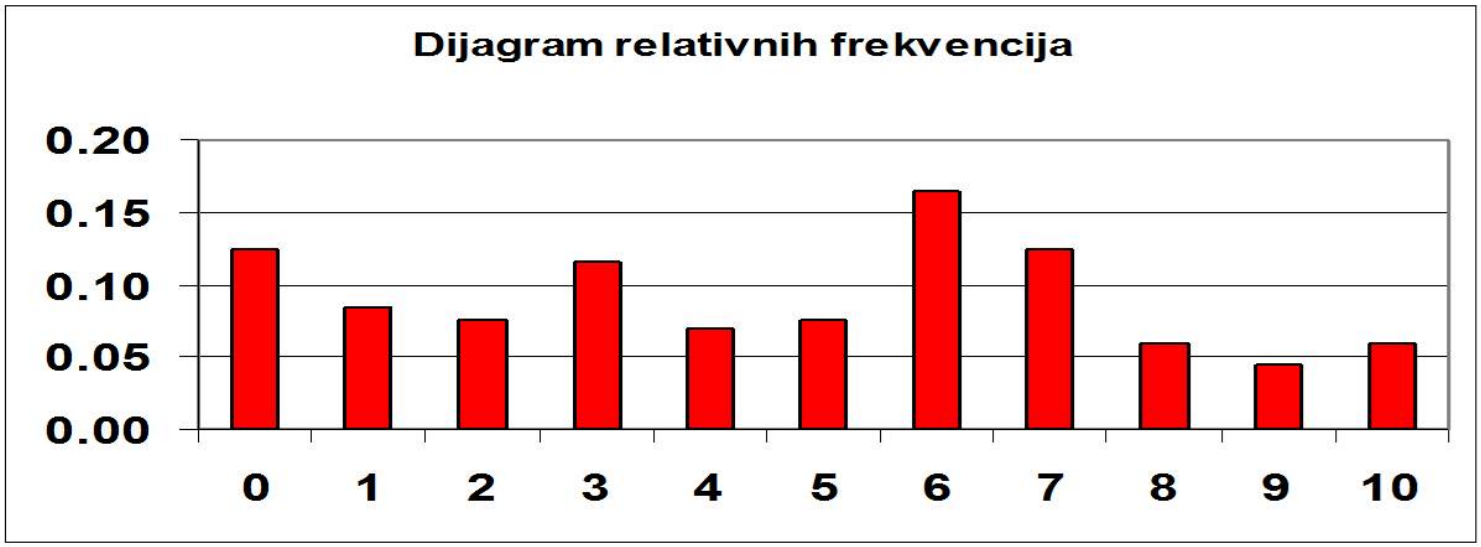
Distribucija frekvencija prikazuje se:

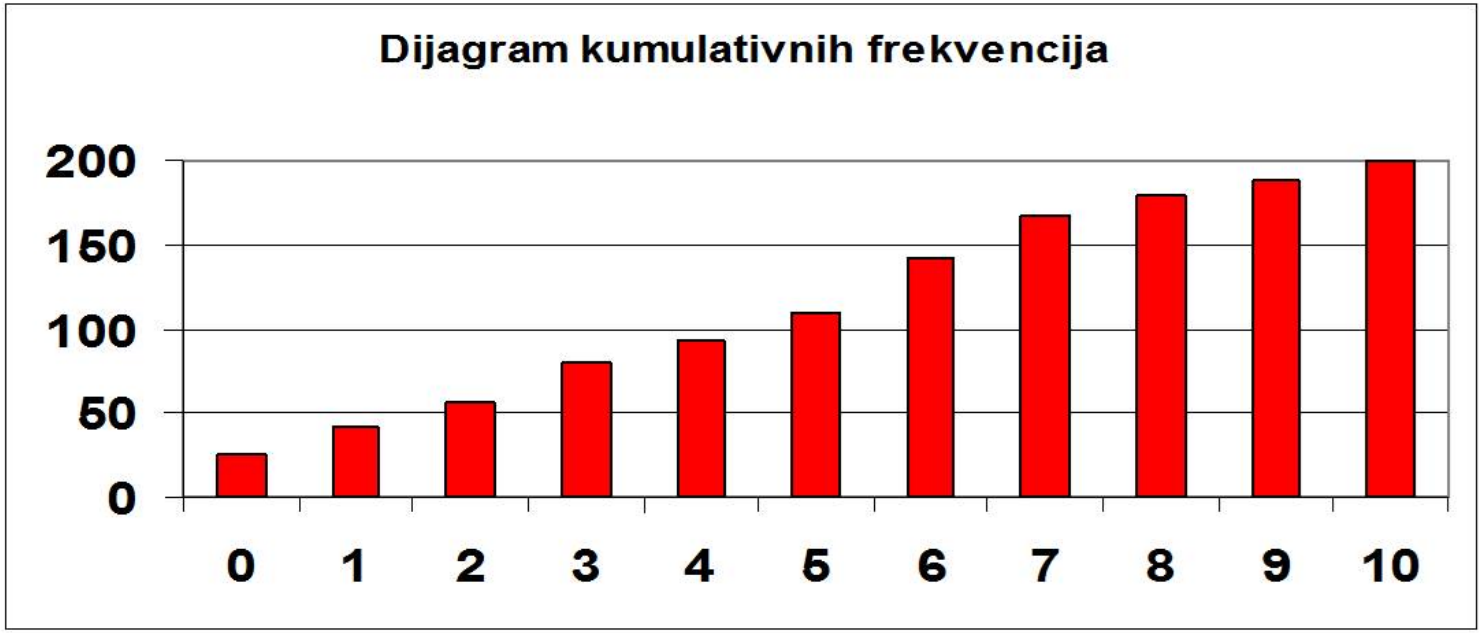
- tablicama
- histogramima - površinski grafikon distribucije frekvencije
- poligonima frekvencija - linijski grafikon
- stepenasti linijski grafikon

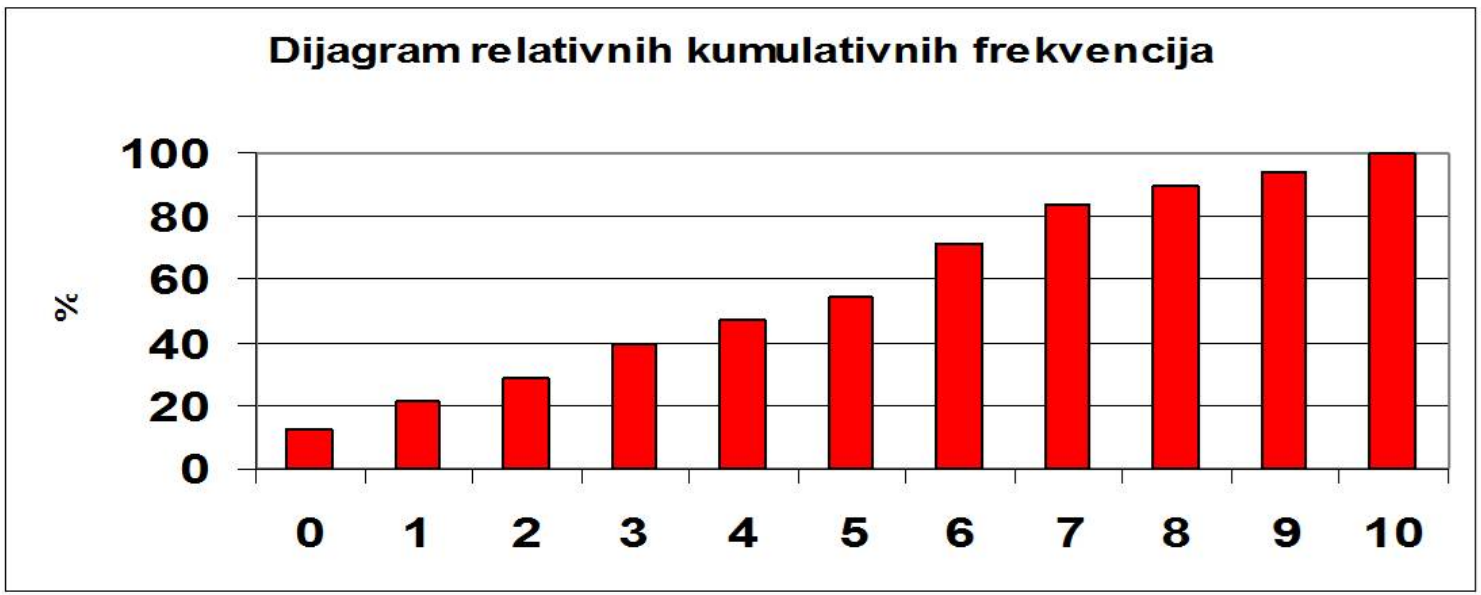
Primjer.

Bodovi na testu	Frekvencija	Kumulativna frekvencija	Relativna frekvencija	Relativna kumulativna frekvencija (%)
0	25	25	0.125	12.5
1	17	42	0.085	21.0
2	15	57	0.075	28.5
3	23	80	0.115	40.0
4	14	94	0.070	47.0
5	15	109	0.075	54.5
6	33	142	0.165	71.0
7	25	167	0.125	83.5
8	12	179	0.060	89.5
9	9	188	0.045	94.0
10	12	200	0.060	100.0
Ukupno	200		1.00	


Dijagram frekvencija
40
30
20
10
0
0 1 2 3 4 5 6 7 8 9 10


Dijagram relativnih frekvencija
0.20
0.15
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10


Dijagram kumulativnih frekvencija
200
150
100
50
0
0 1 2 3 4 5 6 7 8 9 10


Dijagram relativnih kumulativnih frekvencija
%
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10

Primjer.

Bodovi na testu	Kumulativna frekvencija
0	25
1	42
2	57
3	80
4	94
5	109
6	142
7	167
8	179
9	188
10	200

Koliko je studenata dobilo 5 ili 6 bodova?

$142 - 94 = 48$

Primjer.

Bodovi na testu	Relativna kumulativna frekvencija (%)
0	12.5
1	21.0
2	28.5
3	40.0
4	47.0
5	54.5
6	71.0
7	83.5
8	89.5
9	94.0
10	100.0

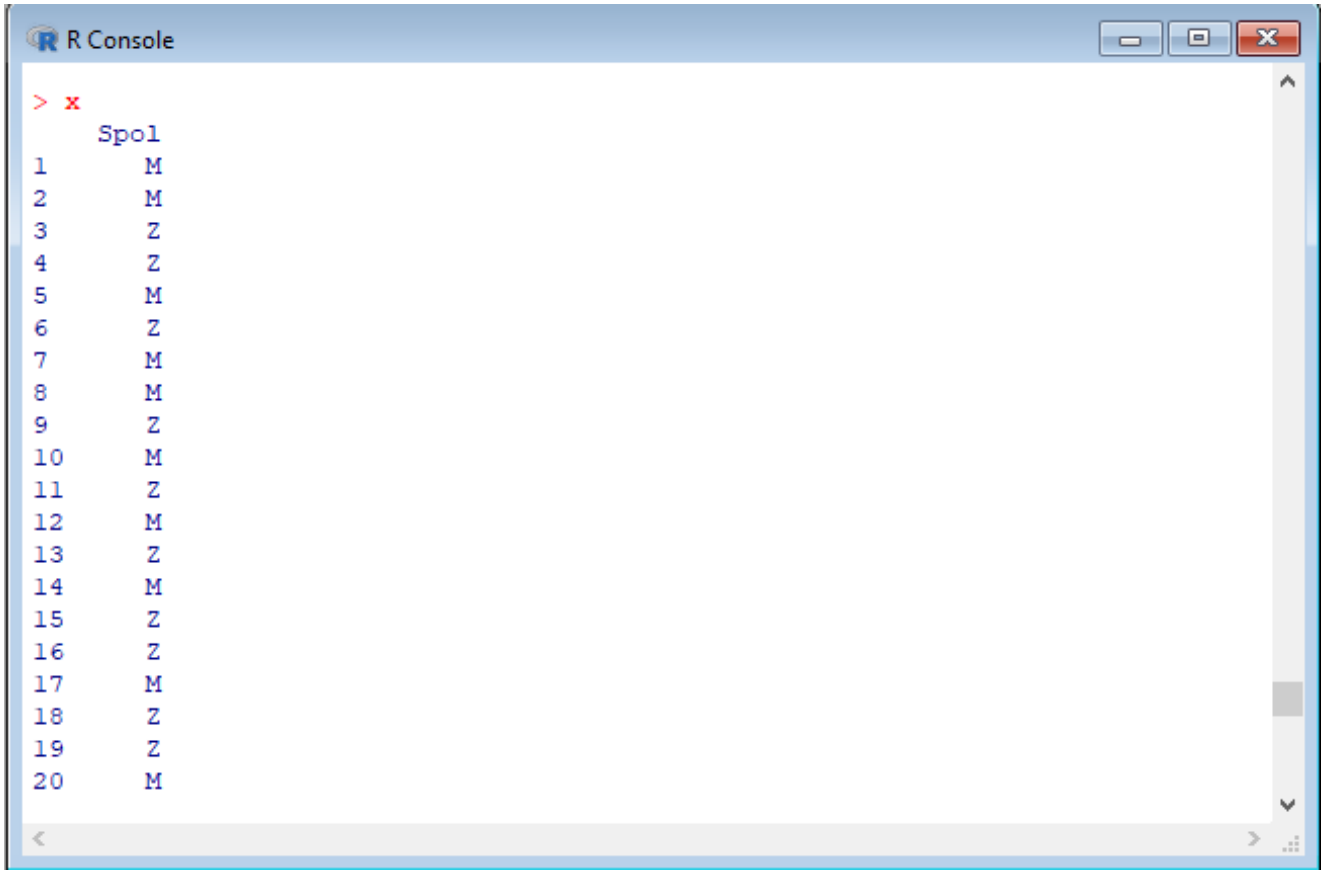
Koliko je studenata dobilo 7, 8 ili 9 bodova?

$94.0\% - 71.0\% = 23.0\%$

Frekvencije u R-u

Podaci su u datoteci data1.csv:

```
> x=read.csv("data1.csv")
```

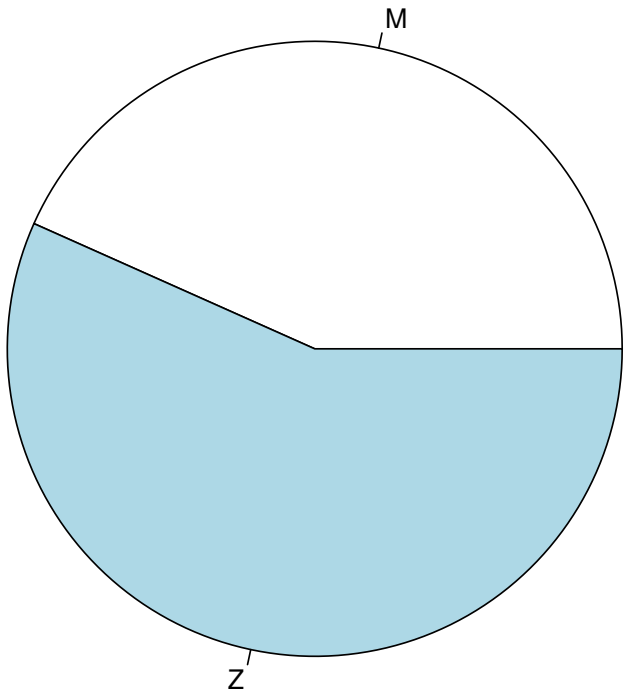


Frekvencije:

```
> tab <- table(x)
> print(tab)
x
  M   Z
172 225
```

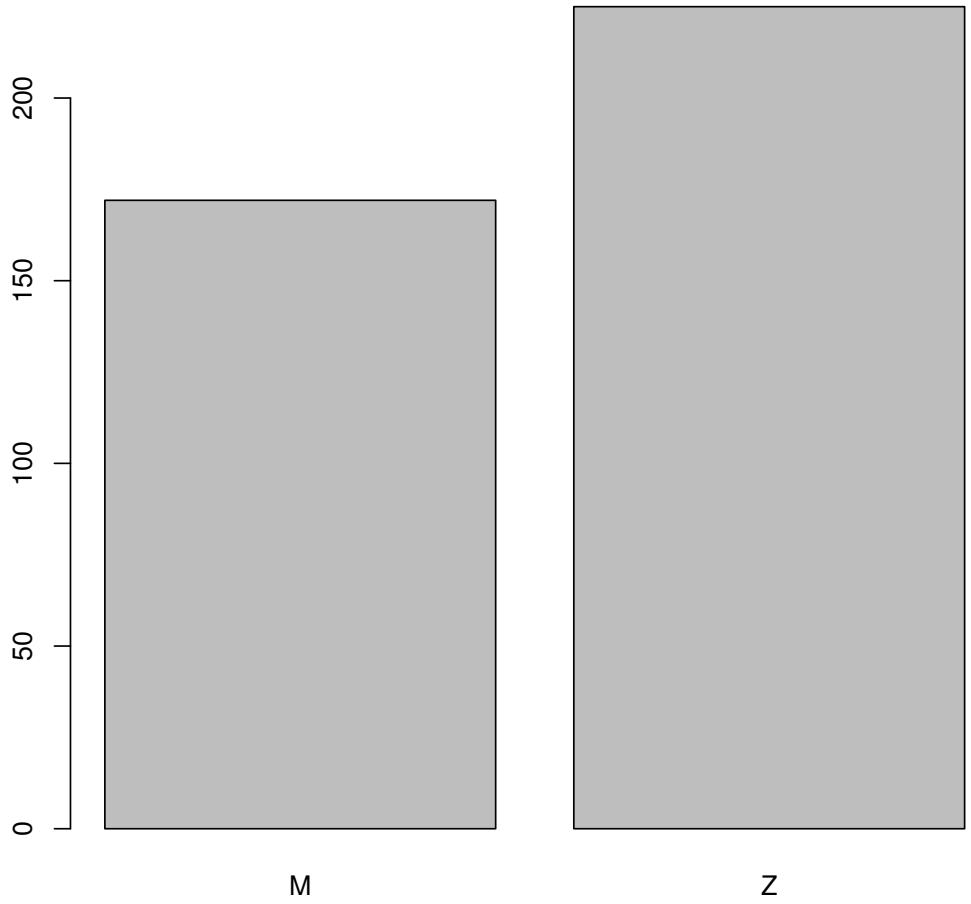
Strukturni krugovi:

```
> pie(tab)
```



Stupčasti grafikon:

```
> barplot(tab)
```



Grupiranje podataka (razredi)

Primjer.

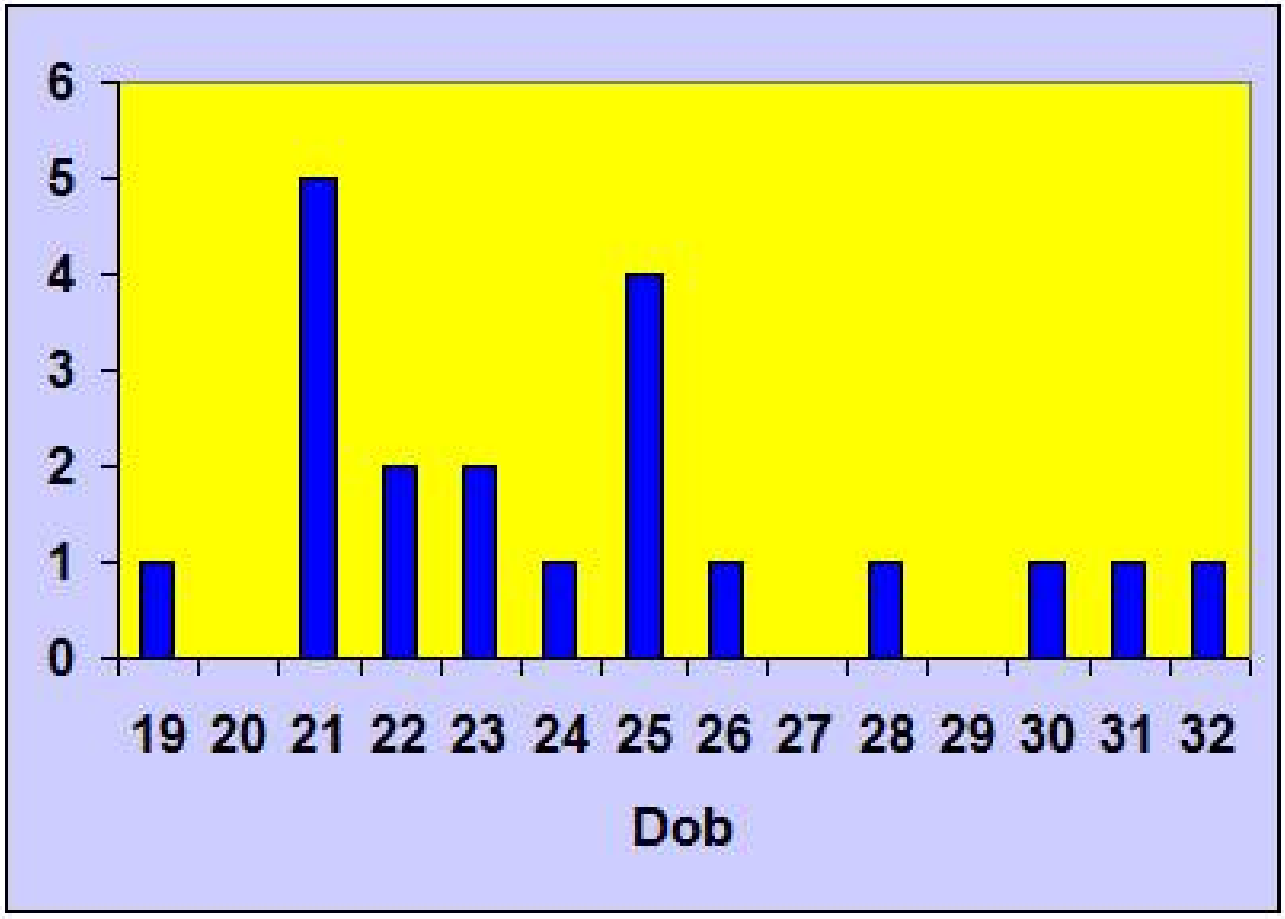
Varijabla DOB

- kvantitativna
- neprekidna

Distribucija frekvencija:

Dob	Frekvencija	Dob	Frekvencija
19	1	26	1
20	0	27	0
21	5	28	1
22	2	29	0
23	2	30	1
24	1	31	1
25	4	32	1

Distribucija frekvencija za varijablu dob:



Ukoliko je broj mogućih vrijednosti velik, podaci se grupiraju u razrede.

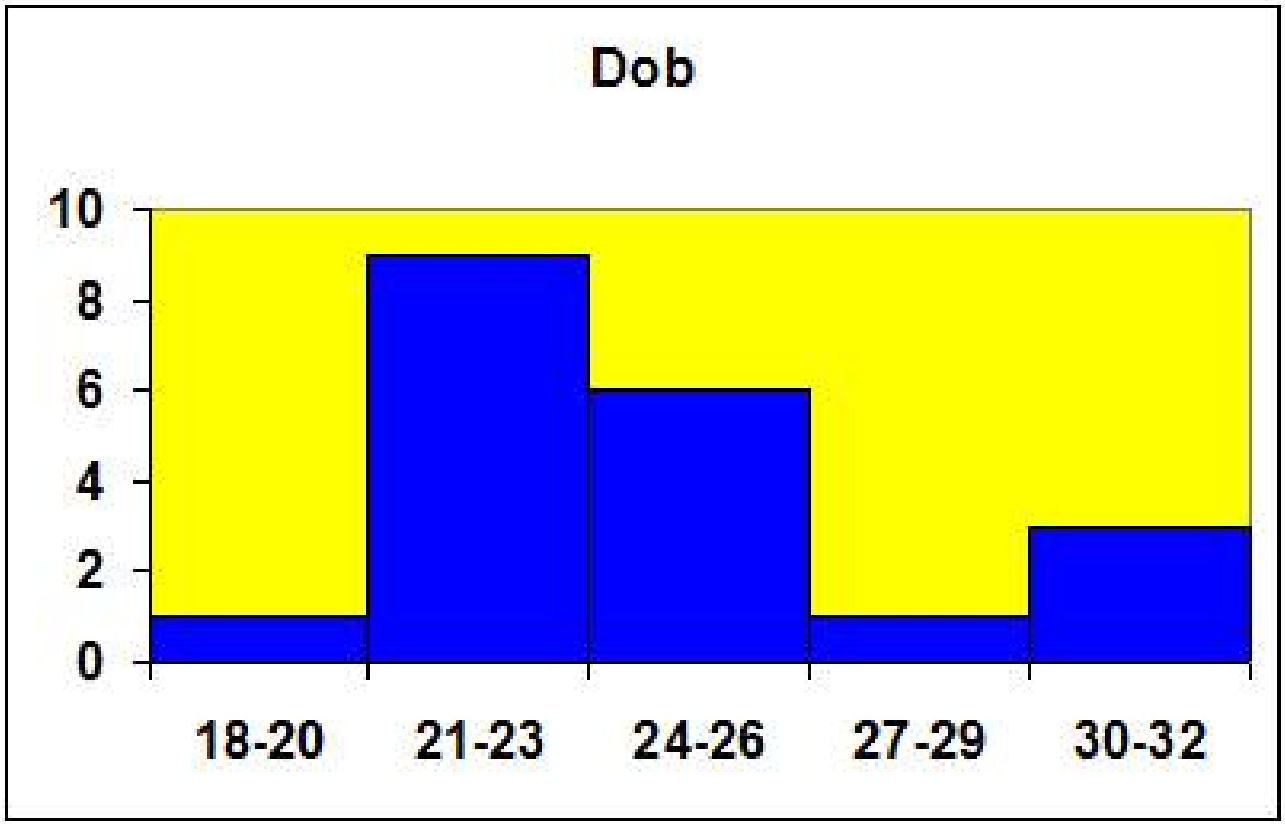
Razredi su intervali koji:

- potpuno pokrivaju cijeli skup mogućih vrijednosti obilježja (**načelo iscrpnosti** - svaki podatak se nalazi u nekom razredu)
- se ne preklapaju (**načelo isključivosti** - svaki podatak se nalazi samo u jednom razredu)

Razredi za varijablu DOB

- 18 – 20
- 21 – 23
- 24 – 26
- 27 – 29
- 30 – 32

Distribucija frekvencija za varijablu dob:



Frekvencija razreda - broj jedinki u razredu

Razred	Frekvencija razreda
18 – 20	1
21 – 23	9
24 – 26	6
27 – 29	1
30 – 32	3

Granice razreda - najmanja i najveća moguća vrijednost koja pripada tom razredu.

Oznaka:

- L_i donja granica i-tog razreda
- U_i gornja granica i-tog razreda

Razred	Donja granica	Gornja granica
18 – 20	18	20
21 – 23	21	23
24 – 26	24	26
27 – 29	27	29
30 – 32	30	32

Granice razreda

- **se ne preklapaju** kod diskretnog obilježja
- **se preklapaju** kod neprekidnog numeričkog obilježja (gornja granica razreda jednaka je donjoj granici sljedećeg razreda)

Ukoliko je podatak jednak granici razreda, dogovorno se pridružuje jednom razredu (zaokruživanje na dolje ili gore)

Širina razreda

Ako se granice ne preklapaju, tada je širina i-tog razreda

$$i_i = L_{i+1} - L_i, i = 1, 2, \ldots, k-1$$

Širina zadnjeg razreda:

$$i_k = U_k - U_{k-1}.$$

ako se granice preklapaju, tada je širina i-tog razreda

$$i_i = U_i - L_i, i = 1, 2, \ldots, k$$

(Napomena: formula vrijedi i za slučaj kada se granice ne preklapaju.)

Razredi mogu biti iste ili različite širine.

Otvoreni razred - razred bez gornje ili donje granice.

Ponekad zadnji razred nema gornje granice. (Rjeđi je slučaj da prvi razred nema donje granice.)

Primjer. Mjesečni prihod.

Razred
0 – 1.000 €
1.000 – 2.000 €
2.000 – 3.000 €
3.000 – 5.000 €
više od 5.000 €

U zadnji razred spadaju sve vrijednosti veće od 5.000 € i gornja granica razreda nije definirana

Sredina razreda - aritmetička sredina gornje i donje granice razreda:

$$X_i = \frac{L_i + U_i}{2}, \quad i = 1, 2, \ldots, k$$

Sredina razreda za otvoreni razred se procjenjuje na temelju poznavanja pojave.

