Rešitve - statistični testi

Nataša Kejžar

Naloga 1 - mladi kadilci

a. mladi kadilci v Sloveniji (do katerega leta?) $\Rightarrow$ Anketo naredijo med npr. 25–30 letniki-kadilci
b. $\mu = 13$, enostavna
c. $\mu \neq 13$, sestavljena, dvostranska
d. Testna statistika naj bo vzorčno povprečje. Ker vemo, da so enote porazdeljene po normalni porazdelitvi, potrebujemo še vrednost σ.

$$\begin{aligned}
T &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
X &\sim N(\mu, \sigma^2) \\
E(T) &= \mu \\
var(T) &= \frac{\sigma^2}{n}
\end{aligned}$$

σ določimo iz podatka, da 80% mladih začne kaditi do 16.leta starosti. Torej

$$\begin{aligned}
P(X > 16) &= 0.2 \\
P(Z > (16-13)/\sigma) &= 0.2 \\
3/\sigma &= 0.84 \quad \# \text{ qnorm}(0.8) \\
\sigma &= 3.56
\end{aligned}$$

e. $(-\infty, 12, 3), (13, 7, \infty)$ Uporabimo

```
meja1 = qnorm(0.025,13,3.56/sqrt(100)) # in
meja2 = qnorm(0.975,13,3.56/sqrt(100))
meja1
```

```
## [1] 12.30225
```

```
meja2
```

```
## [1] 13.69775
```

f.

```
# moč testa
pnorm(meja1,mean=14,sd=3.56/sqrt(100)) +
  pnorm(meja2,mean=14,sd=3.56/sqrt(100),lower.tail=FALSE)
```

```
## [1] 0.8020672
```

g.

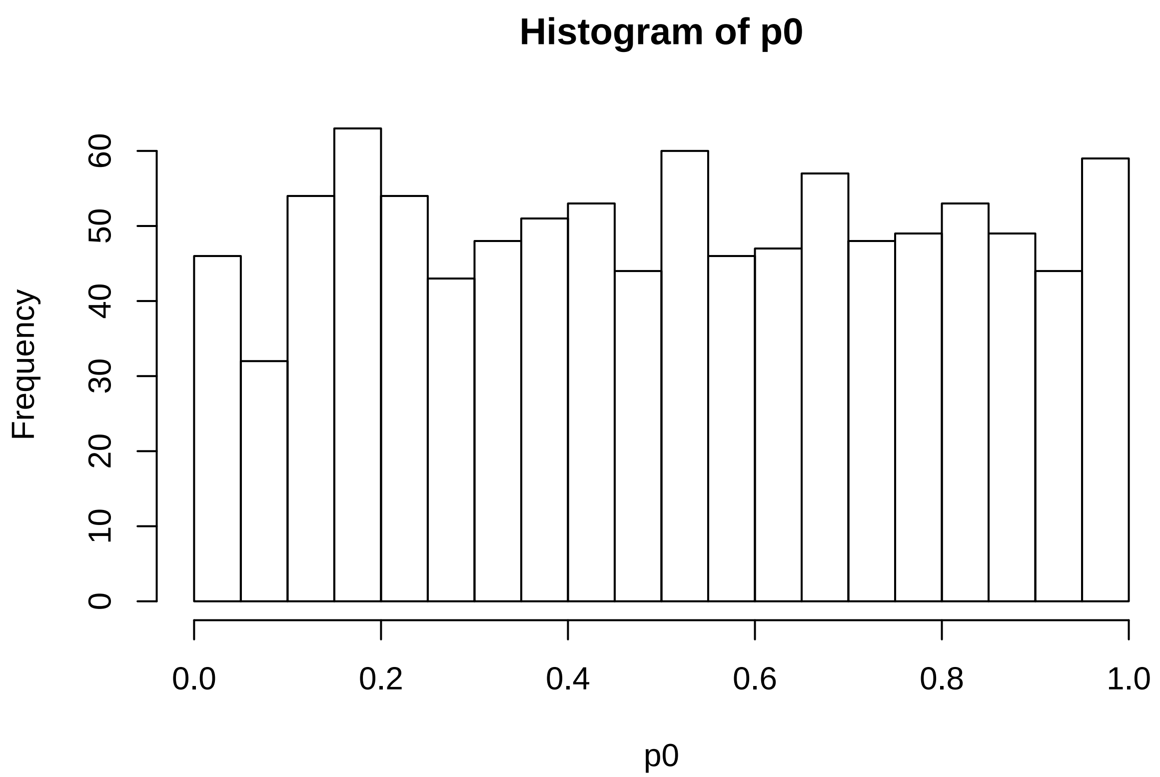
```r
pvrednost <- function(t,mu,SE){
  if(t > mu)
    2*pnorm(t,mu,SE,lower.tail=FALSE)
  else
    2*pnorm(t,mu,SE)
}

# simulacije
simuliraj <- function(ponovi,n,muVzorec,sigmaVzorec,mu0,sigma0){
  pVrednosti = rep(NA,ponovi)
  for(i in 1:ponovi){
    vzorec = rnorm(n,muVzorec,sigmaVzorec)
    pVrednosti[i] = pvrednost(mean(vzorec),mu0,sigma0/sqrt(n))
  }
  pVrednosti
}

s = 3.56
simuliraj(1,100,13,s,13,s)
```

```
## [1] 0.5845143
```

