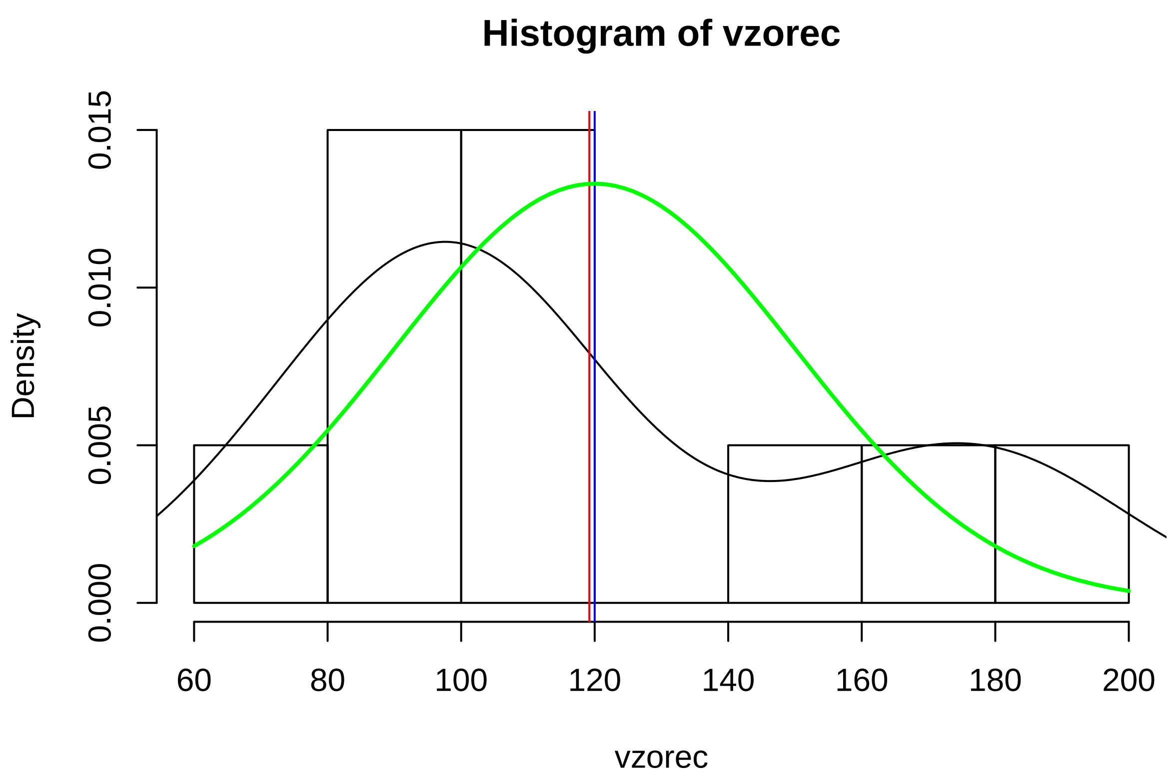
Rešitve - pričakovana vrednost, varianca

Nataša Kejžar

Naloga 1

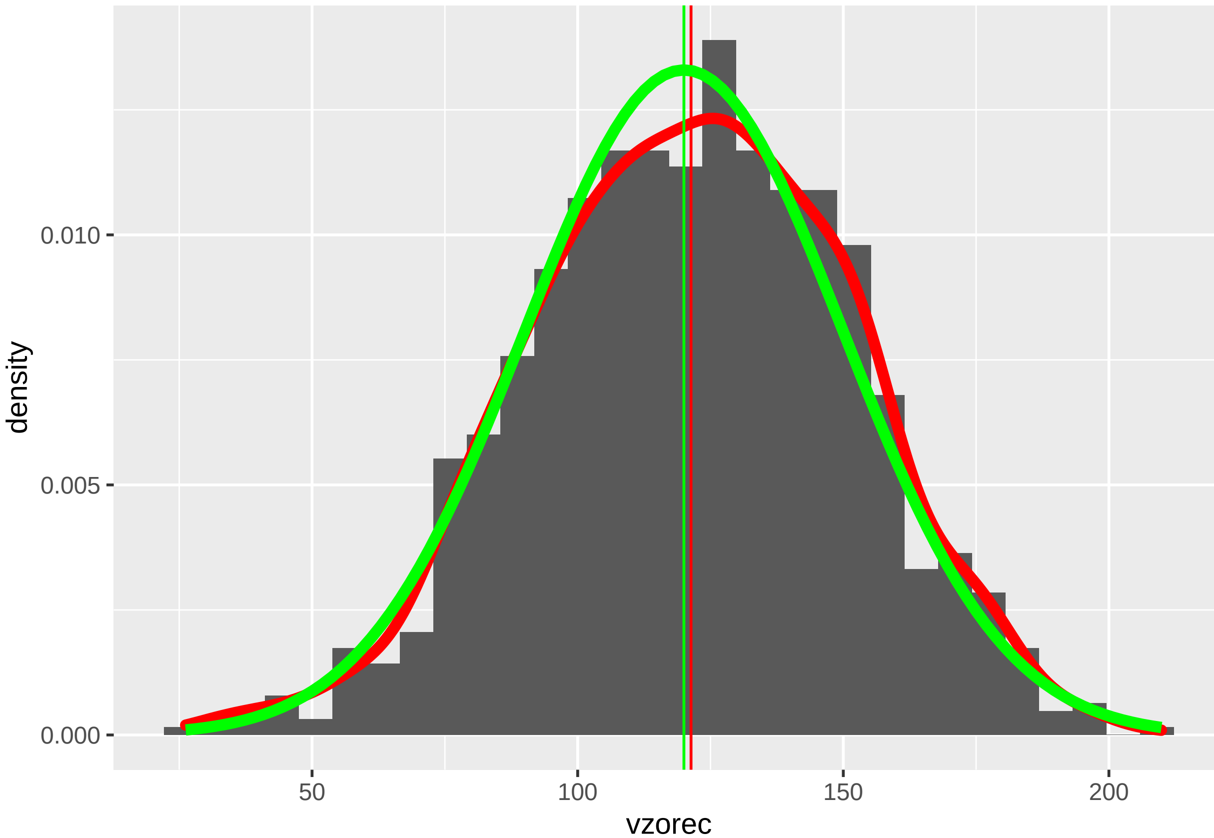
```
vzorec = rnorm(10,mean=120,sd=30)
hist(vzorec,freq=FALSE)
lines(density(vzorec))
abline(v=mean(vzorec),col="red")
abline(v=120,col="blue")
curve(dnorm(x,mean=120,sd=30),add=TRUE,col="green",lwd=2)

# se z ggplot
#install.packages("ggplot2")
library(ggplot2)
```



