# Bootstrap metoda za pouzdane intervale

**Maričić, Tea**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Tea Maričić

# BOOTSTRAP METODA ZA POUZDANE INTERVALE

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, 2022

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

*Hvala Bogu !!!*
*Hvala mojoj obitelji, posebno roditeljima i sestrama Tini i Eni.*
*Hvala mentorici doc.dr.sc Snježani Luburi Strunjak na svoj pomoći tijekom pisanja ovog rada.*
*Hvala svim mojim prijateljima koji su uvijek bili tu za mene, bodrili me, vjerovali u mene i pomagali ma koliko minornom tu pomoć smatrali.*
*Hvala najboljoj ekipi s PMF-a koja mi je uljepšala studentske dane. Hvala vam za sve divne trenutke i vraćanja u 5 ujutro.*
*Hvala mojim curama iz učionice na podršci i svim divnim trenutcima, naravno u pauzama od učenja ;).*
*Hvala svim dragim ljudima u eSTUDENTu koje sam upoznala tijekom proteklih godina i bez kojih ne bih sada bila ovdje gdje jesam. Poklonili ste mi ono najvrjednije, a to je vjera u sebe.*

# Sadržaj

# Uvod

**Definicija** Neka je $\Omega$ neprazan skup. Familija podskupova $\mathcal{F}$ od $\Omega$ zove se $\sigma-$ algebra (ili $\sigma-$ događaja) ako vrijede sljedeća tri svojstva:

($i$) $\Omega \in \mathcal{F}$;

($ii$) Ako je $A \in \mathcal{F}$, onda je i $A^c \in \mathcal{F}$ (zatvorenost na komplement);

($iii$) Ako su $A_j \in \mathcal{F}$, $j \in \mathbb{N}$, onda je i $\bigcup_{j=1}^{\infty} \in \mathcal{F}$ (zatvorenost na prebrojive unije);

Uređen par $(\Omega, \mathcal{F})$ zove se *izmjeriv prostor*.

**Definicija** Neka su $(X, \mathcal{F})$ i $(Y, \mathcal{G})$ izmjerivi prostori. Za funkciju $f : X \to Y$ reći ćemo da je $(\mathcal{F}, \mathcal{G})-$izmjeriva ako je $f^{-1}(B) \in \mathcal{F}$ za sve $B \in \mathcal{G}$.

**Definicija** Neka je $X$ skup, a $\mathcal{U}$ familija skupova koja sadrži $\emptyset$ i $X$ te koja je zatvorena na proizvoljne unije i konačne presjeke. $\sigma-$algebru generiranu famlilijom $\mathcal{U}$ nazivamo Borelovom $\sigma-$algebrom na $X$ i označavamo sa $\mathcal{B}(X)$, njezine elemente nazivamo Borelovim skupovima.

**Definicija** Neka je $\Omega$ neprazan skup i $\mathcal{F}$ $\sigma-$ algebra događaja. Vjerojatnost na izmjerivom prostoru $(\Omega, \mathcal{F})$ je funkcija $\mathbb{P} : \mathcal{F} \to [0, 1]$ koja zadovoljava sljedeća tri aksioma:

($i$) (nenegativnost) $\forall A \in \mathcal{F}, \mathbb{P}(A) \geq 0$;

($ii$) (normiranost) $\mathbb{P}(\Omega) = 1$;

($iii$) ($\sigma-$ aditivnost) Za svaki niz $(A_j)_{j\in\mathbb{N}}$ po parovima disjunktnih događaja $A_j \in \mathcal{F}(A_i \cap A_j = \emptyset, i \neq j)$ vrijedi:

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}\left(A_j\right).$$

Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$ zove se *vjerojatnosni prostor*.

**Definicija** Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ je slučajna varijabla na $\Omega$ ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathcal{B}$, tj. $X^{-1}(\mathcal{B}) \subset \mathcal{F}$.

**Defnicija** Neka je $X$ slučajna varijabla na $\Omega$. Funkcija distribucije od $X$ je funkcija $F_X : \mathbb{R} \to [0, 1]$ definirana sa:

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}\left(X^{-1}(-\infty, x]\right) =$$
$$= \mathbb{P}\{\omega \in \Omega; X(\omega) \leqslant x\} = \mathbb{P}\{X \leqslant x\}, \quad x \in \mathbf{R}$$

gdje je $\mathbb{P}_X$ vjeojatnosna mjera definirana sa

$$\mathbb{P}_X(B) = \mathbb{P}\left(X^{-1}(B)\right) = \mathbb{P}\{\omega \in \Omega; X(\omega) \in B\} =$$
$$= \mathbb{P}\{X \in B\}, \quad B \in \mathcal{B}.$$

Neka je $X$ slučajna varijabla čiju distribuciju proučavamo.

