Vežbe 7 i 8

Deskriptivna statistika

Zadaci

1. Za uzorak:

87	103	130	160	180	195	132	145	211	105
145	153	152	138	87	99	93	119	129	145

   naći Mo, Me, $\bar{x}_n$, $\bar{s}_n = \sqrt{\bar{s}_n^2}$, kurtosis i skewness. Nacrtati uzoračku funkciju raspodele, histogram sa deobnim tačkama $(80, 100, 130, 170, 220)$ i poligon. Naći kvantile, Q-Q plot i Boxplot.

2. Za uzorak dat tabelom:

I_i	[0,1]	(1,2]	(2,3]	(3,5]	(5,10]	(10,20]
f_i	15	11	7	7	6	4

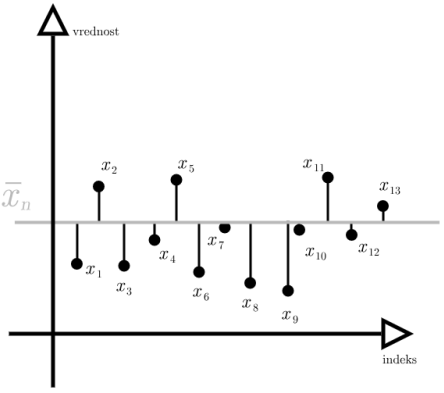
   naći Mo, Me, $\bar{x}_n$, $\bar{s}_n = \sqrt{\bar{s}_n^2}$, kurtosis i skewness. Nacrtati uzoračku funkciju raspodele, histogram sa deobnim tačkama $(80, 100, 130, 170, 220)$ i poligon. Naći kvantile, Q-Q plot i Boxplot.

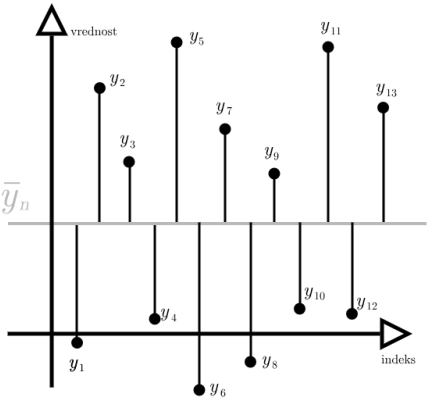
Pregled osnovnih pojmova deskriptivne statistike

Osnovni pojmovi deskriptivne statistike

Kod prostog uzorka $(x_1, x_2, \ldots, x_n)$:

- modus,
  - Vrednost koja ima najveću frekvenciju.
- medijana,
  - Vrednost koja deli uzorak na dva jednaka dela (istog obima).
- aritmetička sredina,
  - $$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
- uzoračka disperzija (nekorigovana!),
  - $$\bar{s}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \bar{x}_n^2.$$





$x \leftarrow$ c(3.5, 5.2, 2.9, 3.5, 5.5, 2.8, 4.2, 2.5, 2.1, 4, 5.1, 4, 4.7)

$\min(x) = 2.1$
$\max(x) = 5.5$
$\text{mean}(x) = 3.8462$

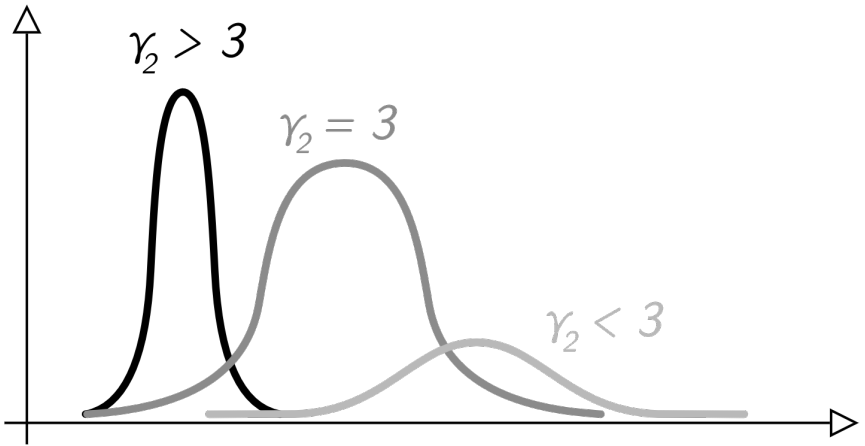
$y \leftarrow$ c(-0.5, 7.6, 5.6, 0.5, 8.1, -2, 7.5, -1, 6, 1, 9.5, 0.7, 7)

