Statistika

SIIT

školska 2020/21

Vežbe 6

Pokretanje R-a

1. Unutar foldera OpenStatistics napraviti svoj folder sa oznakom XXNNYY, gde je XX oznaka smera (SW/IT), NN broj indeksa (7→07) i YY godina upisa (2018→18)

2. Kopirati ikonu R-a u svoj direktorijum

3. Podesiti Properties kopirane ikone: u Start In uneti punu putanju do svog foldera (na primer c:\Users\statistika\Desktop\OpenStatistics\SW0018)

4. Pokrenuti R

5. Isprobati u komandnom promptu

   $x <- \sin(1)$
   $q()$

   Save Workspace Image: Yes
   (biće sačuvano u fajlu .RData, istorija komandi u .Rhistory)

6. Ponovo pokrenuti R

```
objects()
x
```

treba da je x = 0.841471.

7. Izračunati u R-u zadatak 5. a)

```
1-pbinom(21,40,.5)
```

8. Izračunati u R-u zadatak 5. c)

```
1-ppois(21,40*.5)
```

9. Približno rešiti zadatak 2. simulacijom u R-u.

```
n<-10000
s<-numeric(n)
for(k in 1:n){s[k]<-sum(runif(30))<17}
p<-mean(s)
```

Grupisanje podataka

```
x<-c(1,2,4,5,1,9,9,1)
y<-c(rep("a",4),rep("b",3),"c")

#losa ideja:
Dc<-cbind(x,y)          #"spajanje" vektora u matricu po kolonama/vrstama
Dr<-rbind(x,y)          #korekcija tipova podataka

#bolja ideja:
D<-data.frame(x,y)      #data frame moze da sadrzi kolone razlicitih tipova, za razliku od matrica
D<-data.frame(Brojevi=x,Slova=y)

D$Brojevi[7]=2          #izmena vrednosti
D$Brojevi==2            #upit o vrednosti
D[D$Brojevi==2, ]       #selekcija vrsta po kriterijumu

#alternativno:
subset(D,D$Brojevi==2)

D$Brojevi[D$Slova=="a"] = D$Brojevi[D$Slova=="a"] * 2
```

Faktori

```
gender<-factor(c("male", "female", "female", "male"))
typeof(gender)
attributes(gender)
as.character(gender)

f.pomocni<-c(rep("a",4),rep("b",6),rep("c",2))
f<-rep(f.pomocni,length=100)                    #f ima 100 elemenata
f<-as.factor(f)                                 #nivoi faktora: a, b, c

y<-round(rnorm(100),digits=2)
z<-rpois(100,lambda=.5)

D<-data.frame(f,y,z)

D$y[D$f=="a"]
D[D$f=='a', ]                                   #navodnici ili apostrofi
D[D$f=="a" & D$z==0, ]

write.csv(D, file ="frame.csv")  #upisivanje u csv fajl
#write.csv(D, file ="frame.csv",row.names=FALSE)
```

Vežbe 7

Deskriptivna statistika

1. Neka je dat uzorak

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

a) Odrediti modus i medijanu.

```
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)
tx<-table(x)                                  #tabela frekvencija
Mo<-as.numeric(names(tx)[which.max(tx)])      #ispisuje samo prvi modus
Mo<-as.numeric(names(tx)[tx==max(tx)])        #ispisuje sve moduse
```

b) Izračunati aritmetičku sredinu i standardnu devijaciju uzorka.

```
xn<-mean(x)
n<-length(x)
sn<-sqrt(sd(x)^2*(n-1)/n)                     #var(x)=sd(x)^2
```

c) Odrediti koeficijent spljoštenosti i koeficijent asimetrije.

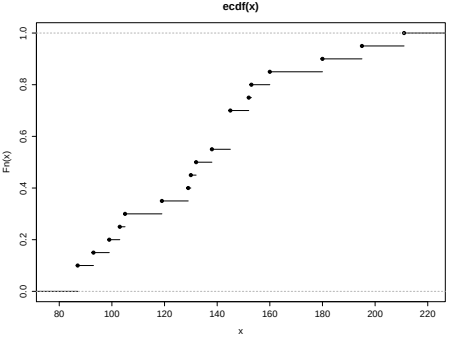
```
mi4<-mean((x-xn)^4)
mi3<-mean((x-xn)^3)
mi2<-mean((x-xn)^2)

Ks<-mi4/mi2^2
Ka<-mi3/mi2^(3/2)

#pomocu funkcija iz paketa e1071
library(e1071)

moment(x, order=5)                  #obican moment reda 5
moment(x, order=5, center=TRUE)     #centralni  moment reda 5

kurtosis(x,type=1)+3                # koeficijent   spljostenosti
skewness(x,type=1)                  # koeficijent   asimetrije
```

d) Nacrtati uzoračku funkciju raspodele.

Uzoračka (empirijska) funkcija raspodele $F_n^* = \frac{N_x}{n}$

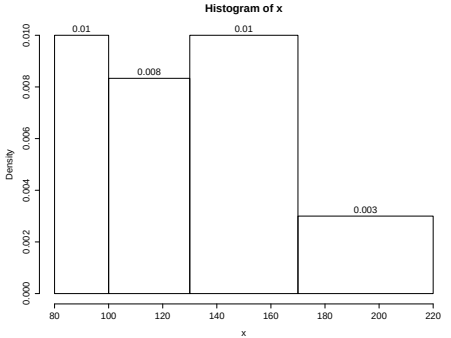
- N_x broj elemenata uzorka koji su $\leq x$
- n obim realizovanog uzorka

plot.ecdf(x)	*#empirical cumulative distribution function*
table(x)/n	*#verovatnoce da se x nalazi u intervalu*
cumsum(**table**(x))/n	*#vrednosti empirijske funkcije raspodele*



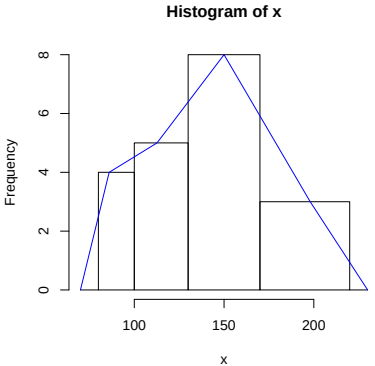
e) Nacrtati histogram i poligon.

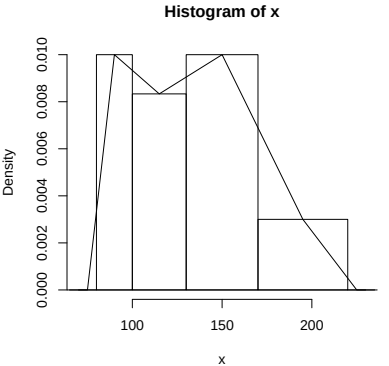
```
hist(x)
mi<-c(80,100,130,170,220)
hist(x,breaks=mi)
hist(x,breaks=mi,probability=TRUE,labels=TRUE)
#poligon rucno docrtamo
```



Za radoznale

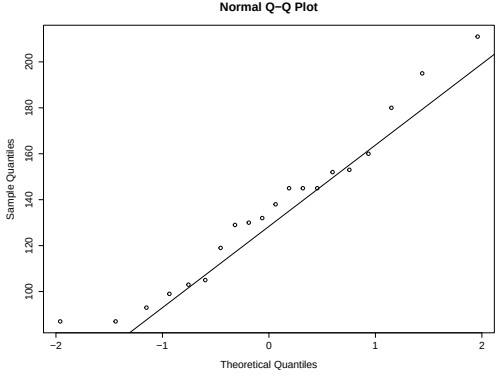
```
h1<-hist(x,breaks=c(70,mi,230),probability=FALSE)
par(new=TRUE)
plot(h1$mids, h1$counts, type="l", col="blue", axes=FALSE, xlab="", ylab="")

library(agricolae)
h2<-hist(x,breaks=c(70,mi,230),probability=TRUE)
polygon.freq(h2,frequency = 3)                    #frequency=1 za obican histogram
```

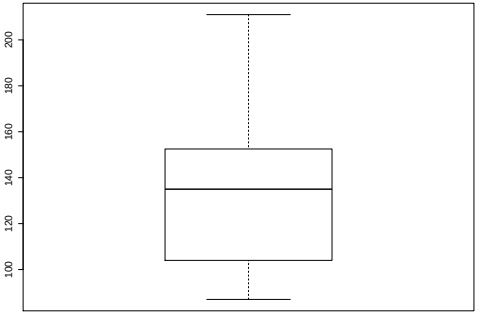




f) Naći kvartile i nacrtati *Q-Q plot* i *Box plot*.

```
summary(x)
IQR(x)
#kvantil proizvoljnog reda p za uzorak x mozemo dobiti naredbom quantile(x,p)
qx<-qnorm((1:n)/n)          #peske crtanje QQ plota
plot(qx,sort(x))
qqnorm(x)                   #pomocu ugradjenih funkcija
qqline(x)
boxplot(x)
```





2. Neka je dat intervalni uzorak

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

a) Nacrtati uzoračku funkciju raspodele, histogram i poligon.

