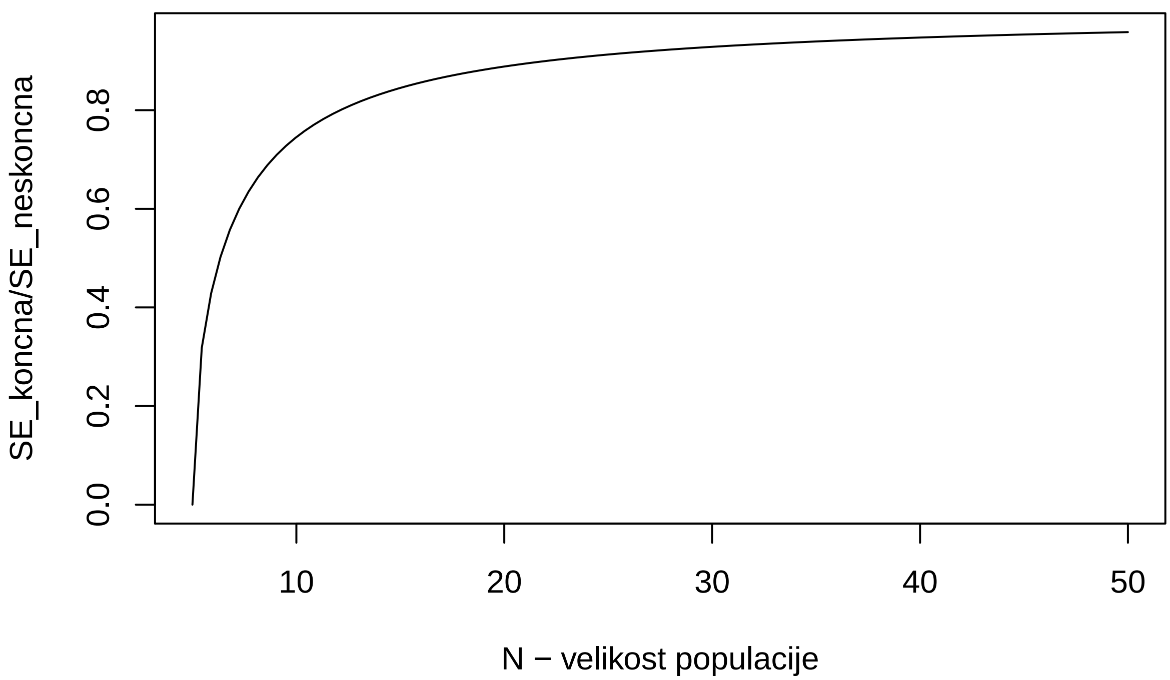
Rešitve - končna populacija, CLI

Nataša Kejžar

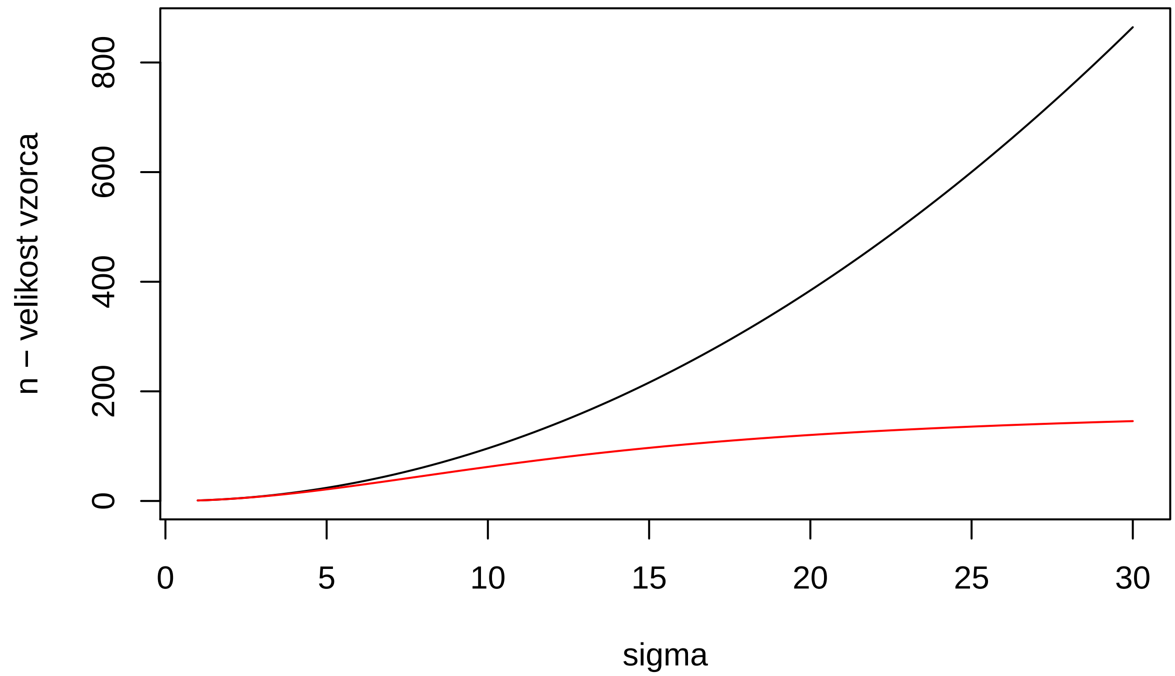
Naloga 1 - razmerje

```
n = 5
relativSE = function(n,N){
  sqrt((N-n)/(N-1))
}
curve(relativSE(n,x),from=5,to=50,
      xlab="N - velikost populacije",
      ylab='SE_koncna/SE_neskoncna')
```



Naloga 2 - MIZŠ

```
koncna = function(sigma,N,alfa=0.05){
  (qnorm(1-alfa/2)*sigma)^2*N/(4*(N-1)+(qnorm(1-alfa/2)*sigma)^2)
}
neskoncna = function(sigma,N,alfa=0.05){
  (qnorm(1-alfa/2)*sigma/2)^2
}
curve(neskoncna(x,175),from=1,to=30,
      xlab='sigma',ylab="n - velikost vzorca")
curve(koncna(x,175),from=1,to=30,add=TRUE,col="red")
```



Naloga 3 - volitve

Vemo, da je $X \sim Ber(\pi)$, $E(X) = \pi$ in $var(X) = \pi(1-\pi)$:

Cenilka:

$$p = \frac{\sum_{i=1}^{n} x_i}{n} \qquad ; n = 10$$

Pričakovana vrednosti n varianca:

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(Y) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \pi$$

$$var(Y) = \frac{1}{n^2} var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} var(X_i)\right] = \frac{1}{n}\pi(1-\pi)$$

Točka d.:

$$1 - P(X > 5) = P(X \leq 4)$$

