Rešitve - ocenjevanje parametrov

Nataša Kejžar

Naloga 1 - prvi primer MM

```
st =  c(-1,1,-1,1,2,2,0,0,-1,0,1,1,1)
mi1 = mean(st)
mi2 = mean(st^2)
(mi2 -1)/2 # a
```

```
## [1] 0.1153846
```

```
(mi1-2*mi2 + 3)/2 # b
```

```
## [1] 0.5
```

Naloga 2 - radio

```
st =  c(6.39,4.33,-0.03,3.79,1.98)
mi1 = mean(st)
mi2=mean(st^2)
mi1-sqrt(3*(mi2-mi1^2)) # a
```

```
## [1] -0.4773564
```

```
mi1+sqrt(3*(mi2-mi1^2)) # b
```

```
## [1] 7.061356
```

Naloga 3

```
st = c(0.4,0.7,0.9)
mi = mean(st)
mi/(1-mi) # metoda momentov
```

```
## [1] 2
```

Naloga 4 - Hardy-Weinberg

a.
$$E(X) = -1 \cdot \theta^2 + 0 \cdot 2\theta(1-\theta) + 1 \cdot (1-\theta)^2 = 1 - 2\theta$$

b. Velja
$$\theta = \frac{1-\mu}{2}$$

Zato je cenilka po metodi momentov enaka

$$\widehat{\theta} = \frac{1-\widehat{\mu}}{2}$$

c.

$$E(\widehat{\theta}) = E\left(\frac{1-\widehat{\mu}}{2}\right) = \frac{1-E(\widehat{\mu})}{2} = \frac{1-\mu}{2} = \theta$$

Da, cenilka je nepristranska.

d.

$$var(X) = E(X^2) - E(X)^2$$

Pri tem je

$$E(X^2) == \theta^2 + (1-\theta)^2 = 2\theta^2 - 2\theta + 1$$

in zato

$$var(X) = 2\theta^2 - 2\theta + 1 - (1-2\theta)^2 = 2\theta^2 - 2\theta + 1 - 1 - \theta^2 + 4\theta = 2\theta(1-\theta)$$

e.

$$\begin{aligned} var\left(\frac{1-\widehat{\mu}}{2}\right) &= \frac{var(\widehat{\mu})}{4} \\ &= \frac{var(X)}{4n} = \frac{\theta(1-\theta)}{2n} \end{aligned}$$

f.

$$\begin{aligned} \widehat{\mu} &= \frac{-1 * n_1 + n_3}{n} \\ \widehat{\theta} &= \frac{n + n_1 - n_3}{2n} = \frac{2n_1 + n_2}{2n} \end{aligned}$$

Naloga 5 - Beta

Cenilki po metodi momentov sta:

$$\hat{\alpha} = \frac{\hat{\mu}_1(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

in

$$\hat{\beta} = \frac{(1-\hat{\mu}_1)(\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

Vzorec velikosti 15 (simulacija)

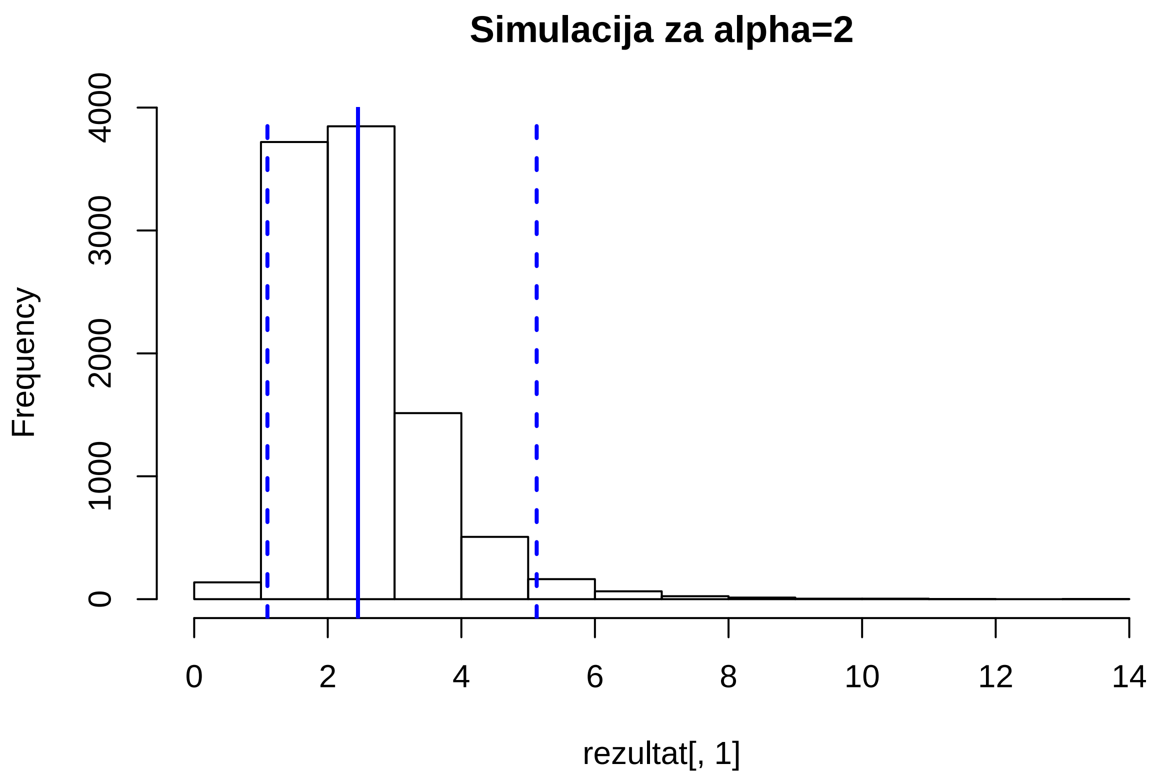
```
simV = function(n,a,b){
  vzorec = rbeta(n,shape1=a,shape2=b)
  mi1 = mean(vzorec)
  mi2 = mean(vzorec^2)
  alfa = (mi1-mi2)*mi1/(mi2-mi1^2)
  beta = (mi1-mi2)*(1-mi1)/(mi2-mi1^2)
  return(c(alfa,beta))
}

rezultat = NULL
for(i in 1:10000){
  rezultat = rbind(rezultat,simV(15,2,5))
}
```