**Definicija** Slučajni uzorak duljine $n$ za $X$ je niz od $n$ nezavisnih i jednakodistribuiranih slučajnih varijabli $X_1, X_2, ...X_n$ koje imaju istu distribuciju kao X.
Za $\omega \in \Omega$ je $x_1 = X_1(\omega), x_2 = X_2(\omega), ..., x_n = X_n(\omega)$ jedna realizacija slučajnog uzorka i zovemo je uzorak. Statistika je funkcija slučajnog uzorka.

Distribucija slučajne varijable $X$ je često opisana parametrima koje pokušavamo procijeniti.

**Definicija** Slučajna varijabla $X : \Omega \to \mathbb{R}$ je apsolutno neprekidna ako postoji $f : \mathbb{R} \to [0, \infty\rangle$ takva da za sve $x \in \mathbb{R}$ vrijedi:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Funkcija $f$ se zove funkcija gustoće od $X$.

# Metoda maksimalne vjerodostojnosti

Neka je $(x_1, x_2, ...x_n)$ opaženi uzorak za slučajnu varijablu $X$ s gustoćom $f(x|\theta)$ gdje je $\theta = (\theta_1, ...\theta_k) \in \Theta \subset \mathbb{R}^k$ nepoznati parametar.

Definiramo **funkciju vjerodostojnosti** $L : \Theta \to \mathbb{R}$ sa:

$$L(\theta) = f(x_1|\theta) \cdot \cdots f(x_n|\theta), \quad \theta \in \Theta.$$

Vrijednost $\widehat{\theta} = \widehat{\theta}(x_1, ...x_n) \in \Theta$ za koju je

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

zovemo **procjena maksimalne vjerodostojnosti**. Statistika $\hat{\theta}(X_1, ..., X_n)$ je **procjenitelj metodom maksimalne vjerodostojnosti** ili kraće **MLE**.

# Karakteristične funkcije

Neka je $F$ ograničena funkcija distribucije na $\mathbb{R}$.

**Definicija** Karakteristična funkcija of $F$ jest funkcija $\varphi : \mathbb{R} \to \mathbb{R}$ definirana sa:

$$\varphi(t) = \int_{-\infty}^{+\infty} e^{itx}\, dF(x) = \int_{-\infty}^{+\infty} cos(tx)\, dF(x) + i \cdot \int_{-\infty}^{+\infty} \sin(tx)\, dF(x).$$

U detalje gornjih oznaka nećemo ulaziti stoga zainteresiranog čitatelja upućujemo na [4].

**Definicija** Neka je $X$ slučajna varijabla s funkcijom distribucije $F_X$. Karakteristična funkcija $\varphi_X$ od $X$ je karakteristična funkcija od $F_X$.

Sljedeći teorem pokazuje nam kako karakteristična funkcija na jedinstven način određuje slučajnu varijablu i obratno. Njega navodimo bez dokaza te zainteresiranog čitatelja pozivamo da za dokaz pogleda [4], poglavlje 14.

**Teorem 0.0.1.** *(Teorem jedinstvenosti)*
*Neka su $F_1$ i $F_2$ funkcije distribucije na $\mathbb{R}$ i neka one imaju istu karakterističnu funkciju, tj. za sve $t \in \mathbb{R}$ vrijedi*

$$\int_{-\infty}^{+\infty} e^{itx}\, dF_1(x) = \int_{-\infty}^{+\infty} e^{itx}\, dF_2(x).$$

*Tada je $F_1 = F_2$.*

**Teorem 0.0.2.** *Ako su $X_1, ..X_n$ nezavisne slučajne varijable, tada vrijedi:*

$$\varphi_{\sum_{k=1}^n X_k}(t) = \prod_{k=1}^{n} \varphi_{X_k}(t).$$

*Dokaz.*

$$\varphi_{\sum_{k=1}^n X_k}(t) = \mathbb{E}\left[e^{it\sum_{k=1}^n X_k}\right] = \mathbb{E}\left[\prod_{k=1}^{n} e^{itX_k}\right] \overbrace{=}^{\text{teorem 11.5,[4]}} \prod_{k=1}^{n} \mathbb{E}\left[e^{itX_k}\right] = \prod_{k=1}^{n} \varphi_{X_k}(t).$$