Razred	Donja granica	Gornja granica	Širina razreda	Sredina razreda
18 – 20	18	20	3	19
21 – 23	21	23	3	22
24 – 26	24	26	3	25
27 – 29	27	29	3	28
30 – 32	30	32	3	31

Kod grupiranja u razrede **poželjno je**:

- koristiti 5 do 20 razreda
- svi razredi su iste širine (osim eventualno prvog i zadnjeg)
- broj razreda je neparan
- Sturgesova formula:

$$\text{broj razreda} = 1 + \lceil \log_2 \text{broj podataka} \rceil$$

($\lceil\ \rceil$ - zaokruživanje na prvi veći cijeli broj)

Za mali skup vrijednosti obilježja:

- Vrijednosti se ne grupiraju
- Uređenjem se ne gubi informacija
- Uvid u sve vrijednosti skupa

Grafički prikaz distribucije frekvencija razreda. Histogram

Stupčasti grafikon - visina stupca proporcionalna je frekvenciji

Histogram - **površina** stupca proporcionalna je frekvenciji

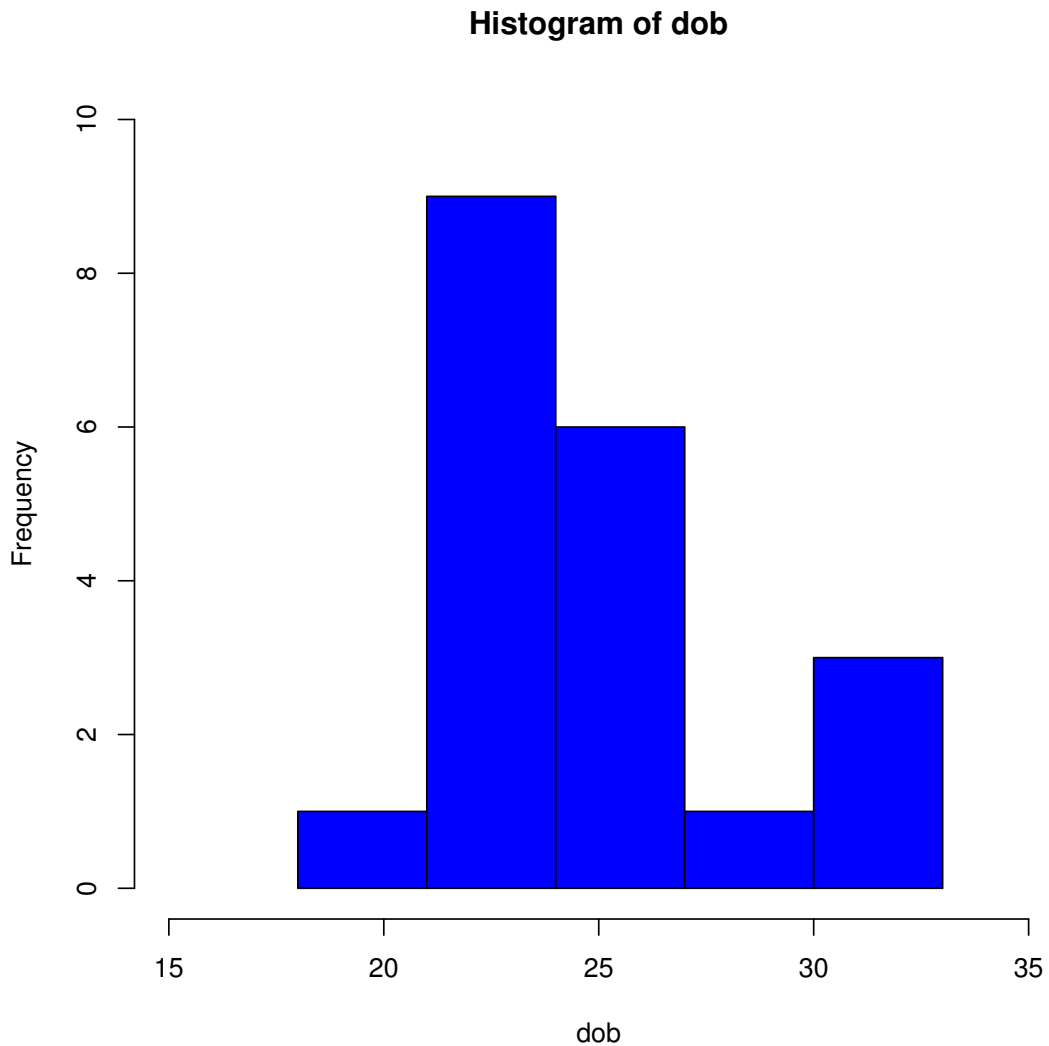
Ukoliko su svi razredi iste širine, stupčasti grafikon i histogram su u biti isti.

Ako su širine razreda različite:

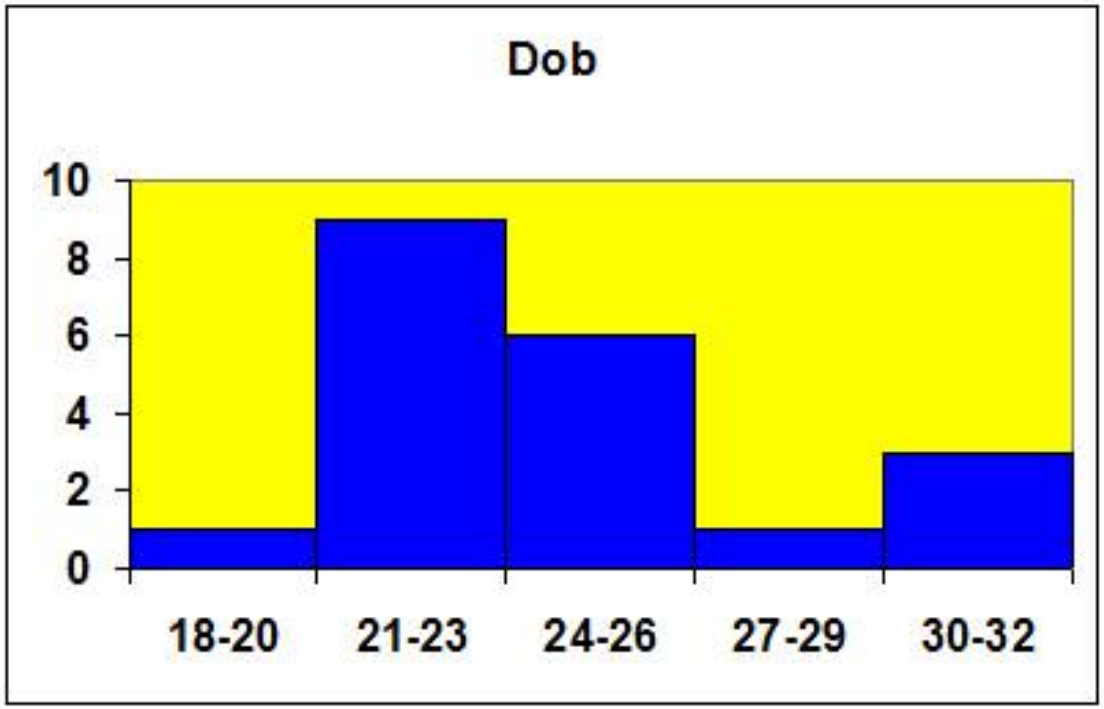
Korigirana frekvencija - omjer frekvencije i širine razreda (ili neke druge prikladne veličine proporcionalne širini razreda)

U **histogramu** je visina stupca proporcionalna **korigiranoj** frekvenciji!

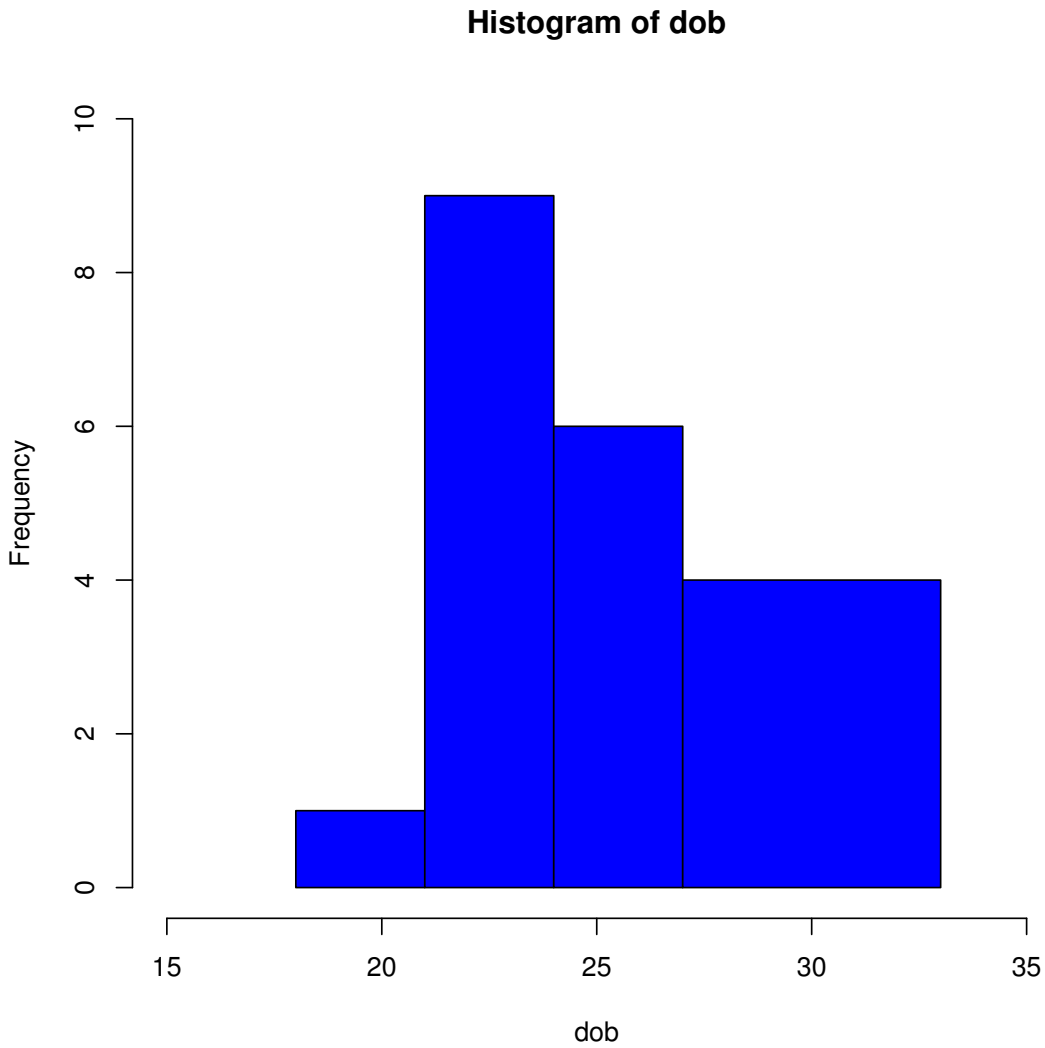
Histogram za varijablu dob.



Usporedite s prethodnom slikom (x-os)!

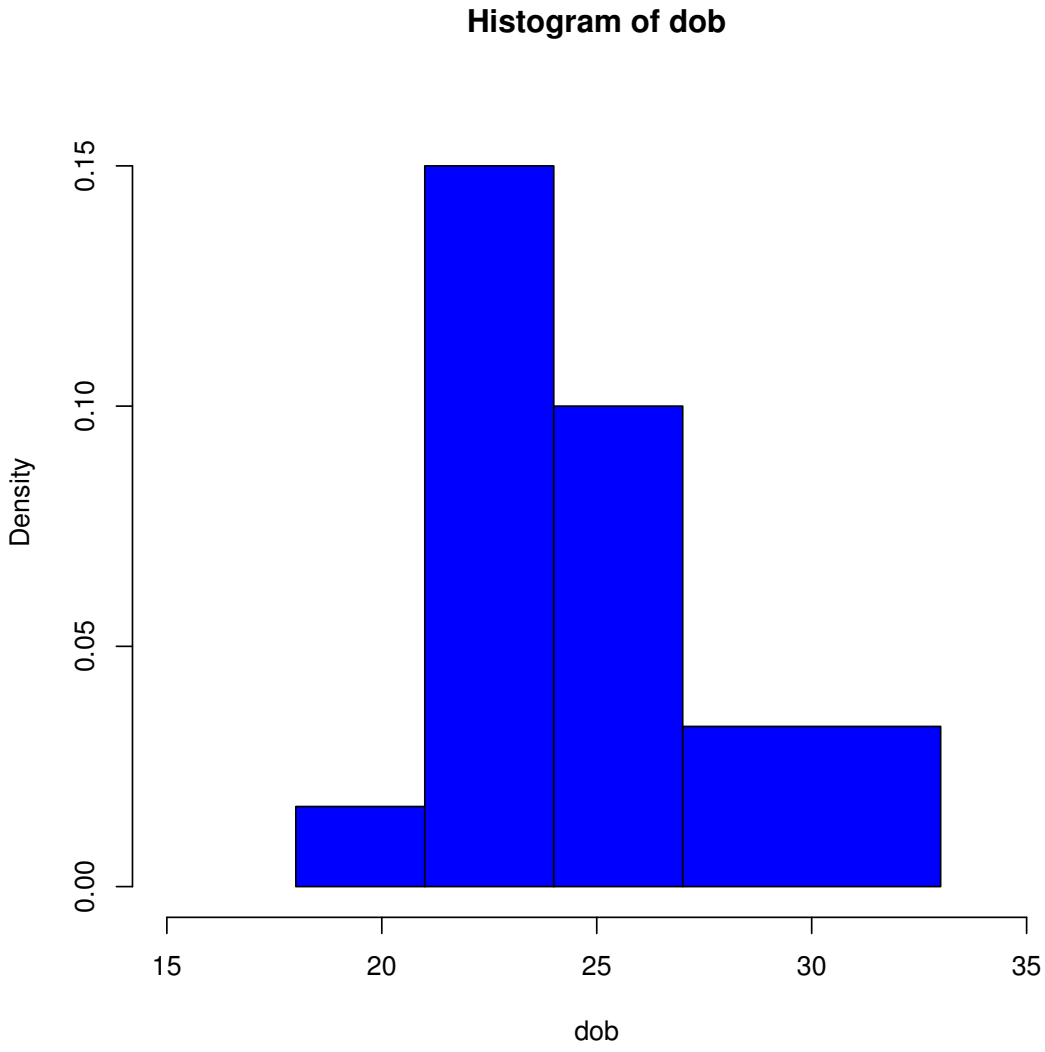


Ako spojimo zadnja dva razreda:



Nije dobro. Visina stupca odgovara frekvenciji.

Ako spojimo zadnja dva razreda:



Sada je dobro.
(Prikazane su relativne frekvencije.)

Histogram:

- za prikaz kvantitativnih podaka (kvalitativni podaci $\rightarrow$ stupčasti grafikon)
- površina stupca proporcionalna je frekvenciji (stupčasti grafikon $\rightarrow$ visina stupca)
- ako se prikazuju relativne frekvencije $\rightarrow$ površina ispod histograma je 1.
- nema razmaka između stupaca (stupčasti grafikon ih može imati)

Poligon frekvencija i ogiva (poligon kumulativnih frekvencija) - samo za neprekidna obilježja

Primjer.

Izmjerena je visina učenika jednog razreda. Dobivene vrijednosti su (izražene u cm):

143	156	156	163	167
142	171	170	169	164
138	158	160	162	164
173	157	158	159	160
138	172	166	166	159
120	125	165	136	168

Grupirajte podatke u 6 razreda i odredite distribuciju frekvencija razreda.
Distribuciju frekvencija razreda prikažite tablično (prikažite apsolutne i relativne frekvencije te kumulativne i relativne kumulativne frekvencije) i grafički koristeći histogram, poligon frekvencija i ogivu.
Odredite širinu i sredinu razreda.

Napomena.

Ukupno je dano 30 podataka:

143	156	156	163	167
142	171	170	169	164
138	158	160	162	164
173	157	158	159	160
138	172	166	166	159
120	125	165	136	168

Sturgesova formula:

$$\text{broj razreda} = 1+\lceil \log_2 \text{broj podataka} \rceil = 1+\lceil \log_2 30 \rceil = 1+\lceil 4.907 \rceil = \mathbf{6}.$$

Rješenje.

Prvo odredimo najmanji i najveći podatak.

143	156	156	163	167
142	171	170	169	164
138	158	160	162	164
173	157	158	159	160
138	172	166	166	159
120	125	165	136	168

Rješenje.

Prvo odredimo najmanji i najveći podatak.

143	156	156	163	167
142	171	170	169	164
138	158	160	162	164
173	157	158	159	160
138	172	166	166	159
120	125	165	136	168

Najmanji podatak: 120

Najveći podatak: 173

Ukupno 173-120+1=54 vrijednosti.

6 razreda $\longrightarrow$ 54:6=9 vrijednosti u razredu.

Razredi: 120–128, 129–137, 138–146, 147–155, 156–164, 165–173.

143	156	156	163	167
142	171	170	169	164
138	158	160	162	164
173	157	158	159	160
138	172	166	166	159
120	125	165	136	168

Frekvencije razreda:

Razred	f
120–128	2
129–137	1
138–146	4
147–155	0
156–164	13
165–173	10

Relativne frekvencije.

Frekvencije razreda dijelimo s 30:

Razred	f	r
120–128	2	0.06667
129–137	1	0.03333
138–146	4	0.13333
147–155	0	0.00000
156–164	13	0.43333
165–173	10	0.33333

Kumulativne frekvencije.

Postupno zbrajamo frekvencije:

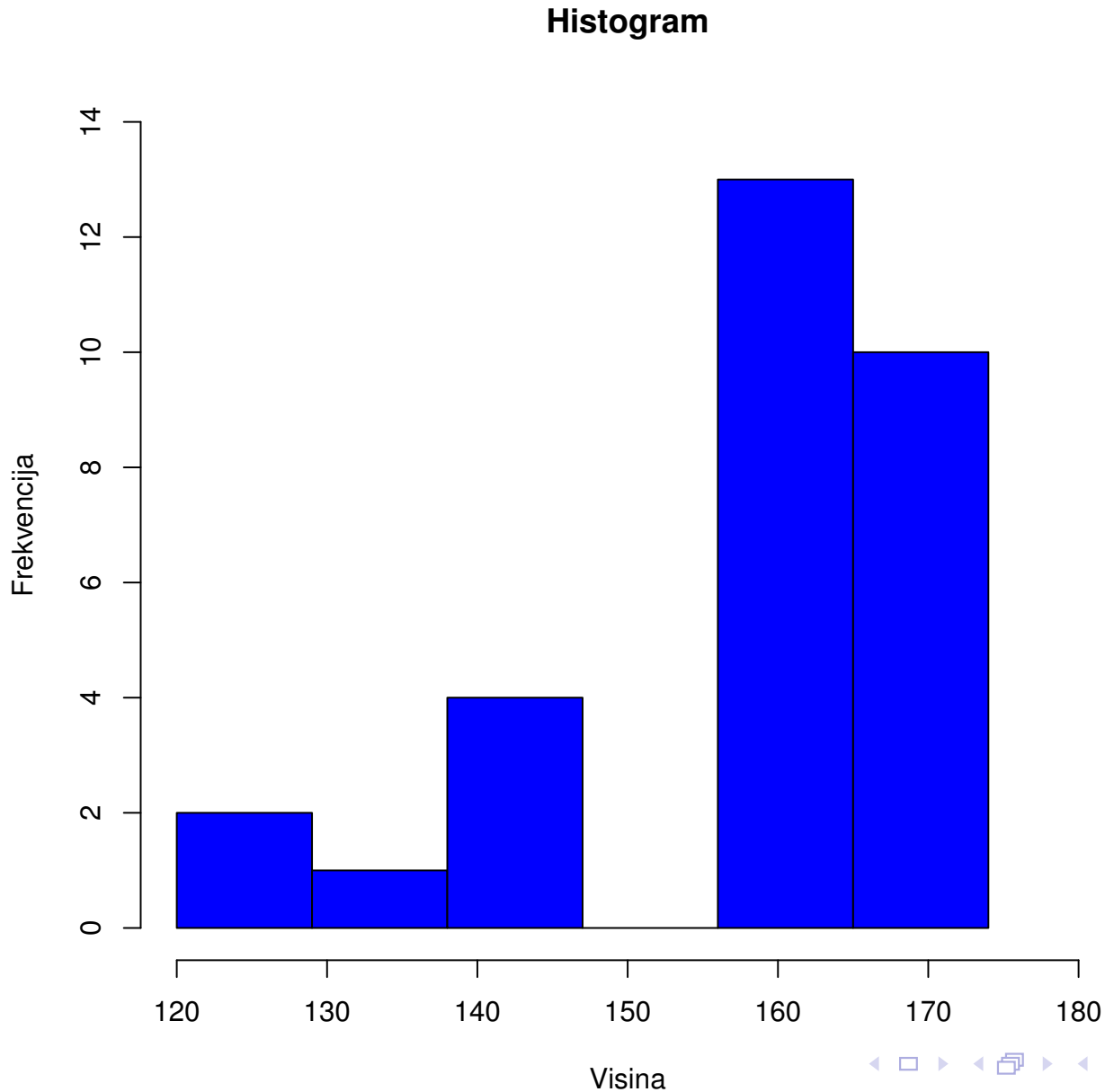
Razred	f	r	c
120–128	2	0.06667	2
129–137	1	0.03333	3
138–146	4	0.13333	7
147–155	0	0.00000	7
156–164	13	0.43333	20
165–173	10	0.33333	30

Relativne kumulativne frekvencije.

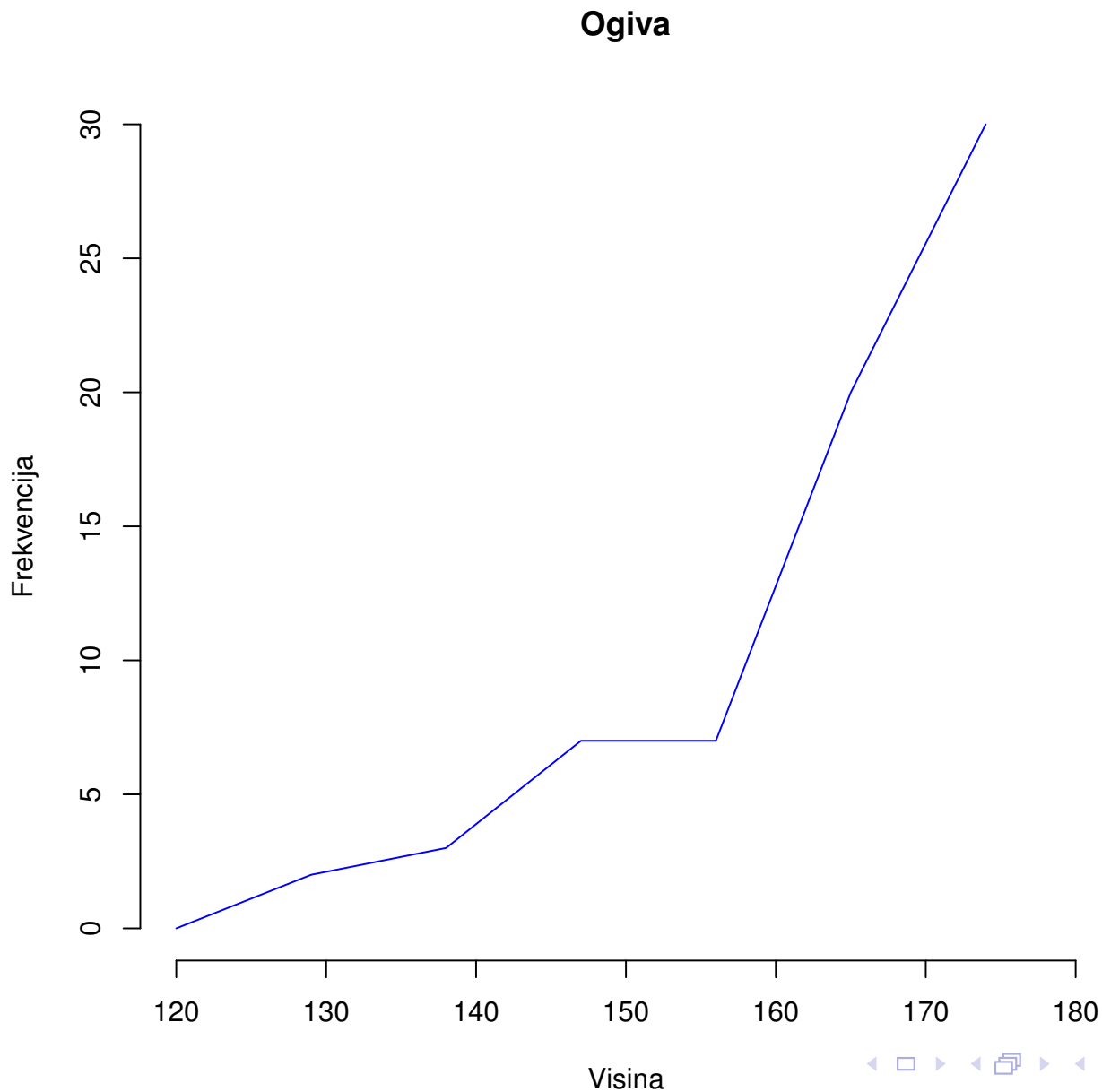
Kumulativne frekvencije razreda dijelimo s 30:

Razred	f	r	c	rc
120–128	2	0.06667	2	0.06667
129–137	1	0.03333	3	0.10000
138–146	4	0.13333	7	0.23333
147–155	0	0.00000	7	0.23333
156–164	13	0.43333	20	0.66667
165–173	10	0.33333	30	1.00000

Histogram


Histogram
Frekvencija
0
2
4
6
8
10
12
14
120
130
140
150
160
170
180
Visina

Poligon frekvencija



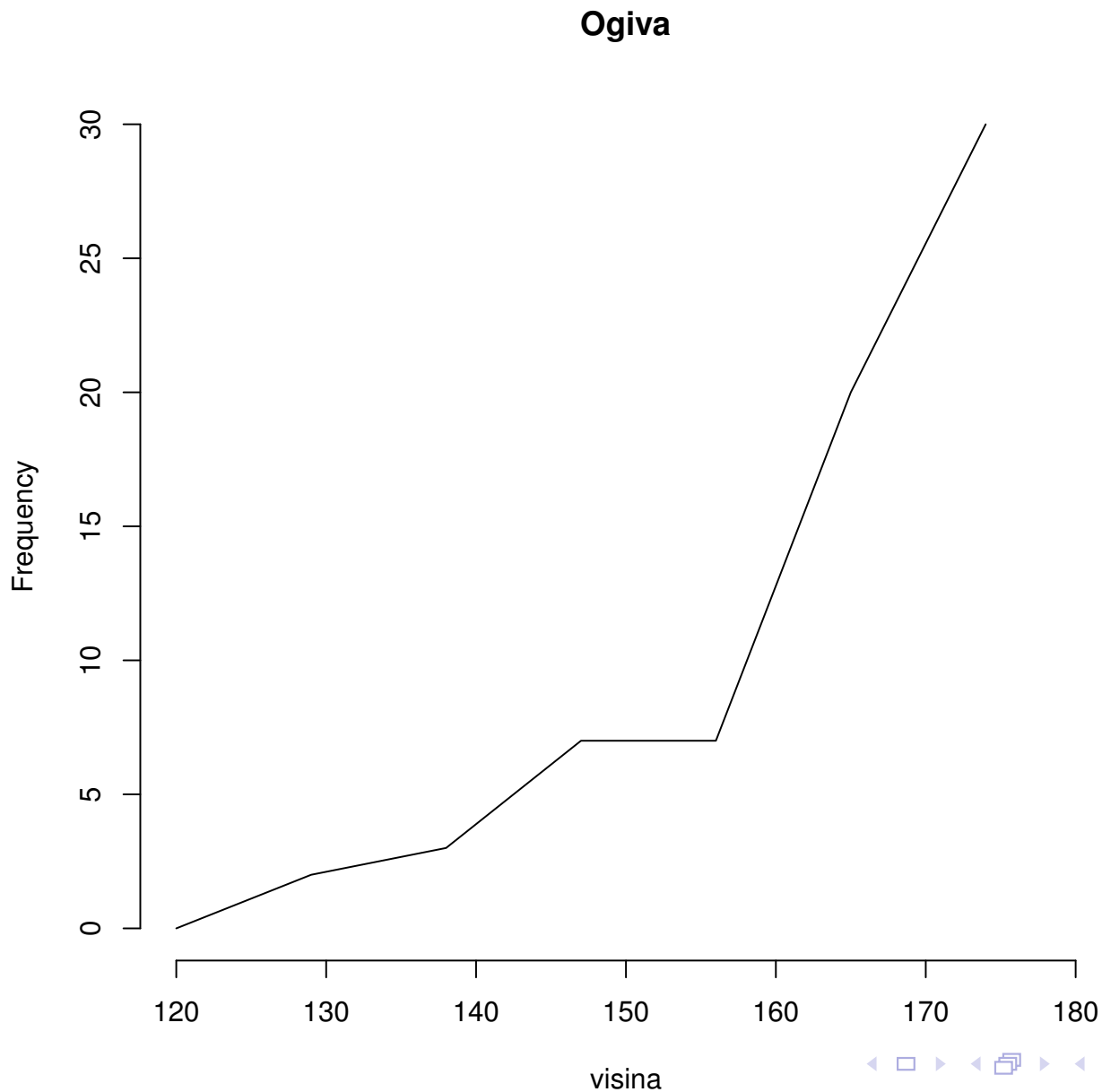
Poligon frekvencija:

- x-os: sredine razreda
- y-os: (korigirana) frekvencija
- Plus:
- donja granica prvog razreda (x-os) i 0 na y-osi
- gornja granica zadnjeg razreda (x-os) i 0 na y-osi

Ogiva - graf (relativnih) kumulativnih frekvencija

- x-os: granice razreda
- y-os: 0 za donju granicu prvog razreda
- y-os: kumulativna frekvencija razreda (uz gornju granicu razreda na x-osi)

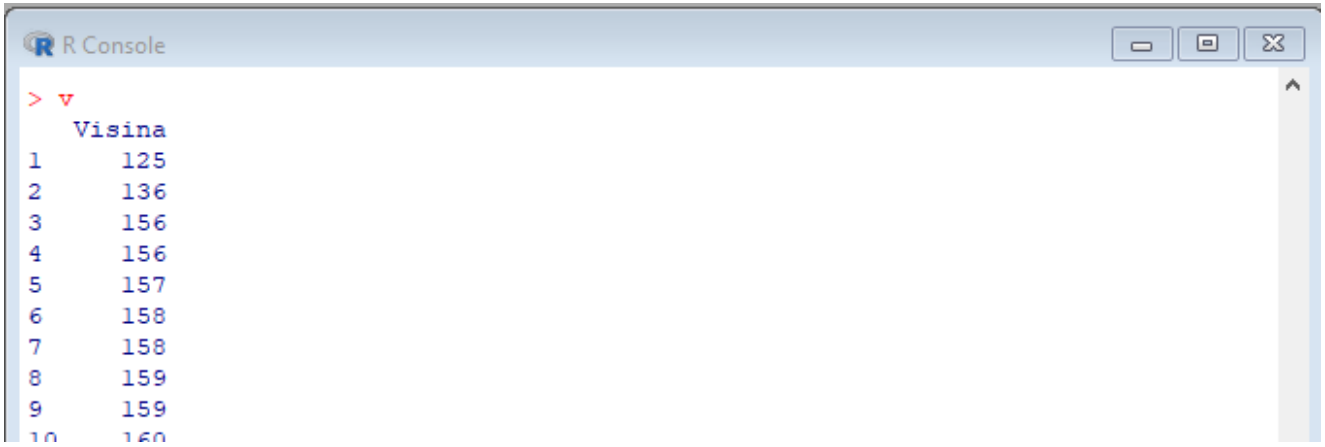
Ogiva:


Ogiva
Frequency
0
5
10
15
20
25
30
120
130
140
150
160
170
180
visina

Rješenje zadatka u R-u

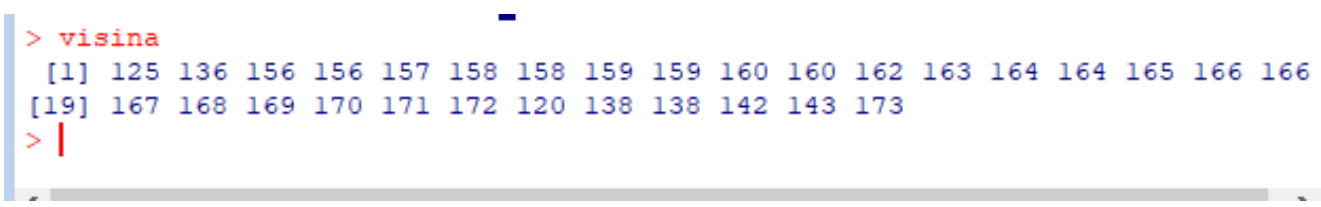
Podaci su u datoteci data3.csv:

```
> v=read.csv("data3.csv")
```



v je baza podataka ('data frame').
Za histogram nam treba **vektor** (niz podataka istog tipa):

```
> visina = v$Visina
```



Histogram:

```
> h <- hist(visina,
breaks=c(120,129,138,147,156,165,174),right=FALSE)
```

breaks=c(120,129,138,147,156,165,174) - granice razreda (osim u zadnjem razredu)

right=FALSE - gornja granica nije uključena u razred

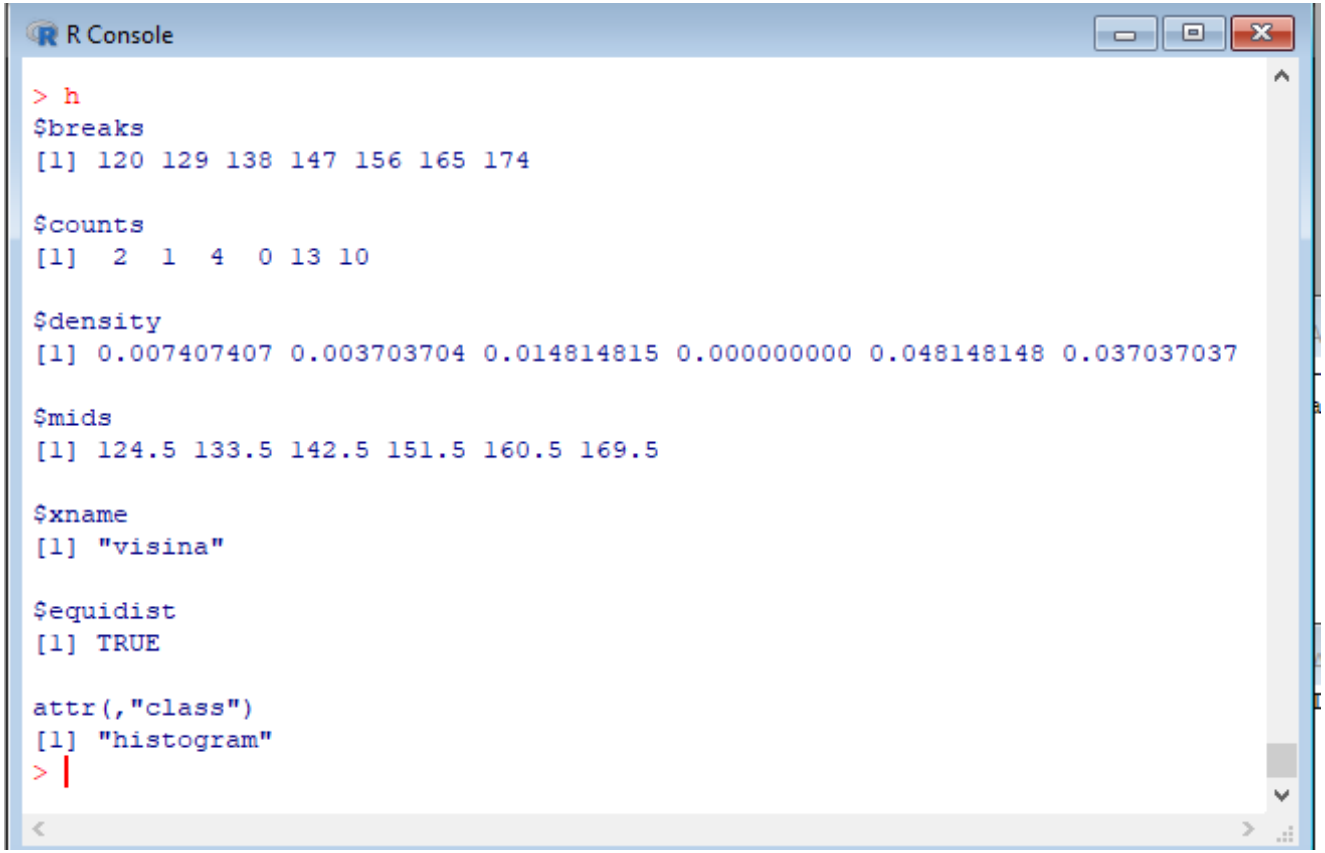
Razredi su 120–128, 129–137, . . .

Ova naredba će nacrtati histogram.

Graf možemo malo prilagoditi (uljepšati):

```
> h <- hist(visina, col = "blue",
breaks=c(120,129,138,147,156,165,174),right=FALSE,
freq=TRUE, main = "Histogram", xlab="Visina",
ylab="Frekvencija", xlim = c(120,180),
ylim = c(0,14))
```

Varijabla h se sastoji od 6 komponenata



- h$breaks - granice razreda
- h$counts - frekvencije razreda
- h$density - korigirane relativne frekvencije razreda
- h$mids - sredine razreda

Frekvencije su u h$counts:

```
> h$counts
```

```
[1] 2 1 4 0 13 10
```

Relativne frekvencije:

$$r_i = \frac{f_i}{N}$$

```
> r= h$counts / 30
```

```
> r
```

```
[1] 0.06666667 0.03333333 0.13333333 0.00000000
0.43333333 0.33333333
```

Kumulativne frekvencije:

```
> cf = cumsum(h$counts)
> cf
[1]  2  3  7  7 20 30
```

Relativne kumulativne frekvencije:

```
> rcf=cf/300
> rcf
[1] 0.06666667 0.10000000 0.23333333 0.23333333
0.66666667 1.00000000
```

Poligon frekvencija

```
> h <- hist(visina, col = "transparent",
border = "transparent",
breaks=c(120,129,138,147,156,165,174),right=FALSE,
freq=TRUE, main = "Poligon frekvencija",
xlab="Visina", ylab="Frekvencija", xlim =
c(120,180), ylim = c(0,14))
```

col = "transparent",border = "transparent" - da se ne vide stupci!

```
> xaxis=c(min(h$breaks),h$mids,max(h$breaks))
> yaxis=c(0,h$counts,0)
> lines(xaxis,yaxis,type='l',col="blue")
```

Ogiva:

```
> h <- hist(visina, col = "transparent",
border = "transparent",
breaks=c(120,129,138,147,156,165,174),right=FALSE,
freq=TRUE, main = "Ogiva", xlab="Visina",
ylab="Frekvencija", xlim = c(120,180), ylim =
c(0,30))
```

```
> cf = cumsum(h$counts))
```

```
> xaxis=h$breaks
```

```
> yaxis=c(0,cf)
```

```
> lines(xaxis,yaxis,type='l',col="blue")
```

Opis kvantitativnih obilježja/podataka

CILJ: sažeto opisati svojstva podataka.

- **Mjere centralne tendencije**
- **Mjere raspršenosti (varijacije)**
- **Mjere asimetrije**
- **Mjere položaja**

Mjere centralne tendencije

- cilj je odrediti broj oko kojeg se grupiraju podaci
  $\rightarrow$ mjere centralne tendencije
- jednim brojem opisujemo skup varijabilnih podataka
- primjeri mjera centralne tendencije: prosječna plaća.

Mjere centralne tendencije:

- **Potpune** - računaju se na temelju svih podataka
- **Položajne** - određene položajem podataka u nizu

Mjere centralne tendencije:

Potpune

- srednja vrijednost (aritmetička sredina)
- geometrijska sredina
- harmonijska sredina

Položajne

- mod
- medijan

Srednja vrijednost

Primjer. Prosječna plaća u tromesječju: 1500 €, 1200 €, 1800 €.

$$\text{Prosječna plaća} = \frac{1500 + 1200 + 1800}{3} = \frac{4500}{3} = 1500.$$

Primjer. Prosječna ocjena. Ocjene: 3, 3, 4, 2, 5, 4.

$$\text{Prosječna ocjena} = \frac{3 + 3 + 4 + 2 + 5 + 4}{6} = \frac{21}{6} = 3.5.$$

Za konačan skup podataka $x_1, x_2, \ldots, x_N$, **srednja vrijednost (μ)** je aritmetička sredina podataka:

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

$\sum$ - oznaka za sumiranje. Još se koristi i $\sum_i x_i$, $\sum x_i$, $\sum x$.

Primjer. Lionel Messi je u Ligi prvaka 2014/15 u jednoj utakmici postigao 3 pogotka, u dvije po 2, u tri po jedan dok u sedam utakmica nije postigao niti jedan zgoditak. Koliko je L. Messi prosječno dao golova po utakmici?

Ukupno je postigao

$$7 \cdot 0 + 3 \cdot 1 + 2 \cdot 2 + 1 \cdot 3 = 0 + 3 + 4 + 3 = 10$$

zgoditaka.

Broj utakmica $= 7 + 3 + 2 + 1 = 13$.

Prosječan broj zgoditaka $= \frac{10}{13} = 0.77$.

Ovdje je zadana distribucija frekvencija:

Broj zgoditaka (x_i)	Broj utakmica (f_i)
0	7
1	3
2	2
3	1

$$\begin{aligned} \text{Broj zgoditaka} &= 7 \cdot 0 + 3 \cdot 1 + 2 \cdot 2 + 1 \cdot 3 = \\ &= f_1 \cdot x_1 + f_2 \cdot x_2 + f_3 \cdot x_3 + f_4 \cdot x_4 = \\ &= \sum f_i x_i \end{aligned}$$

$$\text{Broj utakmica} = 7 + 3 + 2 + 1 = f_1 + f_2 + f_3 + f_4 = \sum f_i$$

Računanje srednje vrijednosti iz distribucije frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $f_1, f_2, f_3, \ldots, f_k$ pripadne frekvencije. Tada je

$$\mu = \frac{\sum f_i x_i}{\sum f_i} = \frac{\sum f_i x_i}{N}.$$

Računanje srednje vrijednosti iz relativnih frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $r_1, r_2, r_3, \ldots, r_k$ pripadne relativne frekvencije. Tada je

$$\mu = \sum r_i x_i.$$

Ovaj se izraz lagano dobije iz prethodne formule:

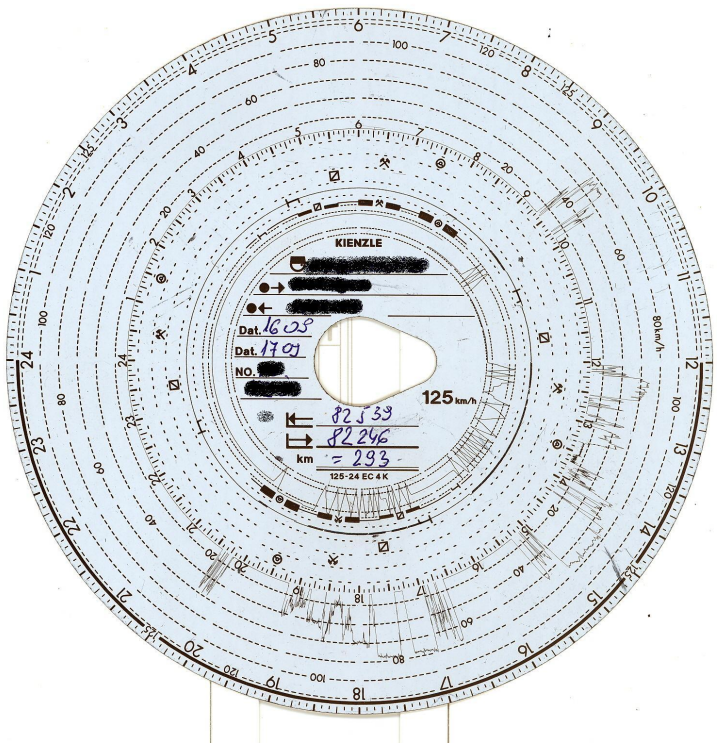
$$\mu = \frac{\sum f_i x_i}{N} = \sum \frac{f_i}{N} x_i = \sum r_i x_i.$$

Primjer. Vozilo je 100 km prešlo za 2 sata. Kolika je prosječna brzina?

$$\text{Prosječna brzina} = \frac{100\ \text{km}}{2\ \text{h}} = 50\ \text{km/h}.$$

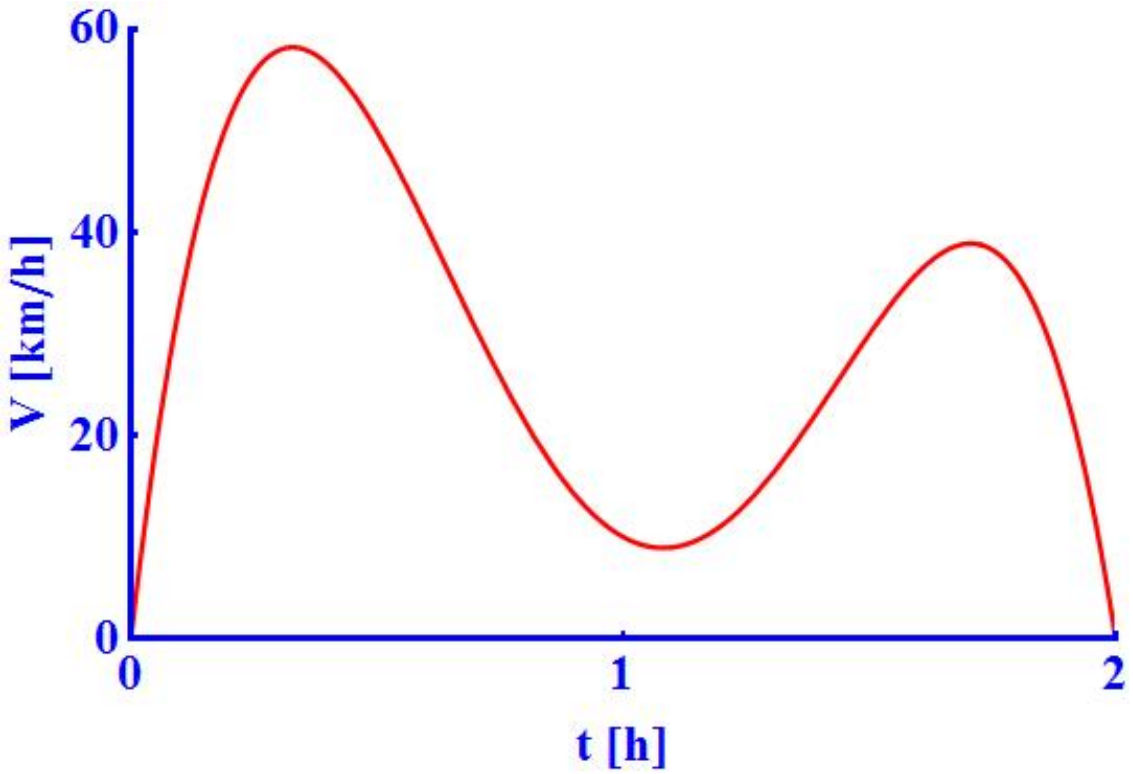
Brzina ne treba biti konstantna!

Kako izračunati prosječnu brzinu ukoliko nam je poznata samo brzina tijekom puta?
Tahograf:

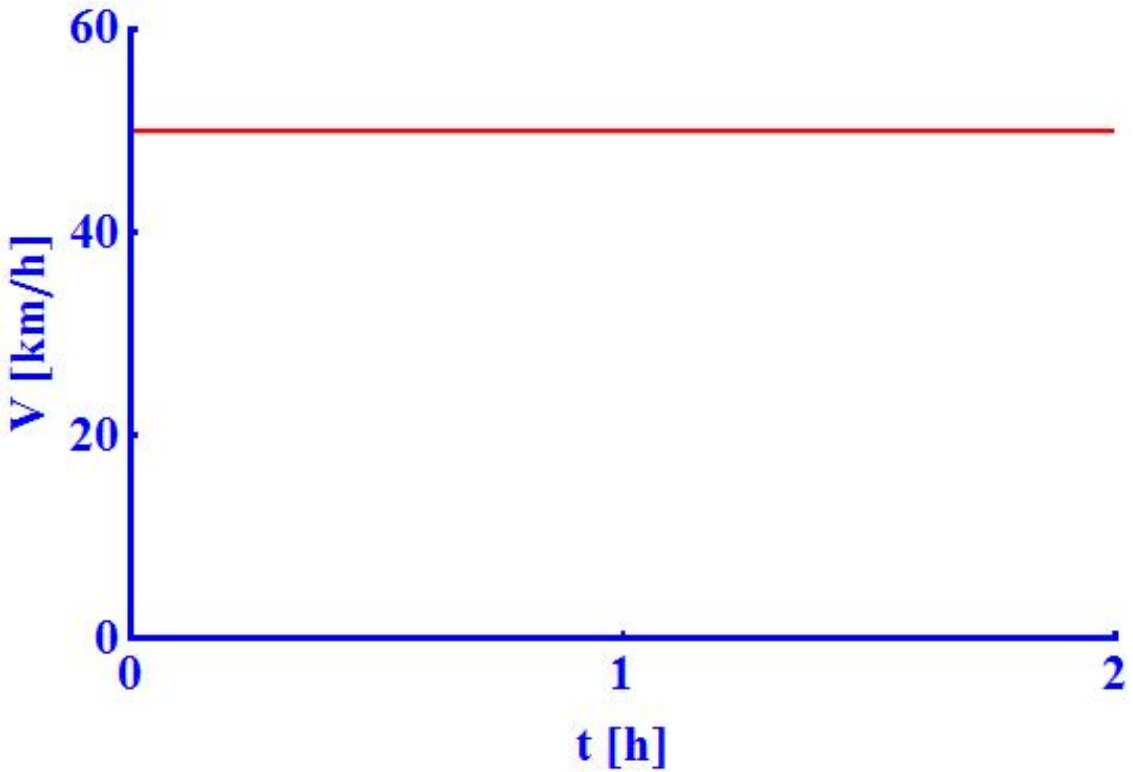


Prosječnu brzinu bismo mogli izračunati ukoliko znamo prijeđeni put.

Koliki je prijeđeni put ukoliko je brzina vozila bila kao na slici?

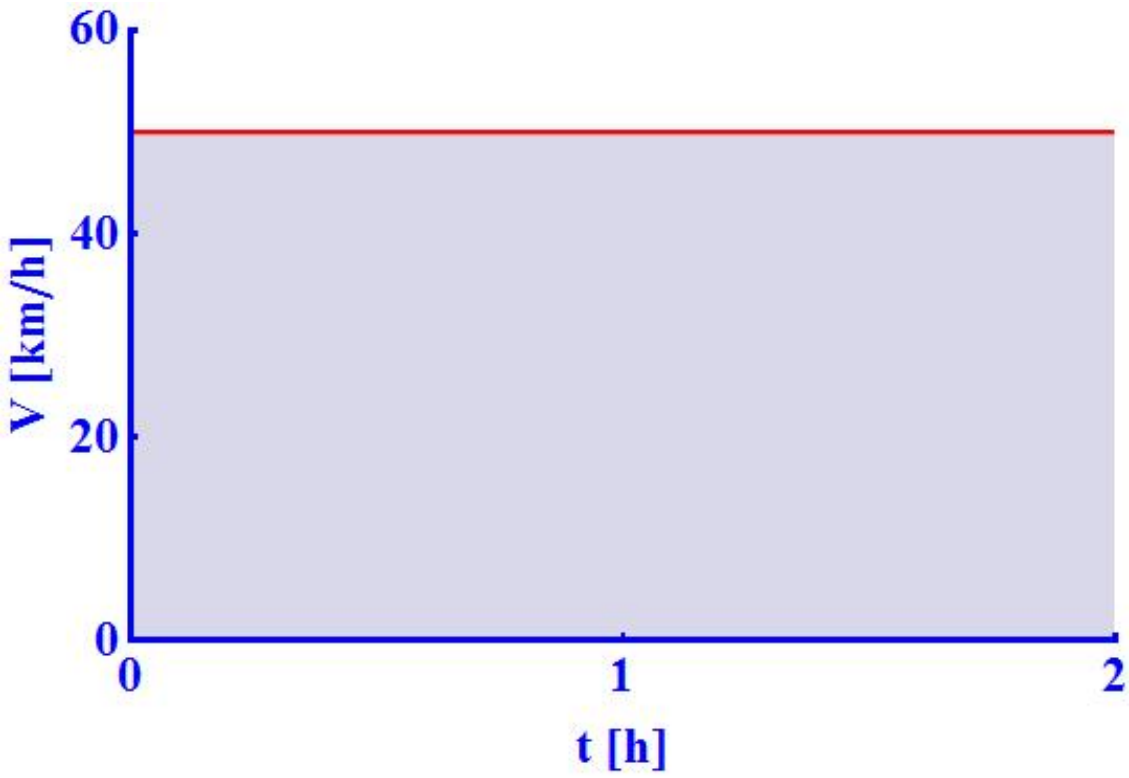


Ukoliko je brzina konstantna:



put je dan s

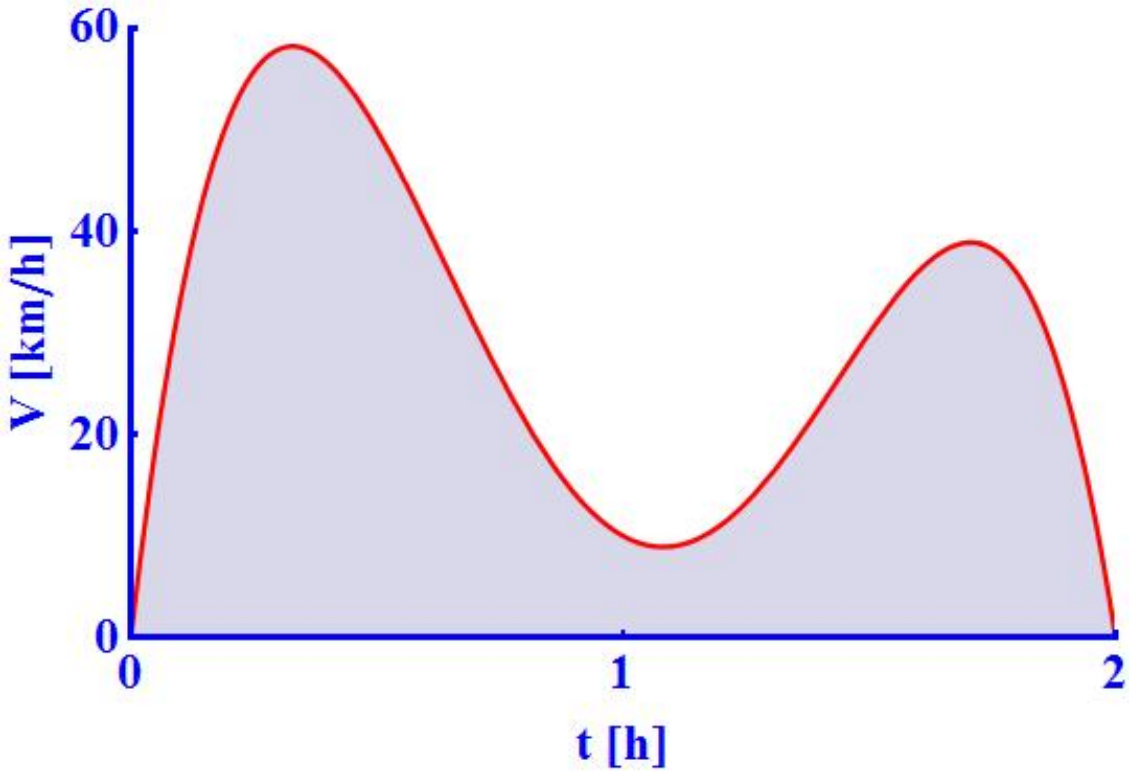
$$s = v \cdot t.$$



$$s = v \cdot t.$$

Prijeđeni put je površina pravokutnika!

Prijeđeni put je površina ispod krivulje:



Kako odrediti površinu?

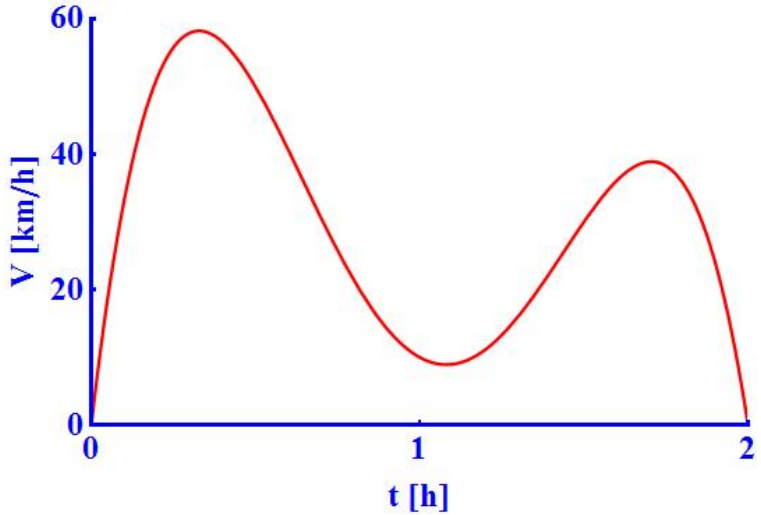
Matematička oznaka za površinu ispod krivulje je:

$$\int_0^2 V(t)\,dt$$

Znači, prosječna brzina kroz vrijeme T je dana s:

$$\bar{V} = \frac{1}{T}\int_0^T V(t)\,dt.$$

Napomena. U ovom primjeru je skup podataka bio beskonačan.



Napomena 2. U ovom primjeru je

$$\text{Put} = \text{Površina} = 59.52$$

$$\text{Prosječna brzina} = \frac{59.52 \text{ km}}{2 \text{ h}} = 29.76 \text{ km/h}$$

Medijan

Medijan je vrijednost koja skup podataka dijeli na dva istobrojna dijela tako da je polovica podataka veća a polovica podataka manja od medijana.