Rezultati za α s simuliranim 95% IZ.

```
# 95% IZ iz simulacij
IZ = quantile(rezultat[,1],probs = c(0.025,0.975))
IZ
```

```
##     2.5%    97.5%
## 1.096730 5.127985
```

```r
hist(rezultat[,1],main="Simulacija za alpha=2")
abline(v=mean(rezultat[,1]),col="blue",lwd=2)
abline(v=IZ[1],col="blue",lwd=2,lty=2)
abline(v=IZ[2],col="blue",lwd=2,lty=2)
```

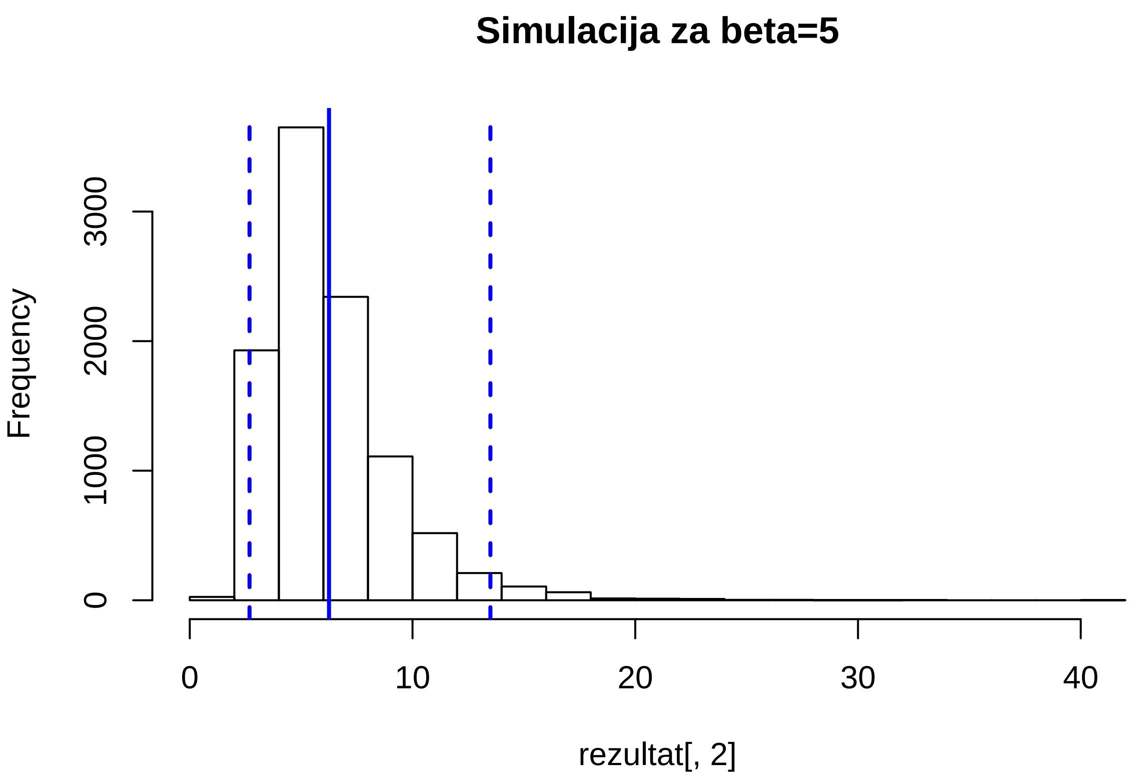


Rezultati za β s simuliranim 95% IZ.

```r
# 95% IZ iz simulacij
IZ = quantile(rezultat[,2],probs = c(0.025,0.975))
IZ
```

```
##      2.5%     97.5%
##  2.683462 13.495989
```

```r
hist(rezultat[,2],main="Simulacija za beta=5")
abline(v=mean(rezultat[,2]),col="blue",lwd=2)
abline(v=IZ[1],col="blue",lwd=2,lty=2)
abline(v=IZ[2],col="blue",lwd=2,lty=2)
```



Opazimo, da sta cenilki zelo verjetno pristranski, precenjujeta pravo vrednost. Da bi lahko preverili, ali sta cenilki dosledni (in tudi, da ocenjujeta pravi količini), povečamo velikost vzorca. Zaradi CLI se sedaj vrednosti ocen porazdeljujejo veliko bolj normalno in tudi cenilki nista več toliko pristranski. Interval zaupanja vsebuje populacijsko količino.

Vzorec velikosti 1000 naredimo podobno.

Naloga 6 - Normalna

```
# simulacije intervala zaupanja za mu in sigma

sim = function(n=100){
  x = rnorm(n)
  return(c(mean(x),sd(x)*(n-1)/n))
}

rezultat = replicate(1000,sim(n=100))

# variabilnost mu s simulacijami
sd(rezultat[1,])
```

```
## [1] 0.1019714
```

```
sd(rezultat[2,])
```

```
## [1] 0.06570656
```

```
# variabilnost mu glede na Fisherjevo informacijo (iz enega vzorca)
n=100
set.seed(1)
x = rnorm(n)
```

```
muMLE = mean(x) # muHat
muSdMLE = sqrt((var(x)*(n-1)/n)/n) # sd po MLE za mu

sigmaMLE = sqrt(var(x)*(n-1)/n) # sigmaHat
sigmaSdMLE = sqrt((var(x)*(n-1)/n)/(2*n)) # sd po MLE za sigma

### MU - 95% interval zaupanja
# po MLE
z = qnorm(0.975)
c(muMLE - z*muSdMLE, muMLE + z*muSdMLE)
```

```
## [1] -0.06627404  0.28404878
```

```
# po teoriji
c(muMLE - qt(0.975,df=n-1)*sd(x)/sqrt(n),muMLE + qt(0.975,df=99)*sd(x)/sqrt(n))
```

```
## [1] -0.06933487  0.28710961
```

```
### SIGMA2 - 95% interval zaupanja
# po MLE
z = qnorm(0.975)
c(sigmaMLE - z*sigmaSdMLE, sigmaMLE + z*sigmaSdMLE)^2
```

```
## [1] 0.5926525 1.0354180
```

```
# po teoriji
c(n*sigmaMLE^2/qchisq(0.975,df=n-1),n*sigmaMLE^2/qchisq(0.025,df=n-1))
```

```
## [1] 0.6219297 1.0887169
```