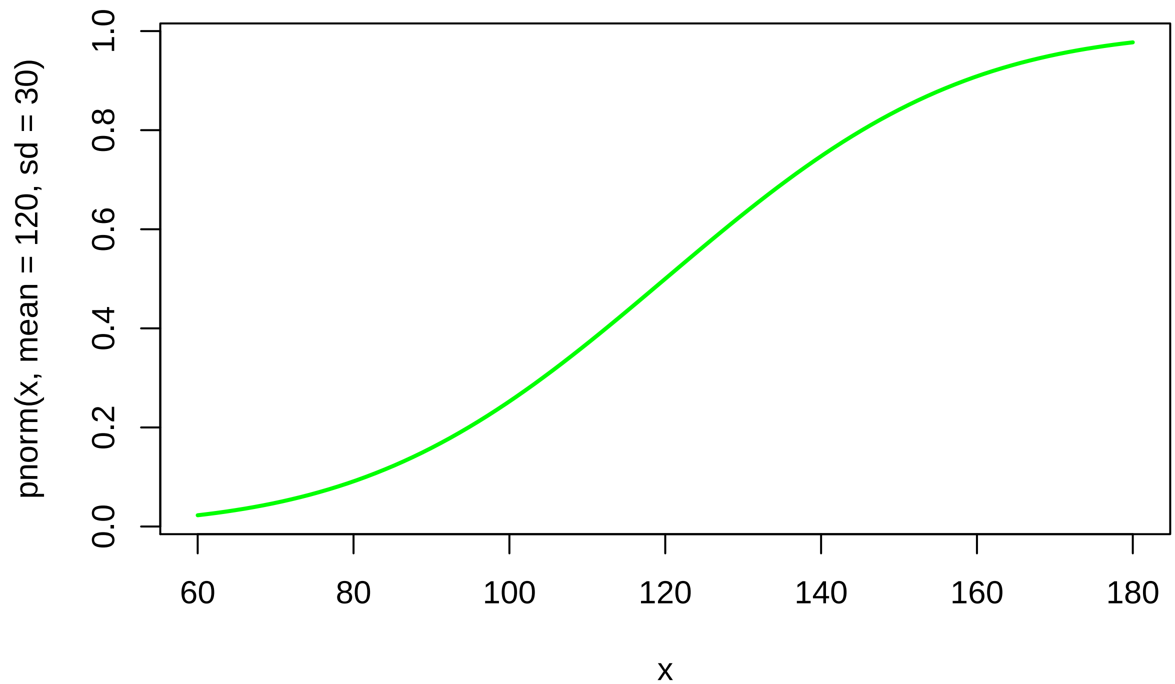

```
vzorec = rnorm(1000,mean=120,sd=30)
df = data.frame(vzorec=vzorec) # za ggplot morajo biti podatki v podatkovnem okvirju

ggplot(df,aes(x=vzorec)) +
  geom_histogram(aes(y=..density..)) +
  geom_density(col="red",size=2) +
  geom_vline(xintercept=mean(vzorec),color="red") +
```

```
  stat_function(fun = dnorm, args = list(mean = 120, sd = 30),colour="green",size=2) +
  geom_vline(xintercept=120,color="green")
```

