

$$EZ[a,b] \Rightarrow Z = \sqrt{\frac{1}{12}(b-a)^2} = \sqrt{\frac{1}{12}2000^2} = 577$$

$$\frac{Z}{\sqrt{n}} = \frac{577}{\sqrt{5}} = 258$$

$$45 = \frac{100}{\sqrt{5}}$$

$$20 = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20 \checkmark$$

Univerza v Ljubljani

Uporabna statistika, Uvod v statistiko

Inštitut za biostatistiko in medicinsko informatiko

3. vaja

Naloga 1 - centralni limitni izrek

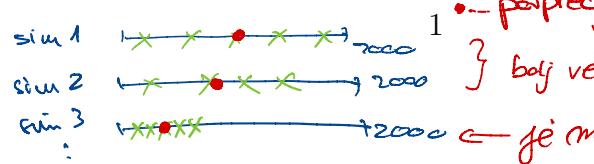
Statistiki si pri preučevanju lastnosti cenilk pogosto pomagamo s simulacijami. Namen te naloge bo s pomočjo simulacij preučiti lastnosti cenilke za populacijsko povprečje. Predpostavimo, da je porazdelitev neke številske spremenljivke v neskončno veliki populaciji normalna z nekim povprečjem ($\mu =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

povprečje vzorčnega povprečja

Oblika porazdelitve v populaciji	Velikost vzorca	Oblika porazdelitve vzorčnih povprečij	Pričakovana vrednost vzorčnih povprečij	Standardna napaka
Normalna $s \sigma = 100$	5	normalna	≈ 1000	45
Normalna $s \sigma = 100$	25	—	—	20
Normalna $s \sigma = 100$	100	—	—	10
Normalna $s \sigma = 50$ <i>mavrska</i>	100	—	—	5 <i>mavrska</i>
Normalna $s \sigma = 250$	100	—	—	25
Enakomerna na $[0,2000]$	$5 = m !!!$	mominalna	1000	≈ 258 (*)
Enakomerna na $[0,2000]$	25	normalna	1000	45
Eksponentna s param. 1	5 $\rightarrow \text{Exp}(1)$	asim., mavj ~ kdt populacija 1	—	0,45
Eksponentna s param. 1	25 $\mu = 1$ $Z = 1$	sim., pridružno normalna	1	0,2

$$\text{Vzorčno povpr.} = \frac{1}{m} \sum_{i=1}^m X_i$$

populacija:



$$\text{CLI: } \bar{X} \approx N(\mu, \frac{\sigma^2}{m})$$

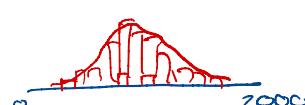
pridružno porazdelitev SE

- večji kol je m, bolj je porazdelitev \bar{X} podobna normalni

• ... povprečje

} bolj verjetni

← je mavj verjetna



pridružno normalno

To je centralni limitni izrek (CLT)!

- 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?

pri $1000 = \mu = \text{populacijsko povprečje}$

Normalna

majša

$$E(\bar{X}) = \mu \quad (\text{fj. } \bar{X} \text{ je nепривидна сењлка za } \mu)$$

pričakovana vrednost

večji $n \Rightarrow$ večja SE

večji vzorec \Rightarrow majša razpršenost
majša (SE)
 \approx
Standard error = st. napaka

- Kako na obliko porazdelitve ocene vzorčnega povprečja vpliva porazdelitev spremenljivke v populaciji?

CLT : \bar{X} porazdeljen pridížno normalno, ne glede na porazdelitev v populaciji.

- Večji $n \Rightarrow$ bolj normalno
- porazdelitev v populaciji bolj simetrična $\Rightarrow \bar{X}$ porazdeljen bolj normalno

- Zapišite formulo za standardno napako.

$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Velja točno, ne glede na porazdelitev v populaciji.

Če je $X \sim \text{normalno}$, potem je $\bar{X} \sim \text{normalno}$.

točno, ne več pridížno.

Koda za simulacijo (datoteka `simulacija.r`)

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

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

##zacni s simulacijo

$n \leftarrow 5$ ##velikost vzorca

$B \leftarrow 1000$ ##stevilo ponovitev simulacije

$Y_{\text{bar}} \leftarrow \text{rep}(\text{NA}, B)$ ##tu se shranijo povprecja v posameznem koraku simulacije
for (i in 1:B) { ##zacni s simulacijo

$\text{id} \leftarrow \text{sample}(1:\text{length}(y), n)$ ##slucajno izberi enote

$Y \leftarrow y[\text{id}]$ ##vrednosti na vzorcu

$Y_{\text{bar}}[i] \leftarrow \text{mean}(Y)$ ##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 $U(0, 2000)$:

$y \leftarrow runif(1000000, min=0, max=2000)$

##prikazi porazdelitev spremenljivke v histogramu:

`par(mfrow=c(2,1))`

`hist(y, xlim=c(0, 2000))`

##zacni s simulacijo

$n \leftarrow 5$

$B \leftarrow 1000$

$Y_{\text{bar}} \leftarrow \text{rep}(\text{NA}, B)$

for (i in 1:B) {

$\text{id} \leftarrow \text{sample}(1:\text{length}(y), n)$

$Y \leftarrow y[\text{id}]$

$Y_{\text{bar}}[i] \leftarrow \text{mean}(Y)$

}

`hist(Y.bar, xlim=c(0, 2000))`

`mean(Y.bar)`

`sd(Y.bar)`

Kaj nas zanimala? μ = PCP. povprečje tipično, podobno kot $S\sigma = 0.95$
 1. $\hat{\mu} = \bar{X}$ verjetnost za μ
 2. Interval zaupanja (IZ) ... Imamo 95% zaupanje, da je μ (prava vrednost) manjša v tem
Naloga 2 - Interval zaupanja za populacijsko povprečje če je X porazdeljena normalno

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

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \xrightarrow{\text{Standardizacija}} \frac{\bar{X} - \mu}{SE(\bar{X})} \sim N(0, 1)$$

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

ne poznamo

$$\frac{\bar{X} - \mu}{SE(\bar{X})} \sim t_{n-1}$$

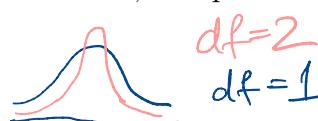
$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

kjer je $SE(\bar{X}) = s / \sqrt{n}$. Simulirajte iz normalne porazdelitve s povprečjem 1000 in $\sigma = 200$. Porazdelitev obih 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: t per. ker 1 parameter, df = degrees of freedom

- Katera krivulja se bolje prilega podatkom v primeru, ko uporabimo pravo standardno napako in katera, ko uporabimo ocenjeno standardno napako?

$N(0, 1)$



t_{n-1}

= stopinje prostosti

- Od česa je odvisna oblika t -porazdelitve?

zaupanje vs verjetnost 2
 $P(\mu \in (\bar{X} \pm t_{n-1} \cdot \frac{s}{\sqrt{n}})) = 0.95 \rightarrow$ verjetnost

$P(\mu \in S \pm 2 \cdot \frac{s}{\sqrt{n}}) = 0$ ali 1 \rightarrow zaupanje je μ med [4, 6]

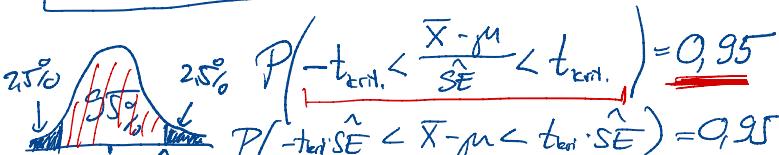
- Kdaj je t -porazdelitev bolj podobna standardni normalni porazdelitvi?

je $n = 100$, f. r. reči kolik je n , kolik je $t_{df=n-1}$
 podoba $N(0, 1)$

- Izpeljite formulo za interval zaupanja za populacijsko povprečje!

$$\frac{\bar{X} - \mu}{SE} \sim t_{n-1}$$

• Zelimo maziti interval, v katerega se nahaja μ s 95% verjetnostjo!



$$P\left(-t_{\text{krit.}} < \frac{\bar{X} - \mu}{SE} < t_{\text{krit.}}\right) = 0.95$$

$$P\left(-t_{\text{krit.}} \cdot SE < \bar{X} - \mu < t_{\text{krit.}} \cdot SE\right) = 0.95$$

$$P\left(-\bar{X} - t_{\text{krit.}} \cdot SE < -\mu < -\bar{X} + t_{\text{krit.}} \cdot SE\right) = 0.95$$

$$P\left(\bar{X} + t_{\text{krit.}} \cdot SE > \mu > \bar{X} - t_{\text{krit.}} \cdot SE\right) = 0.95$$

$$t_{\text{krit.}} = t_{df=n-1; 0.975} \quad P\left(\bar{X} - t_{\text{krit.}} \cdot SE < \mu < \bar{X} + t_{\text{krit.}} \cdot SE\right) = 0.95$$

$$t_{\text{krit.}} = t_{df=n-1; 0.025} \div 2 \quad P\left(\bar{X} - t_{\text{krit.}} \cdot SE < \mu < \bar{X} + t_{\text{krit.}} \cdot SE\right) = 0.95$$

$$2. \text{ možna označa}$$

$$\bullet \text{To NI referenčni interval!}$$

$$\bullet \text{ref. int. je uselovan in } 95\% \text{ srednjih vrednosti porazdelitve - terje } X.$$

