

Prvi čas vežbi ,

Grafičko predstavljanje podataka

Grafičko i (ili) tablično predstavljanje podataka je prvi korak u statističkoj analizi podataka. Tablica sadrži sve moguće vrednosti posmatranog obeležja i njihove frekvencije (apsolutne, relativne) ili sve registrovane vrednosti obeležja. Veoma često se sve moguće vrednosti obeležja predstavljaju intervalno.

Pod grafičkim predstavljanjem podrazumeva se crtnje histograma, box plot, stablo-lišće dijagrama i sl... Pomenuta tri ćemo detaljnije objasniti primerima.

Neka je $(X_1, X_2, \dots, X_n)$ prost slučajni uzorak za posmatrano obeležje X a $(x_1, x_2, \dots, x_n)$ njegova realizacija. Dalje, neka je $(Y_1, Y_2, \dots, Y_n)$ varijacioni niz polaznog uzorka i $(y_1, y_2, \dots, y_n)$ njegova realizacija. U preliminarnoj analizi uzorka pogodno je odrediti realizovane vrednosti sledećih statistika:

1. UZORAČKA SREDINA $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$, realizovana uzoračka sredina je $\bar{x}_n = \frac{1}{n} \sum_{k=1}^n x_k$;
2. UZORAČKA DISPERZIJA $\hat{S}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - m)^2$, ako je poznato matematičko očekivanje obeležja, odnosno $\tilde{S}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$ ako to nije slučaj. Njena realizovana vrednost se označava sa $\tilde{s}_n^2$.
3. POPRAVLJENA UZORAČKA DISPERZIJA $\tilde{S}_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$
4. UZORAČKA MEDIJANA $M_e = \begin{cases} Y_{k+1}, & \text{ako je } n = 2k + 1; \\ \frac{1}{2}(Y_k + Y_{k+1}), & \text{ya } n = 2k. \end{cases}$ Realizovanu vrednost uzoračke medijane obeležavamo sa m_e .
5. RASPON UZORKA $R = Y_n - Y_1$
6. PRVI (Q_1) i TREĆI (Q_3) uzorački kvantil: ako je $n = 2k$ onda je Q_1 medijana niza $(Y_1, \dots, Y_k)$ a Q_3 je medijana niza $(Y_{k+1}, \dots, Y_{2k})$;
ako je $n = 2k + 1$ onda je Q_1 medijana niza $(Y_1, \dots, Y_k)$ a Q_3 je medijana niza $(Y_{k+2}, \dots, Y_{2k+1})$;
Realizovani prvi i treći kvartil se označavaju q_1 i q_3 .

Ako je $\bar{x}_n < (>) m_e$ zaključujemo da je raspodela obeležja pomerena ulevo (udesno), u suprotnom radi se o simetričnoj raspodeli. S obzirom na to da o raspodeli obeležja zaključujemo na osnovu uzorka prilikom zaključivanja treba voditi računa i o vrednosti standardnog uzoračkog odstupanja ($\sqrt{\tilde{s}^2}$).

Primer 1. Tvrdi se da je prosečna minimalna cena bezolovnog benzina u Americi bila 1.35\$. U reklamne svrhe kompanija želi da pokaže kako je njihova cena niža. Da bi potkrepili svoju tvrdnju, statističari iz firme su sakupili sledeće podatke na osnovu slučajnog uzorka.

1.22 1.37 1.27 1.20 1.42 1.41 1.22 1.24
1.28 1.42 1.48 1.32 1.40 1.26 1.39 1.45
1.44 1.49 1.47 1.47 1.24 1.34 1.27 1.35
1.34 1.45 1.49 1.45 1.23 1.20 1.42 1.34
1.43 1.21 1.49 1.36 1.24 1.20 1.45
1.23 1.25 1.24 1.35 1.23 1.39 1.38
1.46 1.48 1.26 1.36 1.22 1.46 1.39
1.22 1.29 1.47 1.24 1.35 1.21 1.21

STABLO-LIŠĆE dijagram

```
12 | 000111222233344444
12 | 5667789
13 | 2444
13 | 5556678999
14 | 00122234
14 | 5556677788999
```

S jedne strane vertikalne linije se piše deo svakog podatka osim zadnje crte i tako se formira grana, a desno od linije se piše, redom, na odgovarajućim mestima, preostala cifra svakog od podataka i tako se dobija lišće. Ako se dobija previše listova i mala grana, dijagram se formira udvajanjem podataka na levoj strani od povučene linije, tako što se na jednoj grani nalaze podaci čija je preostala cifra manja od 5, a na sledećoj oni čija je preostala cifra veća ili jednaka 5. Upravo to je urađeno u ovom primeru.

Realizovana uzoračka sredina je $\bar{x}_n = 1.34$, a $\bar{s}_n^2 = 0.01$, $\tilde{s}_n^2 = 0.01$. Medijanu uzorka dobijamo tako što prvo sortiramo niz. $m_e = \frac{1}{2}(x_{30} + x_{31}) = 1.35$.