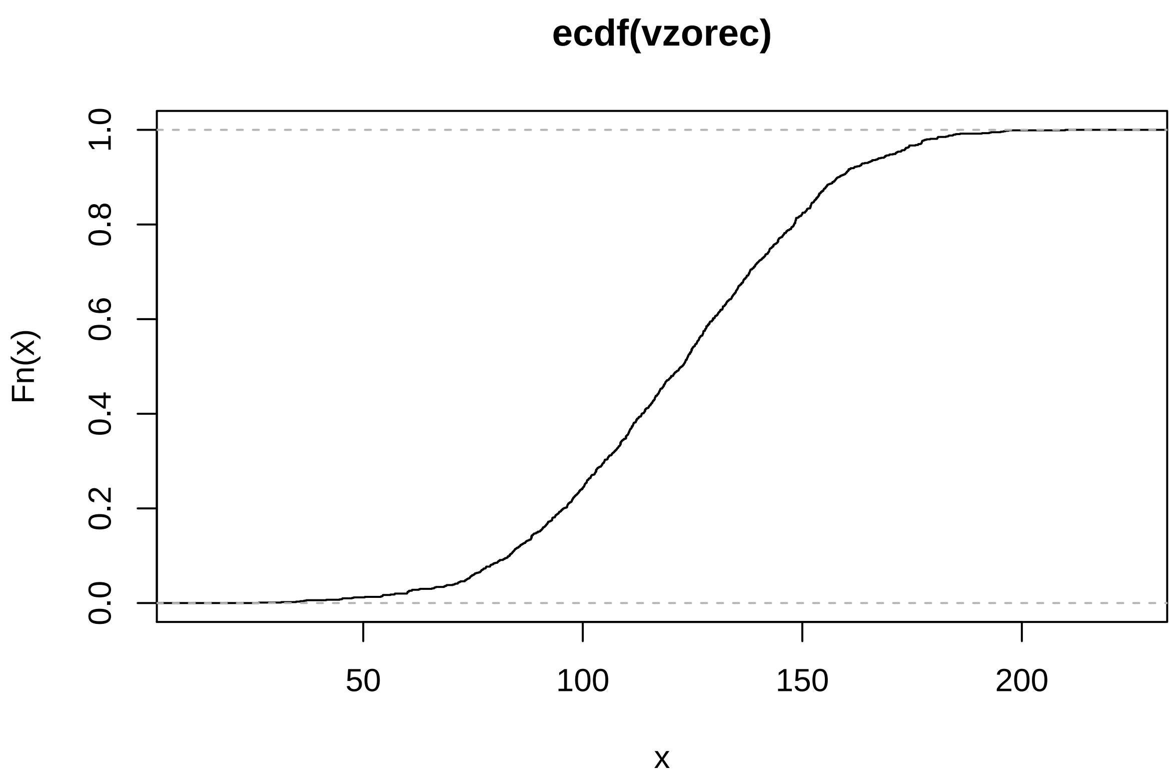
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



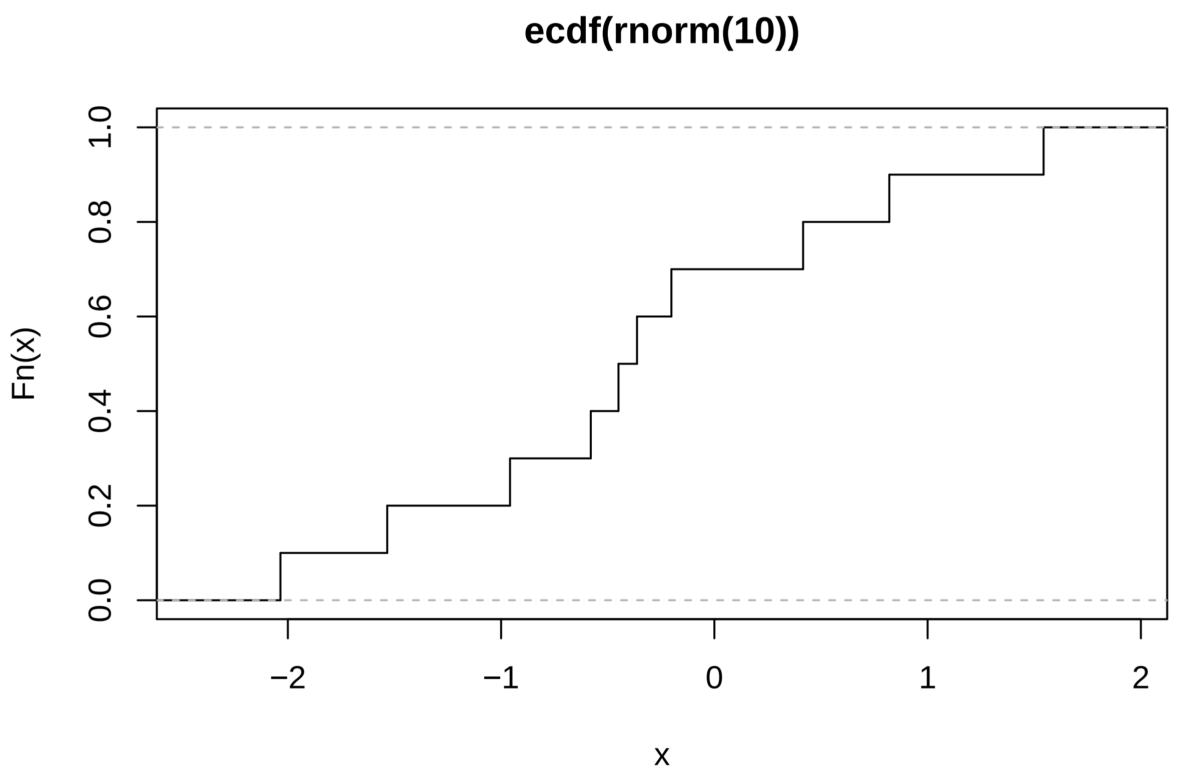

```
# izris kumulativne porazdelitvene funkcije za populacijo
curve(pnorm(x,mean=120,sd=30),col="green",lwd=2,from=60,to=180)
```




```
# za vzorec
x = ecdf(vzorec)
plot(x)
```




```
plot(ecdf(rnorm(10)),verticals=TRUE,do.points=FALSE)
```