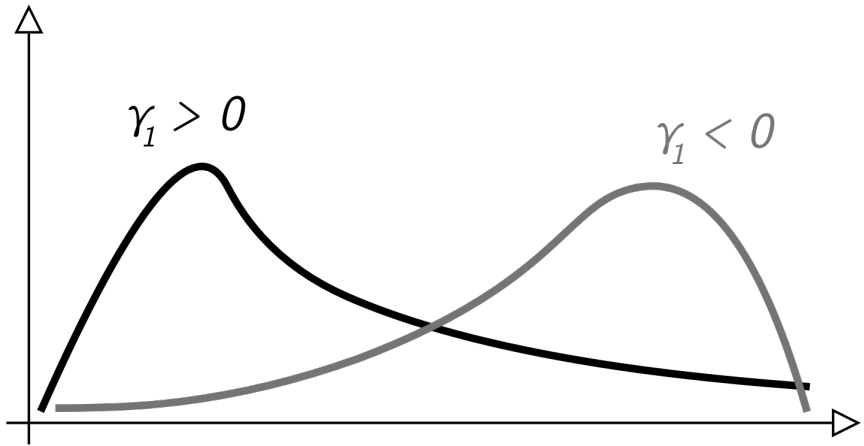
$\min(y) =$ -2
$\max(y) = 9.5$
$\text{mean}(y) = 3.8462$

$$\text{var}(x) = 1.177 \ll 16.743 = \text{var}(y)$$

Osnovni pojmovi deskriptivne statistike - II

- kurtosis (koeficijent spljoštenosti),
  - $$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right)^2},$$
- skewness (Pirsonov koeficijent asimetrije),
  - $$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_n})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x_n})^2\right]^{3/2}},$$
- Q-Q plot,
- Boxplot,
- Histogram,
- Poligon.


$\gamma_2 > 3$
$\gamma_2 = 3$
$\gamma_2 < 3$


$\gamma_1 > 0$
$\gamma_1 < 0$

Zadatak 1

Zadatak 1

1. Za uzorak:

87	103	130	160	180	195	132	145	211	105
145	153	152	138	87	99	93	119	129	145

naći Mo, Me, $\bar{x}_n$, $\bar{s}_n = \sqrt{\bar{s}_n^2}$, kurtosis i skewness. Nacrtati uzoračku funkciju raspodele, histogram sa deobnim tačkama $(80, 100, 130, 170, 220)$ i poligon. Naći kvantile, Q-Q plot i Boxplot.

1 Modus: najčešće se javlja broj 145.

2 Medijana: varijacioni niz je 87, 87, 93, 99, 103, 105, 119, 129, 130, $\boxed{132}$, $\boxed{138}$, 145, 145, 145, 152, 153, 160, 180, 195, 211. (132 i 138 su "u sredini".)
Medijana: $(132 + 138)/2 = 135$.

3 Aritmetička sredina: $\bar{x_n} = \frac{1}{20}(87 + 103 + \cdots + 129 + 145) = \frac{2708}{20} = 135.4$.

4 Uzoračka disperzija: $\bar{s}_n^2 = \frac{1}{20}(87^2 + 103^2 + \cdots + 129^2 + 145^2) - 135.4^2 = 1143.14$.

Zadatak 1 - II

1. Za uzorak:

87	103	130	160	180	195	132	145	211	105
145	153	152	138	87	99	93	119	129	145

naći Mo, Me, $\bar{x}_n$, $\bar{s}_n = \sqrt{\bar{s}_n^2}$, kurtosis i skewness. Nacrtati uzoračku funkciju raspodele, histogram sa deobnim tačkama $(80, 100, 130, 170, 220)$ i poligon. Naći kvantile, Q-Q plot i Boxplot.