□

Sljedeći teorem daje nam vezu između momenata slučajne varijable i karakterističnih funkcija. Ona nam zapravo kaže kako karakterističnu funkciju aproksimirati polinomom konačnog stupnja koristeći momente slučajne varijable. Njega također navodimo bez dokaza te zainteresiranog čitatelja upućujemo na [4], Teorem 13.7.

**Teorem 0.0.3.** *Ako je* $\mathbb{E}[|X|^n] < \infty$ *za neki* $n \in \mathbb{N}$*, tada* $\varphi_X$ *ima* $k-$*tu derivaciju za* $k \leq n$ *i vrijedi:*

$$\varphi_X(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!}\mathbb{E}[X^k] + o(t^n), \quad t \in \mathbb{R},$$

*gdje je*

$$o(t^n) = \frac{(it)^n}{(n-1)!}\int_{-\infty}^{\infty}\int_0^1 x^n(1-y)^{n-1}(e^{itxy}-1)\,dy\,dF_X(x), \quad t \in \mathbb{R}$$

*i* $\lim\limits_{t\to 0}\dfrac{o(t^n)}{t^n} = 0.$

Gornju tvrdnju ćemo koristiti u dokazu centralnog graničnog teorema u prvom poglavlju. Budući da je to područje teorije vjerojatnosti izrazito kompleksno navedimo samo kako ćemo koristiti gornju aproksimaciju za standardnu normalnu slučajnu varijablu polinomom stupnja 2 koja glasi:

$$X \sim N(0,1) \Rightarrow \varphi_X(t) = 1 + it - \frac{t^2}{2} + o(t^2).$$

# Jackknife metoda

Pretpostavimo da imamo slučajni uzorak $\mathbf{x} = (x_1, ..., x_n)$ i procjenu statistike $\hat{\theta}$ dobivenu na temelju tog uzorka.

Jackknife metoda bazira se na *leave one out* principu. Točnije, $\forall i \in \{1, ...n\}$ definiramo novi uzorak duljine $n - 1$ na sljedeći način:

$$\mathbf{x}_{(i)} = (x_1, x_2, ...x_{i-1}, x_{i+1}, ...x_n).$$

Gornji uzorak nazivamo **jackknife uzorak**.

Ponovno, $\forall i \in \{1, ...n\}$ računamo statistiku od interesa. Označimo ju s $\hat{\theta}_{(i)}$.

Jackknife procjena parametra dobivena je kao aritmetička sredina svih gornjih statistika, točnije

$$\hat{\theta}_{jack} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}.$$

Ova metoda smatra se pretečom bootstrap metode što je i tema ovog diplomskog rada i o kojoj ćemo detaljnije diskutirati u sljedećem poglavlju.

# Poglavlje 1

# Pouzdani intervali i Bootstrap metoda

## 1.1 Pouzdani intervali

### Motivacija

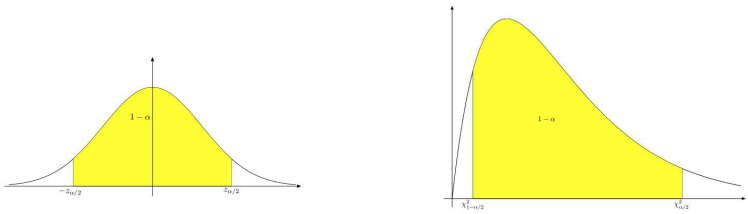
Uz problem određivanja vrijednosti nekog parametra usko je vezan pojam pouzdanih intervala. Naime, kako bismo procijenili neki parametar koristimo uzorak iz neke distribucije. No, da bi procjena parametra bila što preciznija potreban je što reprezentativniji uzorak. Vrlo često taj uzorak nije dovoljno reprezentativan stoga moramo biti u mogućnosti izraziti svoju nesigurnost u vrijednost procijenjenog parametra. Jedan od najklasičnijih pristupa ovom problemu je koristeći MLE procjenitelj na osnovu fiksnog uzorka pronaći dovoljno 'pouzdan' parametar uz pretpostavku da određene statistike imaju unaprijed definiranu razdiobu. Stoga, uvodimo sljedeću definiciju.