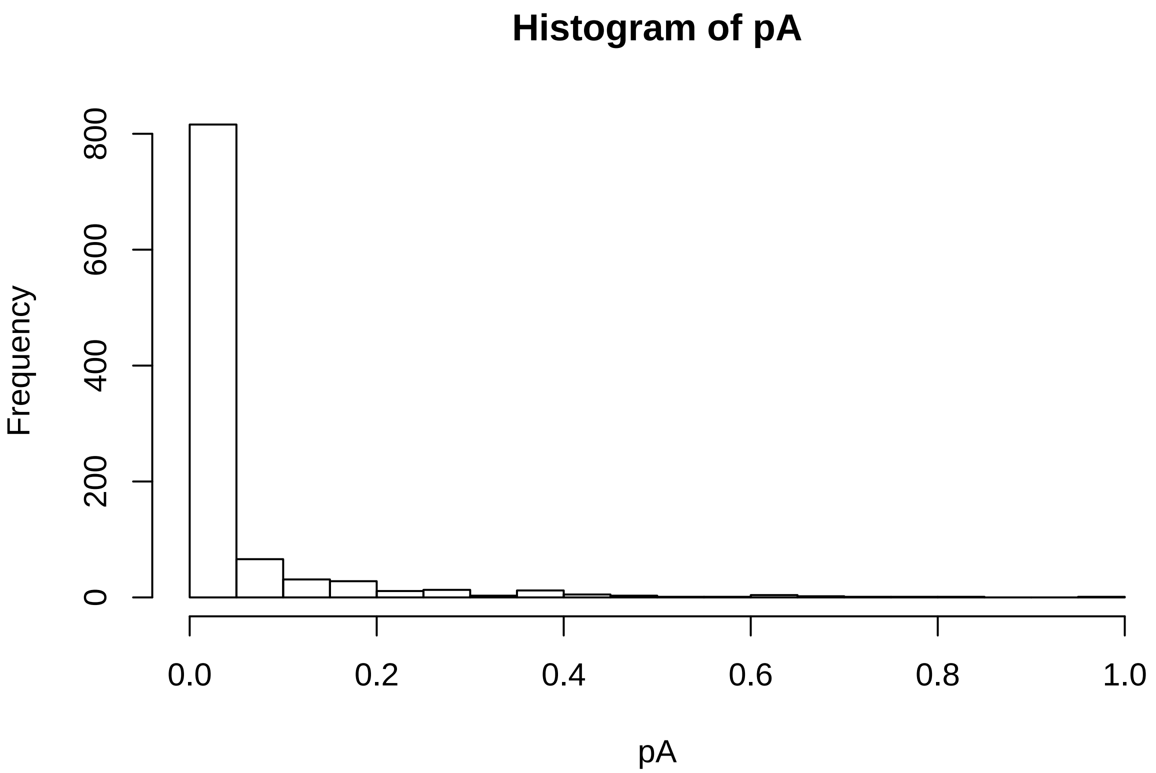
```r
simuliraj(1,100,14,s,13,s)
```

```
## [1] 0.0001513663
```

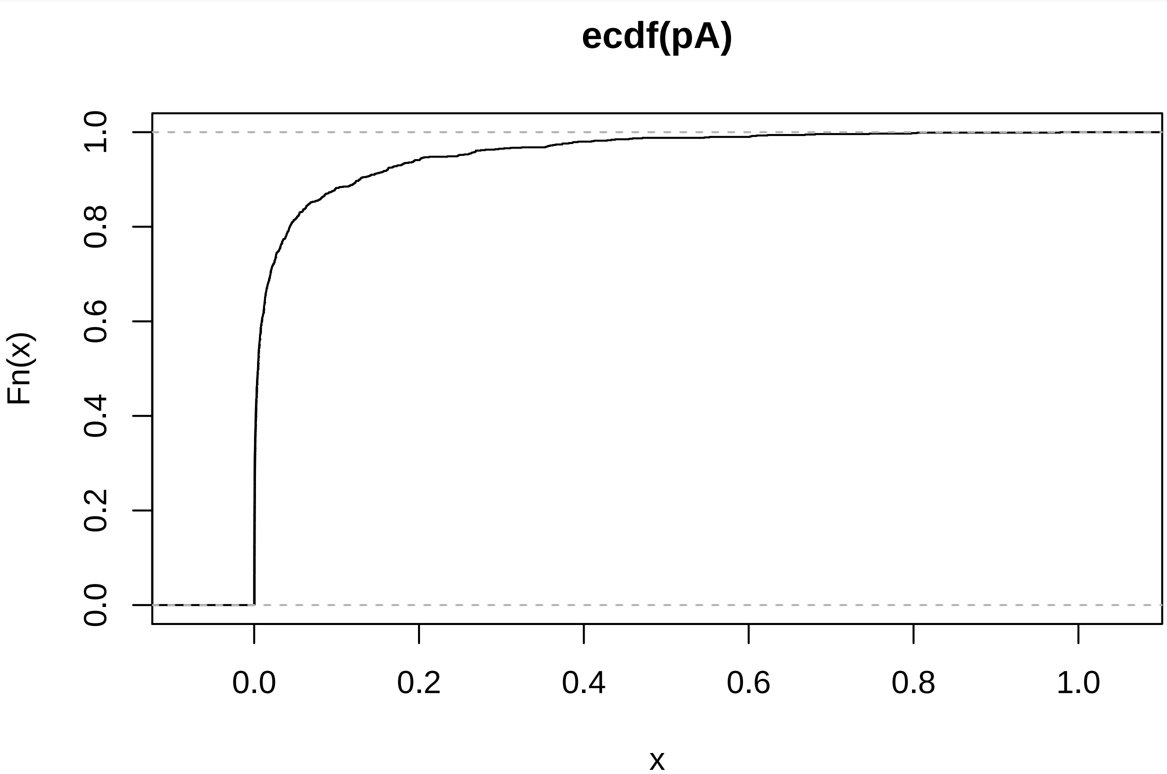
```r
p0 = simuliraj(1000,100,13,s,13,s)
hist(p0,breaks=30)
```




```r
pA = simuliraj(1000,100,14,s,13,s)
hist(pA,breaks=30)
```