Oznaka za medijan: **m, Me**

Postupak računanja medijana.

Podaci se poredaju prema veličini, od najmanjeg do najvećeg (ili od najvećeg do najmanjeg).

- Ako je broj podataka neparan - medijan je središnji član niza podataka.
- Ako je broj podataka paran - medijan je aritmetička sredina dvaju središnjih članova niza podataka.

Primjer. Odredite medijan za podatke

$$7,\ 8,\ 17,\ 9,\ 4,\ 9,\ 10,\ 11,\ 11,\ 3,\ 6,\ 7,\ 11,\ 5,\ 9$$

Podatke poredamo po veličini:

3	4	5	6	7	7	8	**9**	9	9	10	11	11	11	17	podaci
1.	2.	3.	4.	5.	6.	7.	**8.**	9.	10.	11.	12.	13.	14.	15.	redni broj

$m = 9$

Primjer. Odredite medijan za podatke

$$4,\ 2,\ 3,\ 5,\ 2,\ 1,\ 3,\ 4,\ 2,\ 1$$

Podatke poredamo po veličini:

1	1	2	2	**2**	**3**	3	4	4	5	podaci
1.	2.	3.	4.	**5.**	**6.**	7.	8.	9.	10.	redni broj

$$m = \frac{2+3}{2} = 2.5$$

Određivanje medijana iz distribucije frekvencija.

Ukoliko je zadana distribucija frekvencija, medijan određujemo iz kumulativnih frekvencija.

Odreditmo medijan iz sljedećeg primjera.

Dob	Frekvencija	Relativna frekvencija	Kumulativna frekvencija	Relativna kumulativna frekvencija
19	1	0.05	1	0.05
20	0	0.00	1	0.05
21	5	0.25	6	0.30
22	2	0.10	8	0.40
23	2	0.10	10	0.50
24	1	0.05	11	0.55
25	4	0.20	15	0.75
26	1	0.05	16	0.80
27	0	0.00	16	0.80
28	1	0.05	17	0.85
29	0	0.00	17	0.85
30	1	0.05	18	0.90
31	1	0.05	19	0.95
32	1	0.05	20	1.00

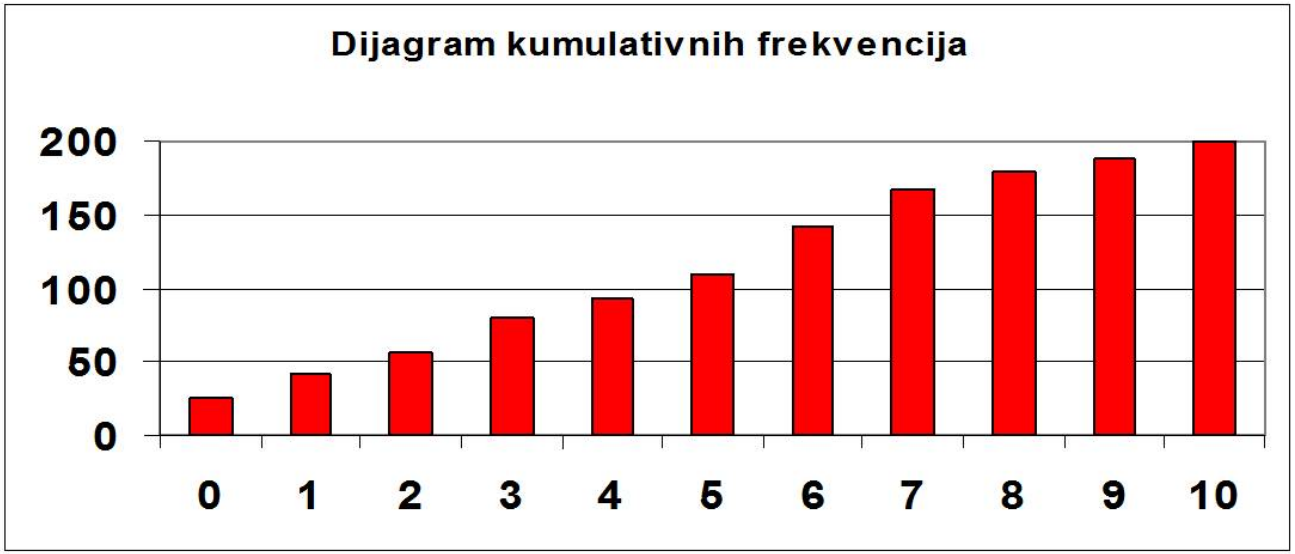
Dob	Frekvencija	Relativna frekvencija	Kumulativna frekvencija	Relativna kumulativna frekvencija
19	1	0.05	1	0.05
20	0	0.00	1	0.05
21	5	0.25	6	0.30
22	2	0.10	8	0.40
23	2	0.10	**10**	**0.50**
24	1	0.05	**11**	**0.55**
25	4	0.20	15	0.75
26	1	0.05	16	0.80
27	0	0.00	16	0.80
28	1	0.05	17	0.85
29	0	0.00	17	0.85
30	1	0.05	18	0.90
31	1	0.05	19	0.95
32	1	0.05	20	1.00

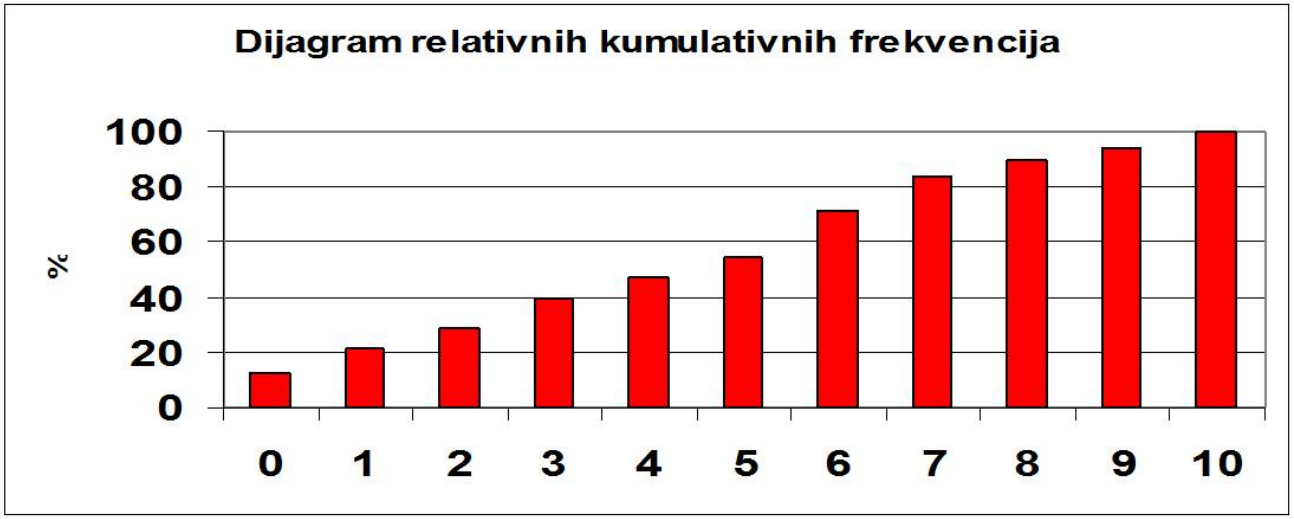
- 10 (50%) podataka je manje ili jednako od 23
- 10 (50%) podataka je veće ili jednako od 24

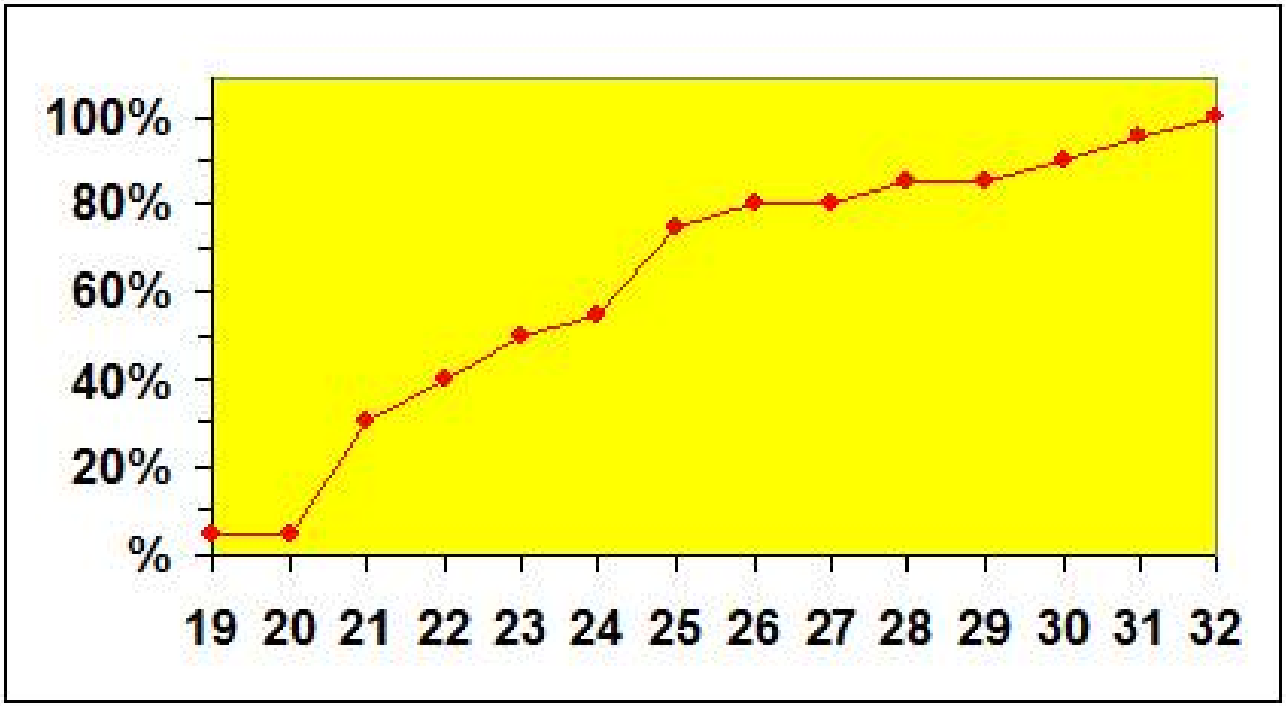
Znači, 23 i 24 su na središnjem mjestu u nizu podataka.

$$m = \frac{23 + 24}{2} = 23.5$$

Primjer. Odredite medijan iz sljedećih grafova kumulativne frekvencije:




Dijagram relativnih kumulativnih frekvencija
%
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10


100%
80%
60%
40%
20%
%
19 20 21 22 23 24 25 26 27 28 29 30 31 32

Druge mjere centralne tenencije

Težinska (ponderirana) srednja vrijednost

Za podatke $x_1, x_2, x_3, \ldots, x_N$ dane su težine (ponderi) $w_1, w_2, w_3, \ldots, w_N$.

Težinska (ponderirana) srednja vrijednost je dana s

$$\mu = \frac{\sum w_i x_i}{\sum w_i}.$$

Primjer. Za podatke 1, 3, 4, 5 i pripadne težine 1/2, 2/3, 2 i 1, težinska srednja vrijednost je

$$\frac{\frac{1}{2}\cdot 1 + \frac{2}{3}\cdot 3 + 2\cdot 4 + 1\cdot 5}{\frac{1}{2}+\frac{2}{3}+2+1} = \frac{\frac{31}{2}}{\frac{25}{6}} = \frac{93}{25} = 3.72$$

Geometrijska sredina

Za podatke $x_1, x_2, x_3, \ldots, x_N$ **geometrijska sredina** je definirana s

$$\sqrt[N]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_N}$$

Primjer. Za podatke 1, 3, 4, 5 geometrijska sredina je

$$\sqrt[4]{1 \cdot 3 \cdot 4 \cdot 5} = \sqrt[4]{60} = 2.78$$

Primjer. Kamatna stopa je u 2010. g. bila 12%, u 2011. 17% i u 2013. g. 10%. Kolika je prosječna kamatna stopa za ove tri godine?

Prosječna kamatna stopa je godišnja stopa rasta po kojoj bi konstantni rast kroz 3 godine bio jednak ostvarenom rastu.

Nije

$$\frac{12+17+10}{3} = 13$$

$$\text{glavnica} \cdot 1.12 \cdot 1.17 \cdot 1.10 = \text{glavnica} \cdot (1+p) \cdot (1+p) \cdot (1+p)$$

t.j.

$$1.12 \cdot 1.17 \cdot 1.10 = (1+p)^3$$

odnosno

$$1+p = \sqrt[3]{1.12 \cdot 1.17 \cdot 1.10} = 1.1296$$

pa je

$$p = 0.1296$$

odnosno, prosječna kamatna stopa je 12.96%.

Harmonijska sredina

Za podatke $x_1, x_2, x_3, \ldots, x_N$ **harmonijska sredina** je definirana s

$$\frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_N}}.$$

Primjer. Za podatke 1, 3, 4, 5 harmonijska sredina je

$$\frac{4}{\frac{1}{1} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}} = \frac{240}{107} = 2.243.$$

Primjer. Vozilo je 100 km vozilo brzinom 40 km/h, sljedećih 100 km brzinom 70 km/h a zadnjih 100 km brzinom 130 km/h. Kolika je prosječna brzina?

Put je $3 \cdot 100$ km a proteklo vrijeme je:

$$\frac{100}{40} + \frac{100}{70} + \frac{100}{130}.$$

Jer je brzina=put/vrijeme, srednja brzina je

$$\bar{V} = \frac{3 \cdot 100}{\frac{100}{40} + \frac{100}{70} + \frac{100}{130}} = \frac{3}{\frac{1}{40} + \frac{1}{70} + \frac{1}{130}} = 63.86$$

Harmonijska sredina!

Mod

Mod je vrijednost varijable s najvećom frekvencijom.

(Najčešća vrijednost.)

Primjer. Odredimo mod za podatke
ACCABACABCAABBCBAACCBAABCCBCAA.

Podatke prvo sortiramo (niz podataka):
AAAAAAAAAAA BBBBBBBB CCCCCCCCCC

Mod je A jer se pojavljuje najviše puta (12).

Obilježje može imati i više modova. Npr. u nizu
1 1 2 2 2 3 4 4 5 5 5 6 7 7
mod je 2 i 5 (dva moda).
U ovom slučaju govorimo o **bimodalnoj** distribuciji frekvencija.

Primjena mjera centralne tendencije

Tip obilježja	μ	g	H	Medijan	Mod
Nominalno	NE	NE	NE	NE	DA
Redoslijedno	NE	NE	NE	DA	DA
Kvantitativno	DA	DA	DA	DA	NE

- Za kvantitativne podatke se mod ne prikazuje (iako se može izračunati)
- ako su podaci jednaki nuli ili manji od nule $\Rightarrow$ ne može se izračunati G ili H
- ako su sve vrijednosti varijable različite $\Rightarrow$ nema moda
- mjere centralne tendencije mogu biti iste ili različite veličine
- $\min\{x_1, \ldots, x_N\} \leq H \leq G \leq \mu \leq \max\{x_1, \ldots, x_N\}$

ZADATAK.

Izračunajte srednju vrijednost, medijan i mod za sljedeći niz podataka:

2	5	5	6	6
8	9	9	11	11
12	13	16	19	20
25	26	26	29	30
30	32	33	34	35
39	39	40	42	43

Rješenje u R-u

Podaci su u datoteci data4.csv:

Medijan:

```
> median(x)
22.5
```

Mod - R nema funkciju za mod.
Možemo izračunati frekvencije:

Sve možemo dobiti jednom funkcijom (summary)

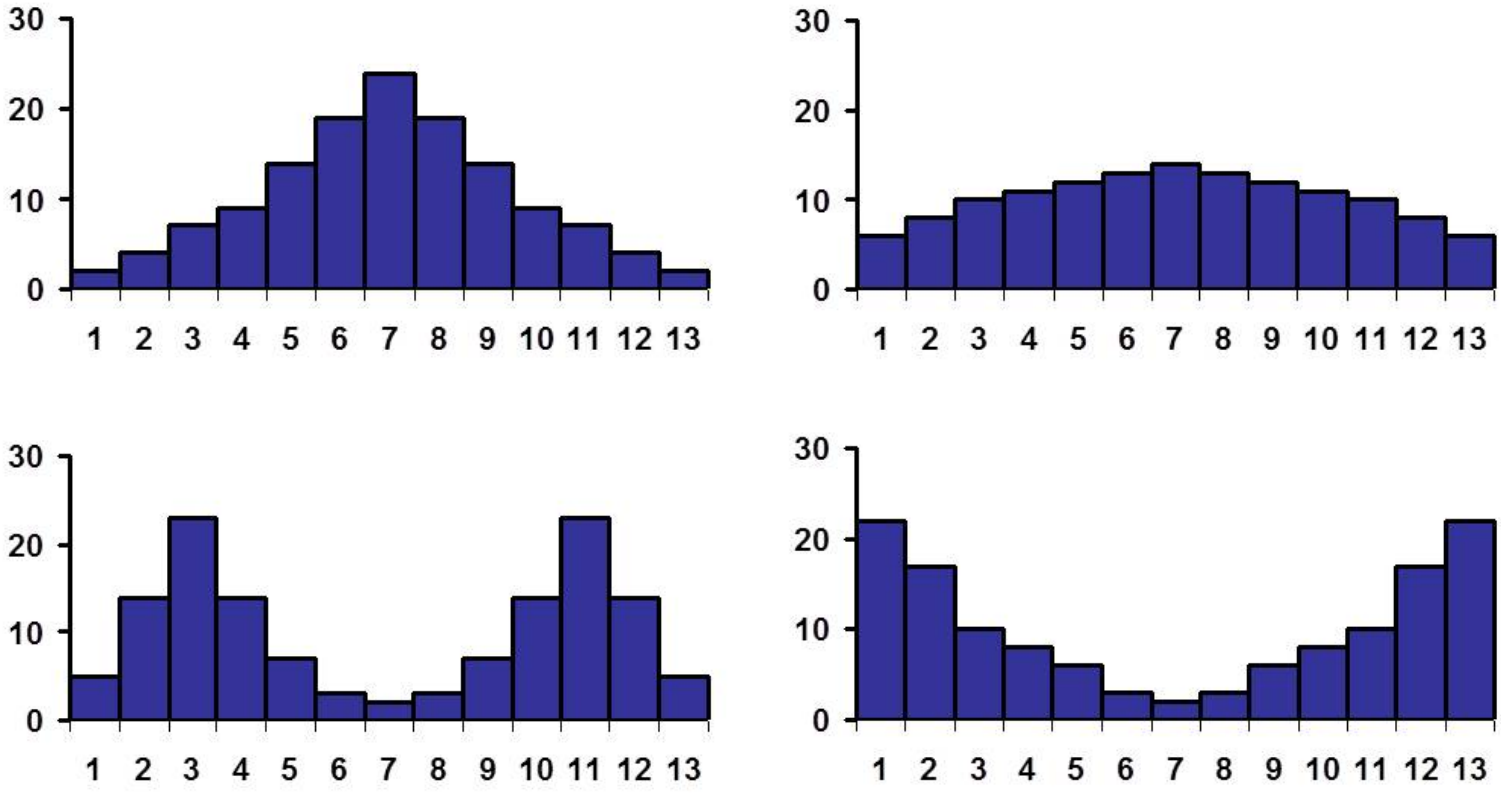
```
> summary(x)
Min.  1st Qu.  Median  Mean 3rd Qu.  Max.
2.00     9.50   22.50 21.83   32.75 43.00
```

Mjere raspršenosti (disperzije)

Usporedimo dva niza podataka:

x_i	10	60	50	30	40	20	$(N = 6,\ \mu = 35)$
y_i	35	45	30	35	40	25	$(N = 6,\ \mu = 35)$

Srednje vrijednosti su iste, ali se podaci ipak razlikuju.
Drugi skup podataka je jače grupiran oko srednje vrijednosti.



U svakom histogramu srednja vrijednost je ista (=7).
Gdje su podaci najviše a gdje najmanje raspršeni?

Varijanca i standardna devijacija

Za skup podataka $x_1, x_2, \ldots, x_N$ definirajmo **odstupanje** i-tog podatka od srednje vrijednosti:

$$x_i - \mu.$$

Loša mjera raspršenosti jer

$$\frac{1}{N}\sum_i (x_i - \mu) = 0.$$

tj., prosječno odstupanje je 0, pa nam to ništa ne govori o ukupnoj raspršenosti podataka.

Kvadratno odstupanje i-tog podatka od srednje vrijednosti:

$$(x_i - \mu)^2 .$$

- Kvadratno odstupanje je uvijek pozitivno.
- Što je udaljenost podataka od srednje vrijednosti veća to je i kvadratno odstupanje veće.
- Srednje kvadratno odstupanje - mjera odstupanja svih podataka.

Varijanca je srednje kvadratno odstupanje podataka od srednje vrijednosti:

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2.$$

σ^2 - oznaka za varijancu.

Alternativna formula za računanje varijance:

$$\sigma^2 = \frac{1}{N}\sum_i x_i^2 - \mu^2.$$

Standardna devijacija je korijen iz varijance:

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_i (x_i - \mu)^2}.$$

σ - oznaka za standardnu devijaciju.

Mjerna jedinica za standardnu devijaciju je ista kao i za obilježje x.

Primjer.
Na farmi je 5 jelena težine:

62.30 kg, 62.28 kg, 69.77 kg, 68.95 kg, 69.88 kg.

Odredite varijancu i standardnu devijaciju za težinu.

Formula za varijancu:

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2.$$

Prvo treba odrediti srednju vrijednost (μ):

$$\mu = \frac{1}{5}(62.30 + 62.28 + 69.77 + 68.95 + 69.88) = \frac{333.18}{5} = 63.636.$$

Varijanca:

$$\begin{aligned}
\sigma^2 &= \frac{1}{5}\Big[(62.30-63.636)^2+(62.28-63.636)^2+(69.77-63.636)^2+ \\
&\quad +(68.95-63.636)^2+(69.88-63.636)^2\Big] = \\
&= \frac{1}{5}\Big[(-4.336)^2+(-4.356)^2+3.134^2+2.314^2+3.244^2\Big] = \\
&= \frac{1}{5}(18.801+18.975+9.822+5.355+10.524) = \\
&= 12.695144
\end{aligned}$$

Standardna devijacija:

$$\sigma = \sqrt{\sigma^2} = \sqrt{12.695144} = 3.563024558$$

Preglednije je koristiti tablicu:

i	x_i	$x_i - \mu$	$(x_i - \mu)^2$
1	62.30	-4.336	18.801
2	62.28	-4.356	18.975
3	69.77	3.134	9.822
4	68.95	2.314	5.355
5	69.88	3.244	10.524
$\sum$	333.18		63.476
$\sum/N$	66.636		12.695

$$\sigma^2 = 12.695$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{12.695144} = 3.563024558$$

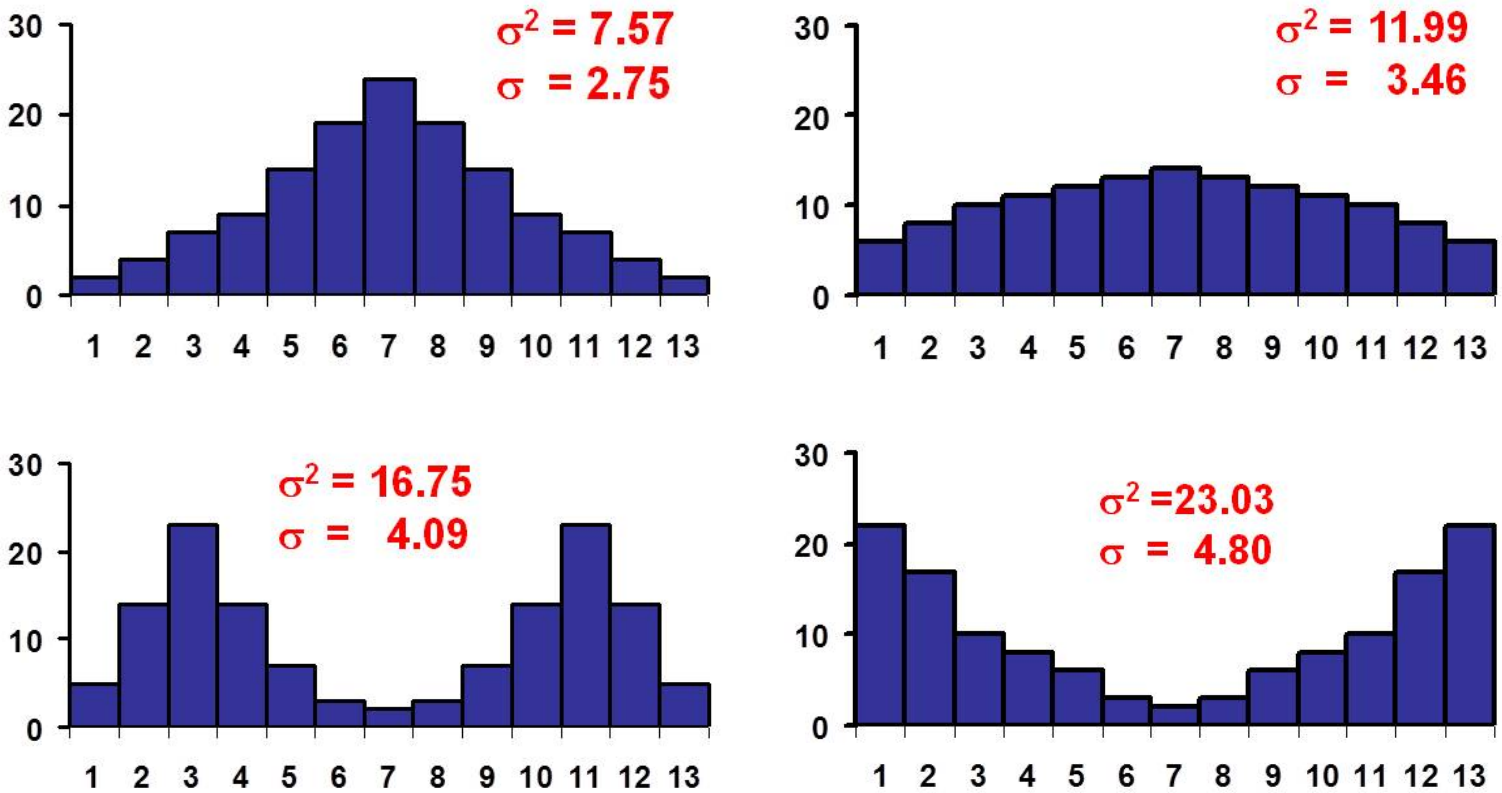
Za alternativnu formulu tablica izgleda ovako:

i	x_i	x_i^2
1	62.30	3881.290
2	62.28	3878.798
3	69.77	4867.853
4	68.95	4754.103
5	69.88	4883.214
$\sum$	333.18	22265.258
$\sum/N$	66.636	4453.052

$$\sigma^2 = \frac{1}{N}\sum_i x_i^2 - \mu^2 = 4453.052 - 66.636^2 = 12.695$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{12.695144} = 3.563024558$$

Varijanca i standardna devijacija za podatke iz uvodnog primjera:



Računanje varijance iz distribucije frekvencija

Neka su $x_1, x_2, x_3, \ldots, x_k$ vrijednosti obilježja i neka su $f_1, f_2, f_3, \ldots, f_k$ pripadne frekvencije. Tada je

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \mu)^2.$$

Alternativna formula za računanje varijance:

$$\sigma^2 = \frac{1}{N} \sum_i f_i x_i^2 - \mu^2.$$

Primjer.

Izračunajte varijancu iz distribucije frekvencija obilježja dob:

Dob	Frekvencija
19	1
20	0
21	5
22	2
23	2
24	1
25	4
26	1
27	0
28	1
29	0
30	1
31	1
32	1

x_i (Dob)	f_i (Frekvencija)	$f_i \cdot x_i$	$f_i \cdot x_i^2$
19	1	19	361
20	0	0	0
21	5	105	2205
22	2	44	968
23	2	46	1058
24	1	24	576
25	4	100	2500
26	1	26	676
27	0	0	0
28	1	28	784
29	0	0	0
30	1	30	900
31	1	31	961
32	1	32	1024
$\sum$	20	485	12013
$\sum / N$		24.25	600.65

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_i f_i x_i^2 - \mu^2 = \\ &= 600.65 - 24.25^2 = \\ &= 12.5875 \end{aligned}$$

Koeficijent varijacije

Varijanca i standardna devijacija ovise o mjernim jedinicama.

- Mjerna jedinica za varijancu je kvadrirana mjerna jedinica varijable.
- Mjerna jedinica za standardnu devijaciju je jednaka mjernoj jedinici varijable.

Primjer. Označimo s y težinu jelena u gramima te izračunajmo varijancu i standardnu devijaciju.