**Definicija** Neka su $L_n = l_n(X_1, \ldots, X_n)$ i $D_n = d_n(X_1, \ldots, X_n)$ statistike slučajnog uzorka $X_1, \ldots, X_n$. Za $[L_n, D_n]$ kažemo da je $(1-\alpha)\cdot 100\%$ pouzdani interval za parametar $\theta$ ako vrijedi

$$\mathbb{P}(L_n \leq \theta \leq D_n) \geq 1-\alpha, \quad \alpha \in \langle 0, 1\rangle.$$

Najčešće vrijednosti parametra $\alpha$ su 0.01 i 0.05 koji redom određuju 90% i 95% pouzdane intervale za dani parametar.

Naravno, gornja procjena je samo gruba aproksimacija traženog intervala koja uvelike ovisi o uzorku.
Prilikom računanja pouzdanih intervala za dane parametre vrlo često su korištene unaprijed definirane distribucije statistika od interesa.



Slika 1.1: 2 Simetrična i asimetrična distribucija parametra

Ako za parametar $\theta$ vrijedi

$$\mathbb{P}(L_n \leq \theta) = \frac{\alpha}{2}, \quad \mathbb{P}(D_n \geq \theta) = \frac{\alpha}{2}$$

radi se o simetričnoj distribuciji parametra iz čega očito slijedi

$$\mathbb{P}(\theta \in [L_n, D_n]) = 1 - \alpha, \quad \alpha \in \langle 0, 1 \rangle.$$

Računanje pouzdanih intervala vrlo često svodi se na distribucije koje su simetrične, no ne mora uvijek biti tako.

Na slici 1.1:2 vidimo primjer simetrične distribucije parametra (lijevo) i asimetrične distribucije parametra (desno).

Pogledajmo koje su najčešće korištene distribucije pri osnovnim tehnikama računanja pouzdanih intervala koje ćemo spominjati i u ovom radu.

## 1.2 Osnovne distribucije parametara

### Normalna distribucija

**Definicija**
Neka su $m, \sigma \in \mathbb{R}, \sigma > 0$. Neprekidna slučajna varijabla $X$ ima **normalnu distribuciju** s parametrima $m$ i $\sigma^2$ i pišemo $X \sim N(m, \sigma^2)$ ako joj je gustoća $f : \mathbb{R} \to \mathbb{R}$ dana sa:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in \mathbf{R}.$$

Ukoliko je $m = 0$ i $\sigma = 1$ kažemo da $X$ ima **standardnu normalnu** distribuciju, tj $X \sim N(0, 1)$.

Najčešće korištena metoda određivanja pouzdanih intervala je svođenje na centralni granični teorem koji kaže prema kojoj distribuciji asimptotski konvergira niz nezavisnih jednakodistribuiranih slučajnih varijabli.

Postoji više verzija, no mi ćemo spomenuti Levyev.

**Teorem 1.2.1.** *(Levy) Neka je* $(X_n, n \in \mathbf{N})$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem* $m$ *i varijancom* $\sigma^2, 0 < \sigma^2 < \infty$ *i neka je* $S_n = \sum_{k=1}^{n} X_k$ $(n \in \mathbf{N})$. *Tada vrijedi*

$$\frac{S_n - n \cdot m}{\sigma \sqrt{n}} \xrightarrow{D} N(0,1) \quad za\ n \to \infty.$$

*Dokaz.* Definirajmo $Z_k = \frac{X_k - m}{\sigma}, (k \in \mathbb{N})$. Tada imamo

$$Y_n = \frac{S_n - m}{\sigma \sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Z_k \quad n \in \mathbb{N}.$$

$(Z_k, k \in \mathbb{N})$ je niz nezavisnih jednakodistribuiranih slučajnih varijabli i

$$\mathbb{E}[Z_k] = 0, \quad \mathbb{E}[Z_k^2] = \mathrm{Var} Z_k = 1.$$

Prema teoremu 0.0.3. vrijedi:

$$\varphi_{Z_k}(t) = 1 - \frac{t^2}{2} + o(t^2).$$

Budući da su $(Z_k)_{k \in \mathbb{N}}$ jednakodistribuirane imaju iste karakterističke funkcije, tj.$o(t^2)$ ne ovisi o $k$.
Koristeći se nezavisnošću imamo:

$$\varphi_{Y_n}(t) = \prod_{k=1}^{n} \varphi_{Z_k}\left(\frac{t}{\sqrt{n}}\right) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{2}\right)\right]^n, \quad t \in \mathbb{R}.$$