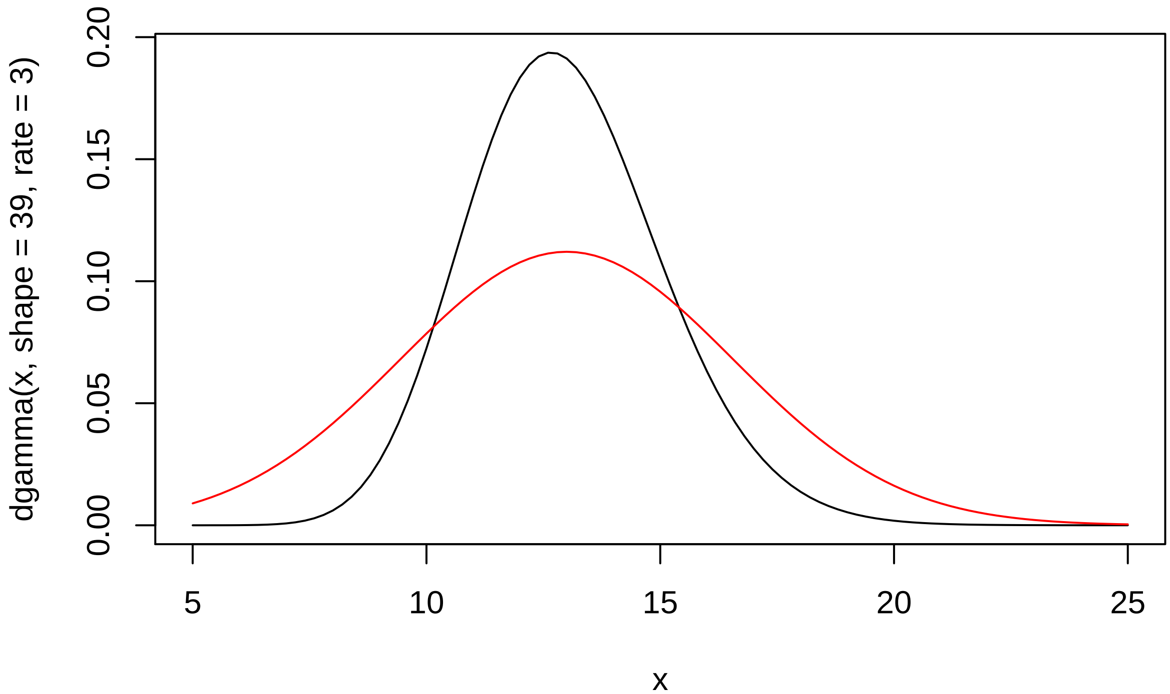

```
plot(ecdf(pA))
```



Naloga 2 - gama porazdelitev

a.

```
curve(dgamma(x,shape=39,rate=3),from=5,to=25)
curve(dnorm(x,mean=13,sd=3.56),add=TRUE,col="red")
```



b. Testna statistika bo spet vzorčno povprečje. Če je vzorec zadosti velik, lahko uporabimo CLI in spet normalno porazdelitev.
c. Izračun za velike n - po normalni porazdelitvi. Za majhne n pa teoretično (težje) ali pa s simulacijami: v for zanki generiramo vzorec velikosti n iz $\Gamma(39, 3)$, izračunamo testno statistiko in jo shranimo. Za vse shranjene testne statistike naredimo interval zaupanja. Meje so ravno kritične vrednosti.

```
### meje zavrnitve - velik n (100)
meja1 = qnorm(0.025,mean=13,sd=sqrt(39/3^2)/sqrt(100))
meja2 = qnorm(0.025,mean=13,sd=sqrt(39/3^2)/sqrt(100),lower.tail=FALSE)
meja1
```

```
## [1] 12.592
```

```
meja2
```

```
## [1] 13.408
```

```
# majhen n (4) - s simulacijami
velikost = NULL; testna = NULL
n=4
for(i in 1:1000){
  vzorec = rgamma(n,39,3)
  testna = c(testna,mean(vzorec))
}
quantile(testna,prob=c(0.025,0.975))
```