I_i	[0,1]	(1,2]	(2,3]	(3,5]	(5,10]	(10,20]
f_i	15	11	7	7	6	4
x_i	0.5	1.5	2.5	4	7.5	15
n_{x_i}	15	26	33	40	46	50
f_n^*	0.3	0.52	0.66	0.8	0.92	1

```
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)

f<-cumsum(fi)/sum(fi)                #vrednosti empirijske funkcije raspodele
plot.ecdf(x)
hist(x,breaks=mi,probability=TRUE,labels=TRUE)
```

Ideja za uzorke većeg obima

```
mi<-c(0,1,2,3,5,10,20)          #deobne tacke
n<-length(mi)

ml<-mi[1:n-1]                   #leve granice intervala
md<-mi[2:n]                     #desne granice intervala

xi<-(ml+md)/2                   #sredine intervala
fi<-c(15,11,7,7,6,4)
x<-rep(xi,fi)
```

b) Odrediti modus i medijanu.

Kod intervalnog uzorka postoje posebne formule za računanje modusa i medijane:

$$M_o = m_{s-1} + h_s \frac{r_1}{r_1 + r_2} = 0 + 1 \cdot \frac{15}{15 + 4} = 0.7895$$

- $I_s = (m_{s-1}, m_s)$ modalni interval
- h_s širina modalnog intervala
- $r_1 = f_s - f_{s-1}$
- $r_2 = f_s - f_{s+1}$

$$M_e = m_{l-1} + h_l \frac{\frac{n}{2} - n_{x_{l-1}}}{f_l} = 1 + 1 \cdot \frac{25 - 15}{11} = 0.90909$$

- $I_l = (m_{l-1}, m_l)$ medijalni interval
- h_l širina medijalnog intervala
- $n_{x_{l-1}}$ kumulativna frekvencija intervala koji prethodi medijalnom

c) Izračunati aritmetičku sredinu i standardnu devijaciju uzorka.

d) Odrediti koeficijent spljoštenosti i koeficijent asimetrije.

e) Naći kvartile i nacrtati *Q-Q plot* i *Box plot*.

```
xn<-mean(x)
n<-length(x)
sn<-sqrt(sd(x)^2*(n-1)/n)
mi4<-mean((x-xn)^4)
mi3<-mean((x-xn)^3)
mi2<-mean((x-xn)^2)
Ks<-mi4/mi2^2
Ka<-mi3/mi2^(3/2)
summary(x)
IQR(x)
qqnorm(x)
qqline(x)
boxplot(x)
```

Domaći rad

Na slučajan način odabrati neke od prvih zadataka iz profesorove zbirke ("Zbirka rešenih zadataka iz verovatnoće i statistike") i kompletno ih rešiti peške* i u R-u. (Poželjno je uraditi jedan zadatak sa običnim uzorkom, jedan sa intervalnim uzorkom i zadatak 144.)

Manipulacija podacima

Pokretanje skript fajla iz R-a:

```
setwd("c:\Users\statistika\Desktop\OpenStatistics\IT0016")
source("zadatak1.R")
```

Čuvanje promenljivih

```
load(".RData")                    #ucitavanje promenljivih iz R formata
save.image("novo.RData")          #cuvanje Work space u fajl novo.RData
save(xn,sn, file ="xnsn.rda")     #cuvanje promenljivih xn i sn u fajl xnsn.rda
```

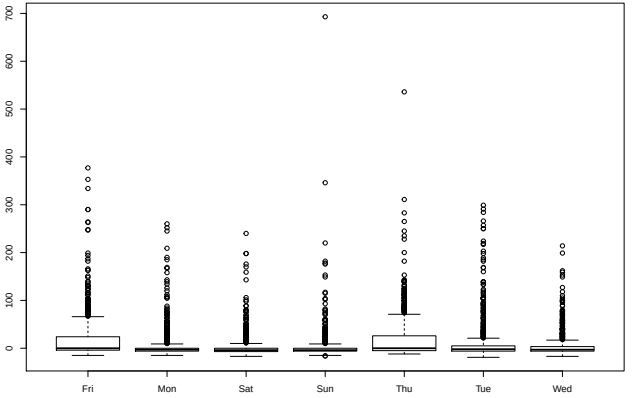
*Nacrtati histogram, boxplot i empirijsku funkciju raspodele bez upotrebe računara

Uvoz podataka u R

```
FlightDelays<-read.csv("FlightDelays.csv")   #ucitavanje csv tabele
summary(FlightDelays)                        #osnovna deskriptivna statistika
FlightDelays<-edit(FlightDelays)             #otvaranje data editora
 str (FlightDelays)                          #vraca strukturu podataka
names(FlightDelays)                          #vraca nazive promenljivih iz tabele
head(FlightDelays)                           #prvih 6 redova
head(FlightDelays,n=10)                      #prvih 10 redova
head(FlightDelays,n=-10)                     #svi redovi osim poslednjih 10
 tail (FlightDelays)                         #poslednjih 6 redova
FlightDelays[1:10,1:3]                       #prvih 10 redova, prve 3 kolone
FlightDelays[1:10,c(1,3,7)]                  #prvih 10 redova, kolone 1, 3 i 7
```

3. Izdvojiti iz kolone Delay vrednosti koje se odnose na ponedeljak i nacrtati *Box plot* za Delay po danima.

```
delay<-FlightDelays$Delay                  #sa $ se pristupa podelementu
indeks<-FlightDelays$Day=="Mon"            #u vektor indeks upisuje TRUE ako je ponedeljak
delaymon<-delay[indeks]                    #izdvaja vrednosti za koje je indeks TRUE
boxplot(Delay~Day,data=FlightDelays)
```



Domaći rad

U posebnom script fajlu napisati funkciju koja za prosleđen uzorak ispisuje deskriptivnu statistiku.

Na primer:

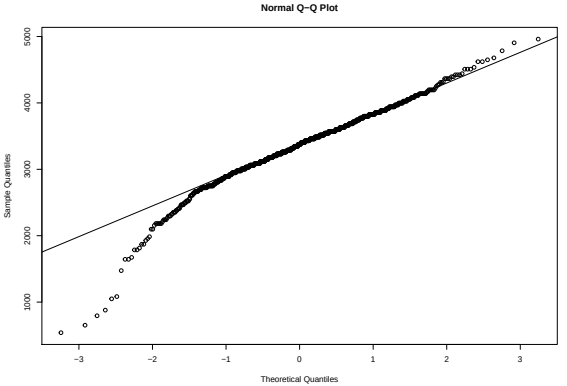
deskstat.R

```
deskstat<-function(x){
  n<-length(x)
  print(paste0("Obim␣uzorka:", n))

   ...              #deskriptivna  statistika  kao u prvom zadatku

  boxplot(x)
}
```

Pronaći kako da se funkciji doda opcioni vektor sa zadatim deobnim tačkama za crtanje histograma i implementirati to u napisanu funkciju.