- x - težina jelena u kilogramima
- y - težina jelena u gramima

i	x_i	x_i^2
1	62.30	3881.290
2	62.28	3878.798
3	69.77	4867.853
4	68.95	4754.103
5	69.88	4883.214
$\sum$	333.18	22265.258
$\sum/N$	66.636	4453.052

i	y_i	y_i^2
1	62300	3881290000
2	62280	3878798400
3	69770	4867852900
4	68950	4754102500
5	69880	4883214400
$\sum$	333180	22265258200
$\sum/N$	66636	4453051600

$$\begin{aligned} \sigma_x^2 &= 4453.052 - 66.636^2 = \\ &= 12.695 \\ \sigma_x &= \sqrt{\sigma_x^2} = \sqrt{12.695144} = \\ &= 3.563024558 \end{aligned}$$

$$\begin{aligned} \sigma_y^2 &= 4453051600 - 66636^2 = \\ &= 12695144 \\ \sigma_y &= \sqrt{\sigma_y^2} = \sqrt{12695144} = \\ &= 3563.024558 \end{aligned}$$

Koeficijent varijacije:

$$CV = \frac{\sigma}{\mu}.$$

Koeficijent varijacije nema mjernu jedinicu.

	Težina jelena	
	kg	g
μ	66.636	66636
σ^2	12.695144	12695144
σ	3.56302	3563.02
CV	0.053	0.053

Varijabilnost (varijanca) duljine hitca u centimetrima je veća, iako se zapravo radi o istim podacima.
Koeficijent varijacije je nepromijenjen.

R - varijanca

```
> pod=read.csv("data5.csv")
> x = pod$Tezina
```

Varijanca:

```
> var(x)
15.86893
```

Mi smo izračunali 12.695144!

U R-u se koristi formula:

$$\frac{1}{N-1}\sum_i (x_i-\mu)^2 \quad \text{umjesto} \quad \frac{1}{N}\sum_i (x_i-\mu)^2.$$

(Uzoračka varijanca - koristi je većina statističkih paketa).

Varijancu u R-u treba korigirati.
Množimo s $(N-1)/N$:

```
> var(x)*4/5
```

```
12.69514
```

Standardnu devijaciju

```
> sd(x)
```

```
3.983583
```

treba korigirati s $\sqrt{(N-1)/N}$:

```
> sd(x)*sqrt(4/5)
```

```
3.563025
```

Interkvartil
Kvartili

Donji (prvi) kvartil (Q_1) je vrijednost od koje je 25% podataka manje a 75% podataka veće.

Gornji (treći) kvartil (Q_3) je vrijednost od koje je 75% podataka manje a 25% podataka veće.

Kvartili dijele niz podataka na 4 jednaka dijela.

Medijan = drugi kvartil

Računanje kvartila

Ako podatke $x_1, x_2, \ldots, x_N$ uredimo po veličini, dobivamo niz

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)},$$

Bilo koji realni s broj između 1 i N možemo prikazati kao

$$s = k + r,$$

tako da je k cijeli broj i $0 \leq r < 1$.

Definiramo

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

Medijan:

- Ako je N neparan, medijan je srednji podatak:

$$\mathrm{Me} = x_{\left(\frac{N+1}{2}\right)}$$

- Ako je N paran, medijan je aritmetička sredina dva srednja podatka:

$$\mathrm{Me} = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} = \frac{1}{2}x_{\left(\frac{N}{2}\right)} + \frac{1}{2}x_{\left(\frac{N}{2}+1\right)} = x_{\left(\frac{N+1}{2}\right)}$$

Uz ovakav zapis je

$$\mathrm{Me} = x_{\left(\frac{N+1}{2}\right)} = x_{\left(\frac{1}{2}(N+1)\right)}.$$

(Bez obzira je li N paran ili neparan.)

Za računanje kvartila nema jednoznačno dogovorene formule kao za medijan!

Računanje kvartila - Metoda 1,2

- Niz podataka podijelimo na dva jednaka dijela.
- Ukoliko je broj podataka neparan:
  - izbacimo medijan. (Metoda 1)
  - medijan dodamo i donjoj i gornjoj polovici. (Metoda 2)
- Donji (prvi) kvartil je medijan prve polovice niza.
- Gornji (treći) kvartil je medijan druge polovice niza.

Računanje kvartila - Metoda 3

Donji (prvi) kvartil (Q_1)

$$Q_1 = x_{\left(\frac{1}{4}(N+1)\right)}$$

Gornji (treći) kvartil (Q_3)

$$Q_3 = x_{\left(\frac{3}{4}(N+1)\right)}$$

Interkvartil (IQ) je udaljenost između donjeg i gornjeg kvartila:

$$IQ = Q_3 - Q_1.$$

- Između donjeg i gornjeg kvartila se nalazi 50% podataka.
- Interkvartil je raspon u kojem se nalazi 50% središnjih članova niza podataka.
- Vrijednost prvih i zadnjih 25% podataka ne utječe na vrijednost interkvartila.

Koeficijent kvartilne devijacije

Koeficijent kvartilne devijacije (VQ) je:

$$VQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

- Relativna mjera raspršenosti
- $0 \leq VQ < 1$

Primjer. Odredite donji i gornji kvartil, interkvartil i koeficijent kvartilne devijacije za sljedeći niz podataka:

1.3 4.1 4.1 4.2 4.4 4.6 5.1 5.2 5.3 5.5 5.5 5.5 5.9 6.1 7.8

Rješenje. (Metoda 1) Prvo određujemo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela. N=15 je neparan.

1.3 4.1 4.1 4.2 4.4 4.6 5.1

5.2

5.3 5.5 5.5 5.5 5.9 6.1 7.8

Donji kvartil (Q1) je medijan prve polovice niza:

1.3 4.1 4.1 **4.2** 4.4 4.6 5.1

Gornji kvartil (Q3) je medijan druge polovice niza:

5.3 5.5 5.5 **5.5** 5.9 6.1 7.8

Donji i gornji kvartil niza podataka:

1.3 4.1 4.1 **4.2** 4.4 4.6 5.1 5.2 5.3 **5.5** 5.5 5.5 5.9 6.1 7.8

Q1 **Q3**

Interkvartil:

$$IQ = Q_3 - Q_1 = 5.5 - 4.2 = 1.3$$

Koeficijent kvartilne devijacije:

$$VQ = \frac{Q3 - Q1}{Q3 + Q1} = \frac{1.3}{9.7} = 0.13402.$$

Rješenje. (Metoda 2) Određujemo samo donji i gornji kvartil.

Niz podijelimo na dva jednaka dijela. N=15 je neparan.

1.3 4.1 4.1 4.2 4.4 4.6 5.1

5.2

5.3 5.5 5.5 5.5 5.9 6.1 7.8

Donji kvartil (Q1) je medijan prve polovice niza:

1.3 4.1 4.1 **4.2 4.4** 4.6 5.1 5.2 $\rightarrow$ $Q_1 = (4.2 + 4.4)/2 = 4.3$

Gornji kvartil (Q3) je medijan druge polovice niza:

5.2 5.3 5.5 **5.5 5.5** 5.9 6.1 7.8 $\rightarrow$ $Q_3 = (5.5 + 5.5)/2 = 5.5$

Rješenje. (Metoda 3) Određujemo samo donji i gornji kvartil.

$N = 15$

Donji kvartil:

$$Q_1 = x_{\left(\frac{1}{4}(N+1)\right)} = x_{\left(\frac{1}{4}(15+1)\right)} = x_{(4)} = 4.2$$

Gornji kvartil:

$$Q_3 = x_{\left(\frac{3}{4}(N+1)\right)} = x_{\left(\frac{3}{4}(15+1)\right)} = x_{(12)} = 5.5$$

U ovom slučaju je rezultat isti kao i iza metodu 1.

Primjer. Isto kao i prije ali bez prvog podatka. Odredite donji i gornji kvartil za sljedeći niz podataka:

4.1 4.1 4.2 4.4 4.6 5.1 5.2 5.3 5.5 5.5 5.5 5.9 6.1 7.8

Rješenje. (Metoda 3) $N = 14$

Donji kvartil:

$$\begin{aligned} Q_1 &= x_{\left(\frac{1}{4}(N+1)\right)} = x_{\left(\frac{1}{4}(14+1)\right)} = x_{\left(3+\frac{3}{4}\right)} = \\ &= x_{(3)} + \frac{3}{4}\left(x_{(4)} - x_{(3)}\right) = 4.2 + \frac{3}{4}(4.4 - 4.2) = 4.35 \end{aligned}$$

Gornji kvartil:

$$\begin{aligned} Q_3 &= x_{\left(\frac{3}{4}(N+1)\right)} = x_{\left(\frac{3}{4}(14+1)\right)} = x_{\left(11+\frac{1}{4}\right)} = \\ &= x_{(11)} + \frac{1}{4}\left(x_{(12)} - x_{(11)}\right) = 5.5 + \frac{1}{4}(5.9 - 5.5) = 5.6 \end{aligned}$$

R - kvartili

Podaci iz 1. primjera

```
> pod=read.csv("data6.csv")
> x = pod$Podaci
```

Kvartili:

```
> summary(x)
Min.   1st Qu.  Median  Mean 3rd Qu.    Max.
1.300    4.300   5.200 4.973   5.500 7.8000
```

Rezultat kao u metodi 2.

Podaci iz 2. primjera (1. primjer bez prvog podatka)

```
> pod=read.csv("data6b.csv")
> x = pod$Podaci
```

Kvartili:

```
> summary(x)
Min.   1st Qu.  Median  Mean 3rd Qu.    Max.
4.100    4.450   5.250 5.236   5.500 7.8000
```

Q_1 različit od rezultata iz svih metoda.

$Q_1 = 4.4$ - Metode 1 i 2, $Q_1 = 4.2$ - Metoda 3

Metoda 3:

$$Q_1 = x_{\left(\frac{1}{4}(N+1)\right)} \qquad Q_3 = x_{\left(\frac{3}{4}(N+1)\right)}$$

R koristi malo drugačiju definiciju:

$$Q_1 = x_{\left(1+\frac{1}{4}(N-1)\right)} \qquad Q_3 = x_{\left(1+\frac{3}{4}(N-1)\right)}$$

medijan je u sredini prvog i zadnjeg podatka. Pozicija medijana (s) je:

$$s = \frac{1+N}{2}.$$

Prvi kvartil je u sredini između 1. podatka i medijana. Pozicija prvog kvartila je

$$s = \frac{1+\frac{1+N}{2}}{2} = \frac{3+N}{4} = 1 + \frac{1}{4}(N-1).$$

Ostale mjere raspršenosti
Srednje apsolutno odstupanje

Umjesto kvadratnog odstupanja podataka od srednje vrijednosti:

$$(x_i - \mu)^2$$

mogli smo promatrati apsolutno odstupanje:

$$|x_i - \mu|.$$

Time dobvamoi mjeru raspršenosti poznatu kao **srednje apsolutno odstupanje**:

$$\text{s.a.o.} = \frac{1}{N} \sum_i |x_i - \mu|.$$

(engl. MAD - mean absolute deviation)

- Može se gledati i srednje apsolutno odstupanje od drugih mjera centralne tendencije.
- Srednje apsolutno odstupanje od srednje vrijednossti.
- Srednje apsolutno odstupanje od medijana.
- Srednje apsolutno odstupanje od moda.
- Umjesto srednje vrijednosti može se koristiti medijan apsolutnih odstupanja $\rightarrow$ medijalno apsolutno odstupanje.
- I ovdje se može promatrati odstupanje od srednje vrijednosti, medijana ili moda.

Napomena. Srednja vrijednost, medijan i standardna devijacija zadovoljavaju:

$$|\mu - m| \leq \sigma.$$

Napomena. Srednja vrijednost (μ) je točka s najmanjim srednjim kvadratnim odstupanjem od podataka.
Drugim riječima, za proizvoljan broj a vrijedi

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \leq \frac{1}{N}\sum_i (x_i - a)^2.$$

Napomena. Medijan (m) je točka s najmanjim srednjim apsolutnim odstupanjem od podataka.
Drugim riječima, za proizvoljan broj a vrijedi

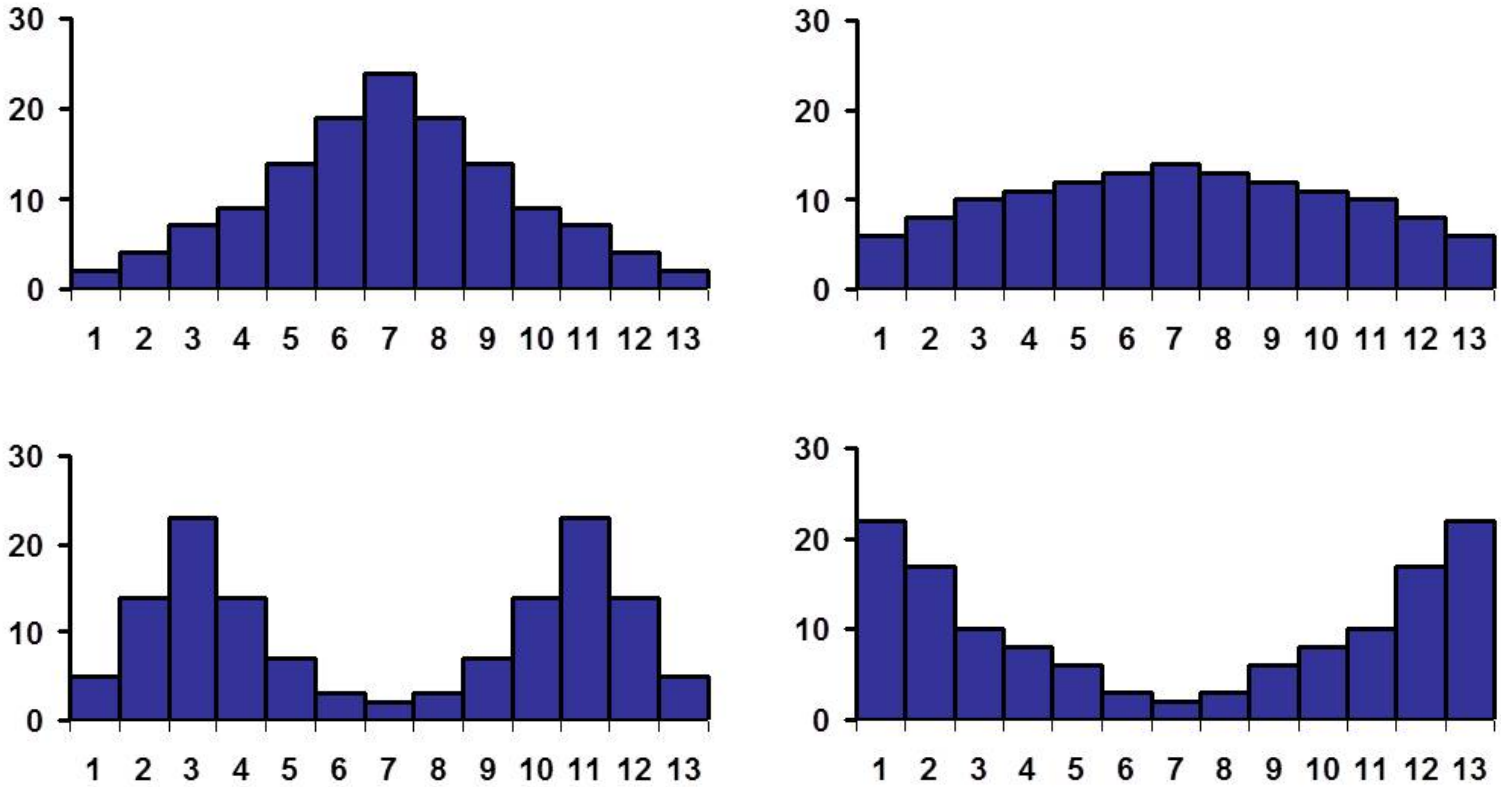
$$\frac{1}{N}\sum_i |x_i - m| \leq \frac{1}{N}\sum_i |x_i - a|.$$

Raspon varijacije

Raspon varijacije je razlika najveće i najmanje vrijednosti u skupu podataka:

$$R_X = x_{\max} - x_{\min}.$$

- najjednostavnija mjera raspršenosti
- temeljena na dva podatka
- iskazan u mjernim jedinicama obilježja
- veći R_X, veća raspršenost, veći stupanj varijabilnosti



U sva 4 primjera je raspon varijacije isti ($R = 13 - 1 = 12$) iako je varijabilnost podataka naočigled različita!

Mjere raspršenosti mogu biti

- u mjernim jedinicama obilježja
- u relativnom iznosu

Mjere raspršenosti temeljene na dijelu podataka:

- raspon varijacije
- interkvartil i koeficijent kvartilne devijacije

Mjere raspršenosti temeljene na svim podacima:

- varijanca, standardna devijacija, koeficijent varijacije
- srednje apsolutno odstupanje

Računanje mjera centralne tendencije i raspršenosti za grupirane podatke

Srednja vrijednost i varijanca

Zadani su razredi i njihove frekvencije (f_i).

Za svaki razred izračunamo sredinu razreda X_i te srednju vrijednost i varijancu računamo kao da svi podaci unutar razreda imaju istu vrijednost (X_i).

Srednja vrijednost:

$$\mu = \frac{\sum f_i X_i}{\sum f_i} = \frac{\sum f_i X_i}{N}.$$

Varijanca:

$$\sigma^2 = \frac{1}{N} \sum_i f_i (X_i - \mu)^2.$$

Medijan i interkvartil

Kod računanja medijana (i kvartila) pretpostavljamo da su unutar svakog razreda frekvencije jednoliko (uniformno) raspoređene (sve vrijednosti imaju istu frekvenciju).

Prvo odredimo **medijalni razred**, t.j. razred u kojem se nalazi medijan.

Medijan računamo iz formule

$$m = \frac{\frac{N+1}{2} - (F+1)}{f_m} w_m + L_m$$

gdje je

- N - veličina skupa
- F - zbroj svih frekvencija razreda do, ali ne uključujući, medijalnog razreda (tj. kumulativna frekvencija razreda prije medijalnog)
- f_m - frekvencija medijalnog razreda
- w_m - širina medijalnog razreda
- L_m - donja granica medijalnog razreda ukoliko se granice razreda preklapaju, ili aritmetička sredina donje granice medijalnog razreda i gornje granice prethodnog razreda ukoliko se granice ne preklapaju

Mod

- U distribuciji frekvencija MOD je vrijednost varijable s najvećom frekvencijom.
- U distribuciji frekvencija s razredima - mod približno određujemo (aproksimiramo ga).
- **Modalni razred** - razred s najvećom korigiranom frekvencijom.
- Mod se ne treba nalazitine nalazi u modalnom razredu.
- modalni razred ne treba biti razred s najvećom frekvencijom.

Napomena. Mjere centralne tendencije i raspršenosti ne možemo egzaktno izračunati iz grupiranih podataka.
Računamo ih približno.

Linearna transformacija varijable

Ako varijabli X pribrojimo konstantu (broj), što se događa sa srednjom vrijednosti i varijancom?

	X	$X+10$	$X+20$
	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
μ	66.64	76.64	86.64
σ^2	12.695	12.695	12.695

Varijabla Y je varijabla X uvećana za b:

$$Y = X + b.$$

Tada je

$$\begin{aligned} \mu_Y &= \mu_X + b \\ \sigma_Y^2 &= \sigma_X^2 \\ \sigma_Y &= \sigma_X \end{aligned}$$

gdje je

μ_X - srednja vrijednost varijable X
σ_X^2 - varijanca varijable X
σ_X - standardna devijacija varijable X
μ_Y - srednja vrijednost varijable Y
σ_Y^2 - varijanca varijable Y
σ_Y - standardna devijacija varijable Y

Ako varijablu X pomnožimo s konstantom (brojem), što se događa sa srednjom vrijednosti i varijancom?

	X	$10 \cdot X$	$100 \cdot X$
	62.30	623.0	6230
	62.28	622.8	6228
	69.77	697.7	6977
	68.95	689.5	6895
	69.88	698.8	6988
μ	66.64	666.4	6664
σ^2	12.695	1269.5	126950

Varijabla Y je varijabla X pomnožena s a:

$$Y = aX.$$

Tada je

$$\begin{aligned} \mu_Y &= a\,\mu_X \\ \sigma_Y^2 &= a^2\sigma_X^2 \\ \sigma_Y &= |a|\,\sigma_X \end{aligned}$$

gdje je

μ_X - srednja vrijednost varijable X
σ_X^2 - varijanca varijable X
σ_X - standardna devijacija varijable X
μ_Y - srednja vrijednost varijable Y
σ_Y^2 - varijanca varijable Y
σ_Y - standardna devijacija varijable Y

Neka je varijabla Y dana s:

$$Y = aX + b.$$

Tada je

$$\begin{aligned} \mu_Y &= a\mu_X + b \\ \sigma_Y^2 &= a^2\sigma_X^2 \\ \sigma_Y &= |a|\,\sigma_X \end{aligned}$$

gdje je

μ_X - srednja vrijednost varijable X
σ_X^2 - varijanca varijable X
σ_X - standardna devijacija varijable X
μ_Y - srednja vrijednost varijable Y
σ_Y^2 - varijanca varijable Y
σ_Y - standardna devijacija varijable Y

Primjer. Na farmi je srednja vrijednost za visinu borova 7.52 m uz standardnu devijaciju od 0.78 m. Kolika je srednja vrijednost i standardna devijacija visine borova izražena u centimetrima?

X - visina u metrima
Y - visina u centimetrima

$Y = 100X$

$$\begin{aligned} \mu_Y &= 100\,\mu_X = 100 \cdot 7.52 = 752 \\ \sigma_Y &= 100\,\sigma_X = 100 \cdot 0.78 = 78 \end{aligned}$$

Srednja vrijednost visine borova je 752 cm uz standardnu devijaciju od 78 cm.

Što se događa s srednjom vrijednošću i varijancom zbroja dvije varijable?

Primjer. Ribolovni obrt posjeduje dva broda. U tablici su prikazani dnevni ulovi (u kg) za svaki brod posebno i ukupan ulov za oba broda kroz period od 10 dana..
Usporedite srednju vrijednost, varijancu i standardnu devijaciju za ova tri ulova.

Skakačica	Ulov 1. brod (X)	Ulov 2. brod (Y)	Ulov Ukupno (Z)
1. dan	122.7	120.7	243.4
2. dan	117.0	122.0	239.0
3. dan	113.3	110.3	223.6
4. dan	114.6	106.6	221.2
5. dan	107.7	111.5	219.2
6. dan	109.6	108.8	218.4
7. dan	112.8	102.1	214.9
8. dan	105.6	108.6	214.2
9. dan	108.3	105.9	214.2
10. dan	105.0	108.4	213.4
Srednja vrijednost (μ)	111.66	110.49	222.15
Varijanca (σ^2)	27.53	35.52	101.68
Standardna devijacija (σ)	5.247	5.960	10.084

Iz primjera vidimo da se srednje vrijednosti zbrajaju dok to ne vrijedi za varijancu i standardnu devijaciju:

X - dnevni ulov 1. broda
Y - dnevni ulov 2. broda
$Z = X + Y$ -ukupni dnevni ulov

Neka je varijabla Z zbroj varijabli X i Y:

$$Z = X + Y.$$

Tada je

$$\mu_Z = \mu_X + \mu_y$$

gdje je

μ_X - srednja vrijednost varijable X
μ_Y - srednja vrijednost varijable Y
μ_Z - srednja vrijednost varijable Z

Primjer.

Za isti ribarski obrt je na godišnjoj razini srednja vrijednost dnevnog ulova iznosila 86.42 kg uz standardnu devijaciju od 5.63 kg. Kolika je srednja vrijednost i standardna devijacija dobiti ukoliko je otkupna vrijednost ribe 10 € po kg a dnevni troškovi iznose 500 €?

Rješenje. Oznaka
U - dnevni ulov
D - dnevna dobit

Dobit se računa prema formuli:

$$D = 10 \cdot U - 500.$$

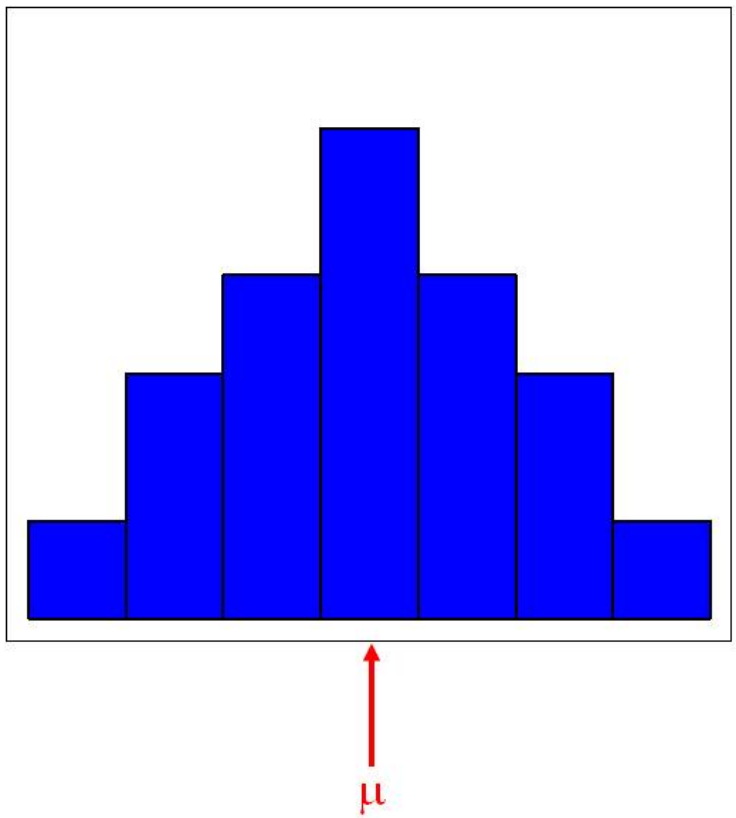
Srednja vrijednost i standardna devijacija su:

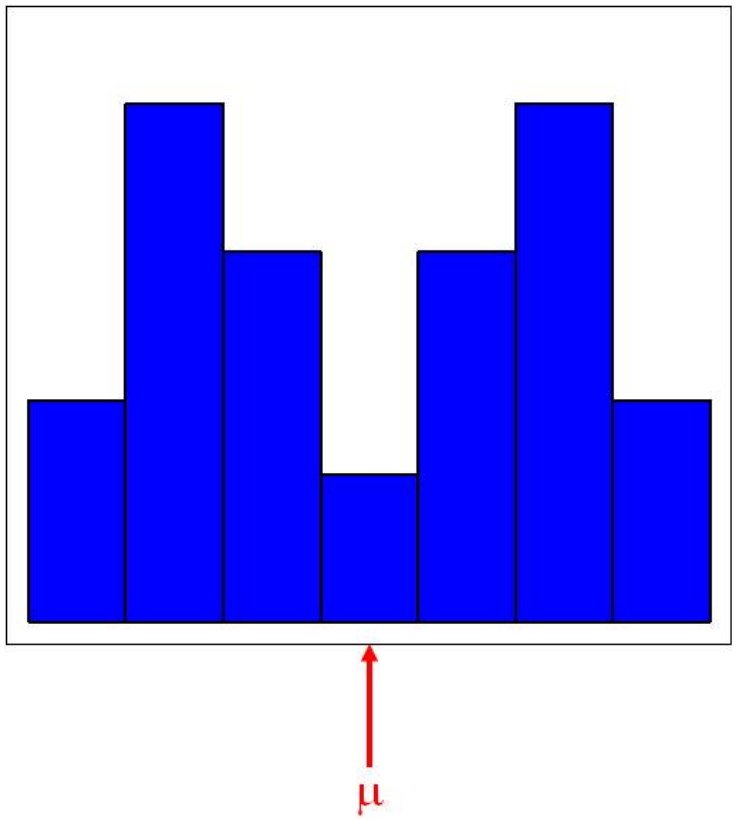
$$\mu_D = 2 \cdot \mu_U - 500 = 10 \cdot 86.42 - 500 = 360.84$$

i

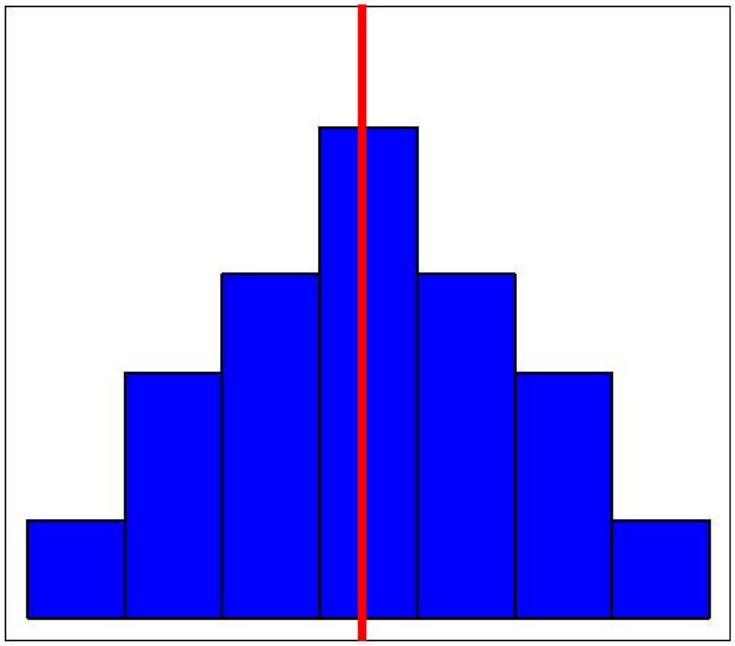
$$\sigma_D = 10 \cdot \sigma_U = 10 \cdot 5.63 = 56.3.$$