koda: `pbinom(4,10,0.55)`

Točka e.:

```r
pbinom(4,size=10,prob=0.55)
```

```
## [1] 0.2615627
```

```r
# s simulacijami
N=1000
stRep=NULL
for(i in 1:N){
  vzorec = sample(0:1,10,replace = TRUE,prob=c(0.45,0.55))
  stRep = c(stRep,sum(vzorec))}
sum(stRep<5)/N # delez vzorcev z manj kot 5 Dem.
```

```
## [1] 0.272
```

```r
#ali
sum(rbinom(N,size=10,prob=0.55)<5)/N
```

```
## [1] 0.28
```

Naloga 4 - Janez o zgodovini

a. Janezove odgovore lahko zapišemo kot Bernoullijeve spremenljivke, odgovor je lahko pravilen ($X = 1$) z verjetnostjo $p = 0,9$, lahko pa napačen ($X = 0$) z verjetnostjo $1 - p = 0,1$. Delež p v populaciji velikosti n lahko zapišemo kot

$$p = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Ker je delež povprečje Bernoullijevih spremenljivk, ga lahko ocenimo z vzorčnim povprečjem, smiselna cenilka je torej

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Eksaktna porazdelitev: Vsota Bernoullijevo porazdeljenih spremenljivk je binomska. Porazdelitev povprečja torej lahko izrazimo z binomsko porazdelitvijo: $P(\hat{p} = k/n) = P(Y = k)$, kjer je Y binomsko porazdeljena spremenljivka.