5 kurtosis: $\gamma_2 = 2.5941$.
6 skewness: $\gamma_1 = 0.4423$.
7 uzoračka funkcija raspodele,
8 histogram, poligon, boxplot,
9 Q-Q plot.

Zadatak 1 - III

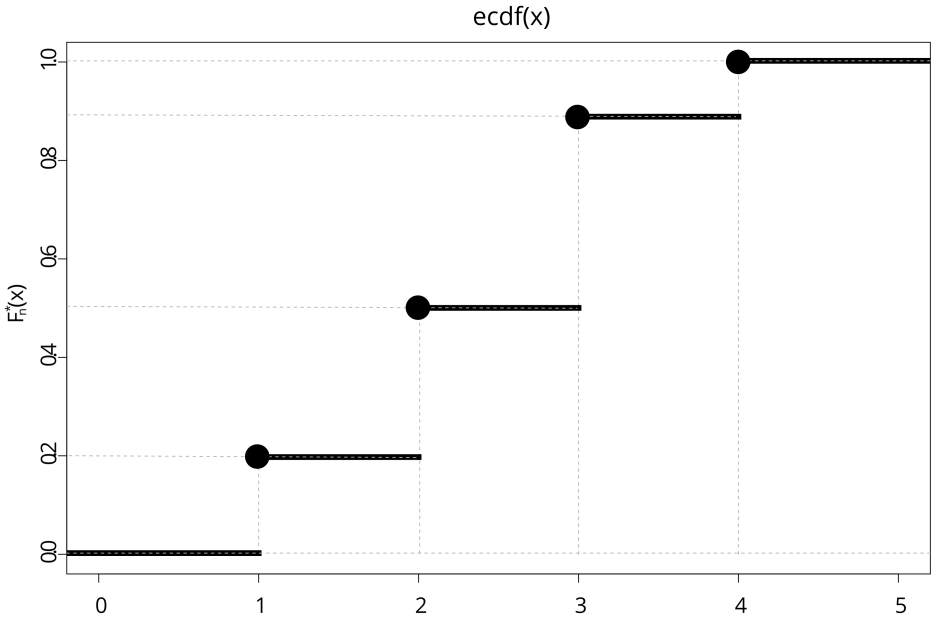
```
x ← c(87, 103, 130, 160, 180, 195, 132, 145, 211,
105, 145, 153, 152, 138, 87, 99, 93, 119, 129, 145)

minimum ← min(x)
maksimum ← max(x)

xn ← mean(x)

n ← length(x)
sn2 ← var(x)*(n-1)/n          #nekorigovana uzoračka disperzija
sn2 ← (sd(x))^2*(n-1)/n       #var(x)=(sd(x))^2
sn ← sqrt(sn2)                #standardna devijacija (nekorig.)
```

Zadatak 1 - IV

```
x ← c(87, 103, 130, 160, 180, 195, 132, 145, 211,
105, 145, 153, 152, 138, 87, 99, 93, 119, 129, 145)

Me ← median(x)                    #probati i median(x, trim=0.2)

tx ← table(x)                                    #frekvencije
Mo ← as.numeric(names(tx)[which.max(tx)]) #jedan modus
Mo ← as.numeric(names(tx)[tx==max(tx)])   #više modusa

library(e1071)
gamma1 ← skewness(x, type=1)                     #type=3 by default
gamma2 ← kurtosis(x, type=1)+3                   #videti domaći
```