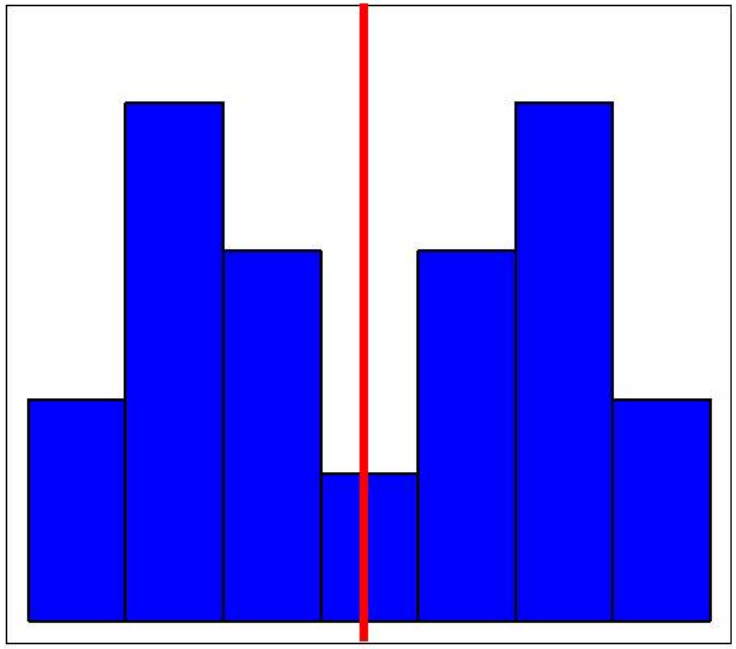
Asimetrija





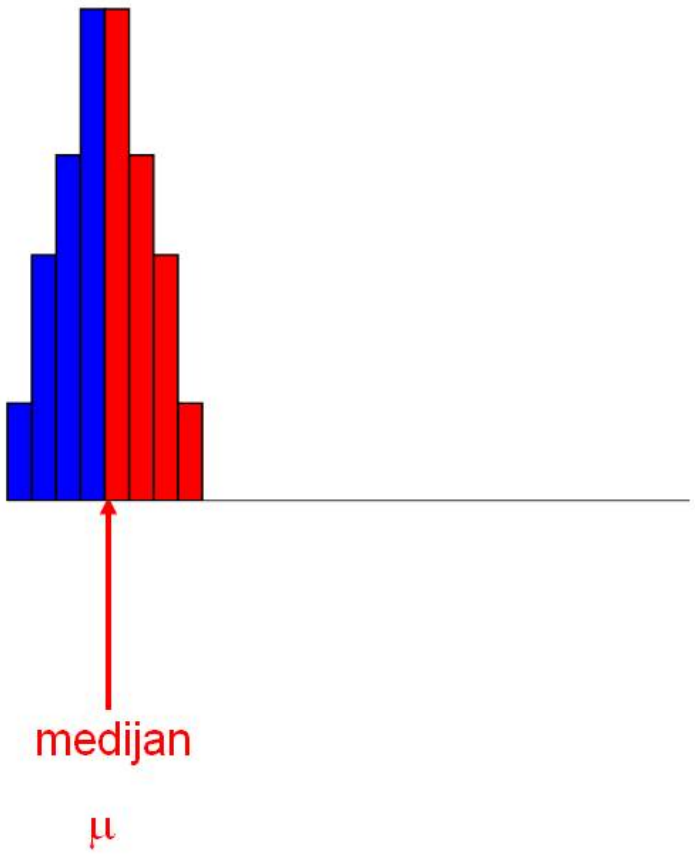
Mjere asimetrije



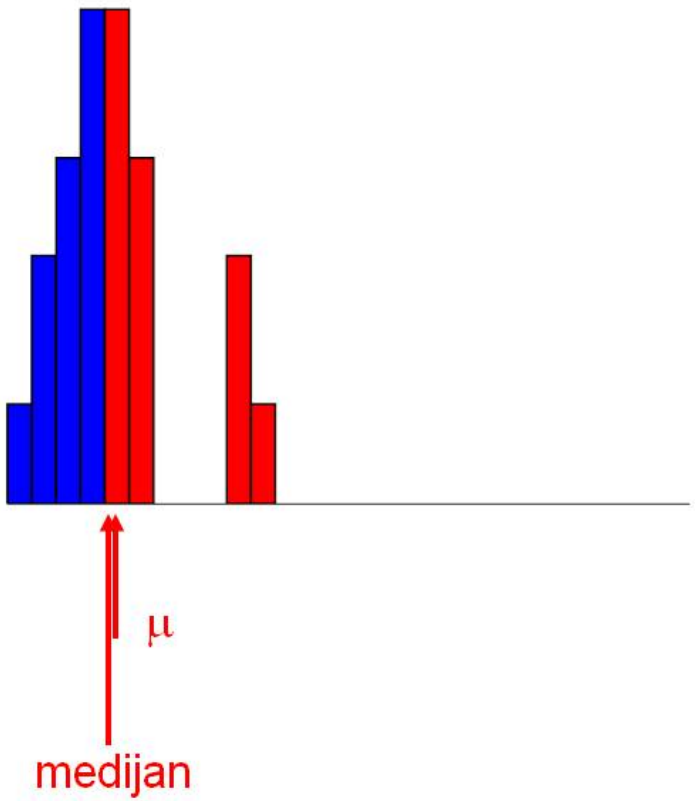


Simetrična razdioba!

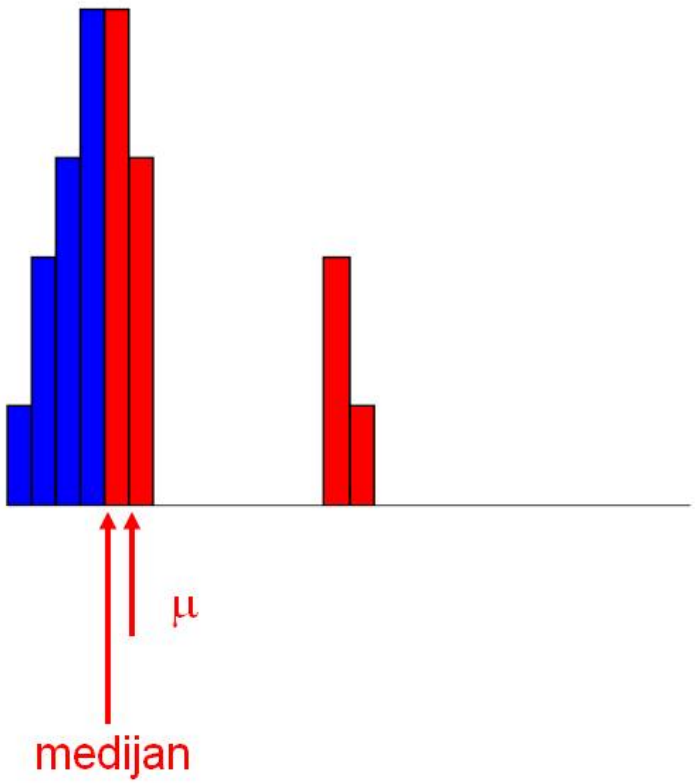
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



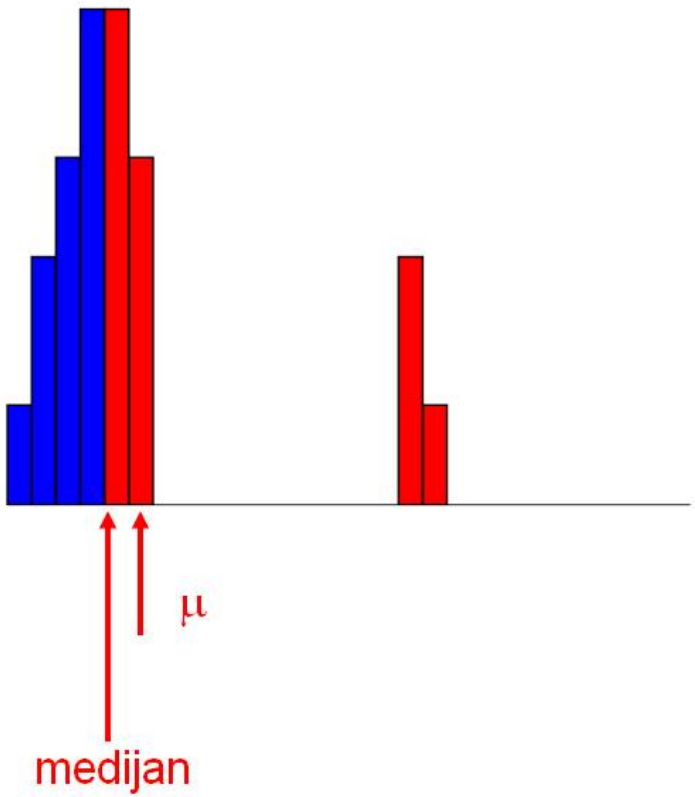
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



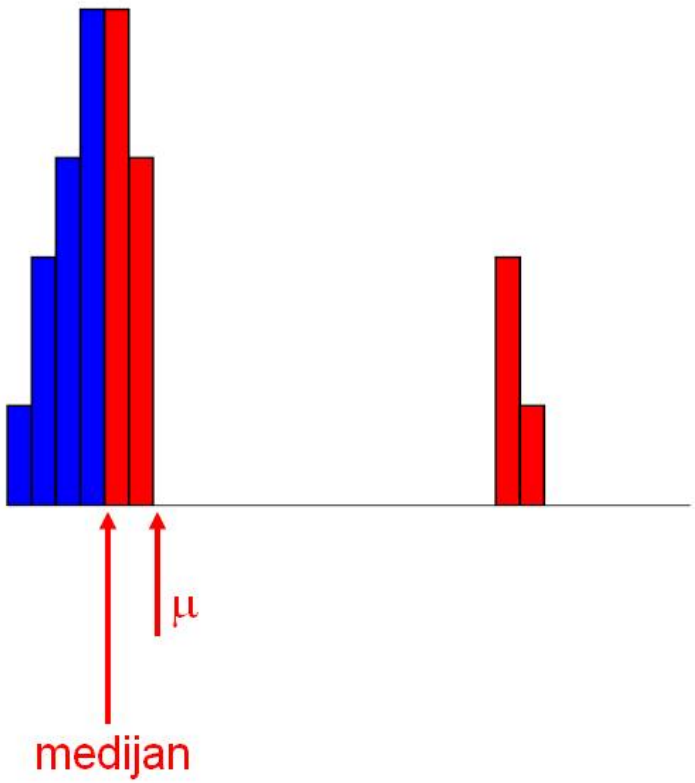
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



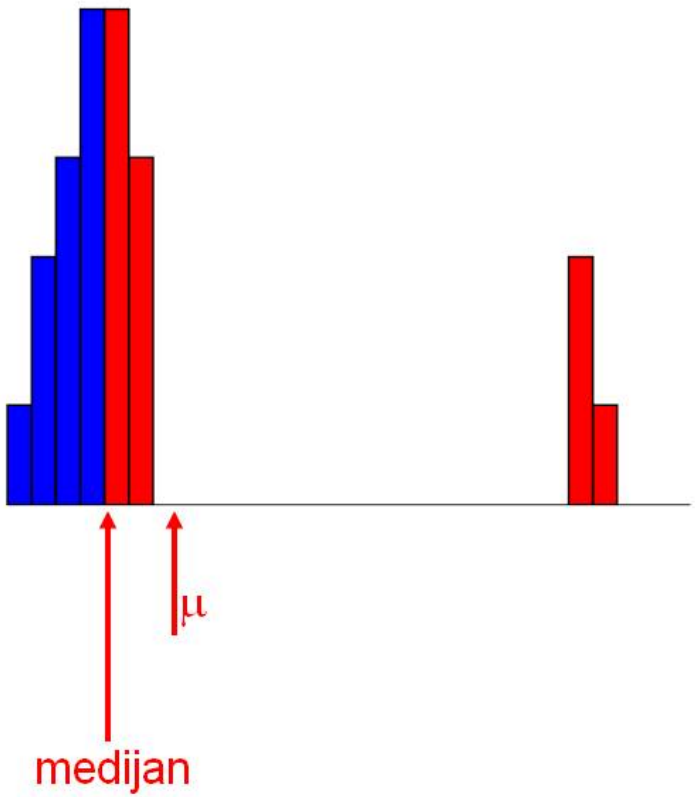
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



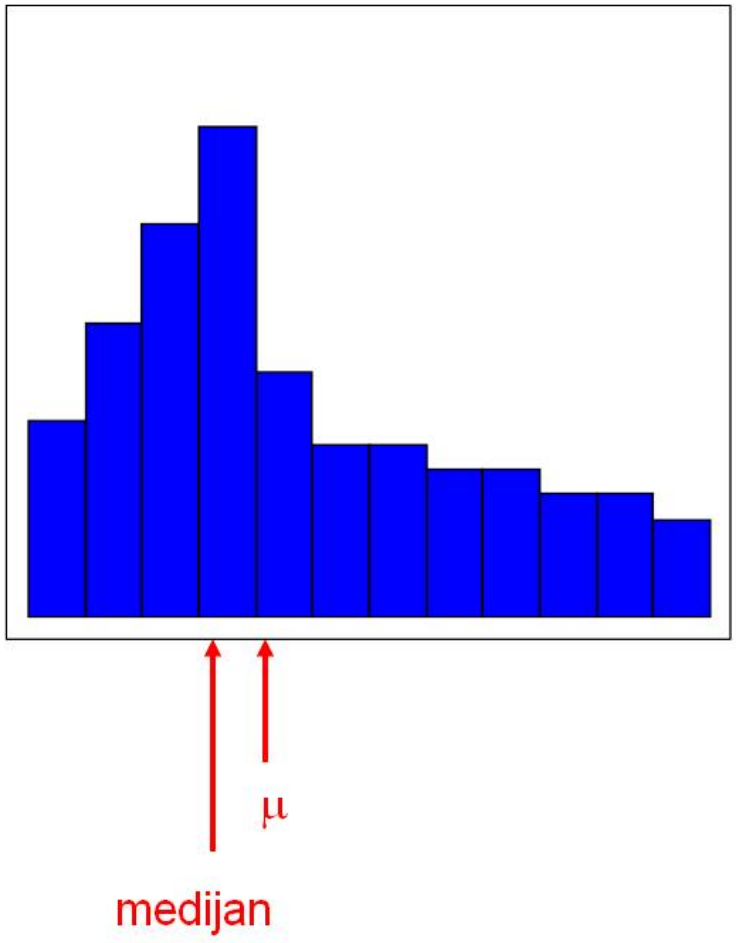
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



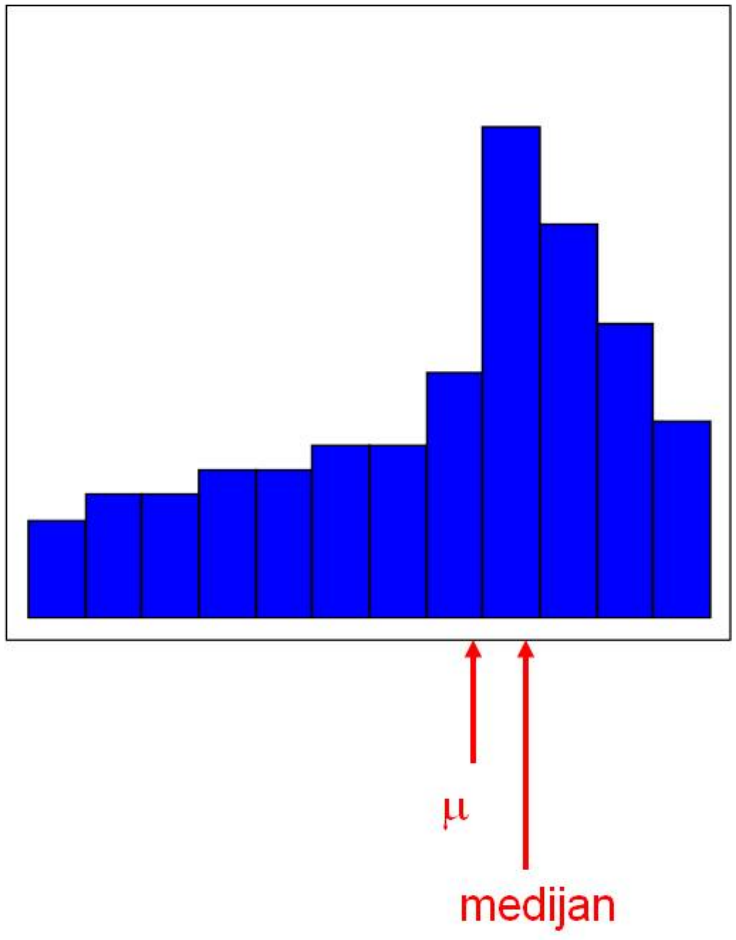
Što se događa sa srednjom vrijednošću i medijanom kada razdioba nije simetrična?



Asimetrične razdiobe:



Pozitivno zakošena



Negativno zakošena

- Nakošenost (engl. skewness)

Pearsonov koeficijent zakošenosti

$$Sk_P = 3\frac{\mu - m}{\sigma},$$

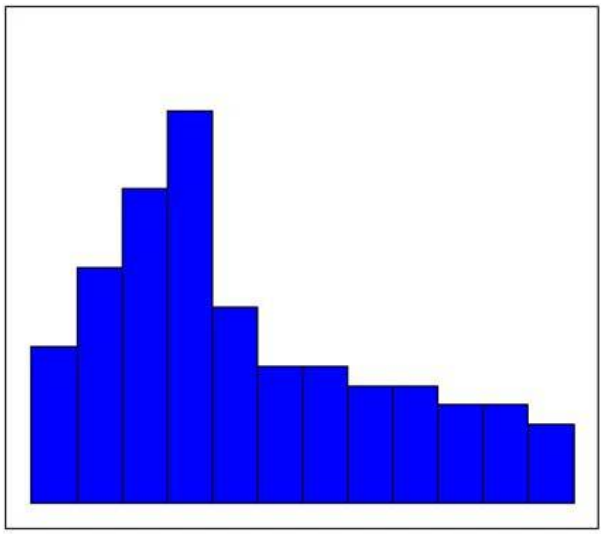
gdje je
μ - srednja vrijednost
m - medijan
σ - standardna devijacija

Zakošenost

$$skew(X) = \sum_i \left(\frac{x_i - \mu}{\sigma} \right)^3$$

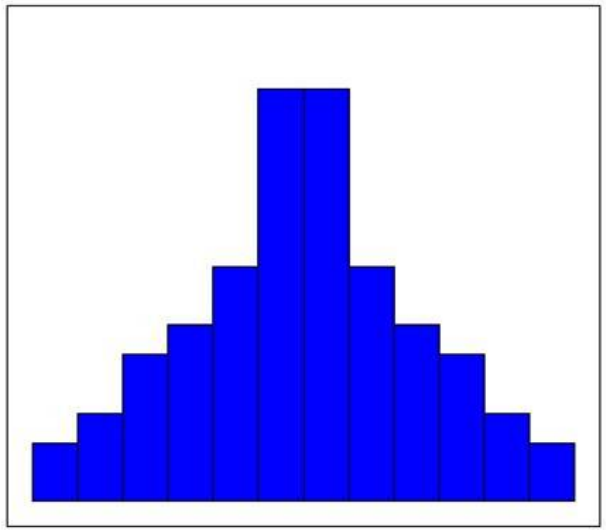
gdje je
μ - srednja vrijednost
m - medijan
σ - standardna devijacija

Primjer.



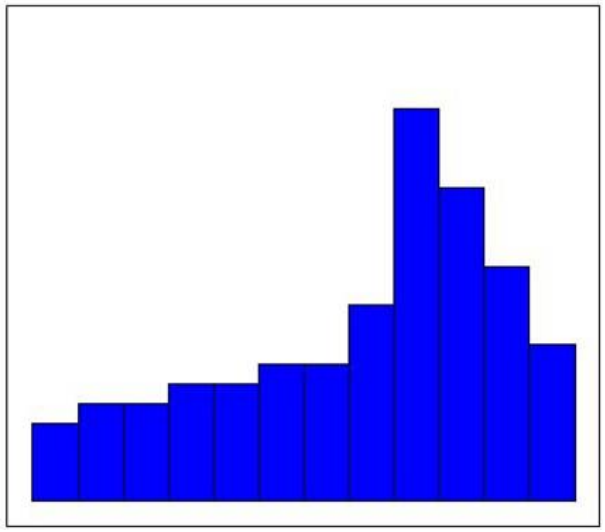
Zakošenost = 0.64

Sk_P = 1.22



Zakošenost = 0

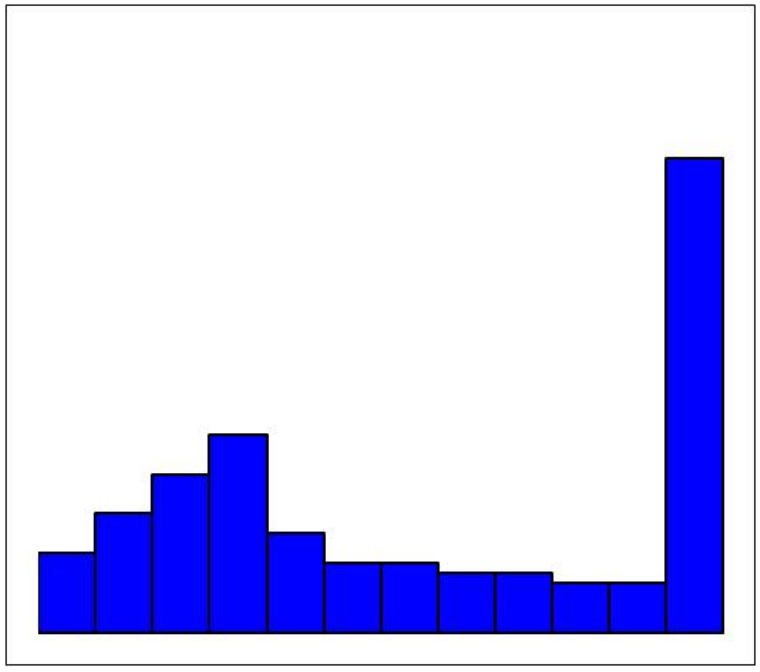
Sk_P = 0



Zakošenost = -0.64

Sk_P = -1.22

Oprez!

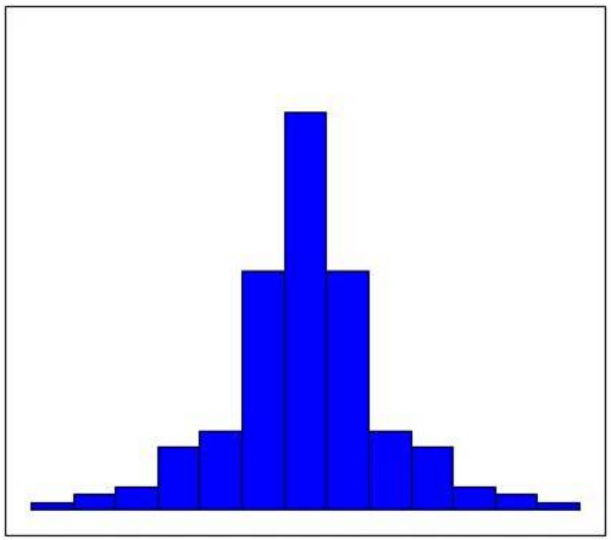


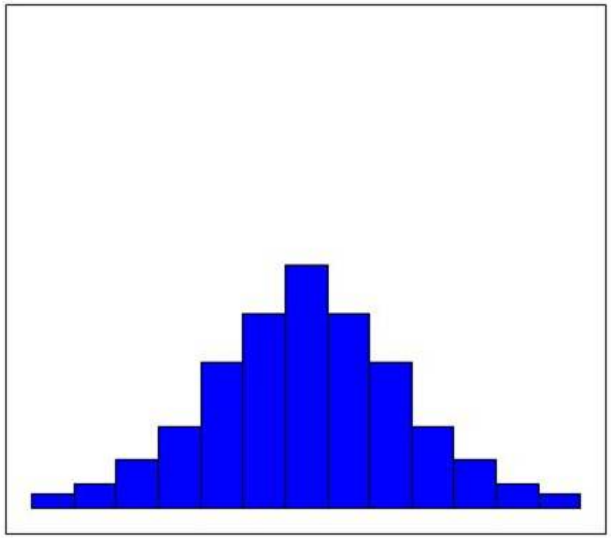
Zakošenost = 0

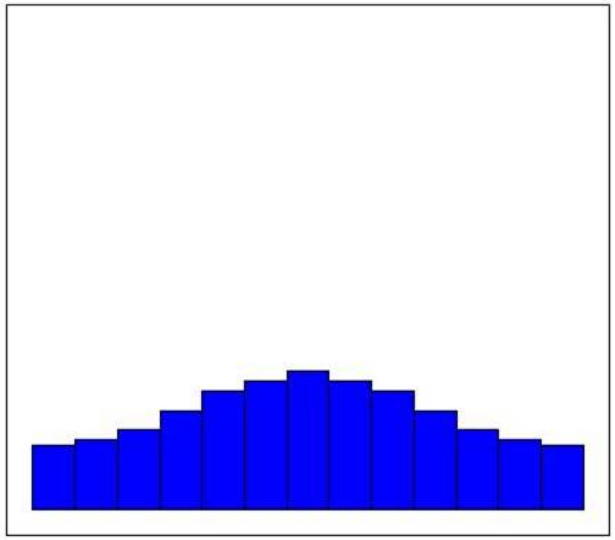
Bimodalna distribucija!

Zakošenost je dobar pokazatelj asimetrije za unimodalne distribucije.

Spljoštenost







Koeficijent spljoštenosti

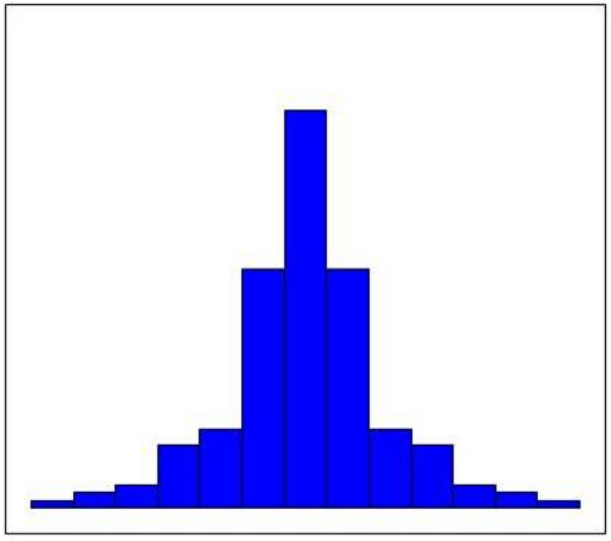
$$kurt(X) = \sum_i \left(\frac{x_i - \mu}{\sigma} \right)^4$$

gdje je

μ - srednja vrijednost
m - medijan
σ - standardna devijacija

Oprez! Koeficijent spljoštenosti se ponekad definira i kao

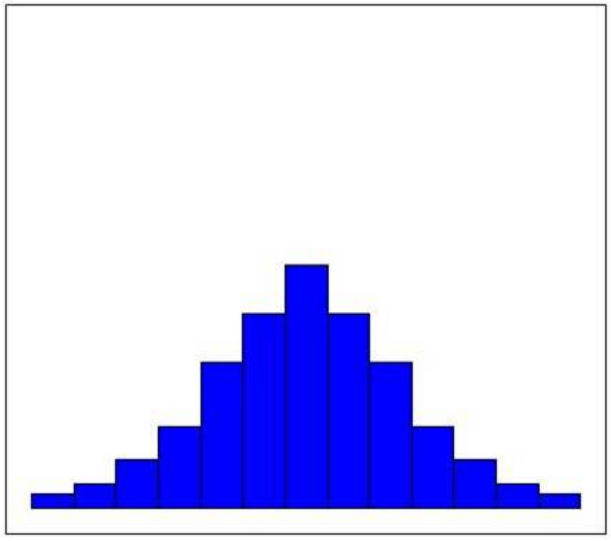
$$kurt(X) = \sum_i \left(\frac{x_i - \mu}{\sigma} \right)^4 - 3$$



$kurt(X) = 4.31$

$kurt(X) > 3$

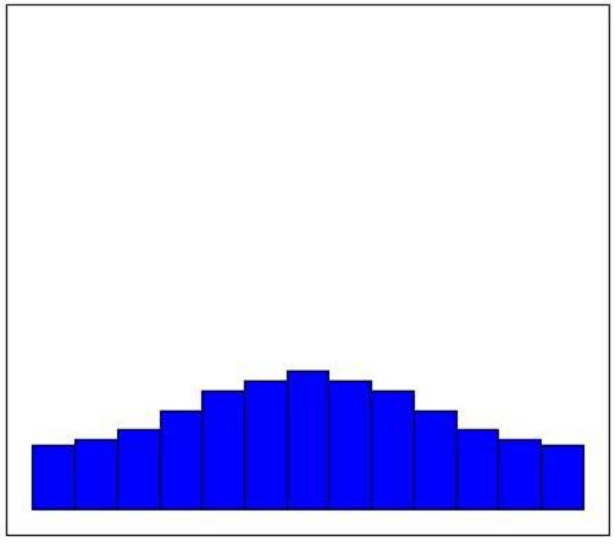
Leptokurtična (izbočena) distribucija



$kurt(X) = 3$

$kurt(X) = 3$

Mezokurtična distribucija



$kurt(X) = 2.13$

$kurt(X) < 3$

Platikurtična (spljoštena) distribucija

Položaj podataka
Rang

Rang je položaj (redni broj) podatka u nizu podataka.
Ukoliko dva ili više podataka imaju istu vrijednost, njihov rang je aritmetička sredina rangova koje bi imali da su vrijednosti različite.