4. Učitati iz fajla TXBirths2004.csv kolonu Weight samo za dečake i napraviti deskriptivnu statistiku.

```
podaci<-read.csv("TXBirths2004.csv")
indeks<-podaci$Gender=="Male"
weightMale<-podaci$Weight[indeks]

source("deskstat.R")
deskstat(weightMale)
```



5. Napraviti deskriptivnu statistiku za kolonu AveKW iz fajla Turbine.csv za mesece april i jun.

```
podaci<-read.csv("Turbine.csv")
indeks<-substr(podaci$Date2010,1,3)=="Apr" | substr(podaci$Date2010,1,3)=="Jun"
podaciAprJun<-podaci$AveKW[indeks]

source("deskstat.R")
deskstat(podaciAprJun)
```

Domaći rad

Pronaći kako se u R-u zapisuju logički operatori $\neq$, $\wedge$, $\neg$.

Domaći rad

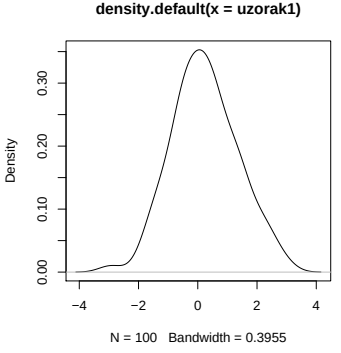
1. Pronaći kako sve možemo upotrebiti funkciju **rep**
2. Pronaći šta radi funkcija **seq**.
3. Šta radi naredba **par(new=TRUE)** sa slajda 7? Pronaći kako da se pomoću funkcije **par** na ekranu prikaže više grafika istovremeno.
4. Šta vraćaju naredbe **h$mids** i **h$counts** za prosleđen histogram h?
5. Proveriti šta radi blok komandi (i uporediti rezultat sa modusom/modusima):

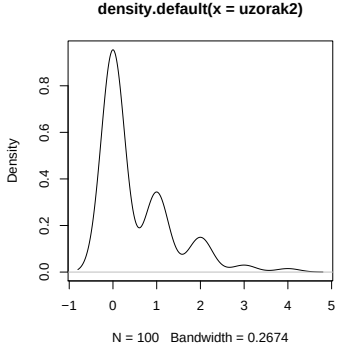
```
y<-tabulate(match(x,unique(x)));
unique(x)[y==max(y)]
```

6. Pronaći razliku između vrednosti **NA**, **NaN**, **Null** i **Inf**. Šta vraća funkcija **is.na**?

6. Posmatraju se dva uzorka obima 100. Prvi uzorak čine brojevi sa $\mathcal{N}(0,1)$ raspodelom (zaokruženi na 2 decimale), a u drugom uzorku su brojevi sa $\mathcal{P}(0.5)$ raspodelom. (Postaviti da "seme" za oba uzorka bude 1111.)

```
set.seed(1111)
uzorak1<-round(rnorm(100),digits=2)
uzorak2<-rpois(100,lambda=.5)

plot(density(uzorak1))                    #grafik funkcije gustine
plot(density(uzorak2))
```

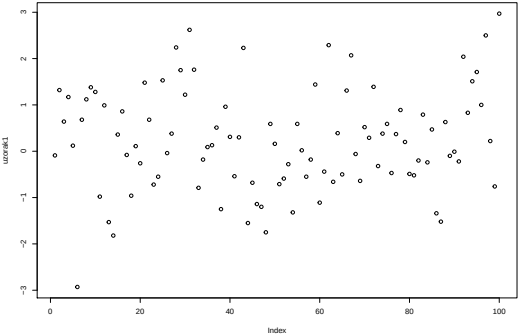


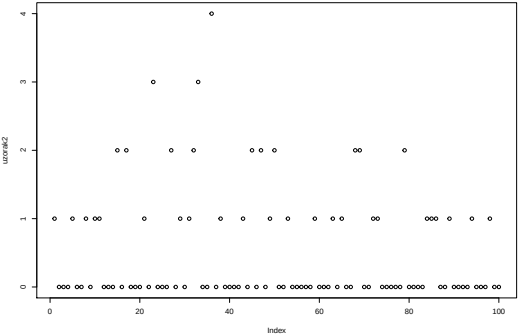


```
cor(uzorak1,uzorak2)                          #korelacija
cov(uzorak1,uzorak2)                          #kovarijansa

mean(uzorak1,trim=0.2)                        #odseca po 20% vrednosti sa obe strane pre izracunavanja

table(uzorak2)
prop.table(table(uzorak2))                    #relativne frekvencije

quantile(uzorak1, probs=seq(0,1,0.05))        #kvantili za uzorak1

scale(uzorak1)                                #z-score za uzorak1
(uzorak1-mean(uzorak1))/sd(uzorak1)           #peske

kvantili <- qchisq(ppoints(100),df=3)         #kvantili Pirsonove raspodele
qqplot(kvantili,uzorak1)
abline(a=0,b=1)                               #prava y=x

kvantili <- qnorm(ppoints(100),5,5)           #kvantili N(5,5) raspodele
qqplot(kvantili,uzorak1)
abline(a=0,b=1)
```

Dijagram rasipanja (*scatterplot*)

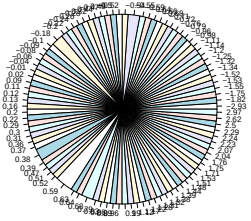
```
plot(uzorak1)
plot(uzorak1,pch=3)              # scatterplot sa +
plot(uzorak2)
```

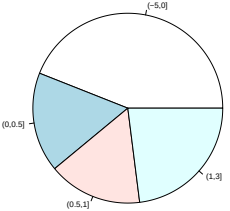


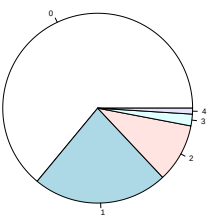


Pita (*pie chart*)

```
pie(table(uzorak1))
x<-cut(uzorak1,breaks=c(-5,0,.5,1,3))
table(x)
pie(table(x))
pie(table(uzorak2))
```

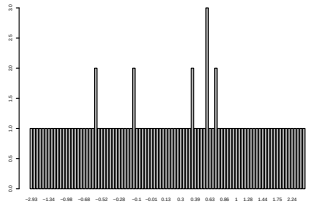


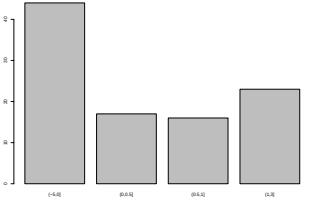


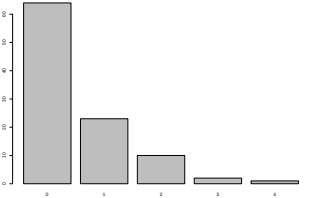


Stubičasti dijagram (*bar chart*)

```
barplot(table(uzorak1))
barplot(table(x))
barplot(table(uzorak2))
#barplot(table(uzorak2),col="cyan")
```







Domaći rad

Šta je razlika između histograma i barplota?

Unos podataka i NA

```
x<-scan()                                  #pojedinacni rucni  unos

x<-c(1,2,3,5,7,9)
data.entry(x)                              #pojedinacna izmena

x<-c(x,NA)
mean(x)                                    #javlja  gresku
mean(x,na.rm=T)                            #pravilno
mean(x[!is.na(x)])                         #pravilno
mean(na.omit(data.frame(x)$x))             #izbegavati posto  brise  cele  vrste
```

7. Napraviti tablicu Gausove raspodele.

```
p<−pnorm(seq(from=0, by=0.01, length=350))
#p<−round(pnorm(seq(from=0, by=0.01, length=350)),digits=4)
dim(p)<−c(10,35)
p<−t(p)
View(p)

#2. nacin
p<−pnorm(seq(from=0, by=0.01, length=350))
p<−matrix(p,35,10,byrow=TRUE)
View(p)
```

8. Napraviti tablicu Studentove raspodele.

```
x<−c(1:30,40,60,120,Inf)
p<−c(.75,.9,.95,.975,.99,.995,.9995)
qt<−t(outer(p,x,"qt"))
View(qt)
```

Domaći rad

Napraviti tablicu Pirsonove χ^2 raspodele

Domaći rad

1. Na slučajan način je izabrano 150 prirodnih brojeva manjih od 1000.

   a) Formirati raspodelu frekvencija sa 10 klasa iste širine i rezultate predstaviti tabelarno.

   b) Odrediti modalni i medijalni interval iz a).

   c) Odrediti aritmetičku sredinu i uzoračku disperziju uzorka.

   d) Odrediti koeficijent asimetrije (skewness) i sedmi centralni momenat uzorka.

   e) Izračunati realizovanu vrednost empirijske funkcije raspodele $f_n^*(615.5)$.

   f) Nactrati histogram i *pie chart* koristeći intervale iz a).