Asimptotska porazdelitev: Centralni limitni izrek nam pove, da z večanjem velikosti vzorca n, porazdelitev cenilke konvergira proti normalni.

b. Vemo, da je nepristranska cenilka za varianco vzorčnega povprečja $\bar{X}$ enaka $\widehat{SE}^2 = \widehat{\sigma}^2/n$, kjer je $\widehat{\sigma}^2$ nepristranska cenilka variance slučajne spremenljivke X. Nepristransko cenilko za varianco cenilke torej zapišemo kot

$$\widehat{var}(\bar{X}) = \frac{\frac{n}{n-1}\hat{p}(1-\hat{p})}{n} = \frac{\hat{p}(1-\hat{p})}{n-1}.$$

c. Interval zaupanja bomo zapisali s pomočjo aproksimativne porazdelitve, torej uporabili bomo normalno porazdelitev. 95% interval zaupanja bo zato enak:

$$\bar{X} \pm 1,96\widehat{SE},$$

kjer je $1,96$ ustrezna vrednost ($z_{1-\alpha/2}$) iz normalne porazdelitve. Velja torej

$$2 \cdot 1,96 \cdot \widehat{SE} = 0,1 \ \Rightarrow\ \widehat{SE} = 0,0255$$
$$0,0255^2 = \frac{0,9 \cdot 0,1}{n-1} \ \Rightarrow\ n - 1 = 138,3$$

Potrebna velikost vzorca za približno 10% širok interval zaupanja je torej 140 vprašanj.

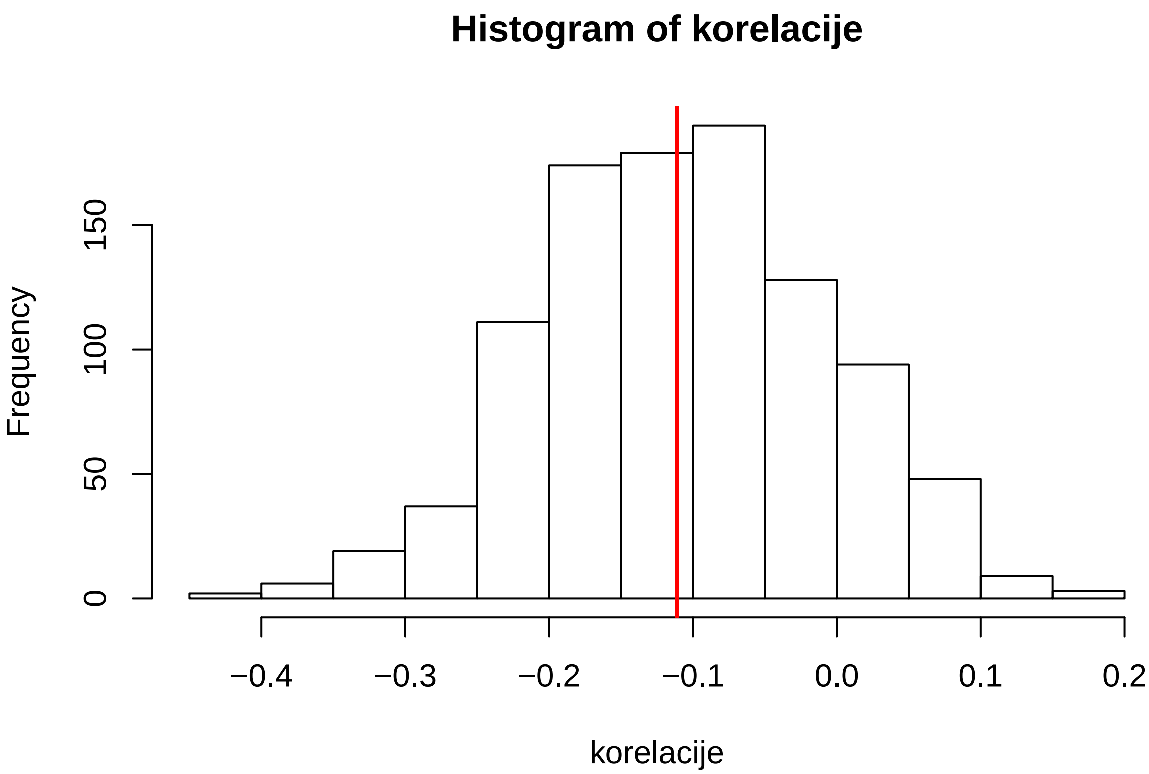
d. Če je vrednost p bližje $0,5$, je standardna napaka večja in interval zaupanja bo širši.