Naloga 2 - novorojenčki

```
sigma = (3300-2500)/qnorm(0.05)
# IQR
qnorm(c(0.25,0.75),mean=3300,sd=486.4)
```

```
## [1] 2971.928 3628.072
```

```
pnorm(2900,mean=3300,sd=486.4)
```

```
## [1] 0.2054336
```

```
# izracun vzorcnega percentila
vzorec = rnorm(50,mean=3300,sd=486.4)
mean(vzorec <= 2900)
```

```
## [1] 0.24
```

```
# IZ za povprecje
vzorec = rnorm(500,mean=3300,sd=486.4)
xbar = mean(vzorec)
se = sd(vzorec)/sqrt(length(vzorec))

# spodnja meja
xbar + qnorm(0.05)*se
```

```
## [1] 3235.505
```

```
# zgornja meja
xbar + qnorm(0.95)*se
```

```
## [1] 3312.961
```

```r
# drugace
qnorm(c(0.05,0.95),mean=xbar,sd=se)
```

```
## [1] 3235.505 3312.961
```

Naloga 3 - teoretične nalogice

Naloga a: izrazimo $var(x) = E[(X - E(X))^2]$, razpišemo in pokrajšamo.

Naloga b: $var(2X) = var(X + X)$, poračunamo varianco in kovarianco.

Naloga c: razpišemo $\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2$, upoštevamo, da pri končni populaciji spremenljivke X velja:

- $\frac{1}{N}\sum_{i=1}^{N} x_i = E(X)$ in $\frac{1}{N}\sum_{i=1}^{N} x_i^2 = E(X^2)$
- $$var(X) = E(X^2) - E(X)^2$$

Naloga 4 - končna populacija diskretne porazdelitve

```r
N = 1500
x = 1:5
px = c(1/15,1/5,4/15,2/5,1/15)

# po definiciji
Ex = sum(x*px)

# cela populacija
populacija = sapply(x,FUN=function(x){rep(x,px[x]*N)})
populacija = unlist(populacija)
mean(populacija)
```

```
## [1] 3.2
```

```r
# varianca na 2 načina
var(populacija)*(N-1)/N
```

```
## [1] 1.093333
```

```r
sum((x - Ex)^2*px)
```

```
## [1] 1.093333
```

```r
# vzorcenje s ponavljanjem na 2 načina
sample(populacija, size=15,replace=T)
```

```
##  [1] 3 4 1 2 4 3 2 4 3 4 3 4 4 5 4
```

```r
sample(x,size=15,replace=T,prob = px)
```

```
##  [1] 3 2 4 2 3 4 1 3 4 4 2 4 4 1 5
```

Naloga 5 - Bernoulli -> Binomska

```r
x=0:1
px=c(0.15,0.85)
sample(0:1,size=100,replace=T,prob = px)
```

```
##   [1] 1 0 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0
##  [36] 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1
##  [71] 1 1 0 1 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1
```