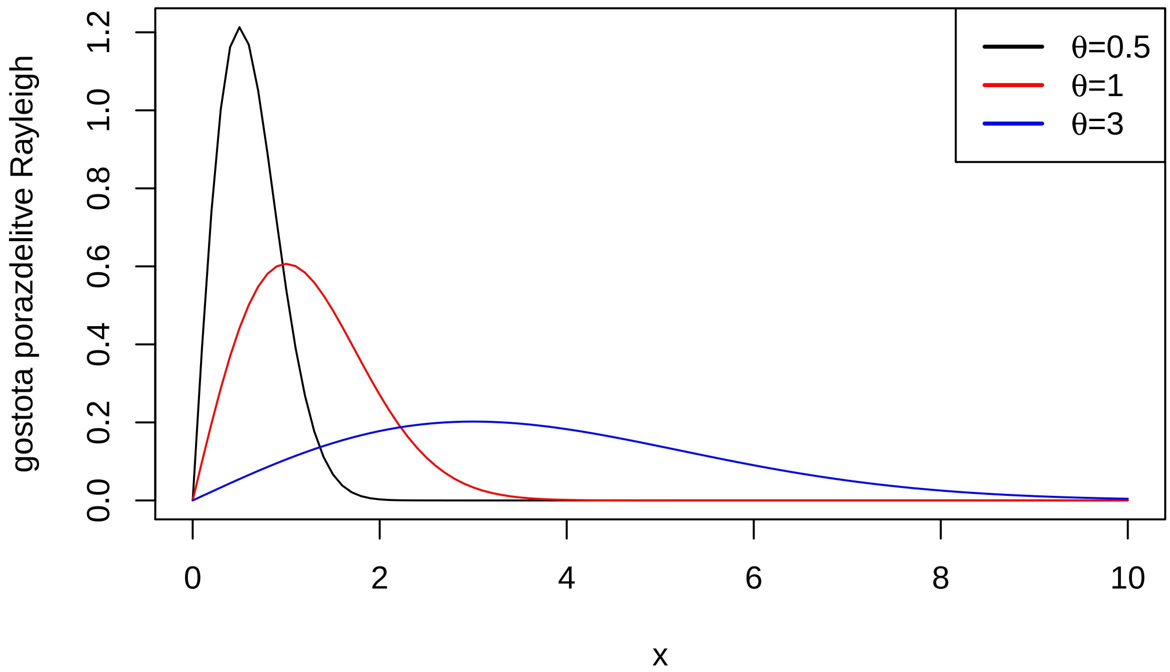
Naloga 8 - Rayleigh

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
### grafična predstavitev porazdelitve Rayleigh
curve(drayleigh(x,0.5),from=0,to=10,ylab="gostota porazdelitve Rayleigh")
curve(drayleigh(x,1),add=TRUE,col="red")
curve(drayleigh(x,3),add=TRUE,col="blue")
legend("topright",legend=c(expression(paste(theta,'=0.5')),
                           expression(paste(theta,'=1')),
                           expression(paste(theta,'=3'))),
       lwd=rep(2,3),col=c("black","red","blue"))
```



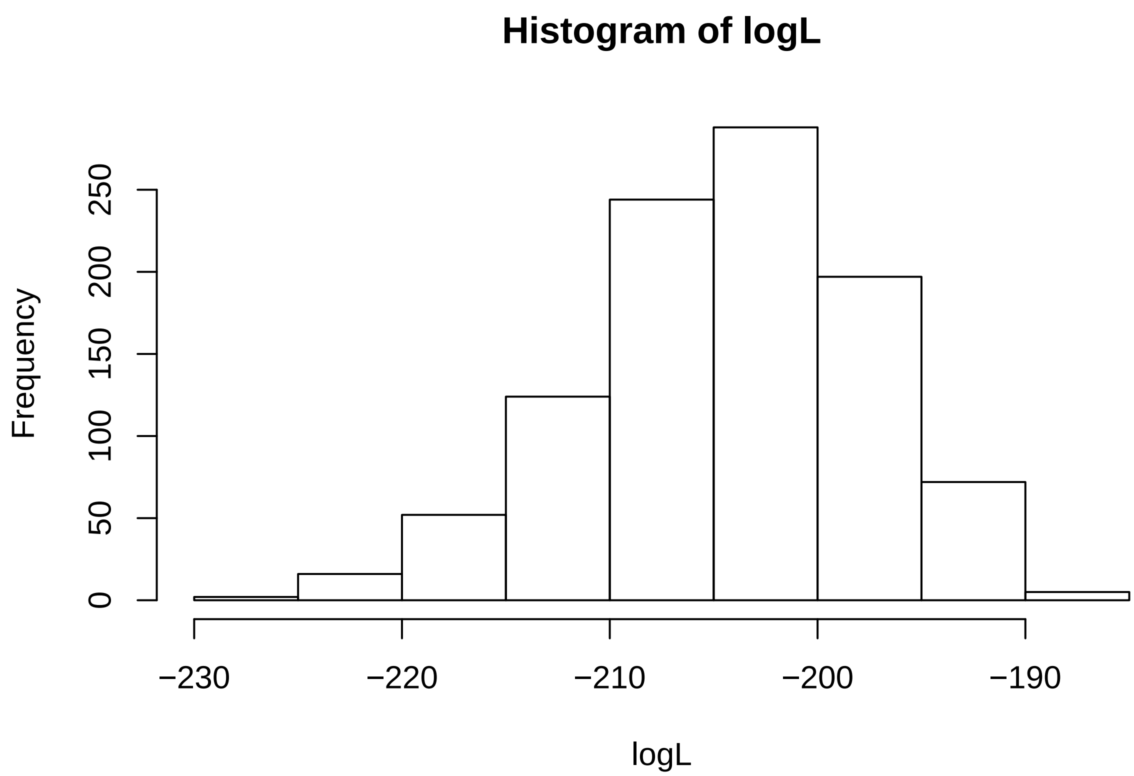

