Intervali poverenja

Danijel Subotic

INTERVALI POVERENJA

Želimo da nađemo takve statistike U_n i V_n da za nepoznati parametar Teta i za vrednost beta (unapred zadatom) važi:

$$P\{U_n <= \Theta <= V_n\} = \beta$$

beta nazivamo nivo poverenja.

(1) Naći 95%-tni interval poverenja za parametar "m", ako je naša slučajna veličina $X \sim \mathcal{N}(m, \sigma^2)$. Poznato nam je σ^2 i n - broj elemenata u prostom slučajnom uzorku. I neka su to $n = 64$, $\overline{x_n} = 118$, $\beta = 0.95$ i $\sigma^2 = 100$.

Rešenje:

Imamo da važi sledeće

$$T = \frac{\overline{X_n} - m}{\sigma}\sqrt{n} \sim \mathcal{N}(0, 1)$$

to T ce da nam sluzi kao pomocna statistika za odredjivanje statistika U_n i V_n.

Kako je T iz normalne raspodele to uvek mozemo naci konstantu c tako da vazi

$$P\{|T| \leq c\} = \beta$$

i to je $c = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$.

Dalje iz nejednakosti u verovatnoci dobijamo sledece

$$U_n = \overline{X_n} - \frac{c \cdot \sigma}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{c \cdot \sigma}{\sqrt{n}} = V_n$$

```
#
n=64
x_sr=118
beta=0.95
sigma_kv= 100 # - poznata vrednost


# Resenje:
c=qnorm((1+beta)/2, 0, 1) # isto sto i qnorm((1+beta)/2) isto sto i qnorm(0.975)
c
```

```
## [1] 1.959964
```

```
interval_1=x_sr-c*sqrt(sigma_kv/n)
interval_2=x_sr+c*sqrt(sigma_kv/n)
interval_1
```

```
## [1] 115.55
```

```r
interval_2
```

```
## [1] 120.45
```

```r
(interval_poverenja=c(interval_1, interval_2))
```

```
## [1] 115.55 120.45
```

(1*) U slučaju da nije poznato da je sigma_kv=100, moramo oceniti disperziju. Pretpostavimo da je vrednost ocenjene disperzije jednaka takodje 100. Naci interval poverenja u tom slucaju.

Rešenje: jedina razlika u izvođenju u odnosu na prethodni zadatak je u tome što sada nije reč o normalnoj raspodeli za pomoćnu statistiku T, već je Studentova sa $n-1$ stepeni slobode.

```r
c = qt(p = (1+beta)/2, df = n-1) # ivverzna funkcija raspodele za studentovu
c
```

```
## [1] 1.998341
```

```r
ocenjena_sigma_kv = 100
interval_1=x_sr-c*sqrt(ocenjena_sigma_kv/n)
interval_2=x_sr+c*sqrt(ocenjena_sigma_kv/n)

c(interval_1, interval_2)
```

```
## [1] 115.5021 120.4979
```

(2) Naci 95%-tni interval poverenja za m na osnovu uzorka obima n ako je $X \sim \mathcal{N}(m, \sigma^2)$, σ^2 - nepoznato. Vrednosti su $n = 25$, $\overline{x_n} = 2.6$ $\overline{s_n} = 170.36$ i $\beta = 0.95$.

Resenje:

```r
n=25
x_sr=2.6
ocena_sigma_kv=170.36
beta=0.95

c=qt((1+beta)/2, n-1) #qt - kvantili studentove raspodele, drugi argument predstavlja broj stepeni slob
print(c)
```

```
## [1] 2.063899
```

```r
#Dok je interval jednak:
(interval_poverenja=c(x_sr-c*sqrt(ocena_sigma_kv/n), x_sr+c*sqrt(ocena_sigma_kv/n) ))
```

```
## [1] -2.787685  7.987685
```

U slučaju da primenimo ovakav interval poverenja i na prvi zadatak. Moramo da dobijemo nesto širi

```r
n=64
x_sr=118
beta=0.95
sigma_kv= 100 # ovde sada smatramo da je to ocenjena vrednost

c1=qt((1+beta)/2, n-1)
c1
```

```
## [1] 1.998341
```

```
interval_1=x_sr-c1*sqrt(sigma_kv/n)
interval_2=x_sr+c1*sqrt(sigma_kv/n)
(interval_poverenja=c(interval_1, interval_2))
```

```
## [1] 115.5021 120.4979
```

```
# I zaista vidimo da je ova malcice siri interval poverenja.
```

U prva dva zadatka smo pominjali intervale poverenja u slucaju normalne raspodele $\mathcal{N}(m, \sigma^2)$ za parametar "m".

To je sve bilo za slucaj kada se ocenjuje parametar m. Dok ako zelimo da intervalno ocenimo σ^2 tu je prica malo drugacija.