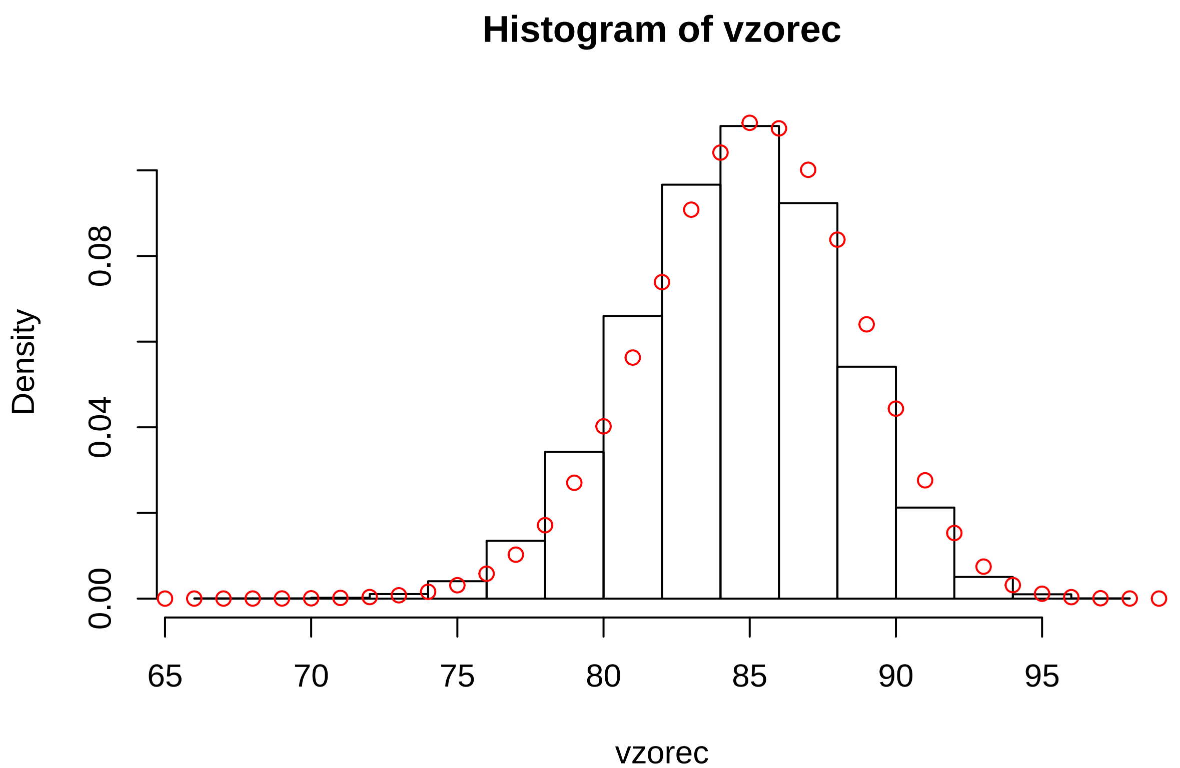
```r
as.numeric(runif(100) < 0.85)
```

```
##   [1] 1 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 1 0 1 1 1 1 1 1 1 1 0 0
##  [36] 1 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1
##  [71] 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1
```

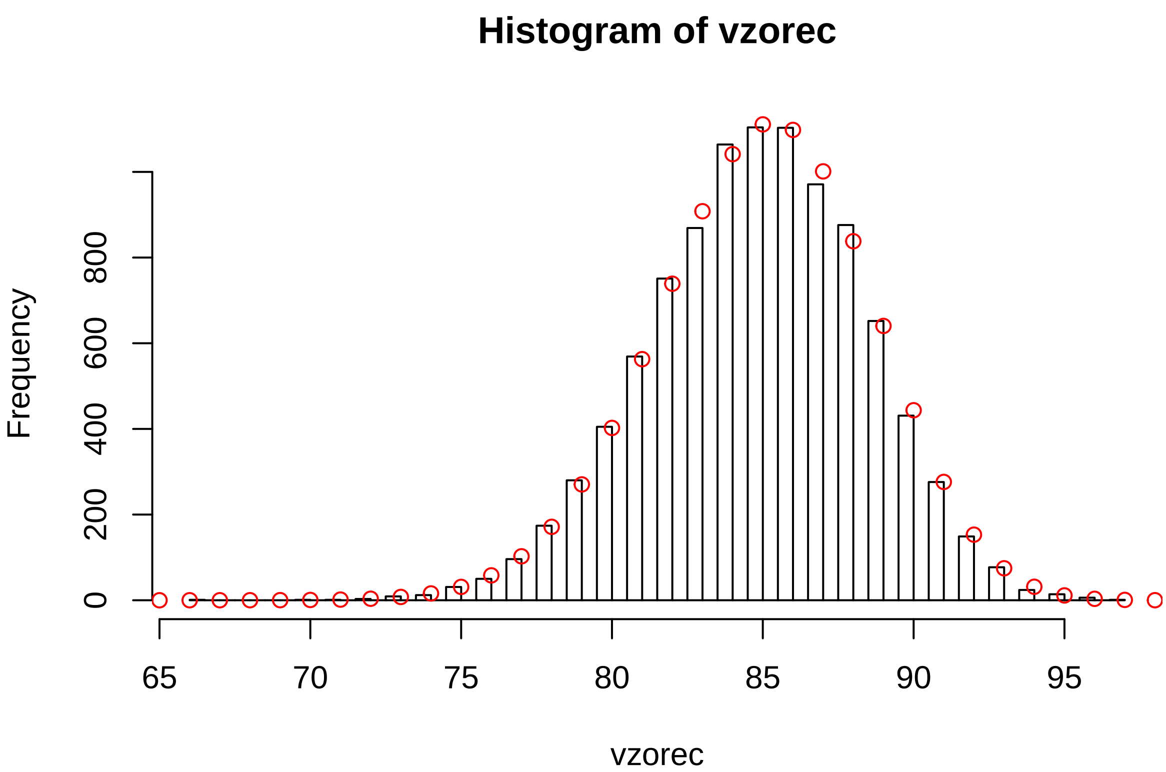
```r
Ex = sum(x*px)
sum((x - Ex)^2*px)
```

```
## [1] 0.1275
```

```r
# graficno
y = function()
{sum(sample(0:1,size=100,replace=T,prob = px))}
vzorec = replicate(10000,y())

# na y osi so "gostote"
x=0:100
hist(vzorec,freq=FALSE)
points(x,dbinom(x,size=100,prob=0.85),col="red")
```




```r
# na y osi so frekvence
hist(vzorec,breaks=100)
points(x,dbinom(x,size=100,prob=0.85)*10000,col="red")
```