Primjer. Odredite rangove za zadani skup podataka:

$$15 \quad 2 \quad 21 \quad 13 \quad 8 \quad 17.$$

Rješenje. Za određivanje rangova podatke moramo prvo posložiti od najmanjeg do najvećeg:

$$2 \quad 8 \quad 13 \quad 15 \quad 17 \quad 21,$$

a zatim im pridružimo rang:

Vrijednost	2	8	13	15	17	21
Rang	1	2	3	4	5	6

Npr. Rang za vrijednost 15 je 4.

Primjer. Odredite rangove za zadani skup podataka:

15 2 21 13 8 17 15 8 15.

Rješenje. Promatramo niz podataka:

2 8 8 13 15 15 15 17 21,

a zatim im pridružimo rang:

Vrijednost	2	8	8	13	15	15	15	17	21
	1	2	3	4	5	6	7	8	9
Rang	1	2.5	2.5	4	6	6	6	8	9

Npr. rang za vrijednost 8 je (2+3)/2 = 2.5 a za 15: (5+6+7)/3=6.

Centili (percentili)

Percentili (centili) dijele niz podataka na 100 jednakih dijelova.

> ***p*-ti percentil** je broj koji niz podataka dijeli tako da je p% podataka manje od p-tog percentila a $(100 - p)$% veće od njega.

Medijan = 50-ti percentil
Donji kvartil = 25-ti percentil
Gornji kvartil = 75-ti percentil

Decili - dijele niz podataka na 10 jednakih dijelova.

Zajednički naziv za percentile, decile, kvartile i medijan je **kvantili** ili **fraktili**.

Standardizirana varijabla

Za varijablu X čija je srednja vrijednost μ i standardna devijacija σ standardizirana varijabla je varijabla

$$Z = \frac{X - \mu}{\sigma}.$$

Vrijednost standardizirane varijable: z-vrijednost, standardizirana vrijednost

Standardizirana varijabla pokazuje koliko je standardnih devijacija vrijednost podatka udaljena od srednje vrijednosti.

Standardizirana varijabla Z zadovoljava:

$$\mu_Z = 0 \quad \text{i} \quad \sigma_Z = 1.$$

(Srednja vrijednost je 0 a standardna devijacija 1).

Jer je

$$Z = \frac{1}{\sigma} X - \frac{\mu}{\sigma},$$

vrijedi

$$\mu_Z = \frac{1}{\sigma}\mu_X - \frac{\mu}{\sigma} = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

i

$$\sigma_Z = \frac{1}{\sigma}\sigma_X = \frac{1}{\sigma}\sigma = 1.$$

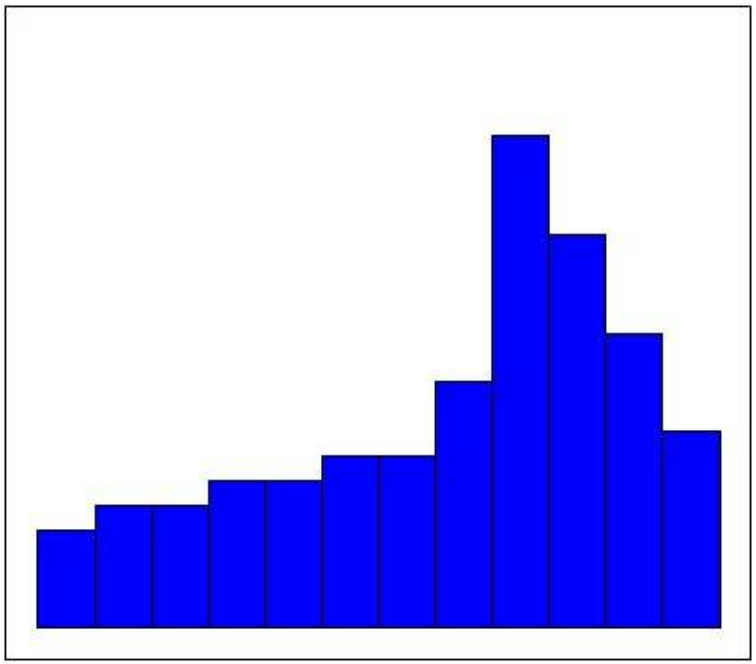
Čebiševljeva nejednakost

Između $\mu - k\sigma$ i $\mu + k\sigma$ (k > 1) nalazi se najmanje

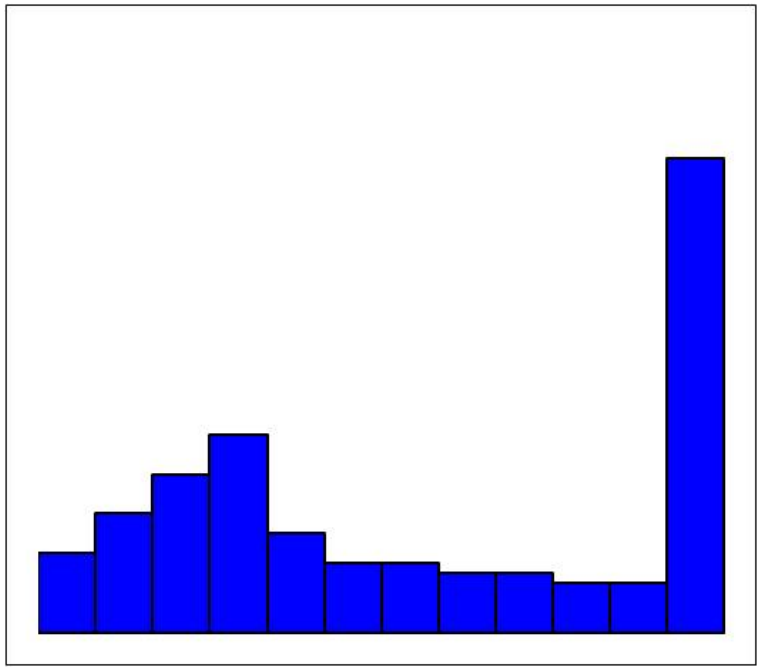
$$\left(1 - \frac{1}{k^2}\right) \cdot 100\%$$

članova populacije.

k	Min % unutar $k\sigma$ od μ	Max % izvan $k\sigma$ od μ
1	0.00	100.00
$\sqrt{2}$	50.00	50.00
1.5	55.56	44.44
2	75.00	25.00
3	88.89	11.11
4	93.75	6.25
5	96.00	4.00
6	97.22	2.78
7	97.96	2.04
8	98.44	1.56
9	98.77	1.23
10	99.00	1.00



Unimodalna distribucija



Bimodalna distribucija

Unimodalna distribucija - samo jedan vrh
Biimodalna distribucija - dva vrha

Vysochanskij-Petuninova nejednakost

Za unimodalne distribucije postoje bolje ocjene od Čebiševljeve nejednakosti.

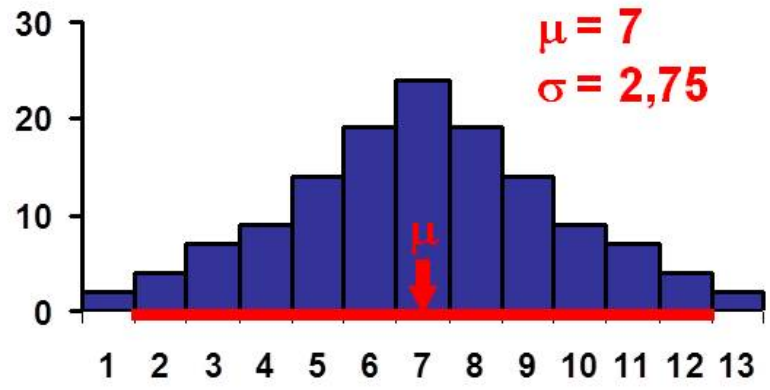
Ukoliko je distribucija varijable unimodalna, tada između $\mu - k\sigma$ i $\mu + k\sigma$ (k > 1) nalazi se najmanje

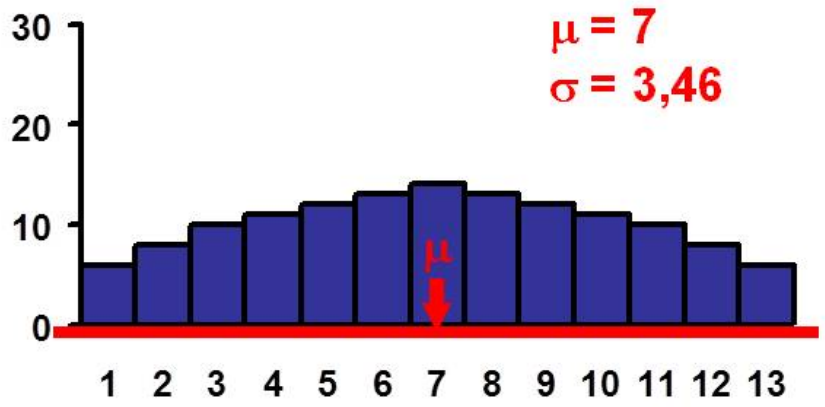
$$\left(1 - \frac{4}{9k^2}\right) \cdot 100\%$$

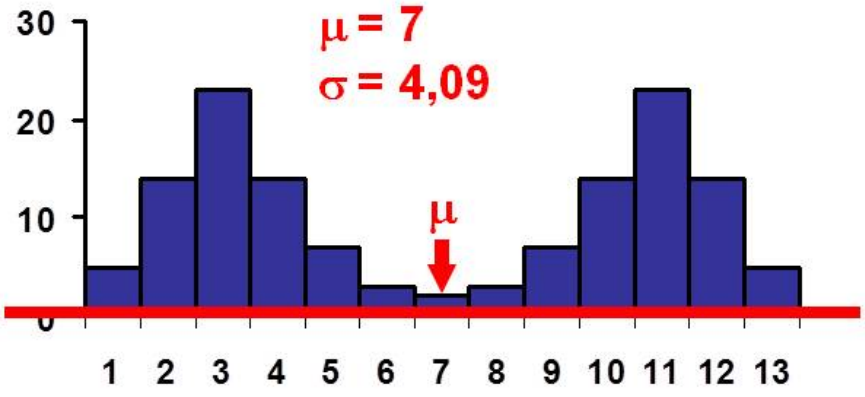
članova populacije.

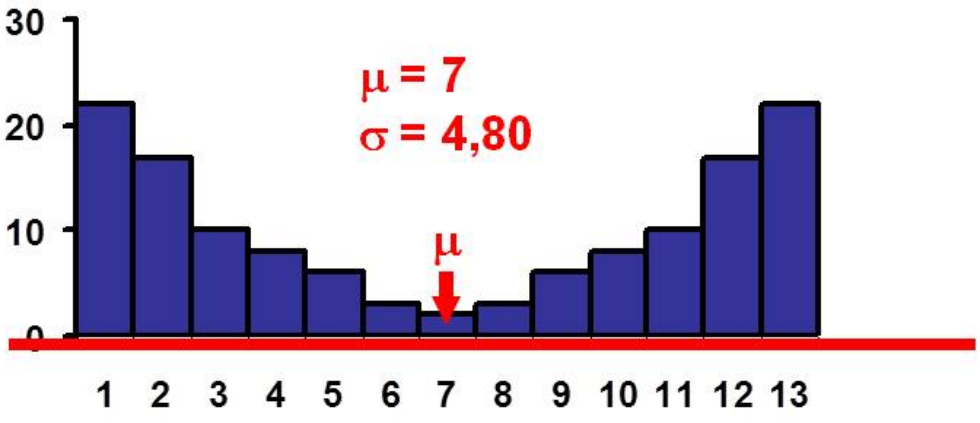
Udio populacije(%) između $\mu - k\sigma$ i $\mu + k\sigma$ (unimodalna razdioba)

k	Čebiševljeva nejednakost	Vysochanskij-Petuninova nejednakost
1	0.00	55.56
$\sqrt{2}$	50.00	77.78
1.5	55.56	80.25
2	75.00	88.89
3	88.89	95.06
4	93.75	97.22
5	96.00	98.22
6	97.22	98.77
7	97.96	99.09
8	98.44	99.31
9	98.77	99.45
10	99.00	99.56


μ = 7
σ = 2,75
μ
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13


μ = 7
σ = 3,46
μ
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13


μ = 7
σ = 4,09
μ
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13


μ = 7
σ = 4,80
μ
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13


μ - 2σ
μ + 2σ

Primjer. Istraživači su na skupini ispitanika primijenili dva kognitivna testa i bilježili vrijeme potrebno za rješavanje testa. Ispitanik A.Z. je u prvom testu postigao vrijeme od 52.22 s a u drugom testu 50.17 s.

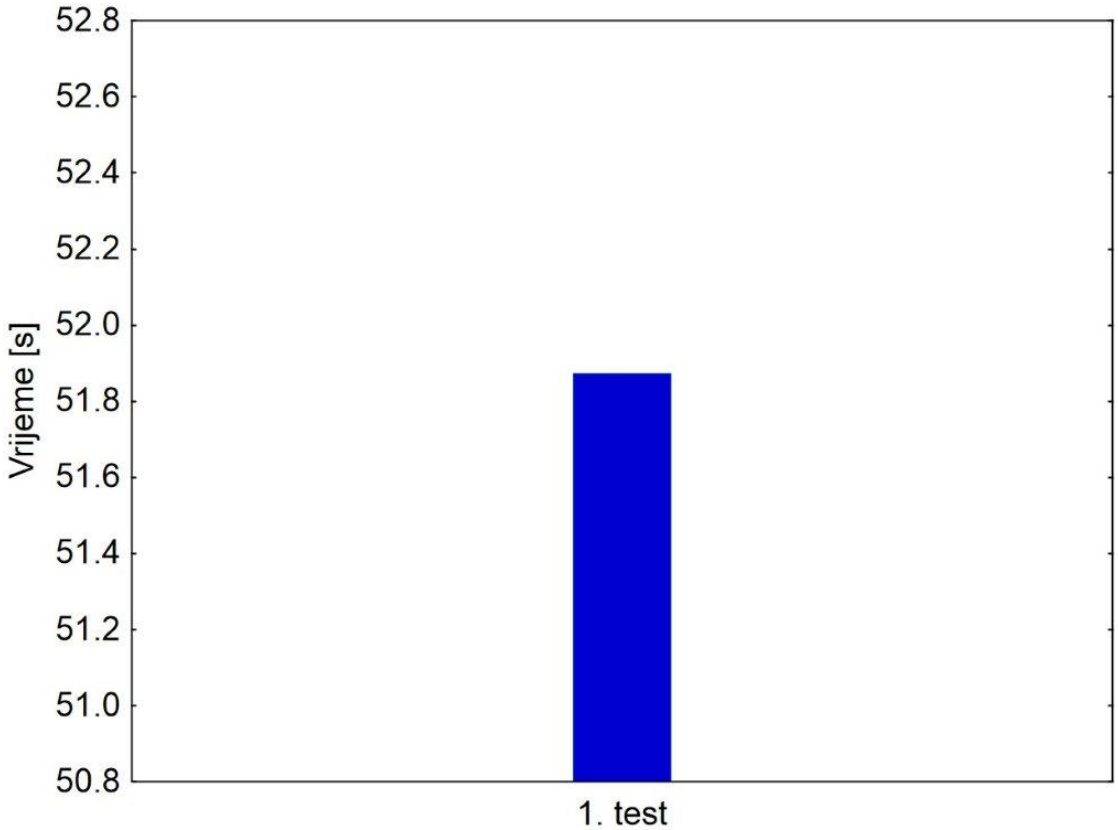
Kakvo je vrijeme postigao u odnosu na druge ispitanike?

	1. test	2. test
Vrijeme	52.22	50.17
Rang	19	1
Relativni rang	63.3 %	3.3 %
z-vrijednost	0.4308	-1.516

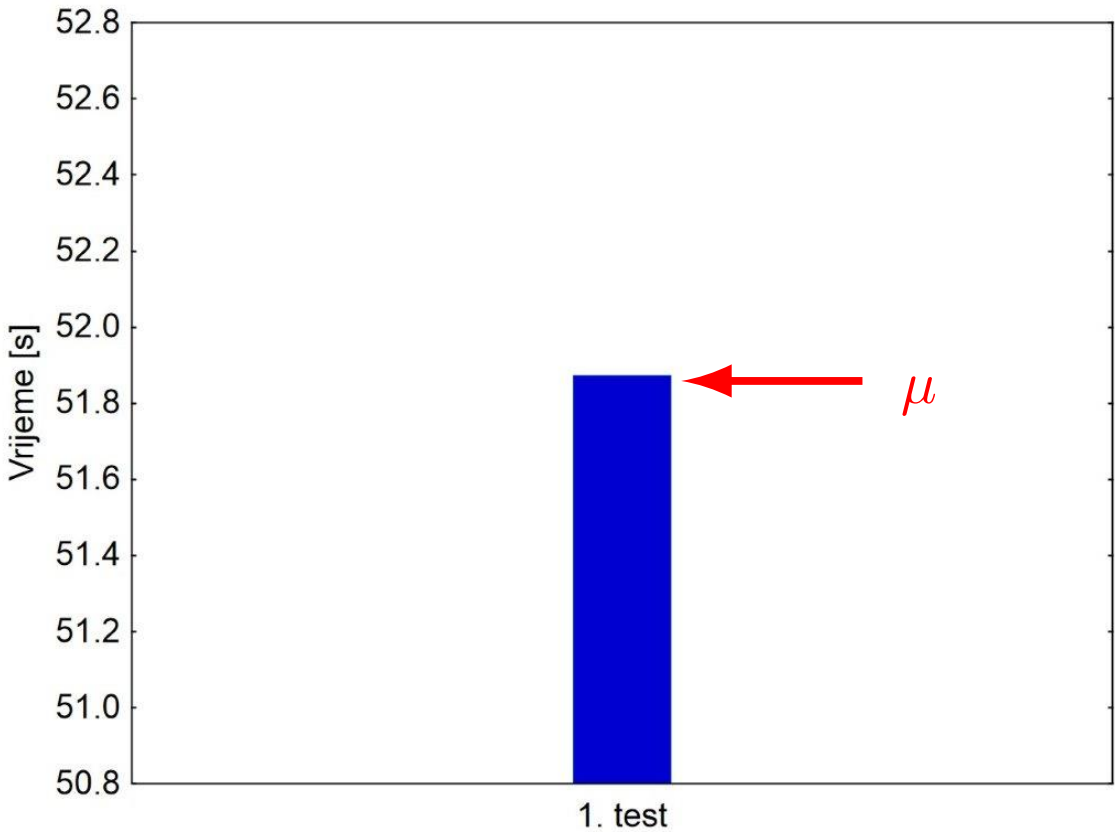
Grafički prikaz mjera centralne tendencije i disperzije

- Stupčasti graf
- Linijski graf
- "Box-whiskers" graf

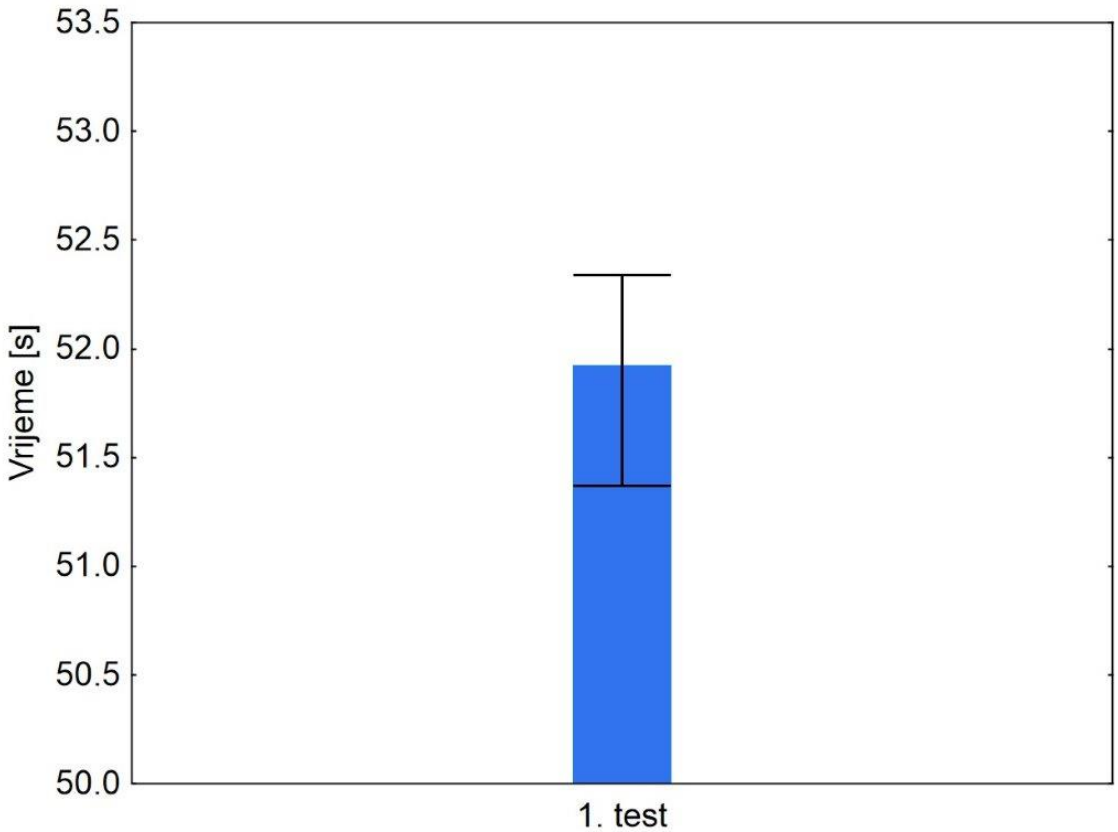
Stupčasti graf (engl. *bars*)



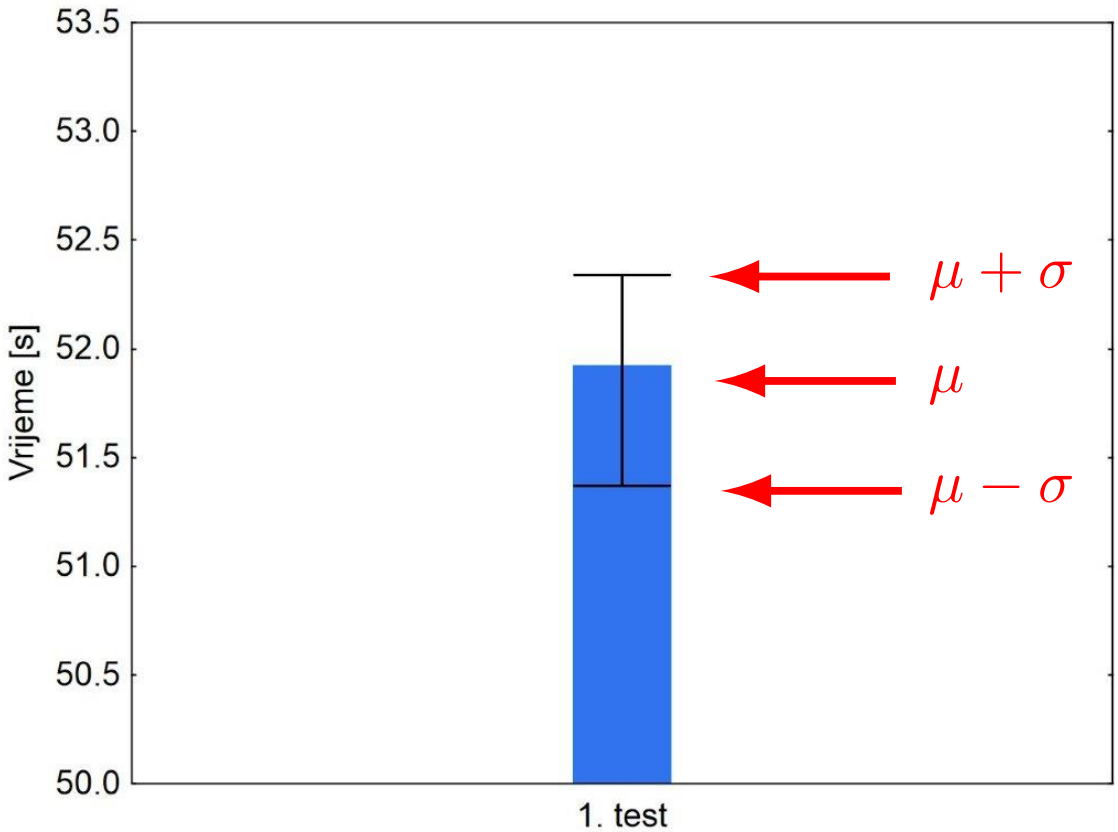
Stupčasti graf (engl. *bars*)



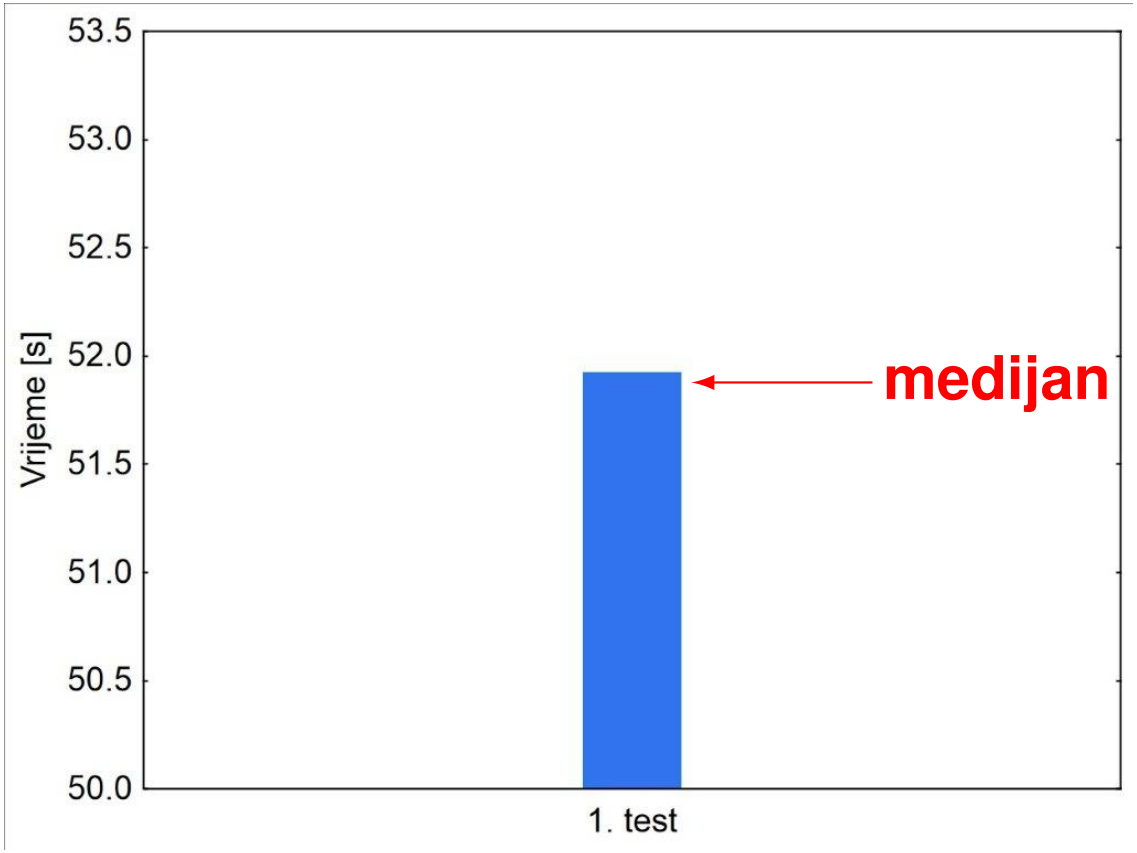
Stupčasti graf (engl. *bars*)