Da li je raspodela pomerena?

Histogrami se dobijaju tako što se na x -osi označe interavali u kojima se nalazi posmatrano obeležje a na y -osi broj elemenata iz uzorka, odnosno relativne frkvencije, i naprave odgovarajući pravougaonici. Za broj intervala(kategorija) se obično uzima $\lceil \log_2 n + 1 \rceil$. Ukoliko je površina pravouganička 1, radi se o histogramu gustine. Minimalan obim uzorka za crtanje uzorka je 16. Preporučuje se da se granice kategorija ne postavljaju tako da se neki elementi uzorka na samoj granici. Prvo se odredi raspon uzorka a zatim se odredi minimalna dužina klase tako da svaki element uzorka pripada nekoj klasi.

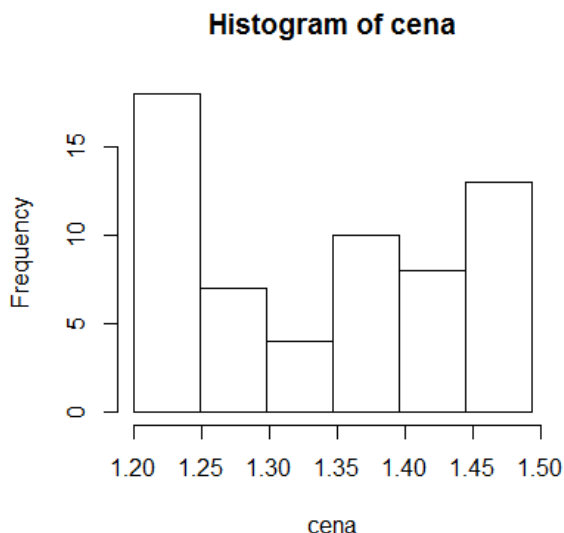
Ukoliko je površina pravouganička 1, radi se o histogramu gustine.

U *Primeru 1* raspon uzorka je 0.29. Obim uzorka je $n = 60$. Broj klasa za crtanje histograma je 6, $0.29/6 = 0.04833333$ odakle zaključujemo da je pogodno uzeti da je veličina svake kategorije 0.049 ili 0.05.

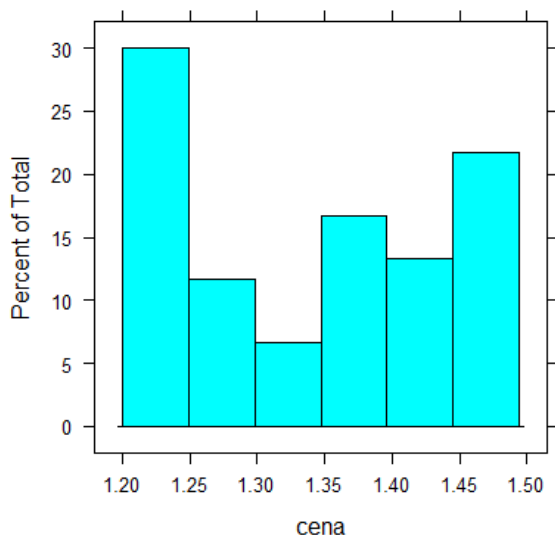
Podaci se mogu predstaviti sledećom tablicom:

vrednosti br.	[1.200, 1.249)	[1.249, 1.298)	[1.298, 1.347)	[1.347, 1.396)	[1.396, 1.445)	[1.445, 1.494)
	18	7	4	10	8	13

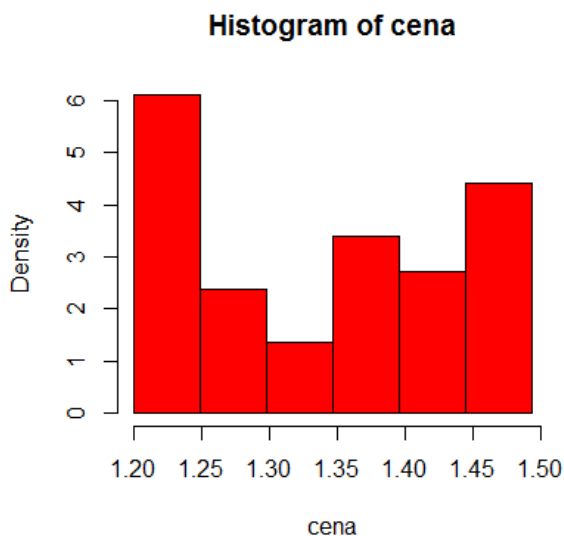
Histogram apsolutnih frekvencija:



Histogram relativnih frekvencija: Visine pravougaonika su n_i/n (može i u procentima) gde je n_i broj elementa u i -toj kategoriji,



Histogram gustine: Ukoliko je veličina i -te kategorije d_i , visina i -tog pravougaonika će biti $h_i = \frac{n_i}{nd_i}$.



Na osnovu histograma gustine možemo pretpostaviti o kojoj se raspodeli obeležja, kao i videti da li je simetrična ili pomerena ulevo ili udesno i sl.