Naloga 6 - korelacija

```r
set.seed(1)
N = 10 # majhna
kor = function(N){
  popul=1:N
  x1=NULL
```

```
  x2=NULL
  for(i in 1:100){
    x = sample(popul,size=2)
    x1[i] = x[1]
    x2[i] = x[2]
  }
  cor(x1,x2)
}

korelacije = replicate(1000,kor(N))
hist(korelacije)
abline(v=-1/9,col="red",lwd=2)
```




```
mean(korelacije) # empiricna vrednost
```

```
## [1] -0.1070252
```

```
-1/(N-1) # teoreticna vrednost
```

```
## [1] -0.1111111
```

Naloga 7 - sistolični krvni tlak

b. Za $X \sim Unif(a, b)$ velja, da je njena gostota enaka $1/(b-a)$, ko $a \leq x \leq b$, 0 pa sicer. Vemo tudi (iz prejšnjih vaj), da je $E(X) = (b-a)/2$. Za naš primer je to 0.

$$
\begin{aligned}
var(X) =& E\left[(X-\mu)^2\right] = E(X^2) - \mu^2 \\
=& E(X^2) \\
=& \int_a^b x^2 f(x) dx \\
=& \frac{1}{b-a} \int_a^b x^2 dx \\
=& \frac{1}{b-a} \frac{x^3}{3} |_a^b \\
=& \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}
\end{aligned}
$$

Ker velja $(b-a)/2 = 0$, je $a = -b$, dobimo, da je $var(X) = \sigma^2 = b^2/3$. Iz tega sledi, da je $b = -a = 50 \cdot \sqrt{3}$.

```
n=100
pnorm(-10,mean=0,sd=50/sqrt(n))
```

```
## [1] 0.02275013
```

```
set.seed(1)
ponovi=1000
cenilka <- function(x){mean(x)< -10}
ocene = NULL
for(i in 1:ponovi){
  vzorec = rnorm(n,mean=0,sd=50)
  ocene = c(ocene,cenilka(vzorec))
}
mean(ocene)
```

```
## [1] 0.013
```

Naloga 8 - voda v podjetju

a. Torej velja, da imamo $n = 100$ enot, vsaka je porazdeljena po: $\mu_X = 6$ in $\sigma_X = 2$. Ker nas zanima spremenljivka $Y = \sum_{i=1}^{n} X_i$, je potrebno ugotoviti, kaj so parametri za asimptotsko porazdelitev te nove spremenljivke.

$$
\begin{aligned}
\mu_Y =& n\mu_X \\
\sigma_Y^2 =& n\sigma_X^2 \\
SE = \sigma_Y =& \sqrt{n}\sigma_X
\end{aligned}
$$

Zanima nas $P(Y > 650) = P(Z > (700 - \mu_Y)/\sigma_Y)$.

```
pnorm(650,mean=600,sd=20,lower.tail=FALSE)
```

```
## [1] 0.006209665
```

b. Ker so količine neodvisne, je naša nova spremenljivka, katere verjetnost nas zanima, porazdeljena po $W \sim Bin(4,p)$, kjer je p enak vrednosti iz prejšnje naloge. Torej nas zanima $P(W > 0)$:

```
verj = pnorm(650,mean=600,sd=20,lower.tail=FALSE)
1-pbinom(0,size=4,prob=verj)
```

```
## [1] 0.02460826
```

c. Podobno kot v prejšnji nalogi torej, $W_2 \sim Bin(365, p)$ in $P(W_2 > 15)$:

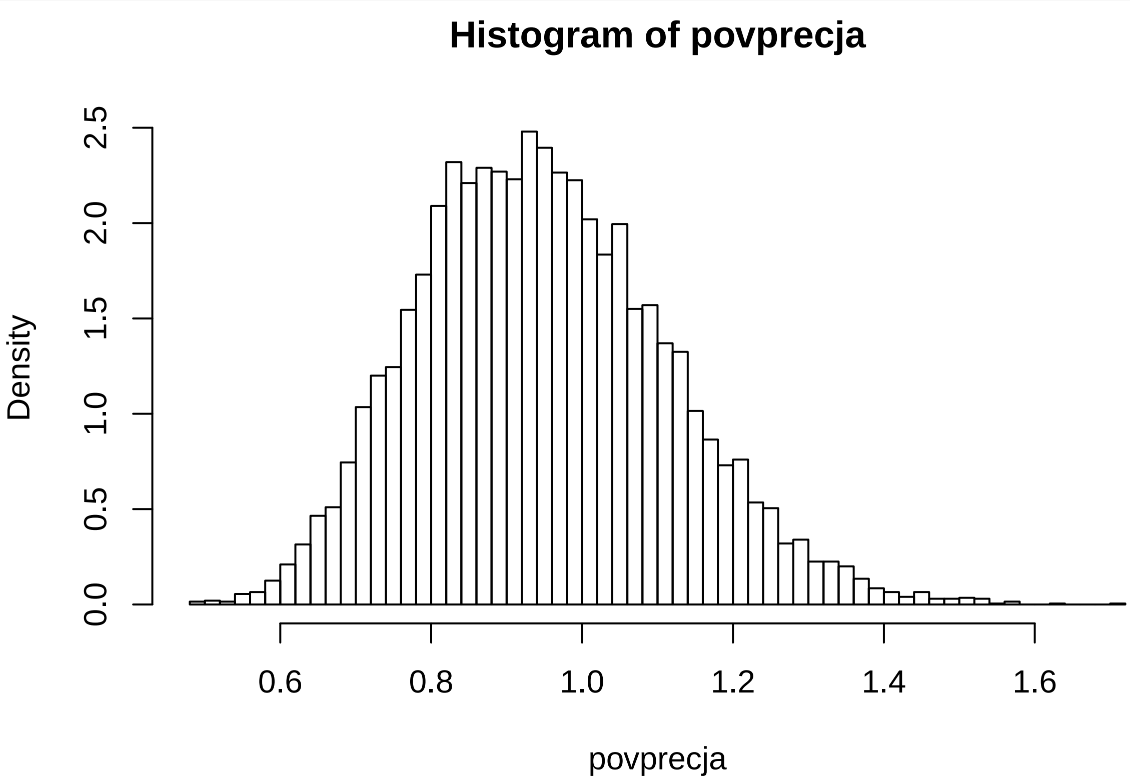
```r
verj = pnorm(650,mean=600,sd=20,lower.tail=FALSE)
1-pbinom(2,size=365,prob=verj)
```

```
## [1] 0.3952867
```