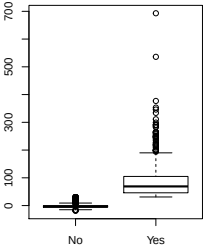
2. Učitati fajl FlightDelays.csv.

   a) Nacrtati *Box plot* za dužinu leta (FlightLength) po mesecu.

   b) Odrediti koliko su u proseku kasnili letovi (Delay) za Denver (Destination: DEN).

   c) Da li je standardna devijacija dužine leta veća kod aviona koji su poleteli između 4 i 8 časova (DepartTime: 4-8am) ili kod onih koji su poleteli između 16 i 20 časova (DepartTime: 4-8pm)?

3. Na osnovu *Box plot*–a uzorka zaključiti da li postoje statistički značajne razlike između srednjih vrednosti elemenata uzorka u ove dve grupe. Odgovor obrazložiti.



4. Naći treći decil za raspodele: $\mathcal{B}(10,\frac{3}{4})$, $\mathcal{G}(\frac{1}{10})$, t_{10}, χ^2_{10}, $\mathcal{N}(15,3)$.

5. Izračunati vrednost funkcije raspodele u tački $x = 7.5$ ako X ima $\mathcal{B}(10,\frac{3}{4})$, $\mathcal{G}(\frac{1}{10})$, t_{10}, χ^2_{10}, $\mathcal{N}(15,3)$ raspodelu.

6. Izračunati $\Gamma(-\frac{3}{2})$.

7. Pronaći šta vraćaju naredbe **dnorm** i **dbinom**.

8. Instalirati paket **ggplot2** i nacrtati nekoliko deskriptivnih grafika koristeći funkcije iz ovog paketa.

Vežbe 9

Intervali poverenja i testiranje hipoteza

Posmatra se uzorak od 100 slučajnih brojeva iz intervala $(0,1)$.

1. Naći 90% interval poverenja za procenat brojeva koji su veći od 0.75.

```
#interval poverenja za nepoznatu verovatnocu
set.seed(12345)
x<-runif(100)

K<-length(which(x>.75))          #broj uspesnih  realizacija
n<-length(x)
z<-qnorm((.9+1)/2)               #kvantil normalne N(0,1) raspodele

a<-n^2+z^2*n
b<--2*K*n-z^2*n
c<-K^2
d<-b^2-4*a*c

x1<-(-b-sqrt(d))/2/a
x2<-(-b+sqrt(d))/2/a
```

Interval poverenja $I = (0.1948498, 0.3377947)$

2. Testirati hipotezu da je procenat brojeva koji su veći od 0.75 jednak 0.3 sa pragom značajnosti 5%.

   a) Koristeći interval poverenja rađen na predavanjima.

```
# testiranje  hipoteze  o  nepoznatoj  verovatnoci
set.seed(12345)
x<-runif(100)
K<-length(which(x>.75))
n<-length(x)
z<-qnorm((.95+1)/2)
a<-n^2+z^2*n
b<--2*K*n-z^2*n
c<-K^2
d<-b^2-4*a*c
x1<-(-b-sqrt(d))/2/a
x2<-(-b+sqrt(d))/2/a
```

$H_0(p = 0.3)$ protiv $H_1(p \neq 0.3)$

$p_0 = 0.3 \in I = (0.184047, 0.3537099) \Rightarrow$ hipoteza H_0 se prihvata

b) Koristeći interval poverenja iz zbirke.

```
p<-K/n
q<-1-p
y1<-p-z*sqrt(p*q/(n-1))
y2<-p+z*sqrt(p*q/(n-1))
```

c) Koristeći `binom.test` iz R-a.

```
>binom.test(K,n,.3,conf.level=.95)

  Exact binomial test

data: K and n
number of successes = 26, number of trials = 100, p-value = 0.4451
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.1773944 0.3573121
sample estimates:
probability of success
                  0.26
```

Učitati fajl prijemni.csv.

3. Naći 90% interval poverenja za uspeh iz srednje škole.

```
#interval poverenja za ocekivanje m, sigma nepoznato
podaci<-read.csv("prijemni.csv")
uspehSkola<-podaci$skola

n<-length(uspehSkola)
xn<-mean(uspehSkola)
s<-sd(uspehSkola)
t<-qt((1+.9)/2,n-1)              #kvantil Studentove t raspodele

x1<-xn-t*s/sqrt(n)
x2<-xn+t*s/sqrt(n)
```

Interval poverenja $I = (28.1033, 30.99209)$

4. Testirati hipotezu da je srednja vrednost uspeha u srednjoj školi jednaka 32 ($\alpha = 0.05$).

```
# testiranje  hipoteze  o  ocekivanju  m, sigma nepoznato
podaci<-read.csv("prijemni.csv")
uspehSkola<-podaci$skola

n<-length(uspehSkola)
xn<-mean(uspehSkola)
s<-sd(uspehSkola)
t<-qt((1+.95)/2,n-1)

x1<-xn-t*s/sqrt(n)
x2<-xn+t*s/sqrt(n)
```

$H_0(m = 32)$ protiv $H_1(m \neq 32)$

$m_0 = 32 \notin I = (27.81335, 31.28203) \Rightarrow$ hipoteza H_0 se odbacuje

5. Naći 95% interval poverenja za varijansu uspeha iz srednje škole.

```
#interval poverenja za varijansu ( disperziju )
podaci<-read.csv("prijemni.csv")
uspehSkola<-podaci$skola

n<-length(uspehSkola)
s<-sd(uspehSkola) ^ 2*(n-1)/n
y1<-qchisq((1+.95)/2,n-1)                    #kvantili Pirsonove raspodele
y2<-qchisq((1-.95)/2,n-1)
x1<-n*s/y1
x2<-n*s/y2
```

Interval poverenja $I = (19.11832, 47.54454)$

6. Testirati hipotezu o jednakosti srednje vrednosti uspeha na prijemnom kod muških i ženskih kandidata ($\alpha = 0.05$).

   Testiranje hipoteze o jednakosti srednjih vrednosti dva obeležja sa nepoznatim varijansama. Nulta hipoteza $H_0(m_1 = m_2)$ protiv alternativne hipoteze $H_1(m_1 \neq m_2)$.

```
podaci<-read.csv("prijemni.csv")
prijemniM<-podaci$prijemni[podaci$pol=="m"]
prijemniZ<-podaci$prijemni[podaci$pol=="z"]

n1<-length(prijemniM)
n2<-length(prijemniZ)
xn1<-mean(prijemniM)
xn2<-mean(prijemniZ)
```

Ako su varijanse jednake koristi se test statistika

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

gde je $\sigma = \sqrt{\frac{n_1\overline{S}_1^2 + n_2\overline{S}_2^2}{n_1+n_2-2}}$, koja ima $t_{n_1+n_2-2}$ raspodelu.

```
#jednake varijanse
sn1<-sd(prijemniM) ^ 2*(n1-1)/n1
sn2<-sd(prijemniZ) ^ 2*(n2-1)/n2
sigma<-sqrt((n1*sn1+n2*sn2)/(n1+n2-2))

t0<-(xn1-xn2)/sigma/sqrt(1/n1+1/n2)
t<-qt((1+.95)/2,n1+n2-2)
```

Nulta hipoteza $H_0(m_1 = m_2)$
$t_0 = -1.739699 \in I = (-2.026192, 2.026192) \Rightarrow$ hipoteza H_0 se prihvata