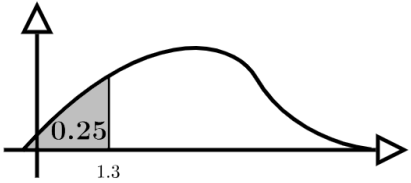
Zadatak 1 - V

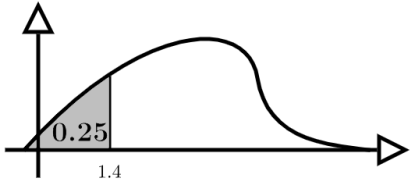
```
x ← c(87, 103, 130, 160, 180, 195, 132, 145, 211,
105, 145, 153, 152, 138, 87, 99, 93, 119, 129, 145)

plot.ecdf(x)          #empirical cumulative relative frequency
                      #videti Zbirku i profesorove folije

qqnorm(x)             #Q-Q plot  (tj. normal probability plot)
                      #videti Zbirku i profesorove folije

qqline(x)             #lakše za porediti
                      #videti domaći
```

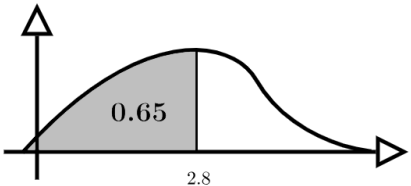

ecdf(x)
$F_n^*(x)$
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5

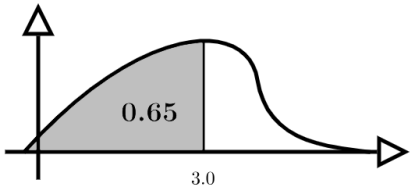
Zadatak 1 - VI

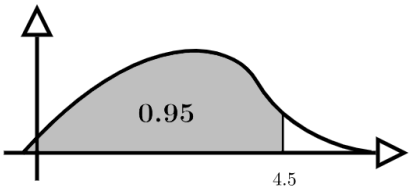
```
summary(x)                  #min, max, medijana, uzoračkI kvartili
IQR(x)                      #interquartile range

quantile(y, probs=seq(0,1,by=0.1))      #kvantili (decili)

boxplot(x)                              #box and whisker plot
boxplot(x, horizontal=TRUE, range=1.5) #videti domaći

cov(y,z)                                #uzoračka kovarijansa
cor(y,z)                                #uzoračka korelacija

hist(x, prob=T)                         #probability scale
lines(density(x))                       #density plot preko h
plot(density(x))                        #density plot (bez h)
```

