

R



Vežbe 5



Podela podataka

- Podaci se uopšteno dele na tri osnovna tipa:
 - kategorijski
 - diskretni numerički
 - kontinualni numerički
- Kategorijski: podaci o određenim kategorijama (pol osobe, pušači/nepušači...)
- Diskretni numerički: imaju tačno određen skup vrednosti (broj članova porodice, broj stanovnika...)
- Kontinualni numerički: uzimaju vrednosti iz skupa realnih brojeva – mogu imati beskonačno mnogo vrednosti (visina, težina, dužina...)



Kategorijski podaci

- Ovi podaci se obično posmatraju kroz tabele, ali se mogu predstavljati i grafički

```
#funkcija table od kategorijske promenljive pravi i prikazuje tabelu
x = c("Da", "Ne", "Ne", "Da", "Da", "Ne", "Ne")
table(x)

#iz paketa mtcars izvuci podatke o broju cilindara vozila i predstaviti
ih tabelarno
cilindri = table(mtcars$cyl)
```



Kategorijski podaci

- Ovi podaci se obično razvrstavaju po nivoima ili kategorijama

```
#funkcija factor vraca koje sve kategorije se nalaze u nekoj listi  
kategorijskih podataka  
factor(x)  
factor(cilindri)
```



Bar chart

```
podaci = c(3, 4, 1, 1, 3, 4, 3, 3, 1, 3, 2, 1, 2, 1, 2, 3, 2, 3, 1, 1, 1, 1,
4, 3, 1)

#za svaki podatak crta jedan stub - nepravilno
barplot(podaci)

#za svaku kategoriju crta po jedan stub
barplot(table(podaci))

#crtanje proporcionalnog udela svake kategorije
table(podaci)/length(podaci)
barplot(table(podaci)/length(podaci))
```



Zadatak 1

- Na bar chart-u predstaviti podatke o broju cilindara automobila iz data frame-a mtcars

```
cilindri = table(mtcars$cyl)

barplot(cilindri, names = c("nizak", "srednji", "visok"), col = c("red", "yellow",
"green"), border = "black")
title(xlab = "Raspodela broja cilindara")
```



Pie chart

```
tabelarni_podaci = table(podaci)

#iscrtavanje sa podrazumevanim imenima i bojama
pie(tabelarni_podaci)

#dodavanje imena pri iscrtavanju
names(tabelarni_podaci) = c("k1", "k2", "k3", "k4")
pie(tabelarni_podaci)

#iscrtavanje sa podesavanjem boja
pie(tabelarni_podaci, col = c("red", "green", "cyan", "white"))
title("Podaci")
```

Zadatak 2



- Na pie chart-u predstaviti odnos broja automatika u manuelnih automobila u data frame-u mtcars

```
auto_podaci = table(mtcars$am)
names(auto_podaci) = c("automatik", "manuelni")
pie(auto_podaci, col = c("red", "blue"))
title("Odnos automatika i manuelnih automobila")
```




Numerički podaci

- Disretni i kontinualni numerički podaci
- Kod ovih podataka se obično traži gde je centrirana njihova distribucija, kao i koliko podaci odstupaju od te vrednosti
- Ovde se koriste ugrađene R funkcije: mean, var, sd, median i summary

Varijansa i standardno odstupanje



- Varijansa:

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

- Standardno odstupanje:

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



```
bonusi = c(12, .4, 5, 2, 50, 8, 3, 1, 4, 0.25)

#srednja vrednost
mean(bonusi)

#varijansa
var(bonusi)

#standardna devijacija
sd(bonusi)

#medijana
median(bonusi)

#kratak opis
summary(bonusi)
```



Kvantili i kvartili

- Kvantili pokazuju ispod koje granice se nalazi određen procenat podataka
- Specijalni slučaj: **kvartili**

```
#25% - prvi kvartil (Q1)
#75% - treci kvartil (Q3)
#50% - MEDIJANA
quantile(bonusi, .30)
quantile(bonusi, .354)
```

Zadatak 3



- U fajlu „masine.csv“ dati su podaci o težini proizvoda koje mašine proizvode u gramima. Naći kolika je srednja težina proizvedenog proizvoda, i objasniti koja mašina je bolja u proizvodnji i zašto.

```
masine = read.csv("masine.csv")  
  
mean(masine$masina1)  
mean(masine$masina2)  
  
sd(masine$masina1)  
sd(masine$masina2)
```