Za različite varijanse koristi se test statistika

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}},$$

koja približno ima $t_{\min\{n_1-1, n_2-1\}}$ raspodelu.

```
# razlicite  varijanse
s1<-sd(prijemniM)^2
s2<-sd(prijemniZ)^2

t0<-(xn1-xn2)/sqrt(s1/n1+s2/n2)          #ako je m1=m2, onda je m1-m2=0
t<-qt((1+.95)/2,min(n1-1,n2-1))
```

Nulta hipoteza $H_0(m_1 = m_2)$
$t_0 = -1.224453 \in I = (-2.446912, 2.446912) \Rightarrow$ hipoteza H_0 se prihvata

Vežbe 10

1. Testirati hipotezu da je uspeh u srednjoj školi raspoređen po normalnoj raspodeli ako su grupe za χ^2-test: 10-25, 25-30, 30-35, 35-40 ($\alpha = 0.05$).

```
podaci<-read.csv("prijemni.csv")
uspehSkola<-podaci$skola

mi<-c(25,30,35)                          #deobne tacke
fi<-hist(uspehSkola,c(10,mi,40))$counts  #frekvencije
n<-length(uspehSkola)

k<-length(fi)                            #broj intervala
s<-2                                     #broj parametara koje treba oceniti
xn<-mean(uspehSkola)
sn<-sd(uspehSkola)

#funkcija raspodele
Fx<-function(x){
  return(pnorm((x-xn)/sn)                #pnorm(x,xn,sn)
}

#teorijske verovatnoce
p<-c(Fx(mi),1)-c(0,Fx(mi))
```

```
#realizovana vrednost   statistike
y0<-sum((fi-n*p)^2/n/p)

#najveca dozvoljena vrednost   statistike
y<-qchisq(1-.05,k-1-s)
```

$y_0 = 1.296183 < y = 3.841459 \Rightarrow$ hipoteza H_0 se prihvata

2. χ^2-testom ispitati saglasnost obeležja čiji je realizovani uzorak (0.12, 0.14, 0.25, 0.05, 0.02, 0.08, 0.03, 0.04, 0.51, 0.07, 0.42, 0.08, 0.33, 0.36, 0.06, 0.23) sa raspodelom:

$$\varphi(x) = \begin{cases} \theta x^{\theta-1}, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}, \theta > 0.$$

```
#uzorak
x<-c(0.12,0.14,0.25,0.05,0.02,0.08,0.03,0.04,0.51,0.07,0.42,0.08,0.33,0.36,0.06,0.23)
xs<-sort(x)                 #sortiran uzorak
fi<-c(6,5,5)                #frekvencije (sami konstruisemo intervale)
mi<-c(0.07,0.23)            #deobne tacke
n<-sum(fi)                  #obim uzorka
xn<-mean(x)
theta<-xn/(1-xn)            #ocena parametra theta
k<-length(fi)               #broj intervala
s<-1                        #broj parametara

p<-c(mi^theta,1)-c(0,mi^theta)      #teorijske verovatnoce
y0<-sum((fi-n*p)^2/n/p)             #realizovana vrednost statistike
y<-qchisq(1-.05,k-1-s)              #najveca dozvoljena vrednost statistike
```

$y_0 = 3.39369 < y = 3.841459 \Rightarrow$ hipoteza H_0 se prihvata

3. Testom Kolmogorov-Smirnov testirati hipotezu o saglasnosti datog uzorka sa normalnom raspodelom $\mathcal{N}(10,4)$ ($\alpha = 0.05$).

$0-5$	$5-15$	$15-20$	$20-30$	$30-40$
3	7	10	15	5

```
fi <- c(3,7,10,15,5)
mi<- c(0,5,15,20,30,40)
ml<- mi[1:length(mi)-1]
md<- mi[2:length(mi)]
xi<- (ml+md)/2
f<- cumsum(fi)
n<- sum(fi)

Fn<- f/n                    #empirijska funkcija  raspodele
Fx<- pnorm(xi,10,4)         #data funkcija  raspodele

sqrt(n)*max(abs(Fn-Fx))     #realizovana vrednost  statistike
```

$2.970034 > 1.36 = \lambda_{0.95} \Rightarrow$ hipoteza H_0 se odbacuje

4. Ispitati nezavisnost obeležja X i Y čije su realizovane vrednosti uzorka:

5	17	9
5	7	19

```
fij <-matrix(c(5,5,17,7,9,19),2,3)

#matrica kontingencije
source("matrica_kontigencije.R")
f<-kont(fij)
print(f)
#f<-matrix(c(5,5,10,17,7,24,9,19,28,31,31,62),3,4)

#obim uzorka
n<-sum(fij)     #n<-f[nrow(f),ncol(f)]

#matrica sa teorijskim frekvencijama
ft<-matrix(NA,nrow(fij),ncol(fij))
for (i in 1:nrow(ft)){
  for (j in 1:ncol(ft)){
    ft[i,j]=f[i,ncol(f)]*f[nrow(f),j]/n
  }
}
```

```
#realizovana vrednost   statistike
y0<-sum((ft-fij)^2/ft)

#najveca dozvoljena vrednost   statistike
y<-qchisq(1-.05,(nrow(fij)-1)*(ncol(fij)-1))
```

H_0 : obeležja X i Y su nezavisna

$y_0 = 7.738095 > y = 5.991465 \Rightarrow$ hipoteza H_0 se odbacuje

matrica_kontigencije.R

```
kont <- function(mat){

  m <- nrow(mat)
  n <- ncol(mat)
  res <- matrix(numeric((m+1)*(n+1)), m+1, n+1)

  for (i in 1:m){
    for (j in 1:n){
      res[i, j] <- mat[i,j]
      res[m+1, j] <- res[m+1, j] + mat[i, j]
      res[i, n+1] <- res[i, n+1] + mat[i, j]
    }
  }

  res[m+1, n+1] <- sum(res[m+1, 1:n])

  return(res)

}
```

Domaći rad

Učitati fajl FlightDelays.csv.

1. Naći 90% interval poverenja dužine leta (FlightLength) za ponedeljak (Day: Mon).
2. Testirati hipotezu da je procenat letova koji su kasnili više od 20 minuta jednak 0.2 sa pragom značajnosti 5%.
3. Ispitati da li postoje statistički značajne razlike između srednjih vrednosti kašnjenja letova tokom maja i juna.
4. Testom Kolmogorov-Smirnov testirati hipotezu da uzorak

   0.16, 0.35, 0.86, 0.05, 0.11, 0.07, 0.04, 0.09, 0.52, 0.38

   ima uniformnu raspodelu $\mathcal{U}(0,1)$.
5. Ispitati nezavisnost obeležja X i Y čije su realizovane vrednosti uzorka:

19	6	12
25	18	16

6. χ^2-testom ispitati saglasnost obeležja čiji je realizovani uzorak

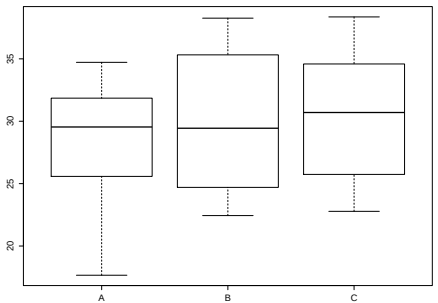
   0.62, 0.54, 0.21, 0.48, 0.71, 0.58, 0.32, 0.48, 0.55, 0.38,
   0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64

   sa eksponencijalnom raspodelom $\mathcal{E}(\lambda)$. Uzorak podeliti na tri intervala koristeći deobne tačke 0.45 i 0.55.

Učitati fajl prijemni.csv.