Naloga 6 - hrčki

```
# kovarianca m.prasickov
n=1000
xEU = rnorm(n,mean=500,sd=100)
xZDA = rnorm(n,mean=500,sd=200)
cov(xEU,xZDA)
```

```
## [1] -510.7997
```

```
# kovarianca je precej variabilna
# zaradi velikih standardnih odklonov
# zato uporabimo funkcijo za korelacijo
cor(xEU,xZDA)
```

```
## [1] -0.02558151
```

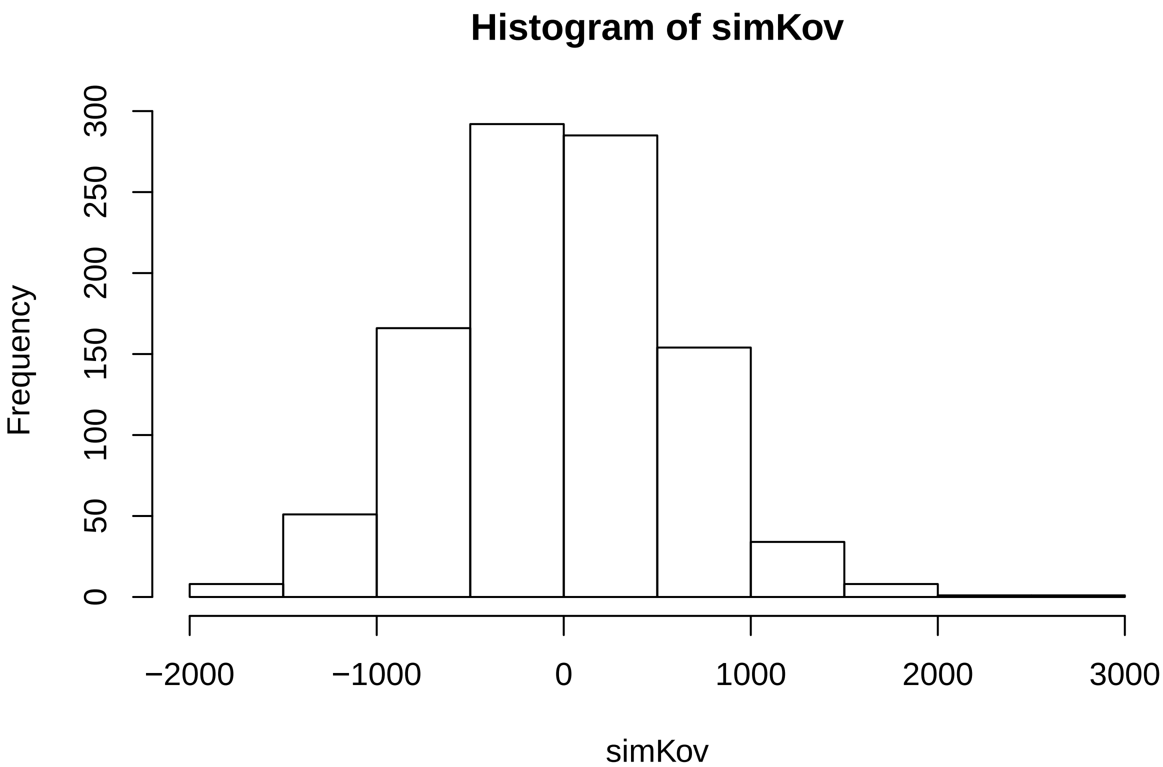
```
cor(xEU,(xEU-xZDA))
```

```
## [1] 0.4591777
```

```
# da kovarianco bolje ocenimo, naredimo simulacijo več vzorcev
# inicializiramo vektor simuliranih kovarianc
simKov = NULL
for(i in 1:1000){
  xEU = rnorm(n,mean=500,sd=100)
  xZDA = rnorm(n,mean=500,sd=200)
  kov = cov(xEU,xZDA) # kov = cov(vzorecX,vzorecY-vzorecX)
  simKov = c(simKov,kov)
}
mean(simKov)
```

```
## [1] -27.72307
```

```r
hist(simKov) # narisemo simulirane vrednosti
```