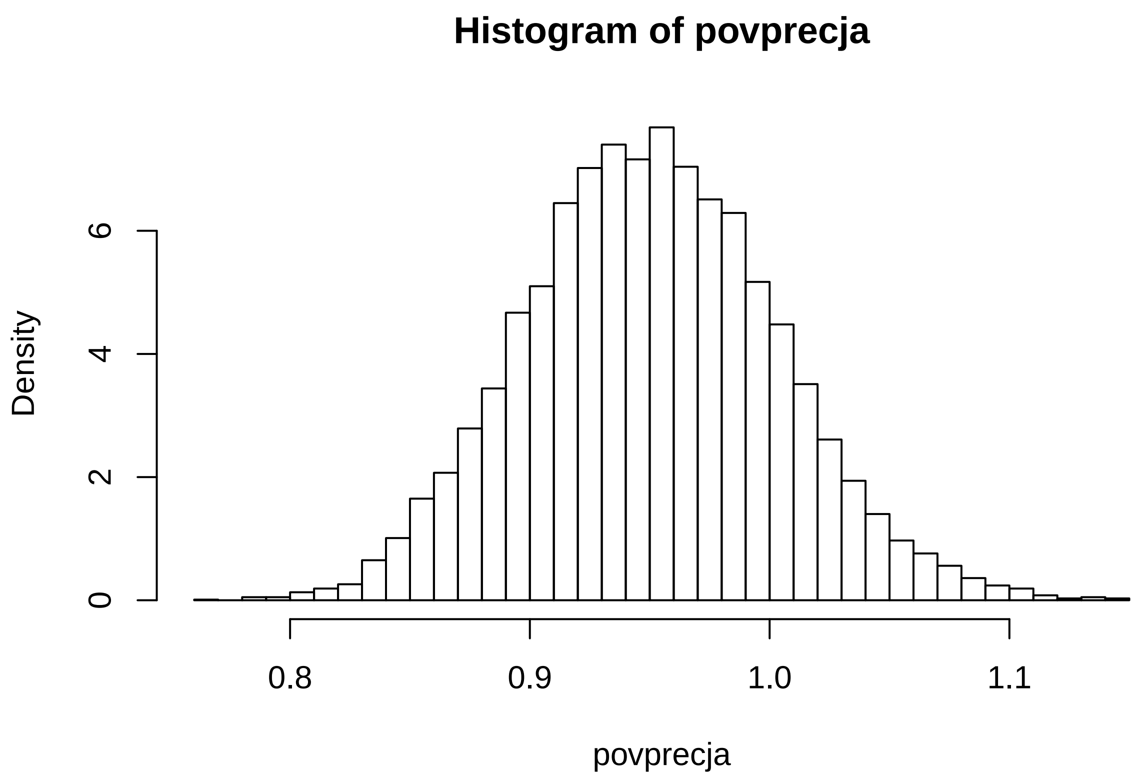
Rezultat je pričakovan, saj je količina vode, ki jo podjetje priskrbi na dan večja kot povprečna vrednost popite vode.

Naloga 9 - pogojna porazdelitev

```r
cli = function(n){
    povprecja = NULL
    for(i in 1:10000){
        # izberemo, ali bo opazovanje iz porazd U(0,1) ali U(0,10)
        izbira = sample(c(1,10),size=n,replace=TRUE,prob=c(0.9,0.1))
        # generiramo opazovanja iz uniformne porazdelitve
        vzorec = sapply(izbira,FUN = function(x){runif(1,min=0,max=x)})
        povprecja = c(povprecja,mean(vzorec))
    }
    hist(povprecja,freq=FALSE,breaks=50)
    # cez histogram lahko dorisemo se normalno krivuljo
}
cli(100)
```




```r
cli(1000)
```



Naloga 12 - CLI

$p(i) = 1/3$ za vsak $i \in \{1, 2, 3\}$.