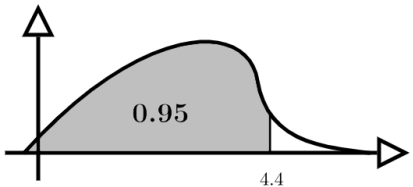

0.25
1.3

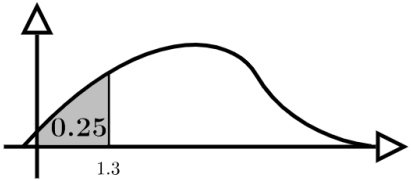

0.25
1.4

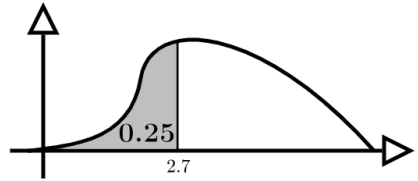

0.65
2.8

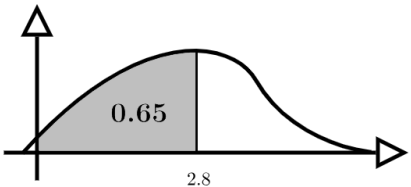

0.65
3.0

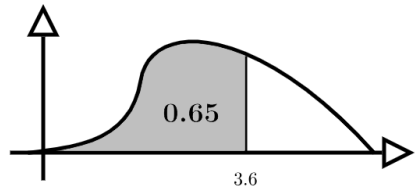

0.95
4.5

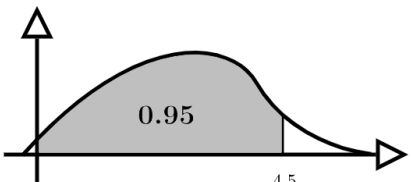

0.95
4.4

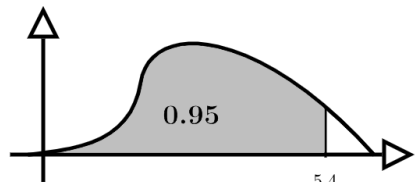

0.25
1.3


0.25
2.7


0.65
2.8


0.65
3.6


0.95
4.5


0.95
5.4

Zadatak 1 - VII

```
set.seed(1111)                              #početna vrednost (PSB)
uzorak ← round(rnorm(100), digits=2)        #zaokruživanje
kvantili ← qchisq(ppoints(100), df=3)

qqplot(kvantili, uzorak)                    #uzorački kvantili
                                            #sa kvantilima χ²₃ rasp.

kvantili ← qnorm(ppoints(100), 5, 5)
qqplot(kvantili, uzorak)                    #uzorački kvantili
                                            #sa kvantilima N(5,5)

abline(a=0, b=1)                            #prava "y=a+bx"
                                            #ovde: y=x
```

Zadatak 1 - VIII

```
set.seed(2020)
x ← rbinom(100, size=100, prob=.5)    #prost uzorak
dbt ← c(35, 50, 57, 65)               #deobne tačke

hist(x)                               #barplot uzorka
hist(x, breaks=dbt)                   #gustina? videti domaći
hist(x, breaks=dbt, freq=T)           #barplot frekvencija

h ← hist(x, breaks=dbt, freq=T)
h$counts                              #frekvencija elemenata

širine ← dbt[2:4]-dbt[1:3]            #širine intervala
h$counts ← h$counts/širine            #menjamo barplot h
plot(h, freq=T)                       #dobijamo histogram
```

Pomoćne funkcije u R-u

```
?distributions                          #implementirane raspodele

dunif(1.9, min=-1, max=2)               # φX(1.9), X:U(-1,2)

pexp(2.7, rate=2)                       # FX(2.7), X:E(2)

qf(0.75, df=1, df=2)                    #kvantil reda 0.75, X:F1,2

set.seed(2020)                          #seed za PSB
rbinom(10, size=100, prob=.5)           #prost uzorak, X:B(100,0.5)

# generisanje uzorka obima 10 iz disk. uniformne raspodele:
sample(c(1,2,3), size=10, p=c(1/3, 1/3, 1/3), replace = T)
```

Zadatak 2

Zadatak 2

2. Za uzorak dat tabelom:

I_i	[0,1]	(1,2]	(2,3]	(3,5]	(5,10]	(10,20]
f_i	15	11	7	7	6	4

naći Mo, Me, $\bar{x}_n$, $\bar{s}_n = \sqrt{\bar{s}_n^2}$, kurtosis i skewness. Nacrtati uzoračku funkciju raspodele, histogram sa deobnim tačkama $(80, 100, 130, 170, 220)$ i poligon. Naći kvantile, Q-Q plot i Boxplot.

```
xi ← c(0.5, 1.5, 2.5, 4, 7.5, 15)    #sredine intervala
fi ← c(15, 11, 7, 7, 6, 4)           #frekvencije
x ← rep(xi,fi)                       #delimična rekonstrukcija
```

Zadatak 2 - II

```
#standardna ideja za uzorke "većeg" obima:
mi ← c(0, 1, 2, 3, 5, 10, 20)
n ← length(mi)

m0 ← mi[1:n-1]                    #leve granice intervala
m1 ← mi[2:n]                      #desne granice intervala

xi ← (m0+m1)/2                    #sredine intervala
fi ← c(15, 11 ,7, 7, 6, 4)        #frekvencije intervala

x ← rep(xi,fi)                    #delimična rekonstrukcija
```