Iz leme 5.1 u [4] zaključujemo da vrijedi

$$\lim_{n \to \infty} \varphi_{Y_n}(t) = e^{-\frac{t^2}{2}}.$$

pa tvrdnja teorema sada slijedi iz teorema neprekidnosti. □

**Korolar 1.2.2.** *Neka je* $(X_n, n \in \mathbf{N})$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem* $\mu$ *i varijancom* $\sigma^2, 0 < \sigma^2 < \infty$. *Vrijedi*

$$\frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} \overset{D}{\sim} N(0,1).$$

*Dokaz.* Dokaz korolara se nalazi u [3], poglavlje 14. □

Koristeći gore navedeni korolar vidimo da za $(1-\alpha)\cdot 100\%$ pouzdani interval za parametar $\mu$ vrijedi sljedeće:

$$\begin{aligned}
-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} \leq z_{\frac{\alpha}{2}} \quad &\Big/ \cdot \frac{\sigma}{\sqrt{n}} \\
-\frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} \leq \bar{X}_n - \mu \leq \frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} & \\
-\frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n \leq -\mu \leq \frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n \quad &\Big/ \cdot(-1) \\
\frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n \geq \mu \geq -\frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n & \\
\Rightarrow \mu \in \left[-\frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n, \frac{\sigma\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n\right]. &
\end{aligned}$$

Da bismo u bolje razumjeli gornji izraz uvodimo sljedeću definiciju.

**Definicija**

Kažemo da je $z_\alpha$ $\alpha - ti$ **kvantil** slučajne varijable $X$ ako vrijedi

$$\mathbb{P}(X \leq z_\alpha) = \alpha.$$

U gornjoj formuli $z_{\frac{\alpha}{2}}$ je pripadni kvantil normalne razdiobe. Naravno, gornji račun je izveden pod pretpostavkom da je naš uzorak iz neke distribucije s poznatom varijancom te ga smatramo asimptotskim. Vrlo često ta pretpostavka nije zadovoljena ili imamo manji uzorak, stoga moramo na drugi način procijeniti pouzdani interval za naš parametar . Na slici 1.2 vidimo graf funkcije gustoće normalne distribucije.

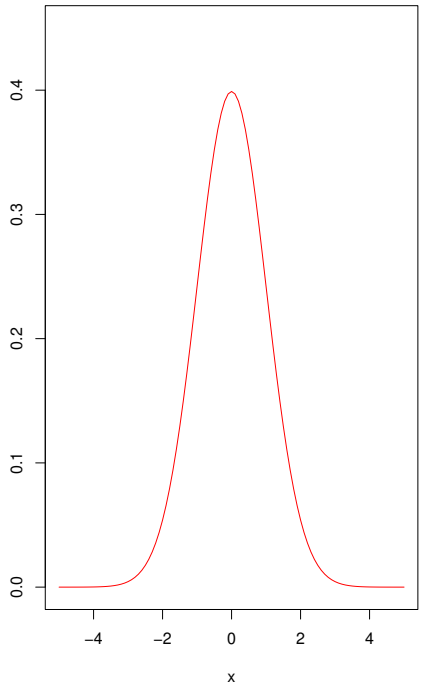
## Studentova t-distribucija

Gornji račun je opravdan pod pretpostavkom da je poznata varijanca te da je uzorak dovoljno velik kako bismo na njega uspjeli primijeniti centralni granični teorem. No, vrlo često nam varijanca uzorka nije poznata te moramo naći alternativni način procjene nekog parametra. S tim u cilju uvodimo sljedeće definicije.

