

## Naloga 1 - centralni limitni izrek

Statistiki si pri preučevanju lastnosti cenilk pogosto pomagamo s simulacijami. Namen te naloge bo s pomočjo simulacij preučiti lastnosti cenilke za populacijsko povprečje. Predpostavimo, da je porazdelitev neke številske spremenljivke v neskončno veliki populaciji normalna z nekim povprečjem ( $\mu = 1000$ ) in standardnim odklonom ( $\sigma = 100$ ). Na podlagi vzorca velikosti  $n$  želimo oceniti  $\mu$ . Izračune bomo opravili s programom R. S pomočjo simulacije izpolnite tabelo in odgovorite na spodnja vprašanja.

Oblika porazdelitve v populaciji	Velikost vzorca	Oblika porazdelitve vzorčnih povprečij	Pričakovana vrednost vzorčnih povprečij	Standardna napaka
Normalna s $\sigma = 100$	5			
Normalna s $\sigma = 100$	25			
Normalna s $\sigma = 100$	100			
Normalna s $\sigma = 50$	100			
Normalna s $\sigma = 250$	100			
Enakomerna na $[0,2000]$	5			
Enakomerna na $[0,2000]$	25			
EkspONENTna s param. 1	5			
EkspONENTna s param. 1	25			

- Oglejte si porazdelitev ocene vzorčnega povprečja ( $\bar{X}$ ). Kje je vrh porazdelitve? Kakšne oblike je porazdelitev? Ali je porazdelitev vzorčnega povprečja bolj ali manj razpršena kot porazdelitev prvotne spremenljivke?
- Kako na razpršenost ocene vzorčnega povprečja vplivata velikost vzorca in razpršenost spremenljivke v populaciji?
- Kako na obliko porazdelitve ocene vzorčnega povprečja vpliva porazdelitev spremenljivke v populaciji?
- Zapišite formulo za standardno napako.

```

Koda za simulacijo (datoteka simulacija.r)

##simuliraj podatke za veliko populacijo iz N(1000,200):
y<-rnorm(1000000,mean=1000,sd=200)

##prikazi porazdelitev spremenljivke v histogramu:
par(mfrow=c(2,1))
hist(y,xlim=c(0,2000))

##zacni s simulacijo

n<-5 ##velikost vzorca
B<-1000 ##stevilo ponovitev simulacije
Y.bar<-rep(NA, B) ##tu se shranijo povprecja v posameznem koraku simulacije
for ( i in 1:B) { ##zacni s simulacijo
  id<-sample(1:length(y),n) ##slucajno izberi enote
  Y<-y[id] ##vrednosti na vzorcu
  Y.bar[i]<-mean(Y) ##povprecje na vzorcu
}
hist(Y.bar,xlim=c(0,2000)) ##narisi porazdelitev Y.bar
mean(Y.bar) ##povprecje porazdelitve
sd(Y.bar) ##razprsenost porazdelitve

##simuliraj podatke za veliko populacijo iz U(0,2000):
y<-runif(1000000,min=0,max=2000)
##prikazi porazdelitev spremenljivke v histogramu:
par(mfrow=c(2,1))
hist(y,xlim=c(0,2000))
##zacni s simulacijo

n<-5
B<-1000
Y.bar<-rep(NA, B)
for ( i in 1:B) {
  id<-sample(1:length(y),n)
  Y<-y[id]
  Y.bar[i]<-mean(Y)
}
hist(Y.bar,xlim=c(0,2000))
mean(Y.bar)
sd(Y.bar)

```

## Naloga 2 - Interval zaupanja za populacijsko povprečje

Zopet bomo uporabili simulacijo, da bomo ugotovili, kako sta porazdeljena izraza

$$\frac{\bar{X} - \mu}{SE(\bar{X})},$$

kjer je  $SE(\bar{X}) = \sigma/\sqrt{n}$  in

$$\frac{\bar{X} - \mu}{\widehat{SE(\bar{X})}},$$

kjer je  $\widehat{SE(\bar{X})} = s/\sqrt{n}$ . Simulirajte iz normalne porazdelitve s povprečjem 1000 in  $\sigma = 200$ . Porazdelitev obeh izrazov prikažite v histogramu, v katerega dodajte še gostoto standardne normalne porazdelitve in gostoto  $t$ -porazdelitve z  $n - 1$  stopinjami prostosti. Najprej naj bo  $n = 5$ , potem pa še  $n = 100$ .