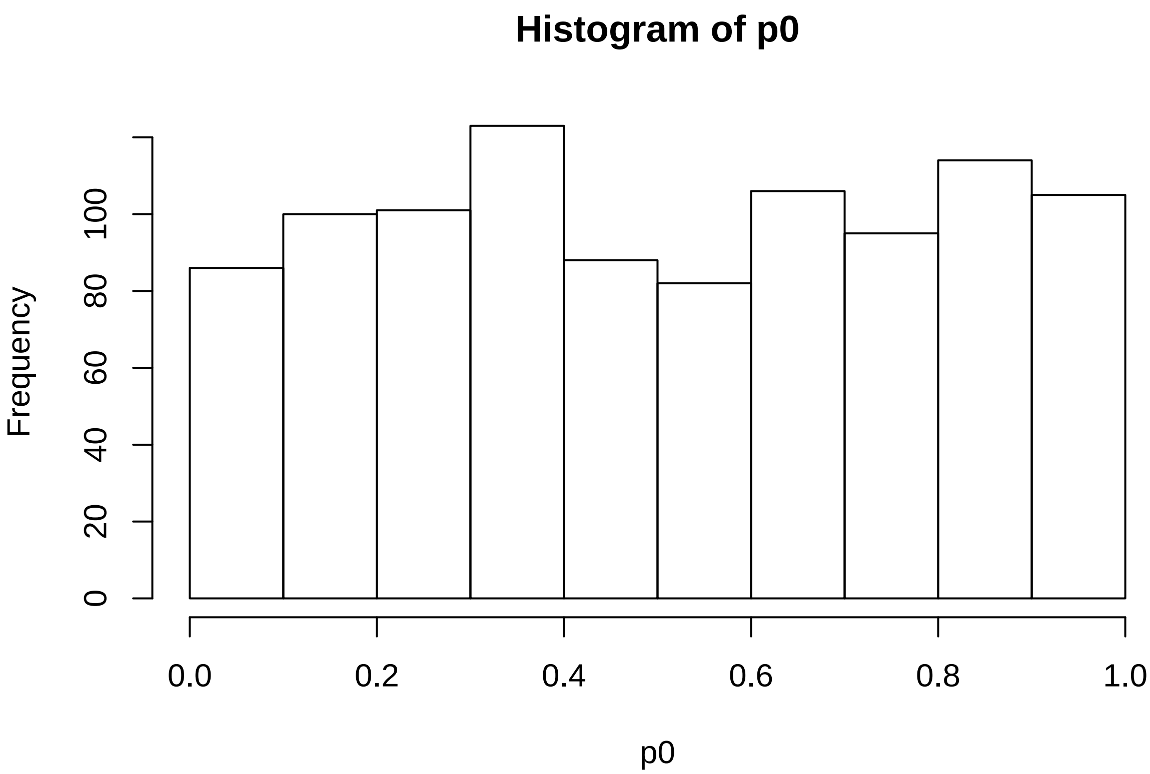
```
##     2.5%    97.5%
## 11.01156 15.15498
```

f.

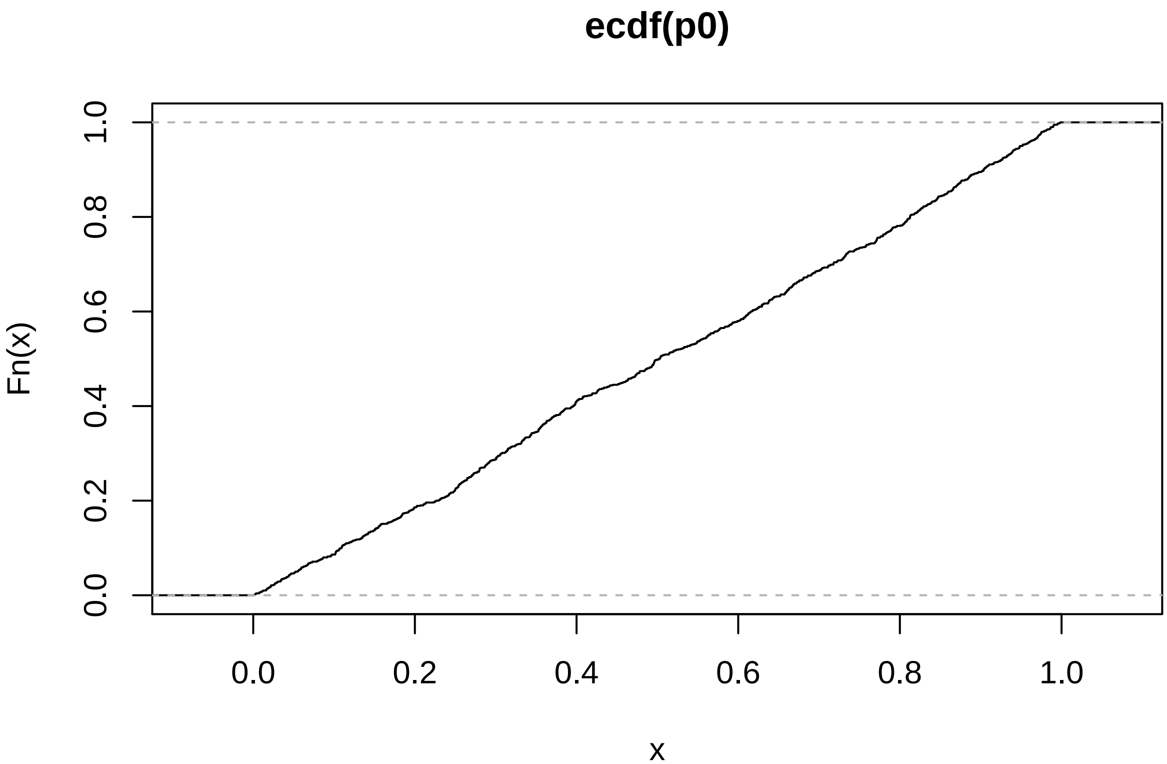
```r
simulirajG <- function(ponovi,n,alfaVzorec,lambdaVzorec,mu0,sigma0){
  pVrednosti = rep(NA,ponovi)
  for(i in 1:ponovi){
    vzorec = rgamma(n,alfaVzorec,lambdaVzorec) # sprememba
    pVrednosti[i] = pvrednost(mean(vzorec),mu0,sigma0/sqrt(n))
  }
  pVrednosti
}
# podatki pod H_0
n=100
p0 = simulirajG(1000,n,39,3,13,sqrt(39/3^2))
table(p0 < 0.05)
```

```
##
## FALSE  TRUE
##   954    46
```

```r
hist(p0)
```