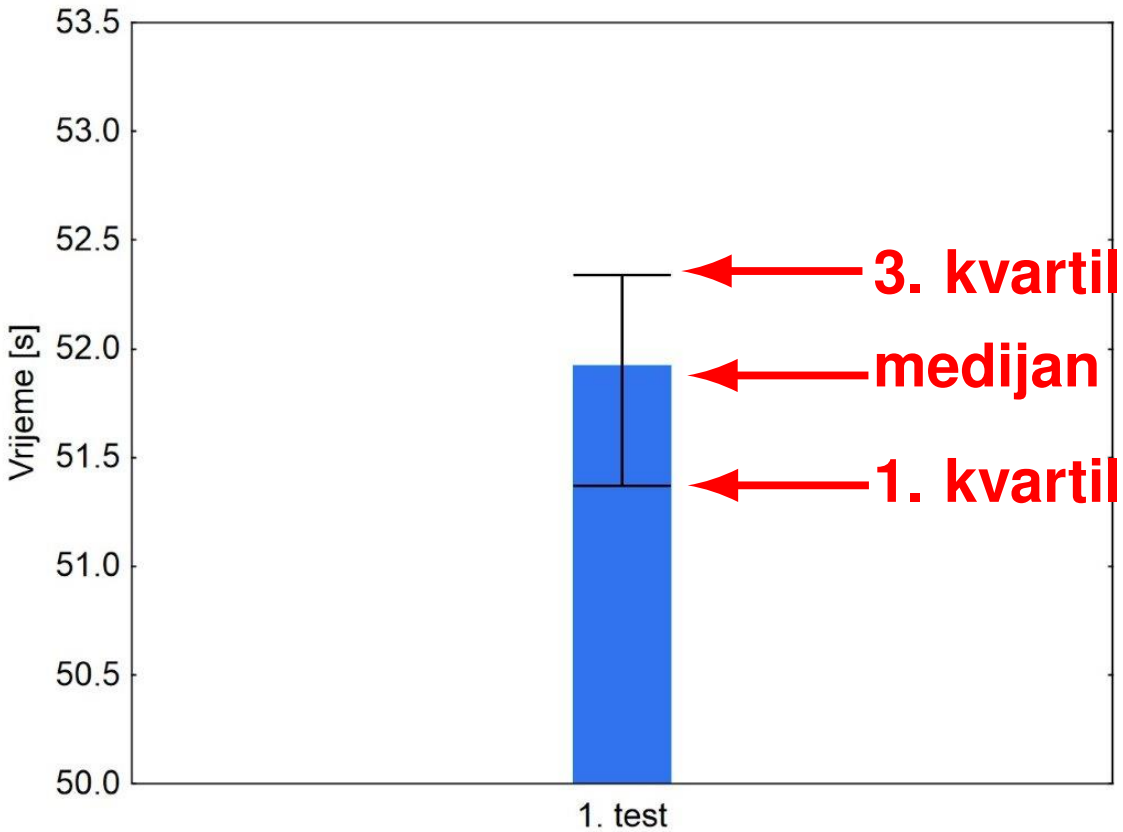
Stupčasti graf (engl. *bars*)



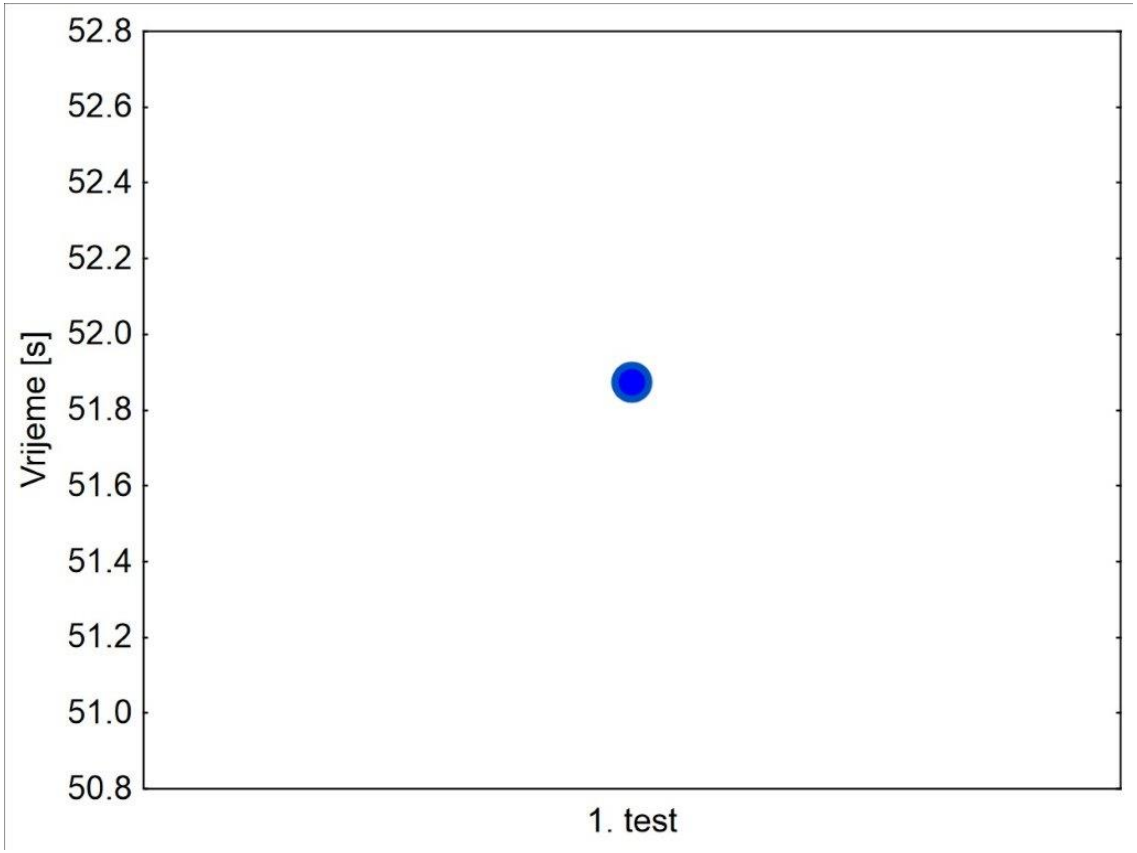
Stupčasti graf (engl. *bars*)



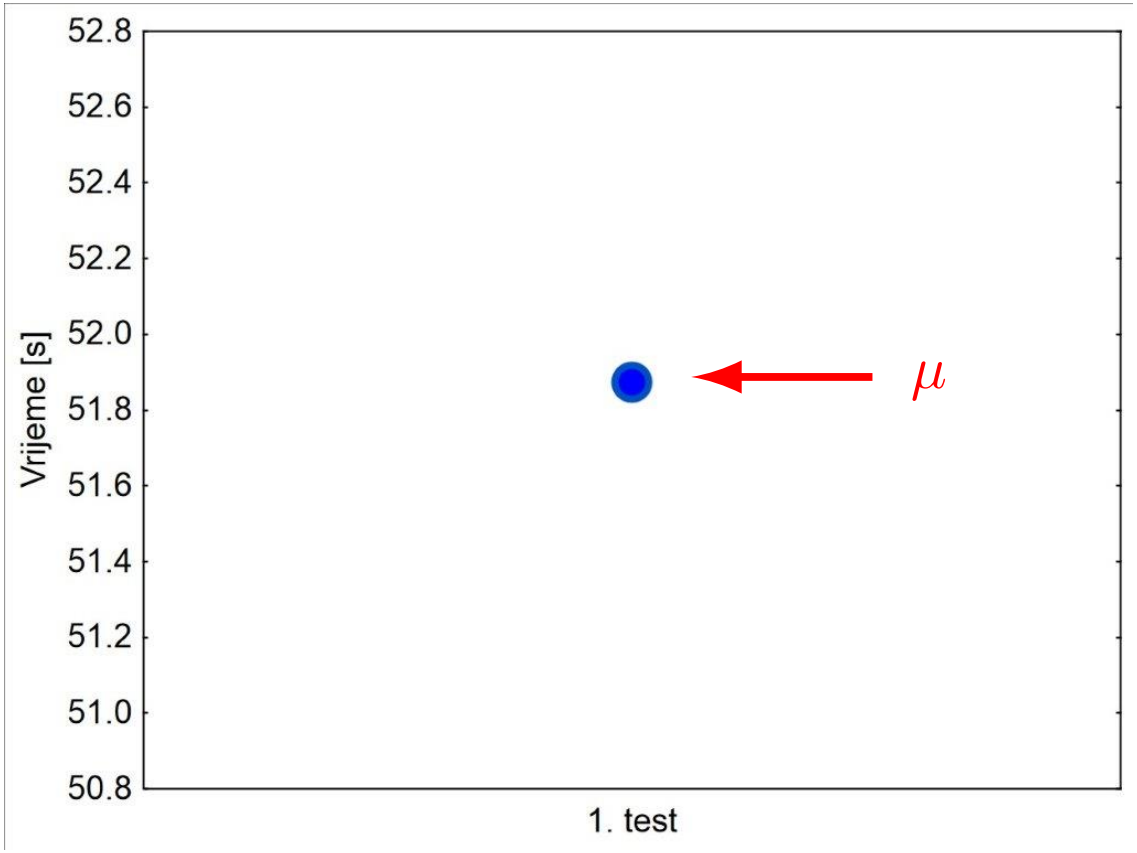
Stupčasti graf (engl. *bars*)



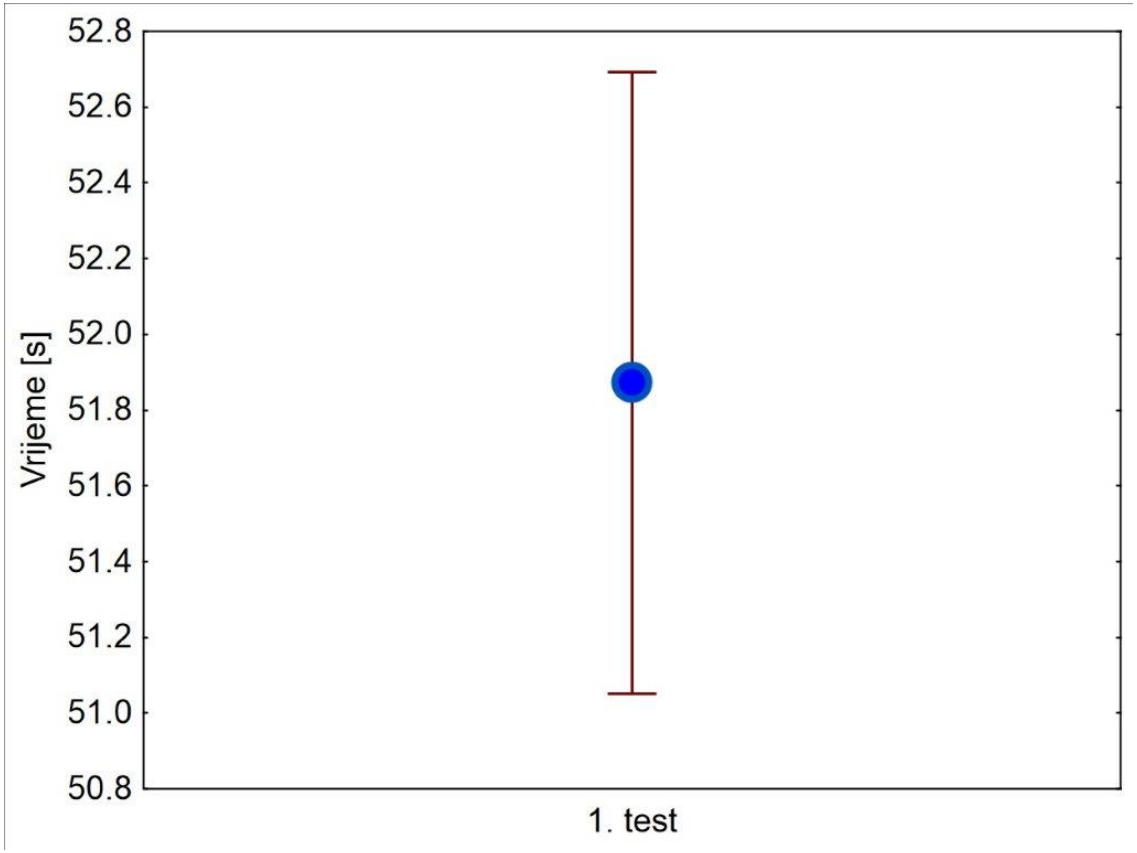
Linijski graf



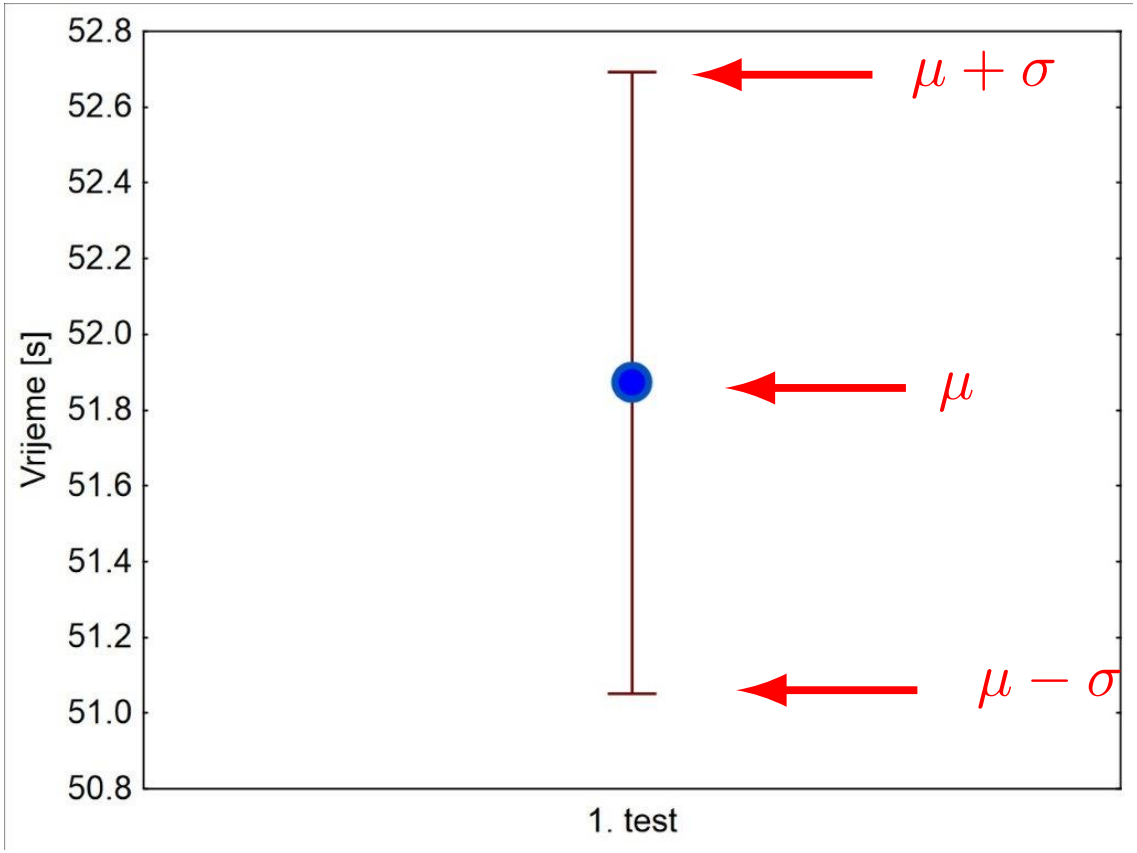
Linijski graf



Linijski graf



Linijski graf

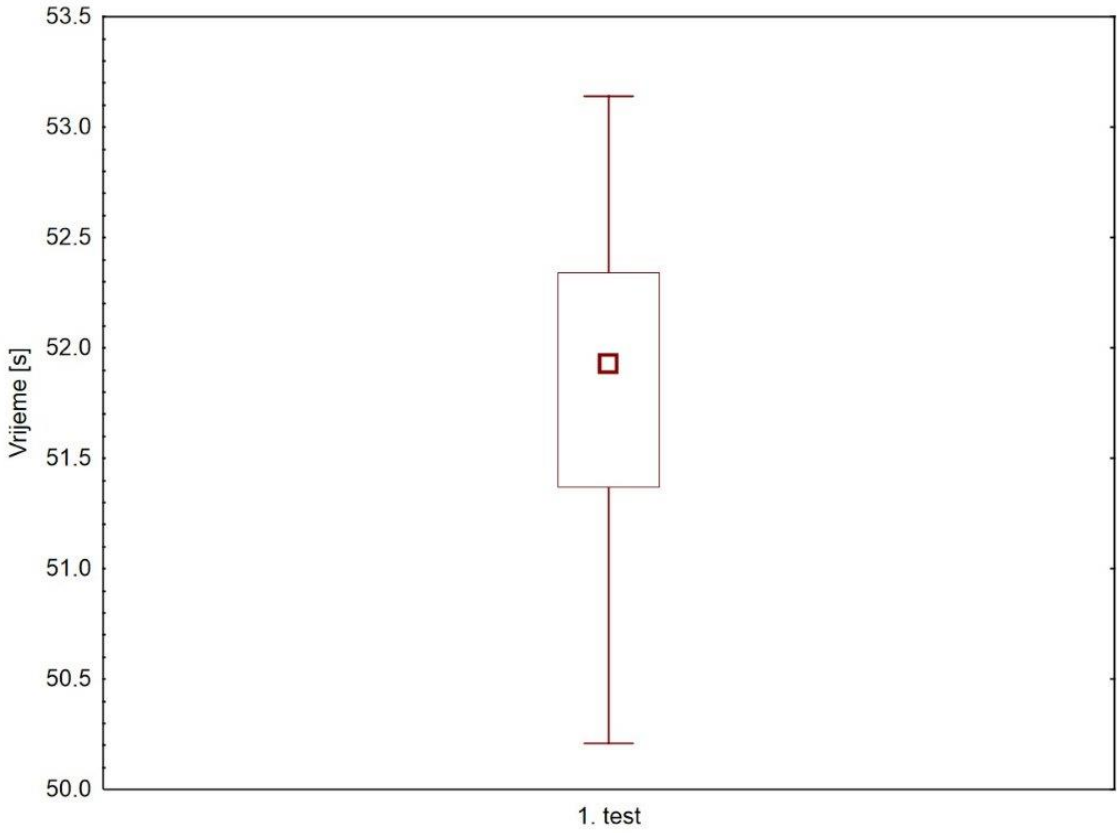


Linijski graf

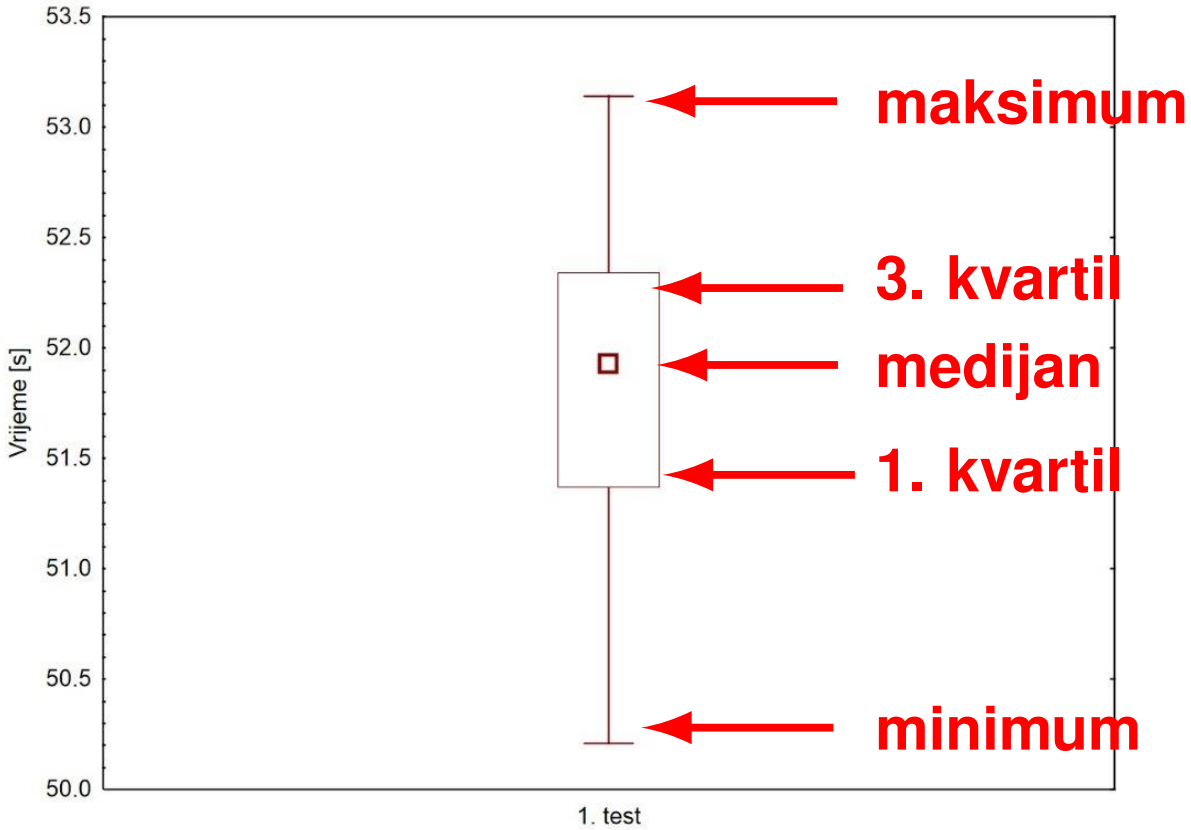
Na isti način se mogu prikazati i

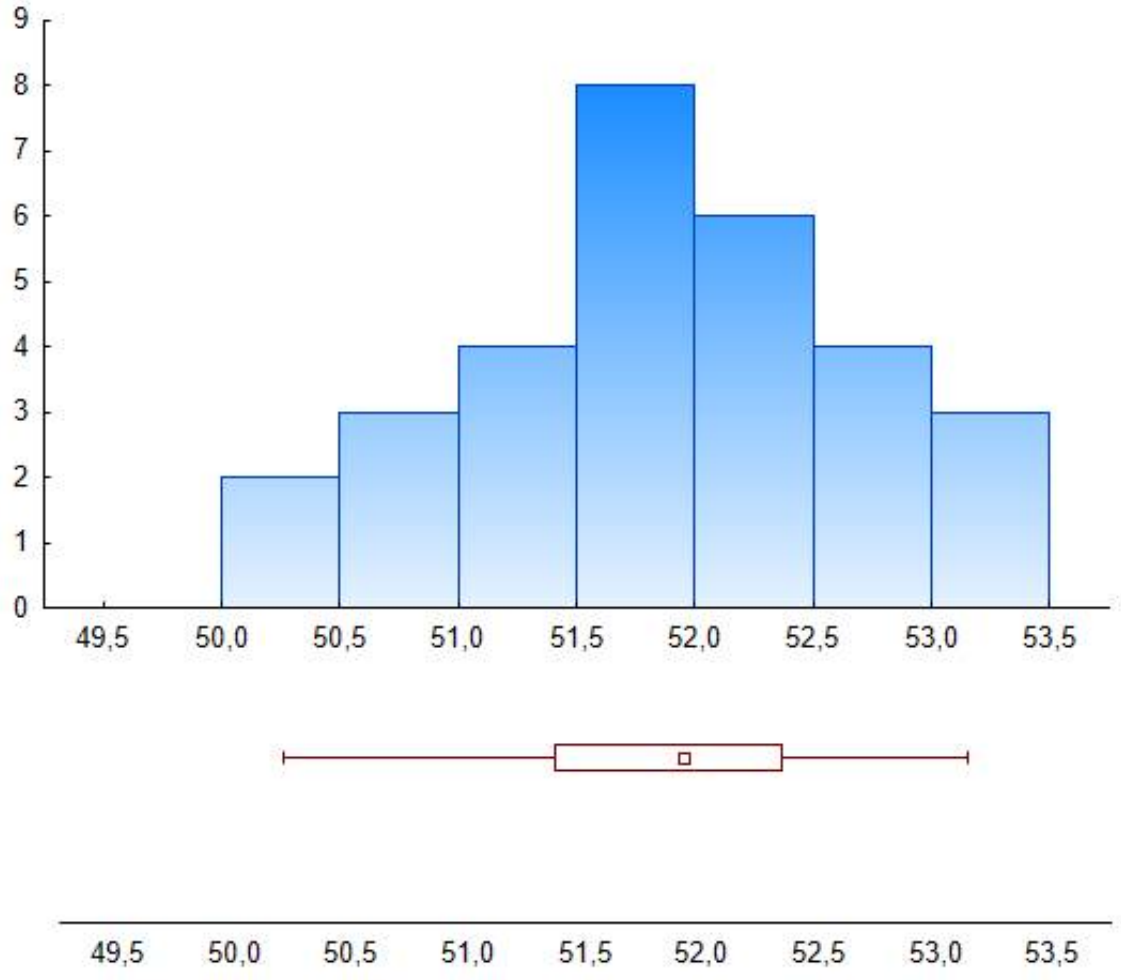
- medijan
- medijan, 1. kvartil, 3. kvartil
- srednja vrijednost, minimum, maksimum
- medijan, minimum, maksimum

"Box-whiskers" graf



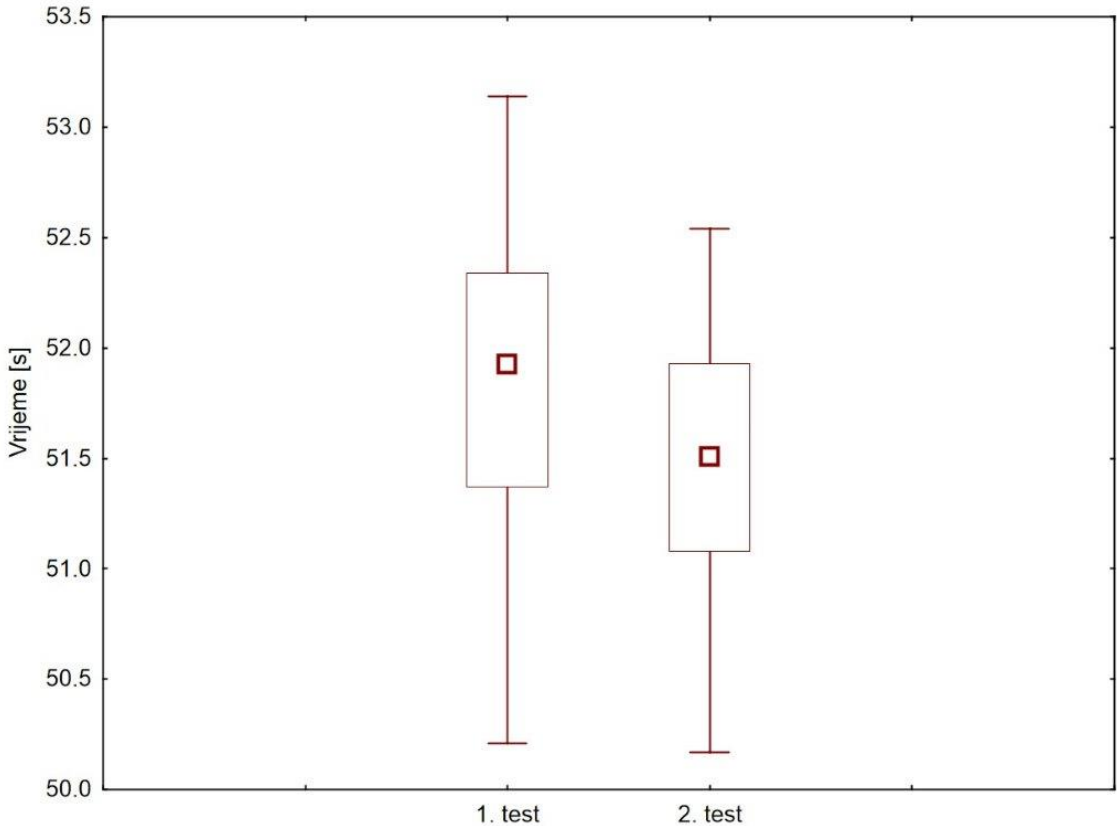
"Box-whiskers" graf




9
8
7
6
5
4
3
2
1
0
49,5
50,0
50,5
51,0
51,5
52,0
52,5
53,0
53,5
49,5
50,0
50,5
51,0
51,5
52,0
52,5
53,0
53,5

Primjer. Rezultati kognitivnog testa.

Slika 1. Rezultati kognitivnog testa (medijan, interkvartilni raspon, minimum, maksimum)



SVAKA ANALIZA ZAPOČINJE UNIVARIJATNOM ANALIZOM!