```r
set.seed(42)
#install.packages("VGAM")
library(VGAM)
n <- 100
vzorec <- rrayleigh(n,3)

#### verjetje za vzorec
L = prod(drayleigh(vzorec,3))
L
```

```
## [1] 2.747877e-94
```

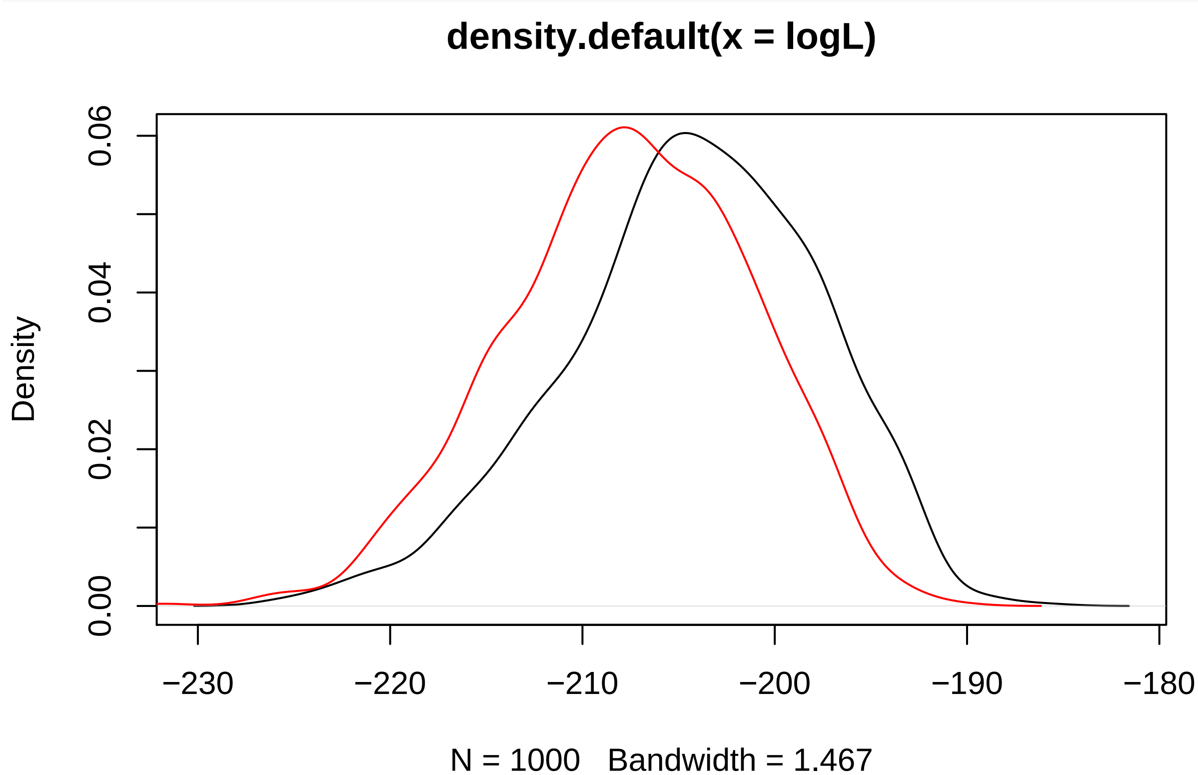
```r
log(L)
```

```
## [1] -215.4322
```

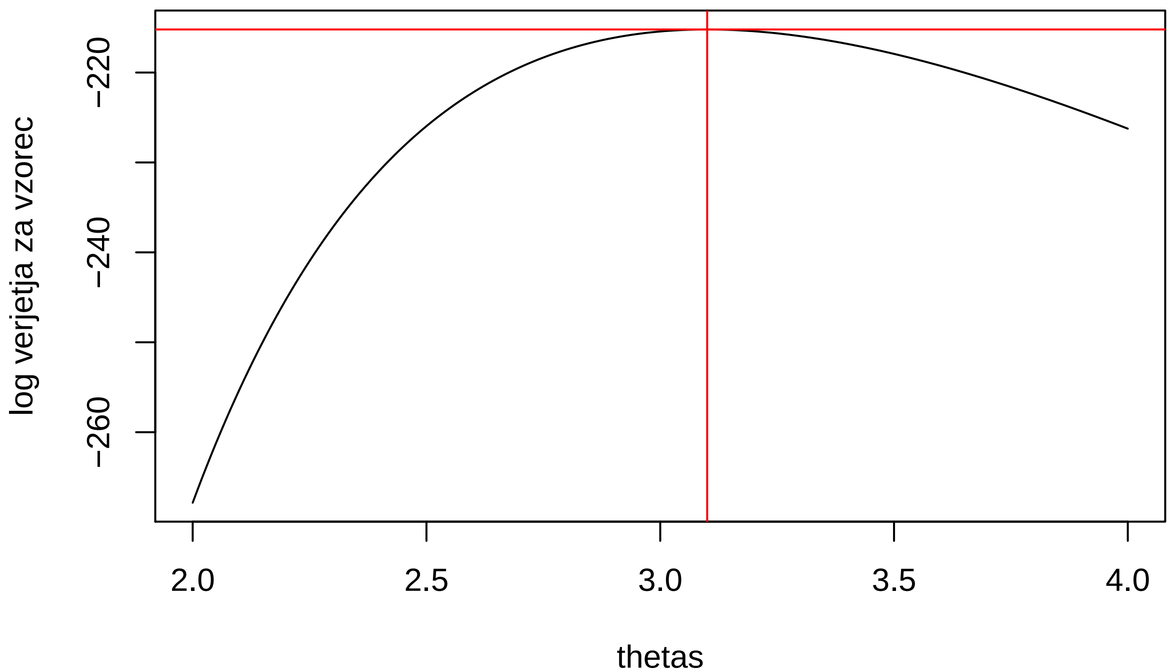
```r
#### izracun ocene theta za "vzorec"
thetaHat = sqrt(sum(vzorec^2)/(2*n))