Odgovorite na spodnja vprašanja:

- Katera krivulja se bolje prilega podatkom v primeru, ko uporabimo pravo standardno napako in katera, ko uporabimo ocenjeno standardno napako?
- Od česa je odvisna oblika  $t$ -porazdelitve?
- Kdaj je  $t$ -porazdelitev bolj podobna standardni normalni porazdelitvi?
- Izpeljite formulo za interval zaupanja za populacijsko povprečje!

Koda za simulacijo:

```
##simuliraj podatke za veliko populacijo iz N(1000,200):
y<-rnorm(1000000,mean=1000,sd=200)
##prikazi porazdelitev spremenljivke v histogramu:
par(mfrow=c(2,1))
hist(y,xlim=c(0,2000))

##zacni s simulacijo

n<-5 ##velikost vzorca
B<-10000 ##stevilo ponovitev simulacije
Y.bar<-rep(NA, B) ##tu se shranijo povprecja v posameznem koraku simulacije
z<-rep(NA,B) ##tu se shrani standardizirano vzorcno povprecje (prava SE)
tt<-rep(NA,B) ##tu se shrani standardizirano vzorcno povprecje (ocenjena SE)
for ( i in 1:B) { ##zacni s simulacijo
  id<-sample(1:length(y),n) ##slucajno izberi enote
  Y<-y[id] ##vrednosti na vzorcu
  Y.bar[i]<-mean(Y) ##povprecje na vzorcu
  z[i]<-(mean(Y)-1000)/(200/sqrt(n))
  tt[i]<-(mean(Y)-1000)/(sd(Y)/sqrt(n))
}

par(mfrow=c(1,2)) ##na isto sliko narisi dva histograma
hist(z,freq=FALSE,main="Prava SE",breaks=100)
##narisi histogram, kjer je na y-osi relativna frekvenca

xx<-seq(from=-4,to=4,by=0.01)
##rabimo zato, da izracunamo gostoto v posamezni tocki
lines(xx,dnorm(xx))
## v sliko dodaj gostoto standardne normalne porazdelitve
lines(xx,dt(xx,df=n-1),col="red")
## v sliko dodaj se gostoto t porazdelitve s n-1 stopinjami prostosti

hist(tt,freq=FALSE,main="Ocenjena SE",ylim=c(0,0.4),breaks=100,xlim=c(-5,5))
##histogram, kjer uporabljena ocenjena SE

lines(xx,dnorm(xx))
## v sliko dodaj gostoto standardne normalne porazdelitve
lines(xx,dt(xx,df=n-1),col="red")
## v sliko dodaj se gostoto t porazdelitve s n-1 stopinjami prostost
```

## Naloga 3 - Podatki iz ankete

S pomočjo podatkov iz ankete bi radi primerjali število ur, ki jih študenti namenijo uporabi interneta med študenti splošne in dentalne medicine (podatki iz anket, koda za izračune v Rju je na drugi strani).

- Izračunajte (vzorčni) povprečji
  - $\bar{x}_{Dent} =$
  - $\bar{x}_{Splosna} =$
- Zapišite formulo in izračunajte standardno napako za vzorce take velikosti (uporabite ocenjeni standardni odklon)
- Izračunajte 95% interval zaupanja (IZ) za populacijski povprečji. Uporabite kodo v R, enega izmed intervalov pa izračunajte tudi na roke.
- Primerjajte ju. Ali se prekrivata? Kaj lahko sklepamo?
- Izračunajte še 99% IZ in ju primerjajte.

- Kaj bi se zgodilo z IZ, če bi v vzorec zajeli več študentov?
- Ali je predpostavka o normalnosti porazdelitve naše spremenljivke smiselna? Kako to vpliva na rezultate?
- Denimo, da raziskavo ponovite 100 krat in vsakič izračunate 95% IZ. V koliko primerih pričakujete, da bo:
  - populacijsko povprečje zajeto v intervalih?
  - vzorčno povprečje zajeto v intervalih?

```
dd<-read.table("Ankete1011.txt",header=T,dec="," ,sep="\t",fill=T)

summary(dd$Internet)

summary(dd$Internet[which(dd$Spol=="moski")])
sd(dd$Internet[which(dd$Spol=="moski")])
t.test(dd$Internet[which(dd$Spol=="moski")],conf.level=0.95)

summary(dd$Internet[which(dd$Spol=="zenski")])
sd(dd$Internet[which(dd$Spol=="zenski")])
t.test(dd$Internet[which(dd$Spol=="zenski")],conf.level=0.95)

##99% IZ
t.test(dd$Internet[which(dd$Spol=="moski")],conf.level=0.99)
t.test(dd$Internet[which(dd$Spol=="zenski")],conf.level=0.99)
```