Zadatak 2 -III

- modus:
  - $$Mo = m_{s-1} + h_s \frac{r_1}{r_1 + r_2},$$
  - m_{s-1} leva granica najfrekventnijeg intervala,
  - $r_1 = f_s - f_{s-1}$,
  - $r_2 = f_s - f_{s+1}$.
- medijana:

$$Me = m_{l-1} + h_l \frac{\frac{n}{2} - n_{x_{l-1}}}{f_l},$$

  - m_{l-1} leva granica medijalnog intervala (interval sa najmanjom kumulativnom frekvencijom većom od $\frac{n}{2}$),
  - $n_{x_{l-1}} = \sum_{i=1}^{l-1} f_i$, kumulativna frekvencija intervala I_{l-1}.

Zadatak 2 -IV

I_i	[0,1]	(1,2]	(2,3]	(3,5]	(5,10]	(10,20]
f_i	15	11	7	7	6	4

- konkretno modus:

$$Mo = m_{s-1} + h_s \frac{r_1}{r_1 + r_2} = 0.789,$$

  - $m_{s-1} = 0, \quad h_s = 1, \quad r_1 = 15 - 0, \quad r_2 = 15 - 11.$
- konkretno medijana:

$$Me = m_{l-1} + h_l \frac{\frac{n}{2} - n_{x_{l-1}}}{f_l} = 1.909,$$

  - $m_{l-1} = 1, \quad h_l = 1, \quad \frac{n}{2} = 25, \quad n_{x_{l-1}} = 15 \quad f_l = 11$

Dataframe-ovi

Dataframe-ovi

```r
#loša ideja:
x ← c(1, 2, 4, 5, 1, 9, 9, 1)
y ← c(rep("a",4), rep("b",3), "c")
Dc ← cbind(x,y)                          #javlja se tzv.
Dr ← rbind(x,y)                          #"coercion of data types"

#bolja ideja:
x ← c(1, 2, 4, 5, 1, 9, 9, 1)
y ← c(rep("a",4), rep("b",3), "c")
Df ← data.frame(x,y)
Df ← data.frame(Slova=x, Brojevi=y)
Df ← data.frame("Slova"=x, "Brojevi"=y)
Df ← data.frame("123abc"=x, "225cde"=y)
```

Dataframe-ovi

```
#dobra ideja: read.csv()        (ali pazi na forward slash)
sluč ← read.csv("C:/ABC/random1.csv")       #opcija sep=";"
drugi ← read.csv("C:/random2.csv", stringsAsFactors=FALSE)

#pristup elementima i podelementima dataframe-a:
sluč$Kolona
sluč$Kolona[1:15]
sluč$Kolona[c(T,F)]
sluč$Kolona[rep(c(T,F),2)]
sluč$Kolona[33:length(sluč$Kolona)]
sluč$Kolona[seq(from=2, to=length(sluč$Kolona), by=3)]
```

Filtriranje (subsetting)

```
#fajl random.csv sa kolonama 'Slova', 'Brojevi'
S ← read.csv("C:/ABC/random.csv")

S$Brojevi=500                    #izmena vrednosti
S$Brojevi==500                   #upit o vrednosti
S[S$Brojevi==500, ]              #selekcija vrsta po kriterijumu

S$Slova=='A'
S[S$Slova=='A', ]                #selekcija vrsta po kriterijumu

#alternativno:
subset(S,S$Brojevi==500)
subset(S,S$Slova=='A')
```

Faktori

1. Način na koji R čuva kategorijalna/kvalitativna obeležja (etnička grupa, boja očiju, pol, krvna grupa itd.);
2. Komanda za pravljenje faktora je `factor()`, za konvertovanje u faktor `as.factor()`, dok R sam dodaje atribut `levels`, koji predstavljaju nivoe posmatranog faktora;
3. R aktivno pokušava da konvertuje stringove u faktore!

```
gender ← factor(c("male", "female", "female", "male"))

typeof(gender)
attributes(gender)

as.character(gender)
```