# simulacija porazdelitve logaritmiranih verjetij za konstanten parameter theta
logVerjetje = function(theta,n){
  vzorec=rrayleigh(n,theta)
  return(log(prod(drayleigh(vzorec,theta))))
}

logL =replicate(1000,logVerjetje(3,100))
hist(logL)
```



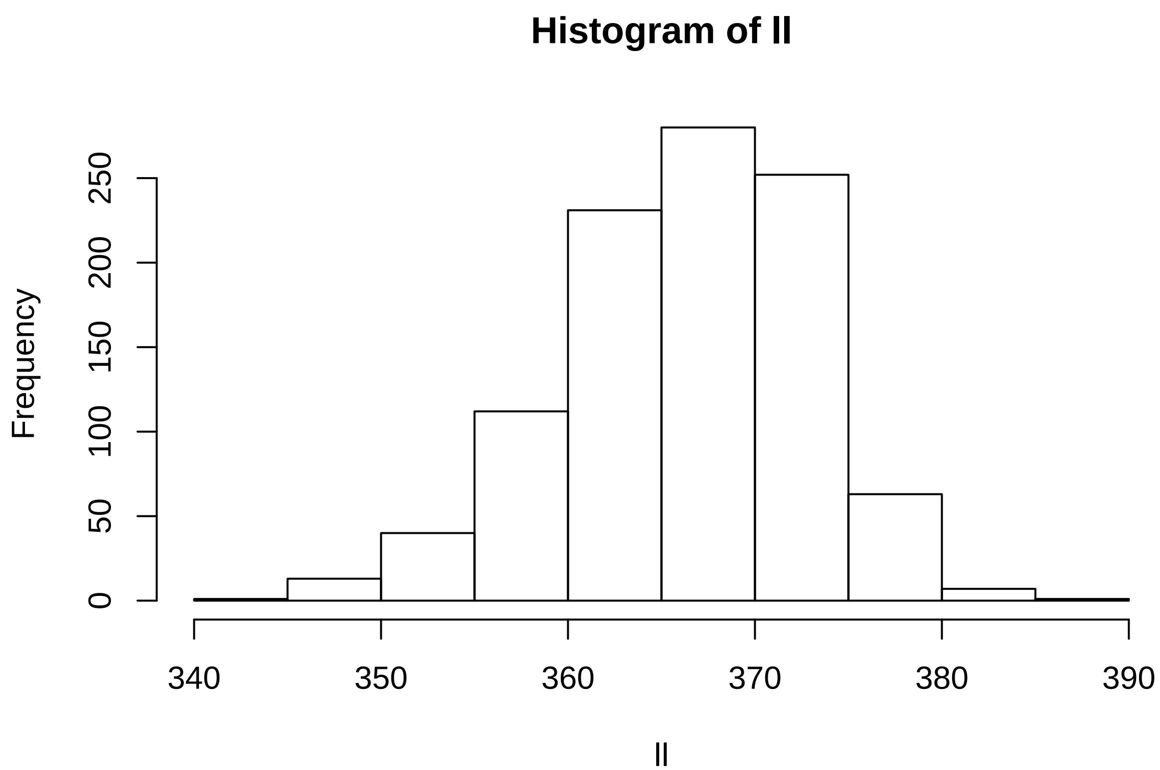

```
plot(density(logL))

# log-verjetja za drug parameter theta (thetaHat = 3.1)
logL1 =replicate(1000,logVerjetje(thetaHat,100))
lines(density(logL1),col="red")
```