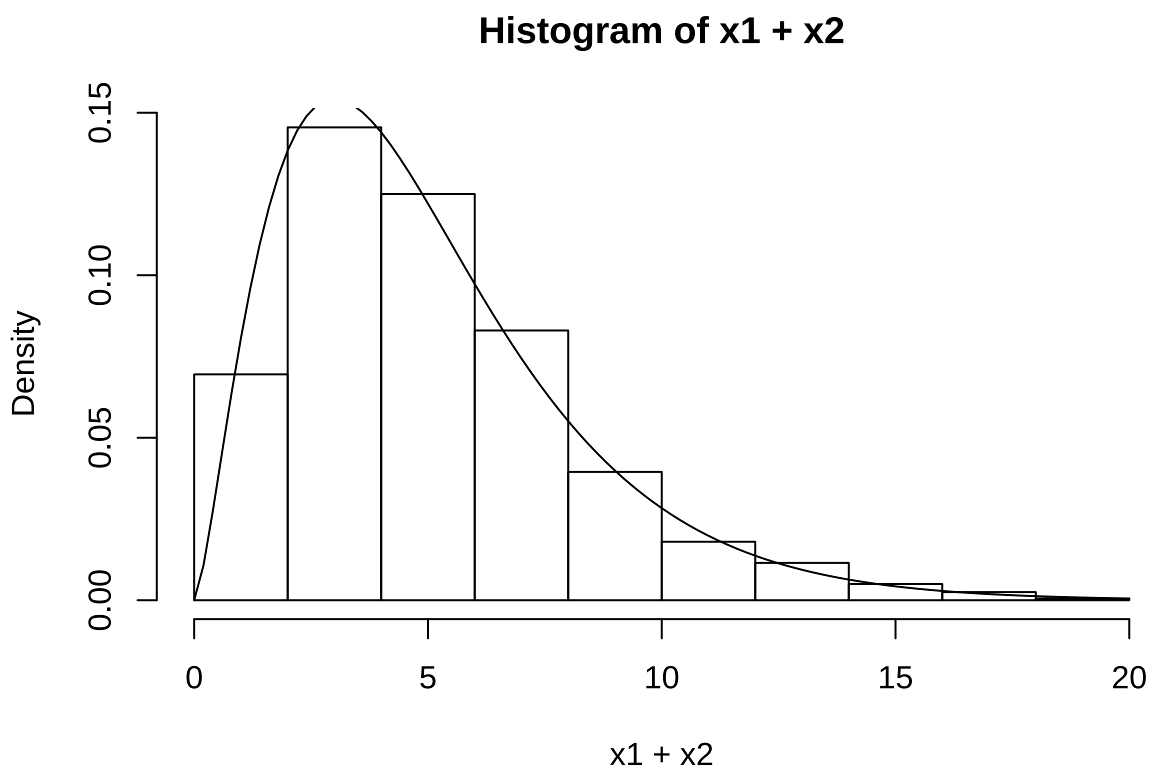


Za teoretični prikaz, da je korelacija neničelna, uporabimo lastnost

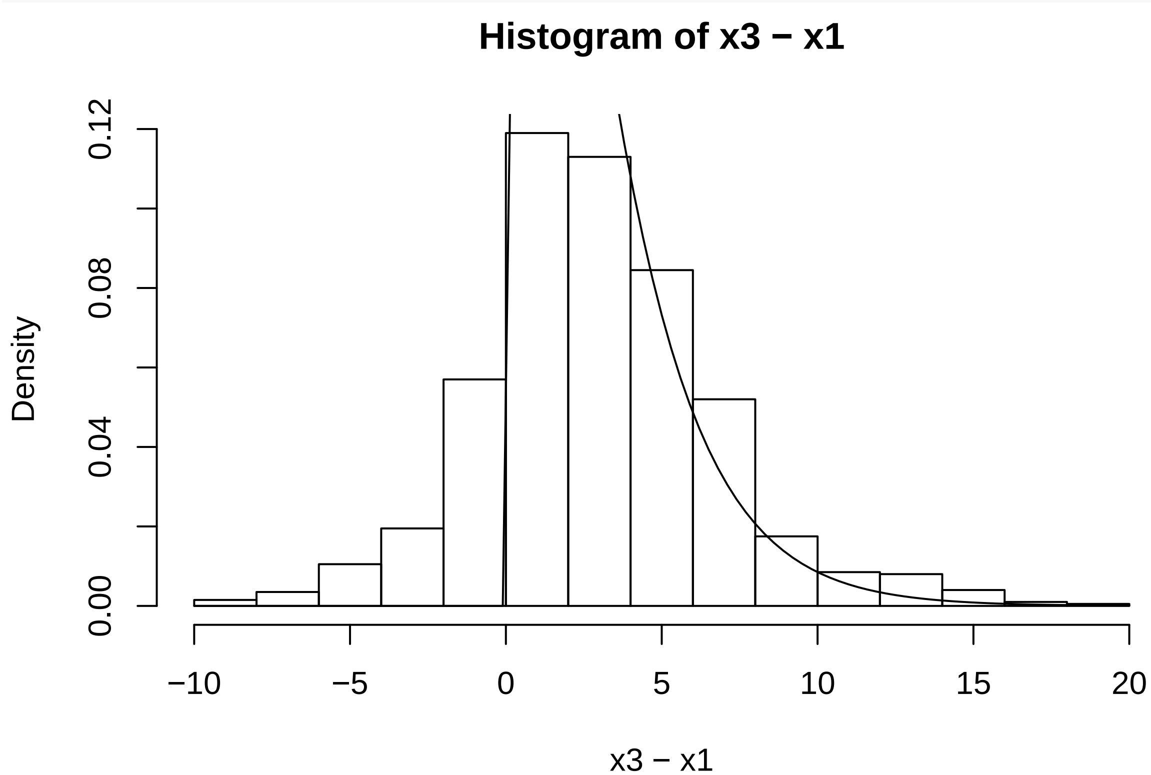
$$cov(X, Y - Z) = cov(X, Y) - cov(X, Z)$$