Primer filtriranja po faktorima

```
ž.pomoćni ← c(rep("a", 4), rep("b", 6), rep("c", 2))
ž ← rep(ž.pomoćni, length=100)          #f ima 100 elemenata
ž ← as.factor(ž)                        #nivoi su mu a, b, c

set.seed(1111)                          #seed za generator PSB
y ← round(rnorm(100), digits=2)         #iz 𝒩(0,1) raspodele
z ← rpois(100, lambda=.5)               #iz 𝒫(0.5) raspodele

Df ← data.frame(y,z,ž)                  #tri kolone
Df$y[Df$ž=="a"]
Df[Df$ž=="a", ]                         #navodnici ili apostrofi
Df[Df$ž=='a' & Df$z==0, ]
```

Unos/uvoz podataka i NA

```
x ← scan()                                  #ručno, pojedinačno

x ← c(1, 2, 3, 5, 7, 9)
data.entry(x)                               #izmena pojedinačnih
edit(x)                                     #unosa

x ← c(1, 2, 3, 5, 7, 9)
x ← c(x, NA)                                #ili npr. x[7]=NA
mean(x)                                     #javlja grešku
mean(x, na.rm=T)                            #pravilno
mean(x[!is.na(x)])                          #pravilno
mean(na.omit(data.frame(x)$x))              #izbegavati, jer
                                            #briše cele vrste
```

Domaći rad

(za radoznale i one koji očekuju iole visoku ocenu)

1. Definisati varijacioni niz kod prostog uzorka.
2. a) Googleovati kako se u R programskom jeziku konstruišu funkcije.
   b) Napisati funkciju koja za zadati vektor x vraća kurtosis od x (bez for/while/repeat petlji).
   c) Napisati funkciju koja za zadati vektor x vraća skewness od x (bez for/while/repeat petlji).
3. Koristeći package `e1071`, komanda `kurtosis(x, type='1')+3` vraća traženu vrednost koeficijenta spljoštenosti uzorka x. Zašto se dodaje broj tri? Šta je "excess kurtosis"? (*hint:* pažljivo pogledati definicije)
4. a) Koliko je $\Gamma(1.5)$? Izračunati u R-u. (*hint:* Google is your friend.)
   b) Napisati kako izgleda funkcija gustine Studentove t_1 i t_2 raspodele. (*hint:* folije)
   c) Da li stepen slobode kod Studentove t-raspodele može biti neceli broj? Navesti bar tri broja koja ne mogu biti stepen slobode ove raspodele. Objasniti.
   d) Da li je funkcija gustine obeležja sa $t_{0.1}$ raspodelom leptokurtična, mezokurtična ili platikurtična?

Domaći rad II

(za radoznale i one koji očekuju iole visoku ocenu)

5 Dato je obeležje X čija funkcija gustine ima pozitivan koeficijent asimetrije i modus u $x = 2$. Da li je veća verovatnoća $P(1.5 \leq X \leq 2.5)$ ili $P(4.5 \leq X \leq 5.5)$?

6 Na slučajan način odabrati neki od prvih zadataka iz statistike (140–150) iz profesorove zbirke ("Zbirka rešenih zadataka iz Verovatnoće i statistike") i kompletno ga odraditi u R-u.

7 Napraviti jedan proizvoljan CSV fajl, importovati ga u RStudio i odraditi deskriptivnu statistiku za jednu kolonu (numeričkih vrednosti).

8 Uzorak obima 10 ima empirijsku funkciju raspodele kao na slici na slajdu 30.

a) Eksplicitno napisati kako izgleda taj uzorak.

b) Naći minimum, maksimum, Q_1, Q_2, Q_3, IQR takvog uzorka.

c) Nacrtati histogram tog uzorka sa tri ekvidistante deobne tačke.