Statistika koja se koristi za to je:

$$W = \frac{(n-1)\tilde{S_n}^2}{\sigma^2} \sim \chi^2_{n-1}$$

tj Hi-kvadrat sa $n-1$ stepeni slobode. gde je: n - obim uzorka, $\tilde{S_n}^2$ - popravljena uzoracka disperzija, σ^2 - parametar koji ocenjujemo.

I sasvim je prirodno da ta statistika, kojom se ocenjuje parametar σ^2 ima neku nenegativnu raspodelu, kao sto je Hi-kvadrat.

Za razliku od ocene za "m", gde su u oba slucaja statistike imale simetricne raspodele oko nule, hi-kvadrat nije simetricna, i stavise nenagativna je. Iz tog razloga na vise nacina se moze prirodno definisati interval poverenja za σ^2. kao na primer:

Napomena: sve granice intervala koje cu da navodim dole, tj (A_n, B_n, U_n, V_n) su zapravo granice za statistiku(!) ne i za parametar. A tek kad se dobiju granice za stistiku - moramo da nadjemo bas Interval poverenja za trazeni parametar σ^2.

(a) Donji interval, tj $(0, A_n]$, tj da pocinje uvek od nule, pa ide do vrednosti gde se dostigne trazena β, gde se A_n nalazi bas iz:

$$P\{W <= A_n\} = \beta \implies A_n = F^{-1}(\beta),$$

gde je F^{-1} inverzna funkcija za funkciju raspodele koju ima statistika W. sto se u R-u poziva na sledeci nacin: $A_n = qchisq(beta, n-1)$ gde je $qchisq()$ - funkcija pomocu koje dobijamo kvantile hi-kvadrat raspodele.

(b) Dvostrani interval: $[U_n, V_n]$, gde se granice biraju tako da je levo i desno od tog intervala ista povrsina koja je jednaka $(1-\beta)/2$, za dato "β". Nalazimo ih pomocu:

$$P\{W <= U_n\} = (1-\beta)/2 \implies U_n = qchisq((1-\beta)/2, n-1).$$

$$P\{W <= V_n\} = (1+\beta)/2 => V_n = qchisq((1+\beta)/2, n-1).$$

Ovde je $(1+\beta)/2$, jer je prosto povrsina koja je pre vrednosti V_n mora da bude $\beta + (1-\beta)/2$ sto je bas jednako $(1+\beta)/2$.

(v) Gornji interval: slicno kao donji, smo sto nije ogranicen sa gornje strane pa je interval $[B_n, +\infty)$. B_n iz formule:

$$P\{W <= B_n\} = 1-\beta \implies B_n = qchisq(1-\beta, n-1).$$

(3) Neka je dato: $X \sim \mathcal{N}(m, \sigma^2)$. $n = 20$, $\tilde{S_n}^2 = 21.12$ - popravljena uzoracka disperzija. Naci dvostrani interval poverenja za parametar σ^2 za nivo poverenja $\beta = 0.99$.

Resenje:

Potrebno je naci interval oblika:

$$W = \frac{(n-1)\tilde{S_n}^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$P\{U_n <= W <= V_n\} = 0.99$$

Prema gore navedenoj teoriji: vrednost za V_n mora za bude vrednost $F^{-1}\left(\frac{1.99}{2}\right)$. $v_n = qchisq(1.99/2, 19)$

Vrednost za U_n mora za bude vrednost $F^{-1}\left(\frac{0.01}{2}\right)$ $u_n = qchisq(0.01/2, 19)$

Odakle dobijamo sledeci izraz iz kojeg moramo da nadjemo interval poverenja.

$P\{6.85 \leq (n-1)\tilde{S}_n^2/\sigma^2 \leq 38.6\}$

$P\{6.85 \leq (n-1)\tilde{S}_n^2/\sigma^2 \leq 38.6\}$

Sto je dalje jednako: $P\{(n-1)*21.12/38.6 <= \sigma^2 <= (n-1)*21.12/6.85\}$

Donja vrednost intervala poverenja: a = (n-1)*21.12/v_n

Gornja vrednost intervala poverenja:b = (n-1)*21.12/u_n

```
n=20

v_n = qchisq(1.99/2, n-1)
u_n = qchisq(0.01/2, n-1)

a = (n-1)*21.12/v_n
b = (n-1)*21.12/u_n

(interval=c(a,b)) # trazeni interval poverenja
```

```
## [1] 10.40064 58.63262
```

Probajte da uradite gornji i donji interval poverenja za ovaj primer.

Jos jedna vrsta intervala poverenja, koju cemo da razmotrimo su intervali poveranja za parametar p u Bernulijevoj raspodeli. Neka je $X \sim Ber(p) = Bin(1, p)$ i zelimo da ocenimo parametar p.