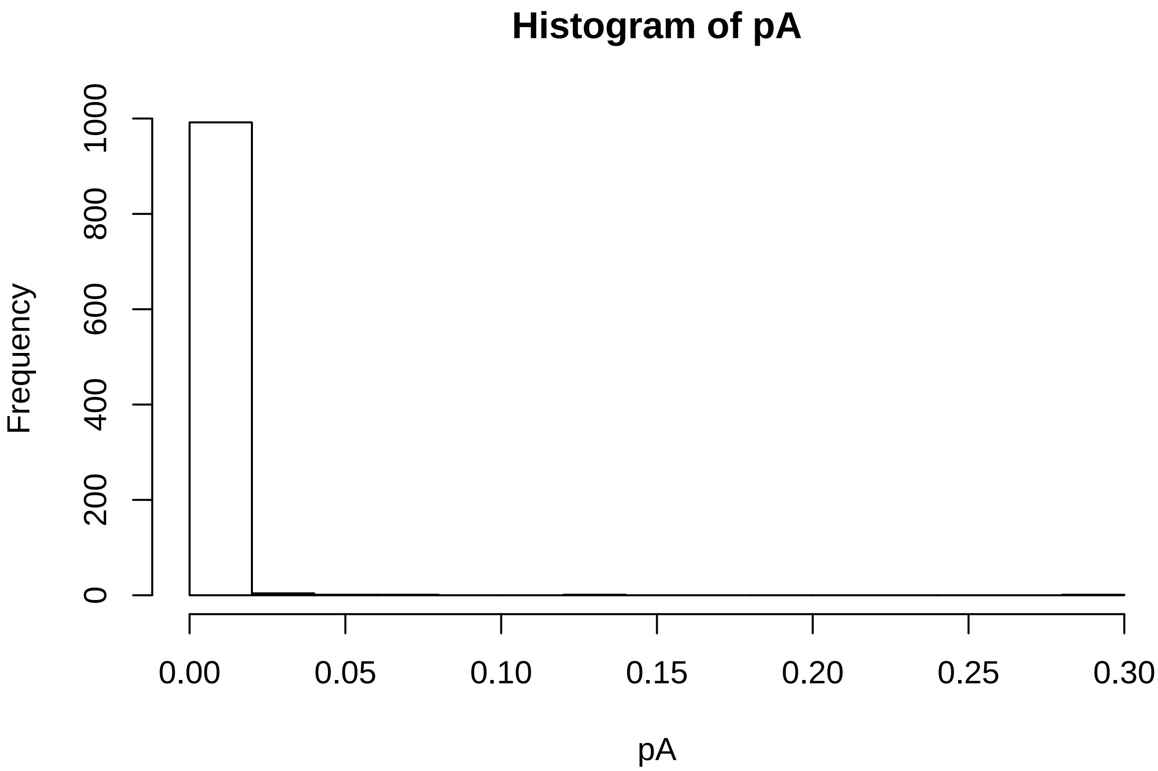

```r
plot(ecdf(p0))
```



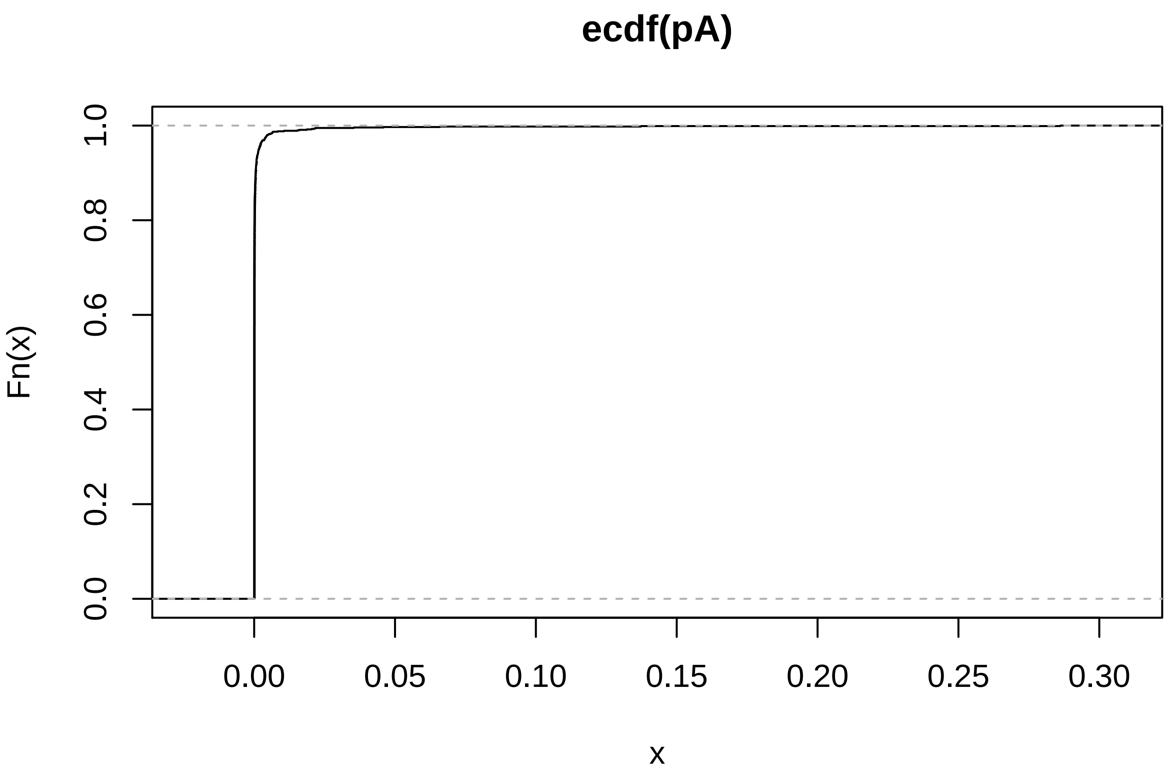

```r
# podatki pod H_A
pA = simulirajG(1000,n,42,3,13,sqrt(39/3^2))
table(pA < 0.05)
```

```
##
## FALSE  TRUE
##     3   997
```

```r
hist(pA)
```