95% IZ za μ : $[\bar{X} - t_{\text{krit.}} \cdot SE, \bar{X} + t_{\text{krit.}} \cdot SE]$

\rightarrow referenčna je ali $X \sim$ normal ali pa je

n dovolj velik (po CLT: $t_{\text{krit.}} = 2.575$)

\rightarrow je kjer zoli v 1/2

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 sharni 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 stopinjam 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 stopinjam prostost
```

Naloga 3 - Podatki iz ankete

POTRAVITI
NAVODILA.

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). Ženskacem, ki so študenti MF in VF.

- Izračunajte (vzorčni) povprečji

$$- \bar{x}_{Dent}^{m\text{en}\ddot{\text{s}}i} = 16,12$$

$$- \bar{x}_{Splošna}^{zenuške} = 14,35$$

- Zapišite formulo in izračunajte standardno napako za vzorce take velikosti (uporabite ocenjeni standardni odklon)

$$\hat{S}_{EM} = \frac{s_m}{\sqrt{n_m}} = 13,89 / \sqrt{43} = 2,12 \quad \left(S_{EM} = \frac{s_m}{\sqrt{n_m}} \text{ ne moremo izračunati!} \right)$$

$$\hat{SE}_{\bar{x}} = 11,7 / \sqrt{97} = 1,19$$

- Izračunajte 95% interval zaupanja (IZ) za populacijski povprečji. Uporabite kodo v R, enega izmed intervalov pa izračunajte tudi na roke.

- 95% IZ za μ_M : $\bar{X}_m \pm t_{\alpha/2} \cdot \hat{S}_{EM} = 16,12 \pm 2,018 \cdot 2,12$
 $t_{\alpha/2} = 42,0975 \rightarrow [11,84 ; 20,39]$

- 95% IZ za μ_Z : $[11,99 ; 16,7]$

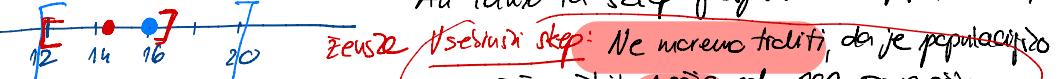
Interpretacija IZ za ženske:

S 95% zaupanjem je povprečno število ur uporabe interneta v populaciji študente MF med 12 urami in 16,7 urami.

- Primerjajte ju. Ali se prekrivata? Kaj lahko sklepamo?

↳ IZ za ženske je ozjivi, ker je jih je doleti več (na strani IZ vpliva razprenočnost in n)

Da. n_M Na vzorcu: povpr. M je večje. Ali lahko ta sklep posledi na pop? NE.



- Izračunajte še 99% IZ in ju primerjajte.

Če $\bar{x}_M = [10,4 ; 21,83]$
 $\bar{x}_Z = [11,23 ; 17,47]$
 Povpr. povpr. M je večje od povpr. povpr. Z.

$$M: [10,4 ; 21,83]$$

6

$$Z: [11,23 ; 17,47]$$

Primerjajte med 95% in 99% IZ: 99% IZ je širši od 95%.

Stopnje zaupanja

- Kaj bi se zgodilo z IZ, če bi v vzorec zajeli več študentov?

zobral si se

*Ali je povprečje? ?
smiselna mera
niničimo IZ za
median*

- Ali je predpostavka o normalnosti porazdelitve naše spremenljivke smiselna? Kako to vpliva na rezultate?

ose porazdelitvi ($\mu \neq \bar{x}$) sta asymetrični

*formula za IZ je lahko OK, ker ima dolg velik n
(formula velja zaradi CL) → zaupanje v izračunanem IZ
je res 95% priblžno.*

- Denimo, da raziskavo ponovite 100 krat in vsakič izračunate 95% IZ. V koliko primerih pričakujete, da bo:

– populacijsko povprečje zajeto v intervalih? *95*

– vzorčno povprečje zajeto v intervalih? *100 (vredno v sredini)*

$$\bar{X} \pm \square$$

```
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)
```

*NA
NASEDNJIH
VAJAH O
SPODNJEM*

Test t → 1 vzorec, μ

$$H_0: \mu = \mu_0 \rightarrow T = \frac{\bar{X} - \mu_0}{\text{SE}}^7$$

$$H_A: \mu \neq \mu_0$$

R: t.test(x)

$$\mu_0 = 0$$

$$H_0: \mu = 0$$

*P ↔ IZ
ODNOS*