BOX-PLOT dijagram Korišćenjem ovog dijagrama se mogu pronaći autlajeri (najverovatnije pogrešni podaci, podaci koje treba proveriti).

$$q_1 = (y_{15} + y_{16})/2 = 1.24 \quad q_3 = (y_{45} + y_{46})/2 = 1.425 \quad (1)$$

Interkvartilno rastojanje je

$$I_{qr} = q_3 - q_1 = 0.185 \quad (2)$$

Unutrašnje granice uzorka su

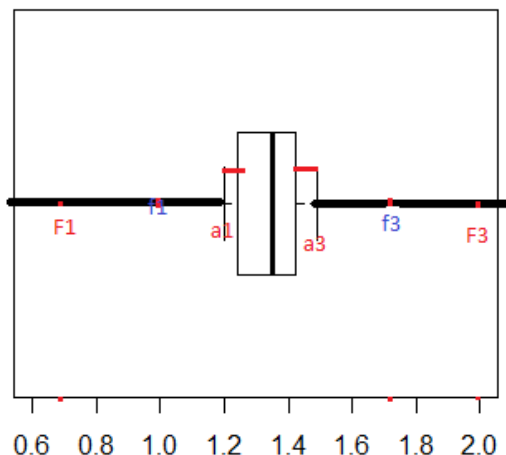
$$f_1 = q_1 - \frac{3}{2}I_{qr} = 0.9625 \quad f_3 = q_3 + \frac{3}{2}I_{qr} = 1.7025 \quad (3)$$

Spolašnje granice uzorka su

$$F_1 = q_1 - 3I_{qr} = 0.685 \quad F_3 = q_3 + 3I_{qr} = 1.98 \quad (4)$$

Najmanji element iz uzorka koji je veći od f_1 označavamo sa a_1 , kao i najveće element iz uzorka koji je manji od f_3 . Njih nazivamo ekstremnim tačkama.

Između F_1 i f_1 kao i između f_3 i F_3 nalaze se blagi autlajeri i na dijagramu obeležavaju sa $\circ$. Oni elementi uzorka koji su manji od F_1 i veći od F_3 su pravi autlajeri.



Nema autlajera.

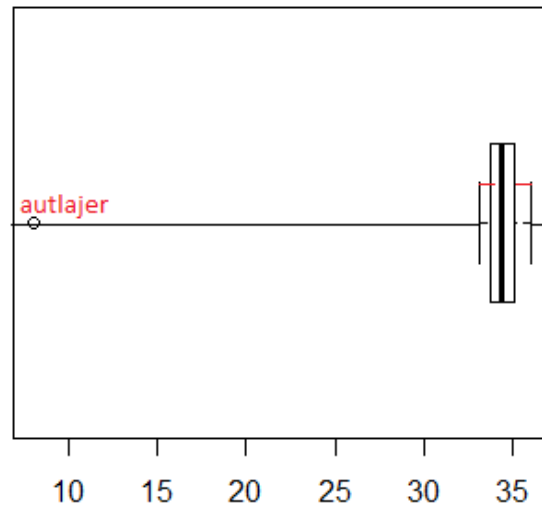
Primer 2. Dati su sledeći podaci

33.10	8.00	34.30	34.60	33.60	33.70	34.00	34.30	34.50	35.10
33.40	33.80	35.60	34.10	34.60	35.80	34.70	33.40	36.10	34.60
34.00	34.70	34.50	33.90	35.20	34.20	35.20	36.00	35.10	33.70

Obim uzorka je $n = 30$. Uzoračka sredina je $\bar{x}_n = 33.50$. Medijana je $m_e = \frac{1}{2}(y_{15} + y_{16}) = 34.4$. Odgovarajući kvartili su $q_1 = y_9 = 33.8$, $q_3 = y_{24}$, interkvartilno rastojanje je $I_{qr} = 1.3$.

Unutrašnje i spoljašnje granice uzorka su $f_1 = 31.85$, $f_3 = 37.05$, $F_1 = 29.9$, $F_3 = 39$. Pravi autlajer je 8. Ovaj podatak je najverovatnije pogrešan (pogrešno izmereno ili zapisano).

Sledi **box-plot** dijagram.



Stablo-lišće dijagram

```
0|8
1|
1|
2|
2|
3|33344444444444
3|5555555555556666
```

S obzirom na raspored listova treba napraviti "udvojen" dijagram.

```
8|0
10|
12|
14|
16|
18|
20|
22|
24|
26|
28|
30|
32|14467789
34|0012335566677112268
36|01
```

Možemo nastaviti sa "udvajanjem" diajgrama. I iz prethodna dva možemo naslutiti da je 8 autlajer. Ukoliko njega izbacimo dobijamo sledeći dijagram.

```
33|144
33|67789
34|001233
34|5566677
35|1122
35|68
36|01
```

Raspodela izgleda simetrično.

Histogarm gustine "novodobijenog" (bez autlajera) uzorka je (raspon uzorka je 3, broj kategorija je 5):