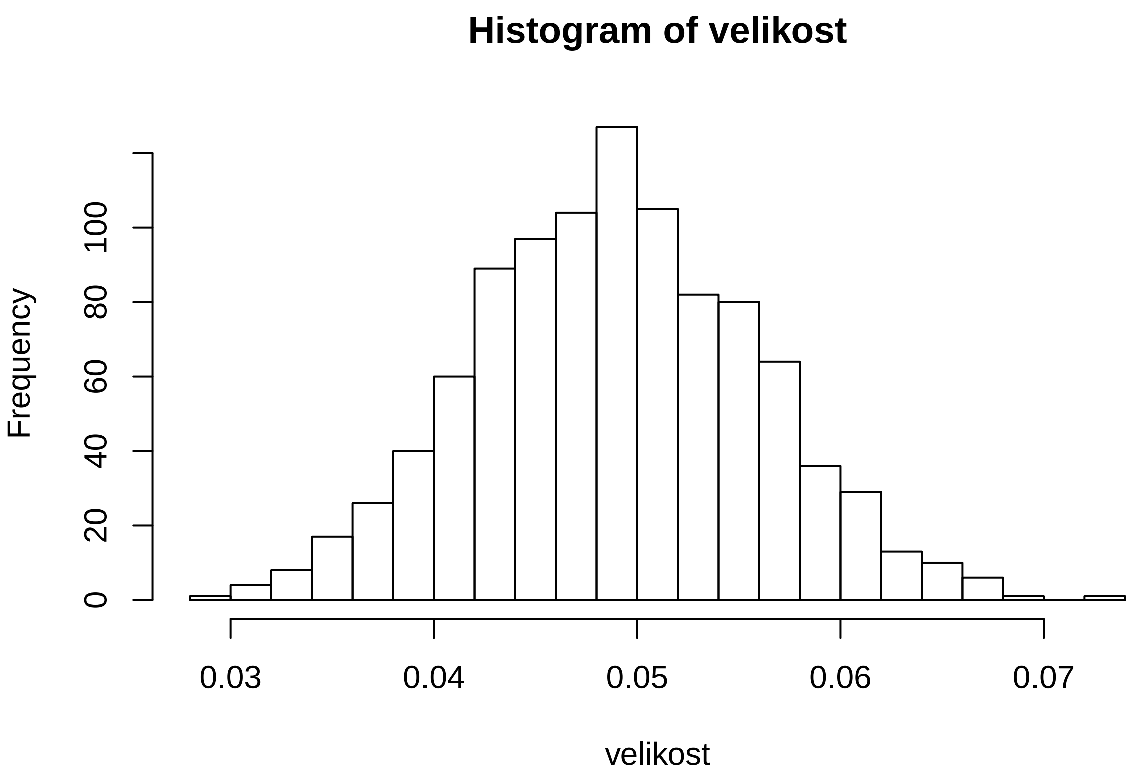

```
plot(ecdf(pA))
```



g.

```
### simulacije za zelo majhne vzorce
velikost = NULL
n=4 # zelo majhen vzorec, CLI ne velja
for(i in 1:1000){
```

```
  p0 = simulirajG(1000,n,39,3,13,sqrt(39/3^2))
  velikost = c(velikost,sum(p0 < 0.05)/1000)
}
hist(velikost,breaks=30)
```



Naloga 3

Lahko bi bila tudi σ drugačna.

Naloga 4 - depresija

a. Otroci z nizkim samospoštovanjem.
b. H_0: v populaciji je povprečen rezultat vprašalnika ≥ 90.
c. $H_A : \mu_A < 90$, enostranska, sestavljena
d. $T = \bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$
   To velja, če: $X_i \sim N(\mu_0, \sigma^2)$, $\mu_0 = 90, \sigma = 14$
   ali pa v primeru, ko je n dovolj velik in X_i porazdeljene dovolj simetrično, varianca ni neskončna, da lahko uporabimo CLI.
e. Ima samo eno mejo, saj gre za enostranski test:

```
qnorm(0.95,mean=90,sd=14/sqrt(length(scores)),lower.tail=FALSE)
```

dobimo torej območje zavrnitve $(\infty, 85.8]$.
Teoretično v tem primeru lahko zapišemo, da začnemo z zavračanjem H_0, če velja

$$P(\bar{X} < meja) = \alpha$$
$$P(Z < \frac{meja - \mu_0}{SE}) = \alpha$$

Ker vemo, da je $\frac{meja-\mu_0}{SE} = z_\alpha$, lahko zapišemo tudi, da zavrnemo, če velja

$$\frac{\bar{X} - \mu_0}{SE} < z_\alpha,$$

oziroma

$$\bar{X} < \mu_0 + z_\alpha \cdot SE = \mu_0 - z_{1-\alpha} \cdot SE$$

To pa je točno naše območje zavrnitve.