5. Nacrtati *Box plot* uspeha iz srednje škole po grupama.

```
podaci<-read.csv("prijemni.csv")
boxplot(skola~grupa,data=podaci)
#plot(podaci$skola~podaci$grupa)
```



6. Testirati hipotezu o jednakosti srednje vrednosti uspeha iz srednje škole po grupama (ANOVA).

```
>podaci<-read.csv("prijemni.csv")
>anova(lm(skola~grupa,data=podaci))
Analysis of Variance Table

Response: skola
           Df  Sum Sq Mean Sq F value Pr(>F)
grupa       2    26.6  13.300  0.4512 0.6404
Residuals  36  1061.2  29.476

#isto dobijamo i sa komandom
summary(aov(lm(skola~grupa,data=podaci)))
```

$H_0(m_1 = m_2 = m_3)$ protiv $H_1(\exists i,j, m_i \neq m_j)$

Tabela ANOVE: $\begin{matrix} Df_1 & SSTR & MSTR & F & \alpha^* \\ Df_2 & SSE & MSE & & \end{matrix}$

$Df_1, Df_2 -$ broj stepeni slobode

$F -$ realizovana vrednost Fišerove test statistike sa Df_1, Df_2 stepeni slobode

$\alpha^* - p$-vrednost

$\alpha^* = 0.6404 > \alpha = 0.05 \Rightarrow H_0$ se prihvata

Napomena:

Faktor po kom se radi ANOVA treba da bude kategorijalno obeležje (faktor). Ukoliko je obeležje numeričko, neophodno ga je prvo transformisati u faktor, pa tek onda raditi ANOVU.

```
>podaci<-read.csv("anova.csv")
#nazivi nivoa faktora iz fajla prijemni.csv su preimenovani u numericke vrednosti 1,2,3 umesto A,B,C
>anova(lm(skola~grupa,data=podaci))
Analysis of Variance Table

Response: skola
          Df  Sum Sq Mean Sq F value  Pr(>F)
grupa      1    22.2  22.204   0.771  0.3856
Residuals 37  1065.5  28.799
#Vidimo da broj stepeni nije dobar, pa onda ni p-vrednost nije odgovarajuca

>podaci$grupa<-as.factor(podaci$grupa)
> anova(lm(skola~grupa,data=podaci))
Analysis of Variance Table

Response: skola
          Df  Sum Sq Mean Sq F value  Pr(>F)
grupa      2    26.6  13.300  0.4512  0.6404
Residuals 36  1061.2  29.476
#Sada smo dobili iste vrednosti kao u zadatku 5.
```

Domaći rad

1. Proveriti šta rade sledeće komande (i dobijene rezultate uporediti sa ANOVOM):

```
prijemni<-read.csv("prijemni.csv")
oneway.test(skola~grupa,data=prijemni)
oneway.test(skola~grupa,data=prijemni,var.equal=TRUE)
```

2. Testirati hipotezu da su srednje vrednosti testa jednake za svih 5 udžbenika (ANOVA).

1 :	82	75	87	76
2 :	67	79	77	81
3 :	91	82	76	79
4 :	66	73	89	84
5 :	82	71	67	76

Prva ideja: "Peške" izračunati realizovanu vrednost test statistike F koristeći formule sa folija. (Koristiti R za brže izračunavanje traženih vrednosti.)

Druga ideja: Transformisati datu tabelu u data frame D i iskoristiti ugrađeni `anova` test iz R-a.

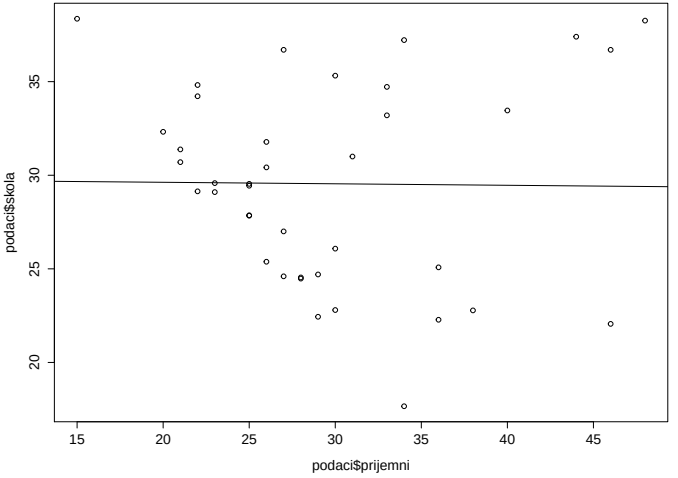
7. Naći koeficijent korelacije uspeha iz srednje škole u zavisnosti od uspeha na prijemnom. Prognozirati kom uspehu u srednjoj školi odgovara 35 bodova osvojenih na prijemnom.

```
>podaci<-read.csv("prijemni.csv")
>lm(skola~prijemni,data=podaci)

Call:
lm(formula = skola ~ prijemni, data = podaci)

Coefficients :
(Intercept)       prijemni
  29.784053     -0.008009
```

Jednačina linearne regresije

$$skola = -0.008009 * prijemni + 29.784053$$

```
plot(podaci$prijemni,podaci$skola)
abline(lm(skola~prijemni,data=podaci))
```



```
# koeficijent   korelacije
cor(podaci$skola,podaci$prijemni)

#predikcija
lm(skola~prijemni,data=podaci)$coefficients%*%c(1,35)
```

Koeficijent korelacije $r = -0.01157393$
$|r| < 0.3 \Rightarrow$ uspeh iz srednje škole i uspeh na prijemnom nisu u korelacionoj vezi

Broju od 35 poena osvojenih na prijemnom odgovara 29.50375 poena iz srednje škole.

8. Koristeći permutacioni test testirati hipotezu o jednakosti srednje vrednosti uspeha iz škole kod muških i ženskih kandidata.

```
prijemni<-read.csv("prijemni.csv")
summary(prijemni) #m:32, z:7
boxplot(skola~pol,data=prijemni)

xn<-tapply(prijemni$skola,prijemni$pol,mean)
#xn<-c(mean(prijemni$skola[prijemni$pol=="m"]),mean(prijemni$skola[prijemni$pol=="z"]))
reg.razlika<-xn[1]-xn[2]                    #registrovana razlika srednjih vrednosti
skola<-prijemni$skola

#ponovno uzorkovanje (resampling)
N<-999
razlika<-numeric(N)
for (i in 1:N){
  indeks<-sample(39,size=32,replace=FALSE)
  razlika[i]<-mean(skola[indeks])-mean(skola[-indeks])
}

#histogram permutacione raspodele
hist(razlika,xlab="skola m - z")
abline(v=reg.razlika,col="blue")            #vertikalna prava u vrednosti originalne razlike
```

```
#p−vrednost
2*(sum(razlika <= reg.razlika) + 1)/(N + 1)

#Jednostrani test: (sum(razlika <= reg.razlika) + 1)/(N + 1)
```

Permutaciona raspodela je promenljiva zbog slučajnog uzorkovanja, pa se p–vrednost i histogram permutacione raspodele takođe menjaju.

$\alpha^* \sim 0.2 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

Napomena 1:

Da smo stavili reg.razlika<-xn[2]-xn[1], p–vrednost bi računali na sledeći način:

2*(sum(razlika >= reg.razlika) + 1)/(N + 1)

Napomena 2:

$p-$vrednost dvostranog testa možemo izračunati i na sledeći način:

(sum(abs(razlika)>=abs(reg.razlika))+1)/(N+1)

Napomena:

Rešenje zadatka korišćenjem T-testa

```
> t.test(prijemni$skola[prijemni$pol=="z"],prijemni$skola[prijemni$pol=="m"])

   Welch Two Sample t-test

data:  prijemni$skola[prijemni$pol == "z"] and prijemni$skola[prijemni$pol == "m"]
t = 1.2648, df = 8.6296, p-value = 0.239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.291815 8.018601
sample estimates:
mean of x mean of y
 31.89714 29.03375
```