9 Da li se funkciji `qqplot` moraju proslediti vektori iste dužine?

10 Proveriti šta radi blok komandi za prost uzorak `x`:

```
y←tabulate(match(x,unique(x)))
unique(x)[y==max(y)]
```

Domaći rad III

(za radoznale i one koji očekuju iole visoku ocenu)

11 Generisati uzorak komandama: `set.seed(2019); x←rpois(100,4);`
   a) Naći prvi kvartil, medijanu, osmi decil i treći percentil uzorka.
   b) Izbrojati koliko ima outlajera u ovom uzorku. Naći i ispitati njihove vrednosti.
   c) Nacrtati histogram (sa četiri ekvidistante deobne tačke) traženog uzorka.
   d) Uz pomoć funkcije `plot( ·, type='l')` nacrtati poligon preko histograma.

12 Šta je tačno "gustina" u `hist(x, breaks=dbt, freq=F)`?

13 Šta radi funkcija dbinom ako diskretne slučajne promenljive nemaju funkciju gustine?

14 Proveriti zašto `rpois(100,-4)` izbacuje poruku upozorenja. Zbog čega nastaje ova greška? Šta su to `NA`?

15 Pronaći komande kojima se određuje vrednost funkcije raspodele i gustine obeležja sa beta raspodelom, gama raspodelom, Košijevom raspodelom, Fišerovom raspodelom, hipergeometrijskom raspodelom, log-normalnom i Vejbulovom raspodelom.

Domaći rad IV

(za radoznale i one koji očekuju iole visoku ocenu)

16 a) Šta je razlika između uzoračkih i teorijskih kvantila?
   b) Kada se koristi `qqnorm`, a kada `qqplot`?
   c) Da li se funkciji `qqnorm` mogu proslediti dva uzorka? Zašto?

17 a) Šta je razlika između barplota i histograma? (*hint:* Zbirka)
   b) Dati primer uzorka čiji se barplot i histogram ne razlikuju.

18 Šta su modus i medijana obeležja sa neprekidnom funkcijom gustine? (*hint:* folije)

19 Realizuje se uzorak od 99 elemenata iz uniformne $\mathcal{U}(0,1)$ raspodele. Ako se tom uzorku element 1000, šta se više menja: aritmetička sredina ili medijana?

20 a) Šta je $P-P$ plot? Koja je razlika u odnosu na $Q-Q$ plot?
   b) Na $Q-Q$ plotu (tačnije na normal probability plot-u) tačke ne leže na pravoj $y=x$. Da li je i dalje moguće da je uzorak dobijen realizacijom obeležja sa normalnom raspodelom? Dati konkretan primer ako jeste.
   c) $P-P$ plot prolazi kroz/sadrži tačku $(0.5, 0.5)$ ako oba obeležja imaju istu medijanu. Dokaži.

Domaći rad V

(za radoznale i one koji očekuju iole visoku ocenu)

21 Googleovati kako se u R-u generišu pie chartovi bez upotrebe nestandardnih package-a. (*hint:* koristiti `table`)

22 Vektor uzorka obima 100 sadrži sve različite vrednosti. Koliko ovaj uzorak ima modusa?

23 Uzorak je generisan komandom `qt(ppoints(100), df=`α`)`, gde je α poznato. Da li je njen kvantil reda 0.005 nužno jednak prvom elementu?

24 Šta tačno radi argument `range` u funkciji `boxplot`? Koja mu je podrazumevana vrednost? Kakve veze ima sa outlier-ima?

25
a) Dokazati da je $\text{cov}(x, \alpha y) = \alpha \,\text{cov}(y, x)$ za svako $\alpha \in \mathbb{R}$.
b) Dokazati da je $\text{cov}(\alpha x, \alpha x) = \alpha^2 \text{cov}(x, x)$ za svako $\alpha \in \mathbb{R}$.
c) Da li uzorci moraju biti jednakog obima da bi bilo moguće izračunati njihovu uzoračku kovarijansu? A uzoračku korelaciju?
d) Dati primer dva (nenula, međusobno različita) uzorka čija je uzoračka kovarijansa jednaka nuli. (*hint:* folije)

(budite
vredni!)