f.

```
pnorm(mean(scores),mean=90,sd=14/sqrt(length(scores)))
```

g. Interval zaupanja v našem primeru je: $(-\infty, \bar{X} + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}]$, kjer je σ populacijski standardni odklon. To je $(-\infty, 92.3]$.
H_0 in IZ imata skupno idejo, tj. povedati, meje (oziroma interval), za katerega velja, da mu zaupamo v 95%. Interval zaupanja v našem primeru vsebuje μ_0, torej lahko rečemo, da ne bomo zavrnili H_0.
h. Ne zavrnemo H_0. Rezultat ni statistično značilen, tj. za populacijo otrok z nizkim samospoštovanjem ne moremo trditi, da so bolj nagnjeni k depresijam kot otroci v splošnem. Rezultat testa je lahko statistično značilen (tj. da v populaciji obstajajo razlike/povezava ipd. med spremenljivkami/skupinami), vendar ni nujno tudi strokovno pomemben (tj. razlika je lahko premajhna, da bi nam strokovno to nekaj pomenilo).

```
scores=read.csv("data/data_depression.csv")
# meja
meja = qnorm(0.05,mean=90,sd=14/sqrt(length(scores$x)))
meja
# p-vrednost
pnorm(mean(scores$x),mean=90,sd=14/sqrt(length(scores$x)),lower.tail=TRUE)
# interval zaupanja: od -infinity do
mean(scores$x) + qnorm(0.95)*14/sqrt(length(scores$x))
```

Naloga 5 - sistolični krvni tlak

a. H_0: $\mu_0 = \mu_{razlika} = 0$. Razlika, vidimo, precej niha, zato ne moremo postaviti enostranske hipoteze.
b. $SE = 20/\sqrt{25} = 4$
c. $\bar{X} = -5$, lahko torej zapišemo IZ: $(-5 - 1.96 \cdot 4, -5 - 1.96 \cdot 4) = (-12.84, 2.84)$, oziroma pnorm(-5,mean=0,sd=4) = 0.106. Oba rezultata govorita, da ne moremo zavrniti H_0.
d.

```
qnorm(0.025,mean=0,sd=4)
```

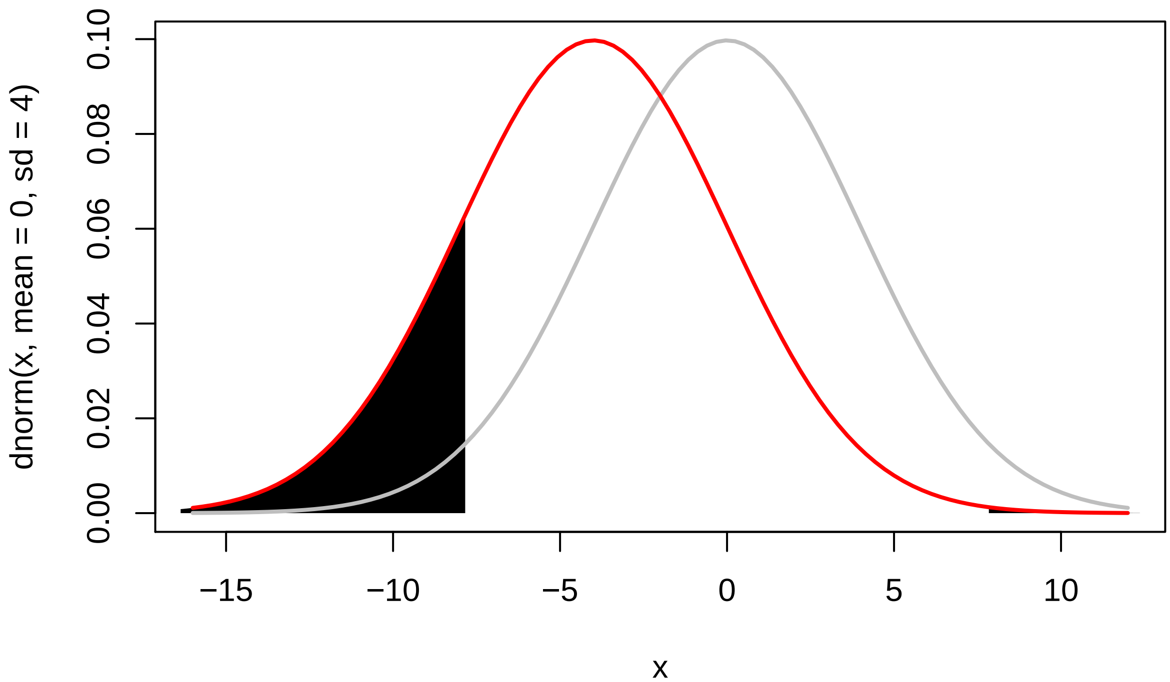
```
## [1] -7.839856
```

oziroma zgornja meja intervala zaupanja ne sme preseči 0:

$$\bar{X} + z_{1-\alpha/2} SE < 0$$
$$\bar{X} < 7.84$$

f.

```
# izris v Rju
curve(dnorm(x,mean=0,sd=4),from = -16,to=12)
# spodnja meja
xx=seq(qnorm(0.001,-4,4),qnorm(0.025,0,4),length.out=1000)
yy=c(0,dnorm(xx,-4,4),0)
xx=c(qnorm(0.001,-4,4),xx,qnorm(0.025,0,4))
polygon(xx,yy,col="black",border=NA)
# zgornja meja
xx=seq(qnorm(0.975,0,4),qnorm(0.999,0,4),length.out=1000)
yy=c(0,dnorm(xx,-4,4),0)
xx=c(qnorm(0.975,0,4),xx,qnorm(0.999,0,4))
polygon(xx,yy,col="black",border=NA)
curve(dnorm(x,mean=0,sd=4),add=TRUE,lwd=2,col="gray")
curve(dnorm(x,mean=-4,sd=4),add=TRUE,lwd=2,col="red")
```




```
# se izracun
pnorm(qnorm(0.975,0,4),-4,4,lower.tail=FALSE) +
  pnorm(qnorm(0.025,0,4),-4,4)
```

```
## [1] 0.170075
```