Naloga 7 in 8

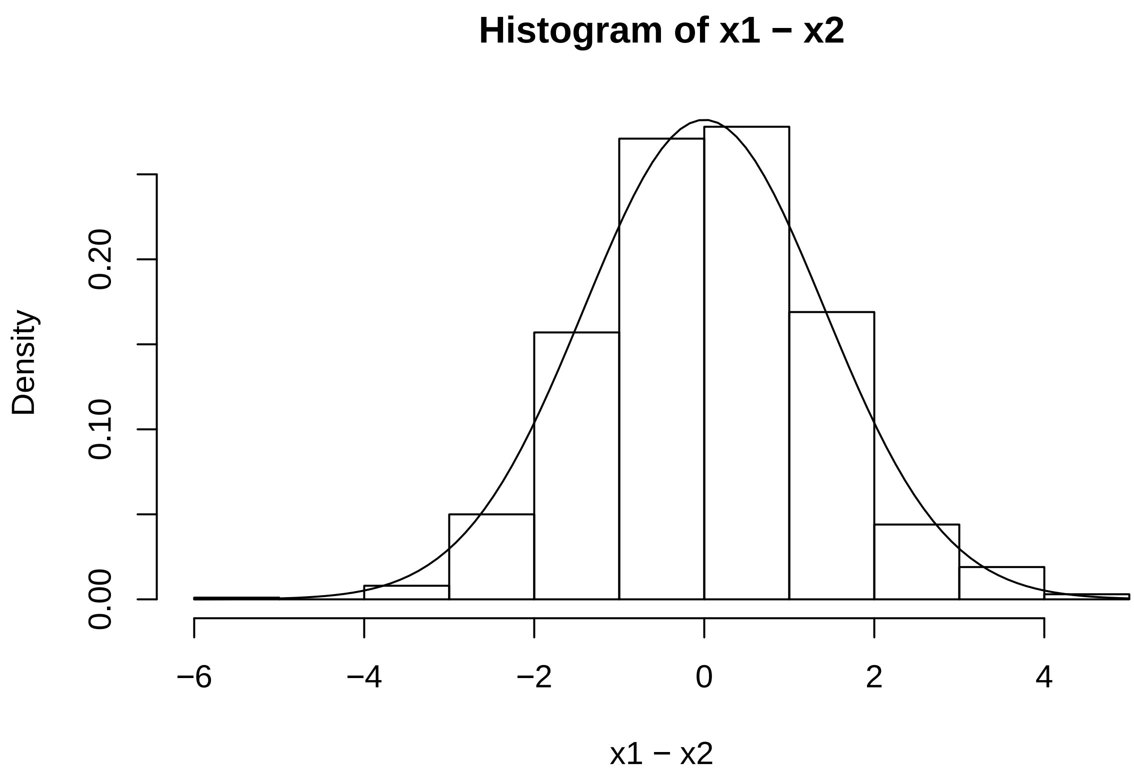
```r
# generiramo 2 spremenljivki, ju narisemo
n = 1000
x1 = rchisq(n, df=2); x2 = rchisq(n, df=3)
hist(x1+x2,freq=FALSE)
curve(dchisq(x,df=5),add=TRUE)
```




```
# izris histograma za razliko dveh spremenljivk
x3 = rchisq(n, df=5)
# ali pa x3 = x1+x2 - KJE JE RAZLIKA?
hist(x3-x1,freq=FALSE)
curve(dchisq(x,df=3),add=TRUE)
```



```r
# histogram razlike dveh normalnih spremenljivk
x1 = rnorm(n); x2 = rnorm(n)
hist(x1-x2,freq=FALSE)
curve(dnorm(x,sd=sqrt(2)),add=TRUE)
```




```r
# simetricno glede na 0, variance neodvisnih spremenljivk se sestevajo
```

Naloga 9

```r
set.seed(1)
 x = sample(x = 1:10, size=10,replace=TRUE)
 N=length(x)
 # (a) povprecje
 povp = 1/N * sum(x)
 # (b) pozitivni odmiki
 odmiki = x - povp
 sum(odmiki[odmiki > 0])
```

```
## [1] 11.6
```

```r
 # negativni odmiki
 sum(odmiki[odmiki < 0])
```

```
## [1] -11.6
```

```r
 # (c) varianca
 varianca=1/N * sum((x-povp)^2)
 1/N*sum(x^2) - povp^2
```

```
## [1] 7.09
```

```r
# (d) vsota kvadriranih odklonov
VKO = function(a,x){
    return(sum((x-a)^2))
}
# (e)
a = 4:10
vko = NULL
for(i in a)
  vko = c(vko,VKO(i,x))
# vko na drugacen nacin: vko = sapply(a,VKO,x=x)
# risanje
plot(a,vko,type="l",lwd=2)
abline(h=N*varianca,col="red")
```