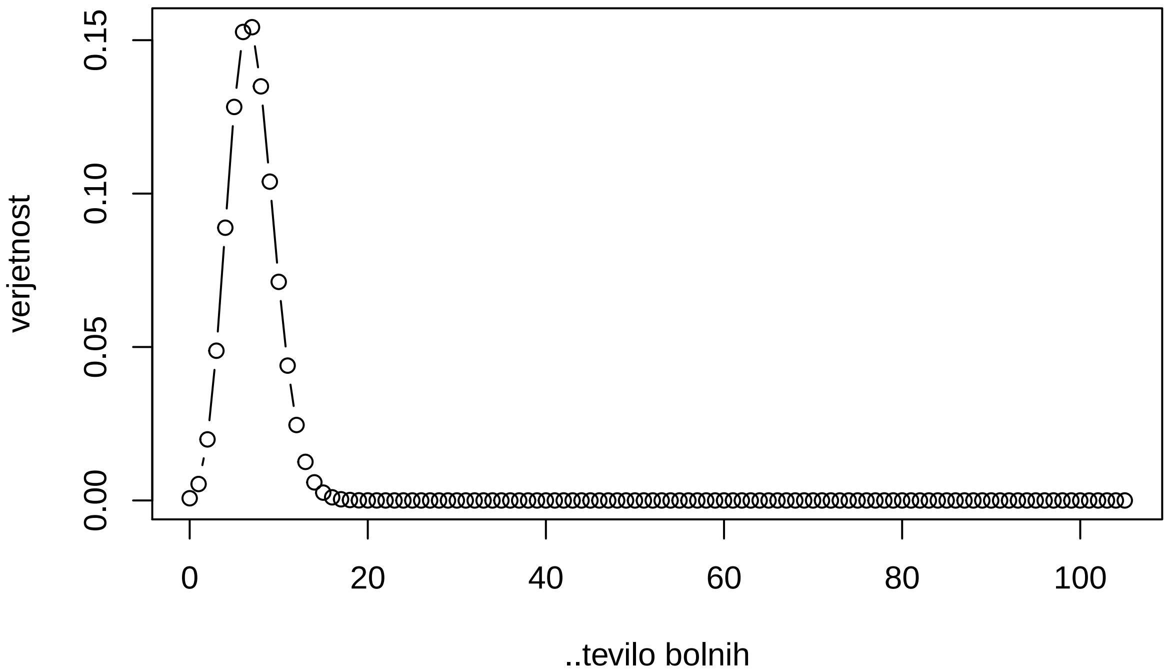
$$
\begin{aligned}
E(X) &= \sum_{i=1}^{3} ip(i) = 2 \\
var(X) &= \sum_{i=1}^{3} (i - E(X))^2 p(i) \\
&= 1/3 \sum_{i=1}^{3} (i-2)^2 = 2/3
\end{aligned}
$$

Naloga 14 - vrtec

```
stBolnih = 0:105
plot(stBolnih, dbinom(stBolnih,size=105,prob=1/15),type="b",
     xlab="število bolnih",ylab="verjetnost")
```

```
## Warning in title(...): conversion failure on 'število bolnih' in
## 'mbcsToSbcs': dot substituted for <c5>

## Warning in title(...): conversion failure on 'število bolnih' in
## 'mbcsToSbcs': dot substituted for <a1>
```




```r
qbinom(0.99,size = 105,prob=1/15) #[1] 14
```

```
## [1] 14
```

```r
ex = 105/15
vx = 105*1/15*14/15
qnorm(0.99,mean=ex,sd=sqrt(vx)) #[1] 12.94623
```

```
## [1] 12.94623
```

```r
# continuity correction
# add +1/2 to the value you get
qnorm(0.99,mean=ex,sd=sqrt(vx)) +0.5 #[1] 13.44623
```

```
## [1] 13.44623
```

Naloga 15 - referendum

Vemo, da v splošnem dobimo nepristransko varianco z

$$
\begin{aligned}
\hat{\sigma^2} =& \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \\
=& \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i-n\bar{X}^2\right) \qquad\qquad X\sim Ber(\pi) \\
=& \frac{n}{n-1}\bar{X}-\frac{n}{n-1}\bar{X}^2 \\
=& \frac{n}{n-1}(\pi-\pi^2) \\
=& \frac{n}{n-1}\pi(1-\pi)
\end{aligned}
$$

```r
p = 0.55
n = 1500
set.seed(1)
ponovi = 1000
delezi = NULL
for(i in 1:ponovi){
  vzorec = runif(n) <= 0.55
  delezi = c(delezi,mean(vzorec))
}
delezi = sort(delezi)
c(delezi[ponovi*0.025],delezi[ponovi*0.975])
```

```
## [1] 0.5266667 0.5773333
```

```r
# ali s pomočjo percentilov
quantile(delezi,probs=c(0.025,0.975))
```

```
##      2.5%     97.5%
## 0.5266667 0.5773500
```

```r
# IZ iz 1 samega vzorca
#(vzamemo zadnji simul.vzorec)
phat = mean(vzorec)
SEhat = sqrt(phat*(1-phat)/(n-1))
c(phat-1.96*SEhat,phat+1.96*SEhat)
```

```
## [1] 0.5415807 0.5917526
```