```r
# izracun log-verjetja za vzorec pri razlicnih thetah
thetas = seq(2,4,0.01)
logLVzorec = sapply(thetas,FUN = function(x)log(prod(drayleigh(vzorec,x))))
plot(thetas,logLVzorec,type="l", ylab="log verjetja za vzorec") # narisemo log. verjetja
abline(v=thetaHat,col="red")
logLVHat = logLVzorec[max(which(thetas<thetaHat))] # pribl. pokazemo, da je thetaHat ekstrem za ta vzor
abline(h=logLVHat,col="red")
```



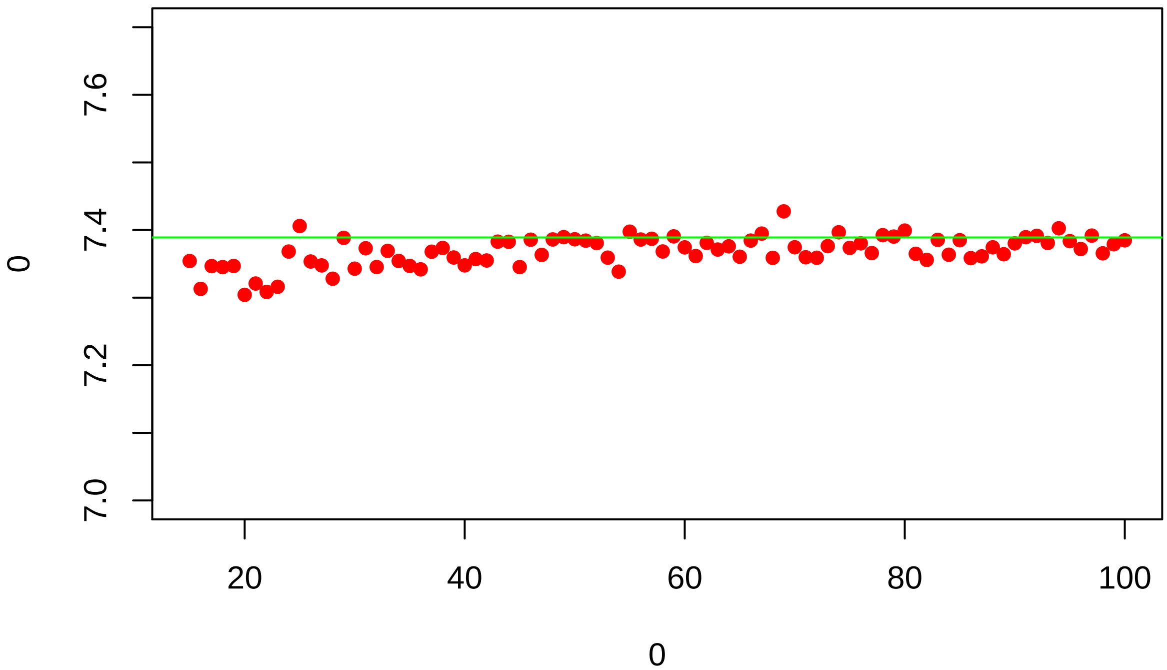

```r
# log-verjetje je lahko pozitivno
# ko so zelo verjetne vrednosti na zelo majhnem intervalu
ll = replicate(1000,logVerjetje(0.01,n))
hist(ll)
```




```r
### interval zaupanja za "vzorec"
SEthetaHat <- sqrt(sum(vzorec^2)/(8*n^2))
#IZ, 95%
c(thetaHat - qnorm(0.975)*SEthetaHat, thetaHat + qnorm(0.975)*SEthetaHat)
```

```
## [1] 2.796541 3.404202
```

```r
# nepristranskost, doslednost cenilke
simulR <- function(n){
  ocene <- NULL
  for(i in 1:1000){
    vzorec <- rrayleigh(n,exp(2))
    ocene <- append(ocene,sqrt(sum(vzorec^2)/(2*n)))
  }
  return(ocene)
}