**Definicija**

Kažemo da neprekidna slučajna varijabla $X$ ima $\chi^2(n)$ distribuciju s parametrom $n$ što označavamo $X \sim \chi^2(n)$ ako joj je gustoća $f : \mathbb{R} \to \mathbb{R}$ dana sa:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x > 0 \\ 0, & x \leqslant 0. \end{cases}$$




Slika 1.2: Funkcija gustoće standardne normalne slučajne varijable

Gdje je $\Gamma(x)$ funkcija definirana sa:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt.$$

**Definicija**
Kažemo da slučajna varijabla T ima Studentovu ili $t$-distribuciju s $n$ stupnjeva slobode ako postoje nezavisne slučajne varijable $X$ i $Y$ t.d. je $X \sim N(0,1)$, $Y \sim \chi^2(n)$ tako da je:

$$T \overset{D}{=} \frac{X}{\sqrt{\frac{Y}{n}}}.$$

Pišemo: $T \sim t(n)$.
Funkcija gustoće dana je sljedećom formulom:

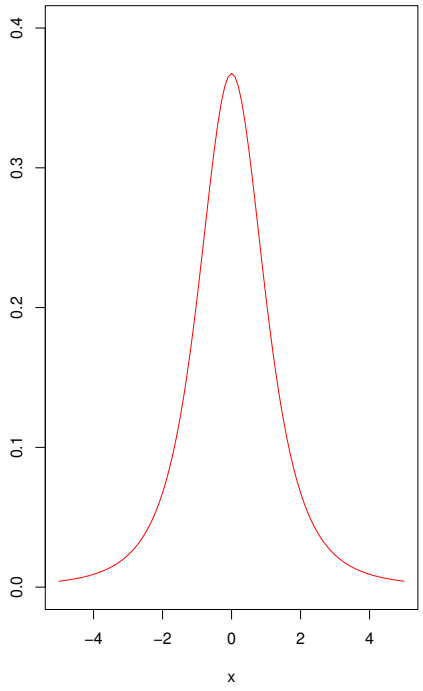
$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{1}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{inače} . \end{cases}$$

Ovdje je važno za primijetiti da se opet radi o simetričnoj distribuciji budući da nam tada treba samo 1 kvantil t-distribucije.

Analogno kao i u normalnom slučaju dolazimo do pouzdanog intervala za očekivanje uzorka:

$$\begin{aligned} -t_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{S_n} \sqrt{n} \leq t_{\frac{\alpha}{2}} \quad & \Big/ \cdot \frac{S_n}{\sqrt{n}} \\ -\frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} \leq \bar{X}_n - \mu \leq \frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} & \\ -\frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n \leq -\mu \leq \frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} - \bar{X}_n \quad & \Big/ \cdot (-1) \\ \frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n \geq \mu \geq -\frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n & \\ \Rightarrow \mu \in \left[ -\frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n, \frac{S_n \cdot t_{\frac{\alpha}{2}}}{\sqrt{n}} + \bar{X}_n \right]. & \end{aligned}$$

Na slici 1.3 vidimo graf funkcije gustoće studentove t-distribucije sa 3 stupnja slobode.




Slika 1.3: Funkcija gustoće studentove razdiobe

## 1.3 Bootstrap metoda

Pretpostavimo sada kako imamo fiksan uzorak $\mathbf{x} = (x_1, x_2, ..., x_n)$ iz nepoznate vjerojatnosne distribucije $F$ te trebamo procijeniti neki parametar od interesa $\theta = t(F)$ na temelju gornjeg uzroka pri čemu je $t$ proizvoljna funkcija spomenutog uzorka iz dane distribucije $F$, npr. aritmetička sredina.

U tu svrhu izračunamo traženi parametar $\hat{\theta} = t(\mathbf{x})$ iz $\mathbf{x}$.
Pitanje koje se nameće ovdje je sljedeće: Koliko je točna procjena $\hat{\theta}$?

Bootstrap metoda uvedena je 1979. godine kao jedan od načina procjene standardne pogreške statistike $\hat{\theta}$. Ova procjena ne zahtijeva nikakve pretpostavke i može se izračunati za svaku statistiku.