Vemo, da velja $Z = \frac{1}{n}\sum(X_{ZA}) - \frac{1}{n}\sum(X_{PROTI}) = \frac{2}{n}\sum(X_{ZA}) - 1$, kjer je $X_{ZA} = X \sim Ber(\pi)$. Torej lahko zapišemo cenilko kot $2\cdot\hat{\pi}-1$, kjer je $\hat{\pi}$ cenilka za delež tistih, ki so ZA referendum.

$$
\begin{aligned}
E(Z) =& 2\pi - 1 \\
var(Z) =& \frac{4}{n^2}\sum_{i=1}^{n} var(X_i) \\
var(Z) =& \frac{4}{n}\pi(1-\pi) \\
\widehat{\sigma_Z}^2 =& \frac{4}{n-1}\hat{\pi}(1-\hat{\pi})
\end{aligned}
$$

Interval zaupanja je torej $2\hat{\pi} - 1 \pm 1.96 \cdot 2\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n-1}}$.

Imamo 95 % zaupanje, da se razlika med deležema ZA–PROTI referendumu nahaja v tem intervalu.

Če naj bi bila razlika deležev pozitivna, mora veljati, da bo spodnja meja 95 % IZ > 0. Za $\hat{\pi}$ vzamemo kar delež 0.55 (ki ga v spodnji izpeljavi označimo kar s π).

$$
\begin{aligned}
2\pi - 1 - z_{\alpha/2} \cdot 2\sqrt{\frac{\pi(1-\pi)}{n-1}} &> 0 \\
\sqrt{\frac{\pi(1-\pi)}{n-1}} &< \frac{2\pi - 1}{2 \cdot z_{\alpha/2}} \\
\frac{\pi(1-\pi)}{n-1} &< \left(\frac{2\pi - 1}{2 \cdot z_{\alpha/2}}\right)^2 \\
\frac{\pi(1-\pi)}{(2\pi-1)^2} 4 \cdot z_{\alpha/2}^2 &< n - 1 \\
n &> 4z_{\alpha/2}^2 \frac{\pi(1-\pi)}{(2\pi-1)^2} + 1
\end{aligned}
$$

```
razlike = NULL
cenilka <- function(x){2*mean(x)-1}
for(i in 1:ponovi){
  vzorec = runif(n) <= 0.55
  razlike = c(razlike,cenilka(vzorec))
}
razlike = sort(razlike)
c(razlike[ponovi*0.025],razlike[ponovi*0.975])
```

```
## [1] 0.05066667 0.15066667
```

```
# za stevilo volilcev
c(razlike[ponovi*0.025],razlike[ponovi*0.975]) * n
```

```
## [1]  76 226
```

```
# IZ iz enega samega vzorca (vzamemo zadnji simul.vzorec)
phat = mean(vzorec)
SErazlika = 2*sqrt(phat*(1-phat)/(n-1))
c(2*phat-1 - 1.96*SErazlika,2*phat-1 + 1.96*SErazlika)
```

```
## [1] 0.06437672 0.16495661
```

```r
# najmanjsi n
p=0.55
4*1.96^2*(p*(1-p)/(2*p -1)^2) + 1
```

```
## [1] 381.3184
```

```r
najmanjsiN <- function(p){
  4*1.96^2*(p*(1-p)/(2*p -1)^2) + 1}
curve(najmanjsiN(x),from=0.53,to=0.8,xlab="delez",ylab="najmanjsi n")
```