plot(0,0,"n",xlim=c(15,100),ylim=c(7,7.7))
for(i in 15:100){
    ocene = simulR(i)
    #points(rep(i,length(ocene)),ocene,pch=".")
    points(i,mean(ocene),col="red",pch=16)
}
abline(h=exp(2),col="green")
```



Naloga 9 - računalnik

Funkcijo MLE zapišemo kot:

$$L(x|p_1,p_2,p_3,p_4) = p_1^{n_1} \cdot p_2^{n_2} \cdot p_3^{n_3} \cdot p_4^{n_4},$$

pri pogoju $\sum_k p_k = 1$. Logaritmiramo L:

$$l(x|p_1,p_2,p_3,p_4) = n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log p_4.$$

Z metodo Lagrangevega multiplikatorja (zaradi pogoja) izračunamo maksimalne cenilke:

$$\begin{aligned}
\Lambda(p_1,p_2,p_3,p_4,\lambda) = &\quad n_1 \log p_1 + n_2 \log p_2 + n_3 \log p_3 + n_4 \log p_4 + \lambda(p_1+p_2+p_3+p_4-1)\\
\frac{\partial \Lambda(p_1,p_2,p_3,p_4,\lambda)}{\partial p_k} = &\quad \tfrac{n_k}{p_k} + \lambda = 0\\
\frac{\partial \Lambda(p_1,p_2,p_3,p_4,\lambda)}{\partial \lambda} = &\quad p_1+p_2+p_3+p_4-1 = 0
\end{aligned}$$

Prve štiri enačbe preoblikujemo takole

$$0 = n_k + \lambda p_k$$

in jih seštejemo, da dobimo

$$0 = n + \lambda \cdot 1.$$

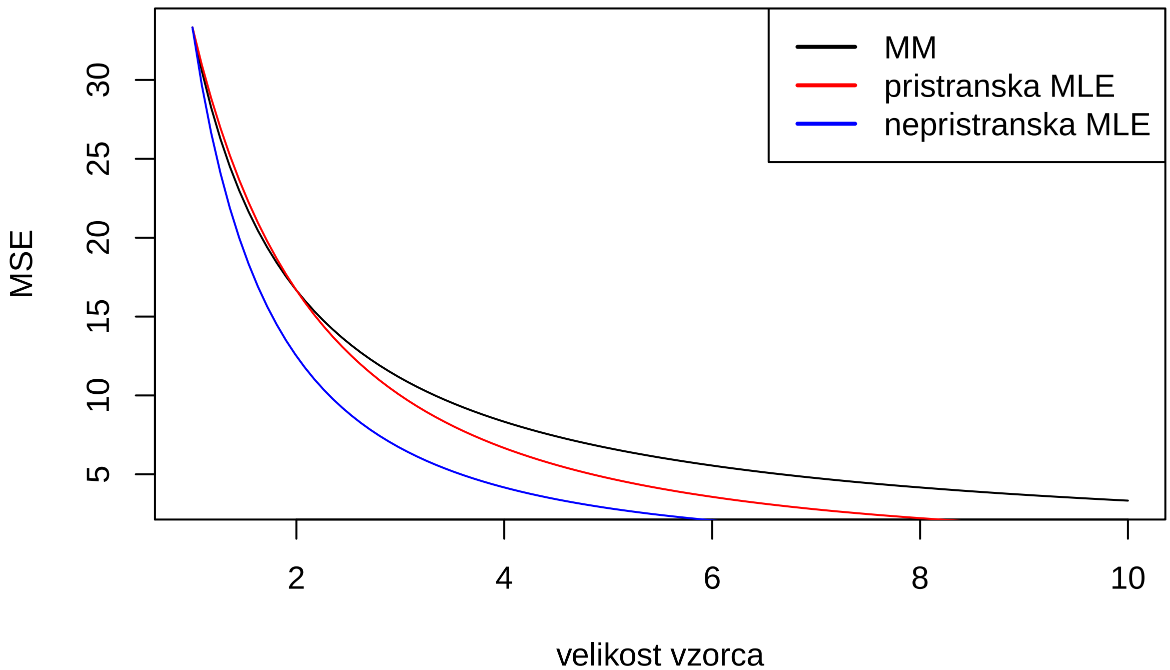
Torej velja:

$$\begin{aligned}
\frac{n_k}{p_k} - n = &\quad 0\\
p_k = &\quad \tfrac{n_k}{n}.
\end{aligned}$$

OPOMBA: Nalogo lahko rešimo tudi brez Lagrangevega multiplikatorja, in sicer tako, da zapišemo $p_4 = 1 - p_1 - p_2 - p_3$.

Naloga 13 - MSE

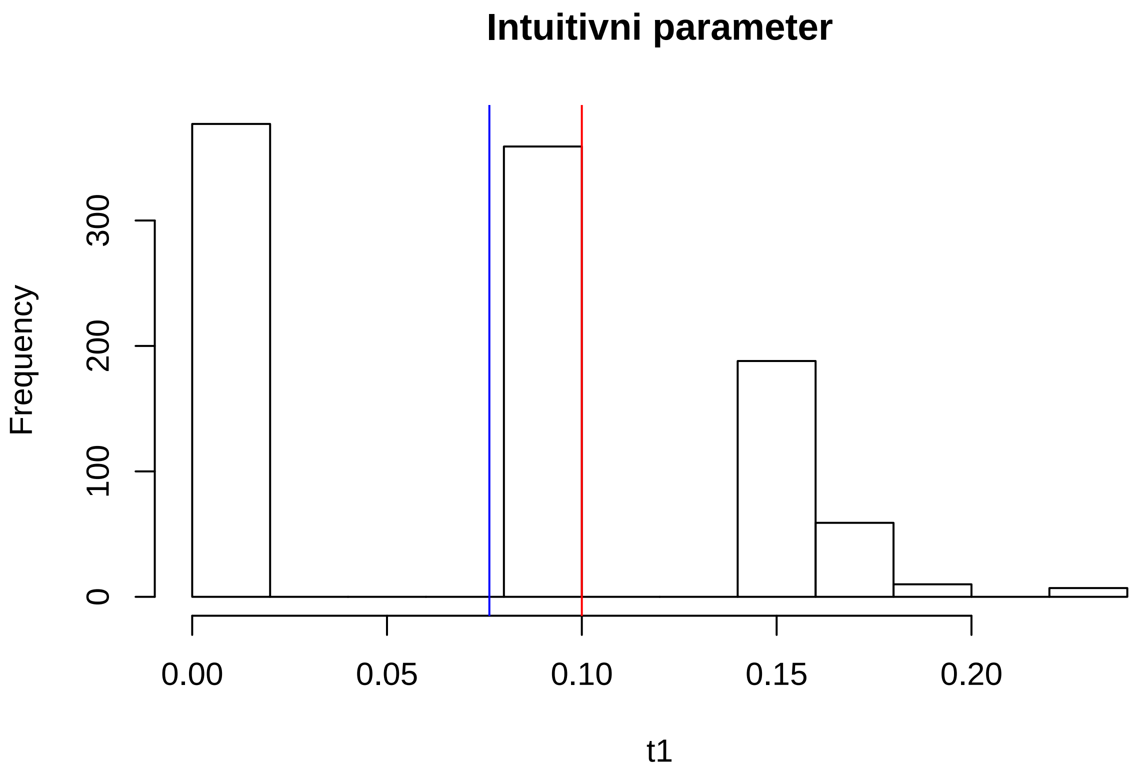
```
a = 10
mseMM = function(a,n){
  a^2/(3*n)}
mseMLEpristranska = function(a,n){
  2*a^2/((n+1)*(n+2))}
mseMLEnepristranska = function(a,n){
  a^2/(n*(n+2))}