$\alpha^* = 0.239 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

Vežbe 10

1. Po jedan uzorak sa dve mašine za pakovanje deterdženta od 10 kg je izmeren na preciznoj vagi.

- 9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95, 11.03, 10.12, 9.33, 9.73, 10.17, 9.48, 10.89, 10.11, 10.30, 8.87, 9.51, 10.42, 10.02, 10.84, 9.96, 10.15, 10.64, 11.30
- 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78, 10.61, 9.49, 7.81, 8.90, 8.60, 8.50, 9.31, 9.97, 8.89, 8.87, 7.23, 7.82, 7.65, 9.11, 8.65, 6.30, 9.38, 8.31, 10.48, 10.56, 9.96, 8.84, 9.10, 11.07, 9.84, 9.75, 9.07, 9.09, 8.96, 8.11, 8.17, 9.73, 9.06, 8.40, 11.12, 9.38, 7.26, 8.69

```
uzorak1<-c(9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95, 11.03, 10.12, 9.33, 9.73, 10.17, 9.48, 10.89, 10.11, 10.30, 8.87, 9.51, 10.42, 10.02, 10.84, 9.96, 10.15, 10.64, 11.30)

uzorak2<-c(9.85, 9.3, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.5, 11.95, 10.92, 9.78, 10.61, 9.49, 7.81, 8.9, 8.6, 8.5, 9.31, 9.97, 8.89, 8.87, 7.23, 7.82, 7.65, 9.11, 8.65, 6.3, 9.38, 8.31, 10.48, 10.56, 9.96, 8.84, 9.10, 11.07, 9.84, 9.75, 9.07, 9.09, 8.96, 8.11, 8.17, 9.73, 9.06, 8.40, 11.12, 9.38, 7.26, 8.69)
```

a) Kolmogorov-Smirnov testom testirati hipotezu o saglasnosti prvog uzorka sa normalnom raspodelom $\mathcal{N}(10, 0.81)$.

```
> ks.test(uzorak1,'pnorm',10,0.81)          #obavezno " ili ""

  One-sample Kolmogorov-Smirnov test

data: uzorak1
D = 0.17074, p-value = 0.3097
alternative hypothesis: two-sided
```

$\alpha^* = 0.3097 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

Napomena:

```
#realizovana vrednost test statistike
ks.test(uzorak1,'pnorm',10,0.81)$statistic*sqrt(length(uzorak1))
```

Vrednost $D = 0.17074$ je realizovana vrednost statistike $D_n = \sup_x |F_n^*(x) - F(x)|$.

Kako je $\sqrt{n_1}D = \sqrt{30} \cdot 0.17074 = 0.9351815 < 1.36$ zaključujemo da se hipoteza H_0 prihvata.

b) χ^2-testom testirati hipotezu o saglasnosti drugog uzorka sa normalnom raspodelom (deobne tačke: 8.0, 8.5, 9.0, 9.5, 10.0).

```
xn<-mean(uzorak2)
sn<-sd(uzorak2)
m<-c(8,8.5,9,9.5,10)
pm<-pnorm(m,xn,sn)
p<-c(pm,1)-c(0,pm)
mi<-c(floor(min(uzorak2)),m,ceiling(max(uzorak2)))
fi<-hist(uzorak2,mi)$counts

y0<-chisq.test(fi,p=p)$statistic
y<-qchisq(.95,length(fi)-1-2)          #ocenili  smo dva parametra

alpha<-1-pchisq(y0,length(fi)-1-2)     #p-vrednost
```

$y_0 = 2.079539 < y = 7.814728 \Rightarrow$ hipoteza H_0 se prihvata

Napomena:

Pošto smo morali da ocenimo 2 parametra, $p-$vrednost koju vraća `chisq.test` nije tačna (ova vrednost bi bila tačna da nije bilo nepoznatih parametara) i moramo peške izračunati α^*.
$\alpha^* = 0.5560632 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

c) Testirati hipotezu da je srednja vrednost prvog uzorka veća od srednje vrednosti drugog.

```
> t.test(uzorak1,uzorak2,alternative='greater')

  Welch Two Sample t-test

data:  uzorak1 and uzorak2
t = 5.3151, df = 77.935, p-value = 4.929e-07
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.7067711       Inf
sample estimates:
mean of x mean of y
 10.10867  9.07960
```

$H_0(m_1 = m_2)$ protiv $H_1(m_1 > m_2)$

$\alpha^* = 4.929 \cdot 10^{-7} < \alpha = 0.05 \Rightarrow$ hipoteza H_0 se odbacuje, tj. prihvata se H_1

d) Testirati hipotezu da je srednja vrednost prvog uzorka veća od srednje vrednosti drugog, pod pretpostavkom da imaju jednake varijanse.

```
> t.test(uzorak1,uzorak2,alternative='greater',var.equal=TRUE)

  Two Sample t-test

data:  uzorak1 and uzorak2
t = 4.6821, df = 78, p-value = 5.892e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6631998       Inf
sample estimates:
mean of x mean of y
 10.10867  9.07960
```

$H_0(m_1 = m_2)$ protiv $H_1(m_1 > m_2)$

$\alpha^* = 5.892 \cdot 10^{-6} < \alpha = 0.05 \Rightarrow$ hipoteza H_0 se odbacuje, tj. prihvata se H_1

Napomena:

$H_0(m_1 - m_2 = \Delta)$ protiv $H_1(m_1 - m_2 \neq \Delta)$

Npr. za $\Delta = 1$ dobijamo

```
> t.test(uzorak1,uzorak2,mu=1)

   Welch Two Sample t-test

data:  uzorak1 and uzorak2
t = 0.15013, df = 77.935, p-value = 0.8811
alternative hypothesis: true difference in means is not equal to 1
95 percent confidence interval:
 0.6436084 1.4145249
sample estimates:
mean of x mean of y
 10.10867  9.07960
```

$\alpha^* = 0.8811 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

e) Testirati hipotezu da je standardna devijacija prvog uzorka jednaka 1.

```
n<-length(uzorak1)
s<-sd(uzorak1)^2*(n-1)/n
y1<-qchisq((1+.95)/2,n-1)
y2<-qchisq((1-.95)/2,n-1)
x1<-n*s/y1
x2<-n*s/y2
```

Nulta hipoteza $H_0(\sigma = 1)$

$\sigma_0^2 = 1 \notin I = (0.2552004, 0.7271325) \Rightarrow$ hipoteza H_0 se odbacuje

Napomena:

Interval poverenja za σ je $I = (\sqrt{x_1}, \sqrt{x_2}) = (0.5051737, 0.8527206)$

f) Testirati hipotezu da su standardne devijacije kod oba uzorka jednake.

```
> var.test(uzorak1,uzorak2)

  F test to compare two variances

data: uzorak1 and uzorak2
F = 0.33427, num df = 29, denom df = 49, p-value = 0.002289
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1776679 0.6653103
sample estimates:
ratio of variances
         0.3342673
```

Nulta hipoteza $H_0(\sigma_1^2 = \sigma_2^2)$

$\alpha^* = 0.002289 < \alpha = 0.05 \Rightarrow$ hipoteza H_0 se odbacuje

Napomena:

Realizovanu vrednost Fišerove test statistike možemo dobiti i na sledeći način

```
n1<-length(uzorak1)
n2<-length(uzorak2)
s1<-sd(uzorak1)^2*(n1-1)/n1
s2<-sd(uzorak2)^2*(n2-1)/n2
F<-n1*s1*(n2-1)/(n2*s2*(n1-1))          #var(uzorak1)/var(uzorak2)
```

$$F = \frac{\frac{n_1\bar{S}^2_{n_1}}{n_1-1}}{\frac{n_2\bar{S}^2_{n_2}}{n_2-1}} = \frac{\hat{S}^2_{n_1}}{\hat{S}^2_{n_2}} : F_{n_1-1,n_2-1}$$ Fišerova test statistika sa n_1-1 i n_2-1 stepeni slobode

Vežbe 9, zadatak 2. a)

```
>prop.test(K,n,.3,conf.level=.95,correct=FALSE)

  1-sample proportions test without continuity correction

data: K out of n, null probability 0.3
X-squared = 0.7619, df = 1, p-value = 0.3827
alternative hypothesis: true p is not equal to 0.3
95 percent confidence interval:
 0.1840470 0.3537099
sample estimates:
   p
0.26
```

$H_0(p = 0.3)$ protiv $H_1(p \neq 0.3)$

$\alpha^* = 0.3827 > \alpha = 0.05 \Rightarrow$ hipoteza H_0 se prihvata

Napomena:

Stavljamo correct=FALSE da ne umanji grešku zaokruživanja koja nastaje kao posledica aproksimacije diskretne slučajne promenljive neprekidnom.