Tada to mozemo oceniti intervalno pomocu statistike:

$$Z = (\overline{X_n} - p)\sqrt{n})/\sqrt{p(1-p)} \sim \mathcal{N}(0, 1).$$

I dalje je ocenjivanje jako slicno kao za parametar "m" sto smo imali, s tim sto ovde imamo "malo vise" javljanja parametra "p" u datoj statistici, nego sto je to bilo u slucaju parametra "m".

Stoga u opstem slucaju moracemo da resavamo kvadratnu nejednacinu i da pazimo da granice ne izadju izvan intervala $(0, 1)$ - jer ocenjujemo parametar p - koji predstavlja verovatnocu nekog dogadjaja!

Dok u nekim specijalnim slucajevima mozemo umesto $p(1-p)$ koristiti $\overline{X_n}(1-\overline{X_n}))$. Za to mora da bude ispunjeno: - parametar p - ne sme da bude blizu 0 ili 1. - $n \geq 100$, tj da imamo dovoljno mnogo elemenata uzorka. U tom slucaju vise nemamo kvadratne nejednacine vec direktne vrednosti koje cemo da ubacimo i brzo izrazimo i dobijemo trazeni interval poverenja.

(4) Od 500 ljudi, 185 se izjasnilo protiv. Naci 90% interval poverenja za verovatnocu da se slucajno odabrana osoba izjasni protiv.

Resenje:

Kako se ta statistika aproksimira normalnom raspodelom, odmah nadjemo trazenu vrednost za "c". Primetimo da je $(1+\beta)/2 = (1+0.9)/2 = 0.95$ Dalje imamo:

$$P\{-c <= (\overline{X_n} - p)/\sqrt{p(1-p)}\sqrt{n} <= c\}$$

mozemo da aproksimiramo sa $\overline{x_n}(1-\overline{x_n})$ jer je n=500>100 i p~0.57

c=qnorm(0.95) c

```
c=qnorm(0.95)
c
```

```
## [1] 1.644854
```

```
n=500
x_sr=185/500
x_sr
```

```
## [1] 0.37
```

```
# dalje nadjemo donju, odnosno gornju vrednost intervala:
a=x_sr-c*sqrt(x_sr*(1-x_sr))/sqrt(n)
b=x_sr+c*sqrt(x_sr*(1-x_sr))/sqrt(n)

# trazeni interval poverenja za "p":
(interval=c(a,b))
```

```
## [1] 0.3344849 0.4055151
```

(5) Istrazivac procenjuje procenat gojaznih osoba sa visokim krvnim pritiskom. Od 25 ispitanika - 20 ima visok pritisak. Naci 95% interval poverenja za parametar p, koji predstavlja verovatnocu da gojazna osoba ima visok krvni pritisak.

Resenje:

Ovde ne mozemo da primenimo aproksimaciju, jer n=25<100. Zato moramo da resimo kvadratnu nejednacinu. Ali pre toga vrednost za c:

```
beta=0.95
c=qnorm((1+beta)/2)
c
```

```
## [1] 1.959964
```

```
#
x_sr=20/25
# P{-c<=(X_{sr}-p)/sqrt(p*(1-p))*sqrt(n)<=c}
# = P{-1.96<=(0.8-p)/sqrt(p*(1-p))*5<=1.96}
# = P{|0.8-p|<=1.96*sqrt(p*(1-p))/5} sto sada kvadriramo:
# = P{0.8^2-1.6*p+p^2<=1.96^2*(p-p^2)/25}
# odakle klasicnim metodama dobijamo:
#(vrednosti su zaokruzene na dve decimale)
a=0.61
b=0.91
(interval=c(a,b))
```

```
## [1] 0.61 0.91
```

(6) Naci 90% i 80% interval poverenja za istrazivanje u kojem je ucestvovalo 100 ljudi i srednja vrednost je 0.45.

Resenje:

Mozemo iskoristiti aproksimacije za parametar p zbog velicine uzorka.

```
X_sr = 0.45
```

```r
# Z = (X_sr - p)/ sqrt[X_sr(1-X_sr)] * sqrt(n)
c_90 = qnorm((1+0.9)/2) # 1.64
c_80 = qnorm((1+.8)/2) # 1.28

(interval_90 = c(0.45 - c_90*sqrt(0.45*.55) / 10, 0.45 + c_90*sqrt(0.45*.55) / 10 ) )
```

```
## [1] 0.3681696 0.5318304
```

```r
(interval_80 = c(0.45 - c_80*sqrt(0.45*.55) / 10, 0.45 + c_80*sqrt(0.45*.55) / 10 ) )
```

```
## [1] 0.3862436 0.5137564
```

(7) Student zeli da istrazi procenat levorukih osoba. Ispitano je 100 ljudi od kojih je samo 5 levorukih. Ako se nadje 90% interval poverenja da li ce on sadrzati vrednost 0.1? Ako ne - koliko-procentni interval poverenja treba da trazimo?