curve(mseMM(a=a,n=x),from=1,to=10,xlab="velikost vzorca",ylab="MSE")
curve(mseMLEpristranska(a=a,n=x),from=1,to=10,add=TRUE,col="red")
curve(mseMLEnepristranska(a=a,n=x),from=1,to=10,add=TRUE,col="blue")
legend("topright",legend=c("MM","pristranska MLE",
                           "nepristranska MLE"),
       col=c("black","red","blue"),lwd=rep(2,3))
```



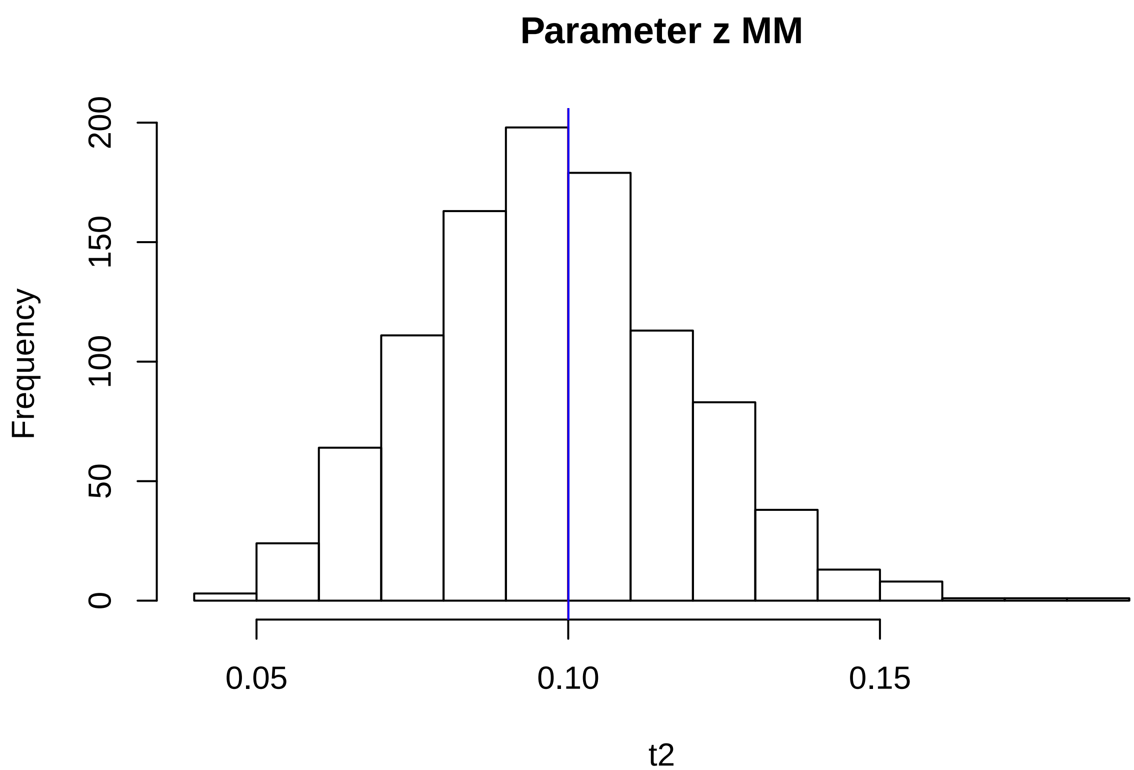
Naloga 17 - HW

```
runs <- 1000
t1 = rep(NA,runs)
t2 = rep(NA,runs)
t3 = rep(NA,runs)
n <- 100
theta=0.1
for(i in 1:runs){
  vzorec <-  sample(-1:1,replace=T,size=n,
        prob=c(theta^2,2*theta*(1-theta),(1-theta)^2))
  t1[i] <- sqrt(sum(vzorec==-1)/n) #intuitivna
  t2[i] <- (1-mean(vzorec))/2 # MM
  t3[i] <- (2*sum(vzorec==-1)+sum(vzorec==0))/(2*n) # MLE
}
#op =par(mfrow=c(2,1))
```

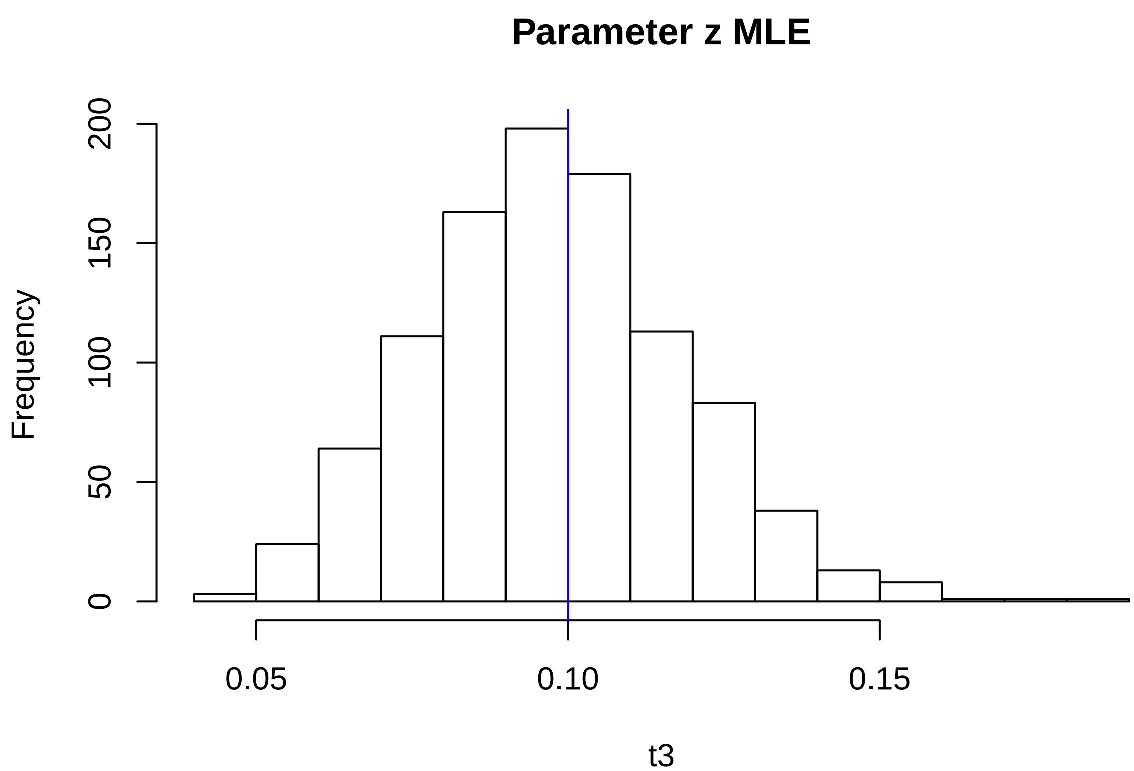
```r
hist(t1,main="Intuitivni parameter")
abline(v=theta,col="red")
abline(v=mean(t1),col="blue")
```




```r
hist(t2,main="Parameter z MM")
abline(v=theta,col="red")
abline(v=mean(t2),col="blue")
```




```r
hist(t3,main="Parameter z MLE")
abline(v=theta,col="red")
abline(v=mean(t3),col="blue")
```




```r
#par(op)
sd(t1)
```

```
## [1] 0.06369227
```

```r
sd(t2)
```

```
## [1] 0.02084427
```

```r
sd(t3)
```

```
## [1] 0.02084427
```