Vežbe 9, zadatak 4.

```
>t.test(uspehSkola,mu=32,conf.level = .95)

   One Sample t-test

data:  uspehSkola
t = -2.8624, df = 38, p-value = 0.006803
alternative hypothesis: true mean is not equal to 32
95 percent confidence interval:
 27.81335 31.28203
sample estimates:
mean of x
 29.54769
```

$H_0(m = 32)$ protiv $H_1(m \neq 32)$

$\alpha^* = 0.006803 < \alpha = 0.05 \Rightarrow$ hipoteza H_0 se odbacuje

Vežbe 10, zadatak 4.

```
>chisq.test(matrix(c(5,5,17,7,9,19),2,3))

  Pearson's Chi-squared test

data: fij
X-squared=7.7381, df=2, p-value=0.02088
```

H_0 : obeležja X i Y su nezavisna

$\alpha^* = 0.02088 < \alpha = 0.05 \Rightarrow$ hipoteza H_0 se odbacuje

Domaći rad

Koristeći ugrađene statističke testove u R–u uraditi

1. zadatke sa vežbi 9 i 10
2. domaći sa vežbi 10

Testovi normalnosti

Nulta hipoteza $H_0(F(x) = \Phi(\frac{x-m}{\sigma}))$ protiv alternativne hipoteze $H_1(F(x) \neq \Phi(\frac{x-m}{\sigma}))$

1. Kolmogorov-Smirnov test (samo ako iz iskustva znamo vrednosti parametara m i σ)
2. Shapiro-Wilk test

```
> prijemni<-read.csv("prijemni.csv")
> shapiro.test(prijemni$skola)

  Shapiro-Wilk normality test

data:  prijemni$skola
W = 0.9675, p-value = 0.3138
```

Napomena:

Paket **nortest** sadrži nekoliko testova normalnosti, od kojih su najpoznatiji Anderson-Darling test (**ad.test**) i Cramér-von Mises test (**cvm.test**).

t-test parova

10 osoba je izmereno pre i nakon 10 nedelja dijete i zabeleženi su sledeći rezultati:

pre	76.05	82.56	72.97	85.96	56.80	79.96	54.98	109.30	66.07	87.64
posle	76.46	76.65	70.82	83.73	60.22	74.53	50.38	110.38	64.68	85.19

Testirati hipotezu da je razlika srednjih vrednosti težina pre i posle dijete veća od 0.

```
> wgt1<-c(76.05,82.56,72.97,85.96,56.80,79.96,54.98,109.30,66.07,87.64)
> wgt2<-c(76.46,76.65,70.82,83.73,60.22,74.53,50.38,110.38,64.68,85.19)
> t.test(wgt1, wgt2, paired=TRUE, alternative="greater")

  Paired t-test

data:  wgt1 and wgt2
t = 2.0536, df = 9, p-value = 0.0351
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2067164       Inf
sample estimates:
mean of the differences
                 1.925
```

Tabela kontigencije

Ispitati nezavisnost obeležja pol i grupa iz fajla prijemni.csv.

```
> prijemni<-read.csv("prijemni.csv")
> table(prijemni$pol, prijemni$grupa)

    A  B  C
  m 14 13 5
  z  1  4  2
> chisq.test(table(prijemni$pol, prijemni$grupa))

  Pearson's Chi-squared test

data:  table(prijemni$pol, prijemni$grupa)
X-squared = 2.1923, df = 2, p-value = 0.3342

Warning message:
In chisq.test(table(prijemni$pol, prijemni$grupa)) :
  Chi-squared approximation may be incorrect
```

Napomena:

Greška koju R javlja je posledica činjenice da su neke frekvencije iz tabele veoma male.

Multipla regresija

Učitati fajl MnGroundwater.csv. Pronaći linearnu zavisnost pH vrednosti vode od količine aluminijuma (Al), arsena (As), hlora (Cl) i olova (Pb).

```
> D<−read.csv("MnGroundwater.csv")
> model1<−lm(pH~Aluminum+Arsenic+Chloride+Lead,data=D)
> model1

Call:
lm(formula = pH ~ Aluminum + Arsenic + Chloride + Lead, data = D)

Coefficients :
(Intercept)     Aluminum      Arsenic     Chloride         Lead
  7.289e+00    5.646e−04   −3.342e−03   −5.487e−08   −5.906e−04
```

$$pH = 5.646 \cdot 10^{-4} Al - 3.342 \cdot 10^{-3} As - 5.487 \cdot 10^{-8} Cl - 5.906 \cdot 10^{-4} Pb + 7.289$$

1. Da li je dobijeni model statistički značajan?
2. Da li je dobijeni model koristan?
3. Da li su (svi) koeficijenti statistički značajni? Ukoliko nisu, koji prediktor možemo izbaciti iz razmatranja?

```
> summary(model1)

Call:
lm(formula = pH ~ Aluminum + Arsenic + Chloride + Lead, data = D)

Residuals:
     Min       1Q   Median       3Q      Max
-5.7889  -0.2535  -0.0388   0.2128   2.1267

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  7.289e+00  1.830e-02  398.261   < 2e-16 ***
Aluminum     5.646e-04  1.168e-04    4.835  1.57e-06 ***
Arsenic     -3.342e-03  1.502e-03   -2.225    .0264 *
Chloride    -5.487e-08  9.168e-08   -0.599    0.5496
Lead        -5.906e-04  1.739e-03   -0.340    0.7343
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4793 on 890 degrees of freedom
Multiple R-squared: 0.03171, Adjusted R-squared: 0.02736
F-statistic: 7.286 on 4 and 890 DF, p-value: 8.957e-06
```

1. Da li je dobijeni model statistički značajan?

   $p-$vrednost: $8.957 \cdot 10^{-6} < 0.05 \Rightarrow$ model je statistički značajan
   (Što je $p-$vrednost manja, model je značajniji.)

2. Da li je dobijeni model koristan?

   Koeficijent determinacije r^2 : 0.03171
   Podešeni koeficijent determinacije: 0.02736
   (Što je koeficijent determinacije bliži 1, model je korisniji. Kod multiple regresije je bolje posmatrati podešeni koeficijent determinacije.)

3. Da li su (svi) koeficijenti statistički značajni? Ukoliko nisu, koji prediktor možemo izbaciti iz razmatranja?

   Koeficijenti koji stoje uz hlor i olovo nisu statistički značajni (velike $p-$vrednosti) i te prediktore možemo zanemariti u razmatranju.

   model2<-lm(pH∼Aluminum+Arsenic,data=D)

$$pH = 5.614 \cdot 10^{-4} Al - 3.316 \cdot 10^{-3} As + 7.287$$

Da li postoje statistički značajne razlike između modela 1 i modela 2?

```
> anova(model1, model2)
Analysis of Variance Table

Model 1: pH ~ Aluminum + Arsenic + Chloride + Lead
Model 2: pH ~ Aluminum + Arsenic
  Res.Df    RSS  Df  Sum of Sq      F  Pr(>F)
1    890 204.48
2    892 204.59  -2  -0.10992 0.2392 0.7873
```

$Df = -2$: model 2 ima 2 prediktora manje od modela 1

$p-$vrednost: $0.7873 \Rightarrow$ ne postoji statistički značajna razlika između ova dva modela

Predikcije

$Al : 25,\ As : 2,\ Cl : 1350,\ Pb : 0.5 \Rightarrow pH = ?$

```
#1. nacin
model1$coefficients%*%c(1, 25, 2, 1350, 0.5)

#2. nacin
v<-data.frame(Aluminum=25, Arsenic=2, Chloride=1350, Lead=0.5)
predict(model1, newdata = v)
```

$$\begin{aligned} pH &= \mathbf{25} \cdot 5.646 \cdot 10^{-4} - \mathbf{2} \cdot 3.342 \cdot 10^{-3} - \mathbf{1350} \cdot 5.487 \cdot 10^{-8} - \mathbf{0.5} \cdot 5.906 \cdot 19^{-4} + 7.289 \\ &= 7.296325 \end{aligned}$$

Napomena:

Data frame mora imati kolone nazvane kao prediktori.