Da li 0.1 pripada (a, b) , ako P(a<=p<=b)=0.9?

Resenje: $p \sim 5\%$, sto jeste blizu nule. Ovde ne smemo da koristimo aproksimaciju za disperziju uzorka.

Test statistika nam je

$$Z = \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}}$$

tj. dalje imamo:

$$P\{a <= p <= b\} = 0.9$$

i

$$P\{-c <= Z <= c\} = 0.9$$

Odakle imamo

$$c = \Phi^{-1}(0.95)$$

Odnosno imamo

$$P\{|\overline{X_n} - p| <= c\sqrt{p(1-p)/n}\} = P\{\overline{X_n}^2 - 2p\overline{X_n} + p^2 <= c^2 p(1-p)/n\}$$

I resavanjem ove kvadratne nejednacine po p dobijamo trazene granice intervala poverenja.

```r
a = 0.025
b = 0.099
(interval_1 = c(a, b) )
```

```
## [1] 0.025 0.099
```

```r
# ne upada 0.1

# ali za beta = 0.91 vec upada jer je c tada
c = qnorm((1+0.91)/2) # 1.69
# i interval je (0.024, 0.10001) otprilike.
```

(8) Od poslednjih 10 puta kada je tostirani sendvic pao - devet puta je pao na "ukusnu stranu". Ako je uzorak iz Bernulijeve raspodele sa parametrom p, naci 80% interval poverenja da sendvic pada na "ukusnu stranu".

Resenje:

$\overline{X_{10}} = 0.9$ sto nije dovoljno daleko od 1, a i $n = 10$ je bas malo. Test statistika je

$$Z = \frac{\overline{X_{10}} - p}{\sqrt{p(1-p)/10}}$$

I opet resavanjem odgovarajuce kvadratne nejednacine dobijamo granice intervala poverenja za p.

Kod:

```
#P{a<=p<=b} = P{-c <= Z <= c} = 0.8
c = qnorm(1.8/2) # 1.28
a = 0.72
b = 0.97
interval = c(a, b)
```

Ukoliko ipak primenimo aproksimaciju:

```
(interval = c(0.9 - c*sqrt(0.1*0.9/10),0.9 + c*sqrt(0.1*0.9/10) ) )
```

```
## [1] 0.7784213 1.0215787
```

```
# Moramo da izbacimo ove vrednost preko 1 iz intervala poverenja.
(interval = c(max(0, 0.9 - c*sqrt(0.1*0.9/10)), min(1, 0.9 + c*sqrt(0.1*0.9/10))))
```

```
## [1] 0.7784213 1.0000000
```

```
# Dobijamo skroz drugacije intervale.
```

(9) Izvrseno je ispitivanje 250 ljudi, ocenjeno p odgovora je 0.45. Isto istrazivanje je ponovljeno na 1000 ljudi i dobijen isti odnos. Naci 95% interval poverenja u oba slucaja. Koliko puta je drugi manji od prvog? Koliko veliko treba da bude n u istrazivanju da bi 95% interval poverenja bio duzine ne vece od 0.01.

Resenje:

$$Z = \frac{\overline{X_n} - p}{\sqrt{\overline{X_n}(1-\overline{X_n})/n}}$$

```
c_95 = qnorm(0.975) # 1.96


interval_250 = c(0.45 - c_95*sqrt(0.45*0.55/250), 0.45 + c_95*sqrt(0.45*0.55/250))
# (0.39, 0.51)
diff(interval_250) # 0.12334
```

```
## [1] 0.1233377
```

```
interval_1000 = c(0.45 - c_95*sqrt(0.45*0.55/1000), 0.45 + c_95*sqrt(0.45*0.55/1000))
interval_1000
```

```
## [1] 0.4191656 0.4808344
```

```
# (0.42, 0.48)
diff(interval_1000) # 0.06167
```

```
## [1] 0.06166883
```

```r
# Bolje u obliku funkcije da se zapise.
duzina_intervala <- function(n)
{
  interval_1000 = c(0.45 - c_95*sqrt(0.45*0.55/n), 0.45 + c_95*sqrt(0.45*0.55/n))
  return(diff(interval_1000))
}
duzina_intervala(38031)
```

```
## [1] 0.009999927
```

```r
# Trazeno n nalazimo na sledeci nacin:
(duzina_intervala(250)/0.01)^2*250
```

```
## [1] 38030.44
```

```r
# ~38031
```