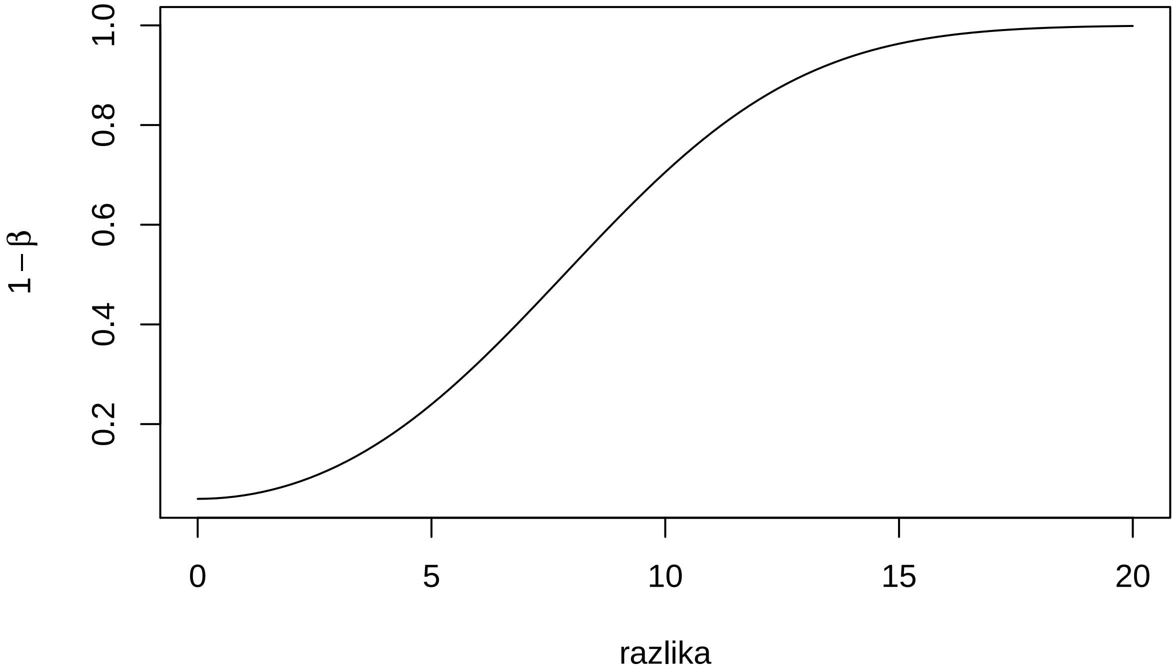
```
pnorm(qnorm(0.975,0,4),-6,4,lower.tail=FALSE) +
  pnorm(qnorm(0.025,0,4),-6,4)
```

```
## [1] 0.3230412
```

h. $\alpha = 0.05$
i. $\beta = 1 - power = 1 - 0.17 = 0.83$
j. Verjetnost zavrnitve (če je razlika večja), naraste.

```
# graf - populacijska razlika - moc
moc <- function(razlika){
  pnorm(qnorm(0.975,0,4),razlika,4,lower.tail=FALSE) +
```

```
    pnorm(qnorm(0.025,0,4),razlika,4)}

curve(moc,from=0,to=20,xlab="razlika",ylab=expression(1-beta))
```



k. α se zmanjša, β se zveča (moč testa je manjša).
l. Testna statistika pri manjšem vzorcu ima večjo variabilnost, zato bo napaka 2.vrste večja. β pada s korenom števila pacientov.

```
# graf - velikost vzorca - beta
n2 <- function(n,razlika=4){
  1-(pnorm(qnorm(0.975,0,20/sqrt(n)),razlika,20/sqrt(n),lower.tail=FALSE) +
       pnorm(qnorm(0.025,0,20/sqrt(n)),razlika,20/sqrt(n)))}

curve(n2,from=10,to=500,xlab="velikost vzorca",ylab=expression(beta))
```

m. Ne, to pomeni, da razlika v krvnem tlaku ni bila tako velika, da bi jo s statističnim testom pri tej velikosti vzorca lahko zaznali.

n. Povzetek - oceniti/poznati je potrebno

- pričakovano razliko (učinek)
- pričakovano razpršenost (variabilnost podatkov)
- α
- željeno moč testa $(1-\beta)$

Naloga 6 - sistolični krvni tlak, velik vzorec

a. $SE = 20/\sqrt{10000} = 0.2$
b. $T = \bar{X} = -0.47$

```
qnorm(0.02/2,mean=0,sd=0.2,lower.tail=TRUE)
```

```
## [1] -0.4652696
```

c. $[-0.86, -0.08]$
d. Populacijska razlika je statistično značilna, vendar zelo majhna. Verjetno ni strokovno pomembna.