### Neparametarski Bootstrap

Ponovno pretpostavljamo kako imamo fiksan uzorak $\mathbf{x}$ duljine $n$ s empirijskom funkcijom distribucije $\hat{F}$ pri čemu pretpostavimo da vrijedi sljedeće:

$$\mathbb{P}(X_i = x_i) = \frac{1}{n} \quad \forall i \in \{1, , , n\}\,.$$

Bootstrap uzorak definiran je kao slučajan uzorak veličine $n$ iz distribucije $\hat{F}$ pri čemu pretpostavljamo kako taj uzorak više nije isti kao naš početni, već se radi o uzorku nastalom metodom ponovljenog uzorkovanja iz $\mathbf{x}$. Kako bi razlikovali ta dva pojma uvodimo sljedeću notaciju za uzorak dobiven ponovljenim uzorkovanjem:
(1.1) nam definira novi poduzorak $\mathbf{x}^*$, a (1.2) nam označava kako taj novi poduzorak dolazi iz iste distribucije $\hat{F}$ kao i naš početni uzorak $\mathbf{x}$

$$\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*) \tag{1.1}$$

$$\hat{F} \rightarrow (x_1^*, x_2^*, ..., x_n^*). \tag{1.2}$$

Ovdje je potrebno naglasiti kako bootstrap metoda ponovljenog uzorkovanja rezultira novim, randomiziranim uzorkom iste duljine kao i početni, ali s ponavljanjem. Točnije, sasvim legitimna je situacija u kojoj imamo 2 ili više istih vrijednosti u uzorku $\mathbf{x}^*$.

Iz gore definiranog bootstrap uzorka računamo statistiku od interesa, onu istu koju smo računali na početnom uzorku, točnije tražimo:

$$\hat{\theta}^* = t(\mathbf{x}^*).$$

Bootstrap metoda daje odgovor na pitanje procjene standardne pogreške procjenitelja $\hat{\theta}$. Budući da ne postoji zatvorena formula ovaj algoritam smatra se dobrom numeričkom aproksimacijom iste.

Suština bootstrap metode bazira se na ponavljanju gornjeg procesa $B$ puta. Procjenu standardne pogreške dobivamo iz standardne devijacije svakog poduzorka nastalog ponovnim uzorkovanjem iz početnog.

Algoritam je dan na sljedećoj slici:

---

**Algorithm 1** Bootstrap metoda za procjenu standardne pogreške od $\hat{\theta}$

---

1) Izaberi B nezavisnih bootstrap uzoraka ($\mathbf{x}^{*1}$,...,$\mathbf{x}^{*B}$) od kojih se svaki sastoji od $n$ podataka dobivenih ponovnim uzorkovanjem s ponavljanjem iz $\mathbf{x}$;

2) Izračunaj statistiku od interesa na svakom od B poduzoraka

$$\widehat{\theta^*}(b) = t(\mathbf{x}^{*b}), \quad b \in \{1, 2..., B\};$$

3) Procijeni standardnu grešku $se_F(\hat{\theta})$ iz standardne devijacije statistika izračunatih B puta:

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[ \widehat{\theta^*}(b) - \widehat{\theta^*}(.) \right]^2 \right\}^{\frac{1}{2}},$$

pri čemu je

$$\widehat{\theta^*}(.) = \sum_{b=1}^{B} \widehat{\theta^*}(b)/B.$$

---

Iz gornjeg algoritma definiramo bootstrap grešku procjene parametra $\hat{\theta}$ sa $se_F(\hat{\theta})$, preciznije

$$\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}} = se_{\hat{F}}(\hat{\theta}^*).$$

Činjenica da se $\hat{se}_B$ približava $se_{\hat{F}}$ kako $B \to \infty$ ekvivalentna je tome da se empirijska standardna devijacija približava populacijskoj.

Bootstrap procjena $se_{\hat{F}}(\hat{\theta}^*)$ i njezina aproksimacija $\hat{se}_B$ zovu se *neparametarske Bootstrap procjene* jer dolaze iz nama nepoznate distribucije.

Pogledajmo primjer simulacije neparametarskim bootstrapom. Pretpostavimo kako imamo neki uzorak:

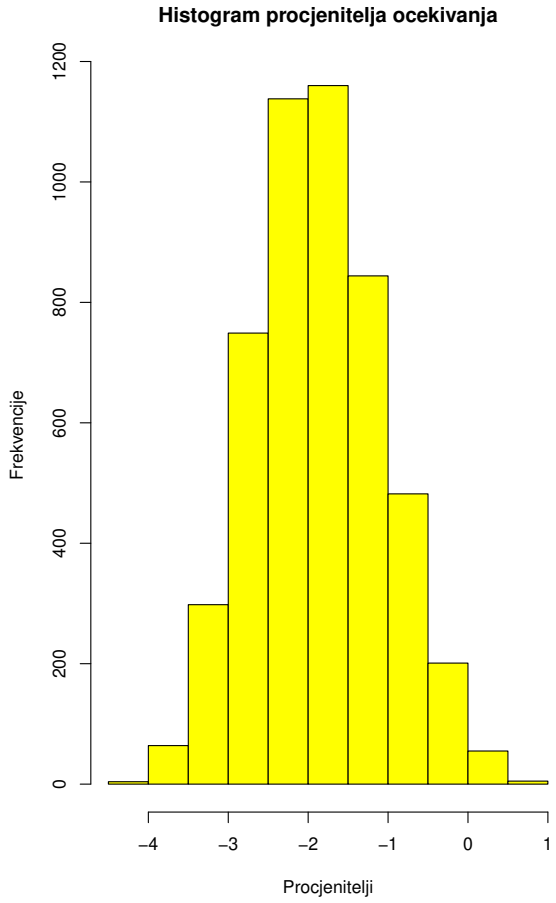
$$\mathbf{x} = (-3.543, -3.973, 3.091, -0.607, -1.204, -4.983, -1.655, -3.042, 1.483, -4.322)$$

$$\Rightarrow \overline{\mathbf{x}} = -1.8754.$$

Radimo neparametarski bootstrap uzorkovanjem 5000 puta iz gornjeg uzorka. Koristeći gornji algoritam dobivamo kako je naša neparametarska bootstrap procjena te procjena standardne pogreške procjenitelja $\hat{se}_B$ :

$$\hat{\theta} = -1.866455, \quad \hat{se}_B = 0.8.$$

Na slici 1.4 vidimo histogram procjena očekivanja neparametarskog bootstrapa dobivenih uzorkovanjem 5000 puta iz zadanog uzorka:




Slika 1.4: Histogram neparametarskih bootstrap procjena očekivanja populacije

Ova procjena potkrijepljena je R-kodom u [6.1].

Kasnije ćemo vidjeti iz koje je distribucije naš uzorak te ćemo dati usporedbu s drugom procjenom.

U slučaju da nam je poznata distribucija $\hat{F} = \mathbf{F}$ početnog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ kažemo kako se radi o *parametarskoj procjeni*.

## Parametarski Bootstrap

Pretpostavimo sada kako imamo fiksan uzorak $\mathbf{x}$ duljine $n$, ali sada nam je poznata klasa distribucija $F$ iz koje dolazi. Nadalje, pretpostavimo kako ta distribucija ovisi o nepoznatom parametru $\theta$. Stoga, da bismo što bolje odredili $F$ procjenjujemo ga sa $F_{\hat{\theta}}$ gdje je $\hat{\theta}$ procjenitelj za $\theta$, recimo MLE procjenitelj. Preciznije,

$$\hat{\theta} = t(\mathbf{x}),$$

pri čemu je $t$ ponovno neka statistika od interesa, npr. aritmetička sredina.

Nakon pronalaska odgovarajuće procjene $\hat{\theta}$ za $\theta$ daljnji postupak pronalaska bootstrap procjenitelja nastavljamo s $F_{\hat{\theta}}$.

Ako nam je $F$ poznata u potpunosti nemamo potrebe računati procjenitelj te samo nastavljamo cijeli postupak uzorkujući iz $F$.

Cijeli postupak dan je sljedećim algoritmom.

---

**Algorithm 2** Bootstrap metoda za procjenu standardne pogreške od $\hat{\theta}$

---

1) Realiziraj slučajni uzorak $(x_1^*, ..., x_n^*)$ duljine $n$ iz $F$ (ako je poznata) ili $F_{\hat{\theta}}$ ;

2) Izračunaj statistiku od interesa na svakom od B poduzoraka

$$\widehat{\theta^*}(b) = t(\mathbf{x}^{*b}), \quad b \in \{1, 2..., B\} ;$$

3) Procijeni standardnu grešku $se_F(\hat{\theta})$ iz standardne devijacije statistika izračunatih B puta:

$$\widehat{se}_B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left[ \widehat{\theta^*}(b) - \widehat{\theta^*}(.) \right]^2 \right\}^{\frac{1}{2}},$$

pri čemu je

$$\widehat{\theta^*}(.) = \sum_{b=1}^{B} \widehat{\theta^*}(b)/B.$$

---

Pogledajmo primjer jedne simulacije parametarskog bootstrapa za očekivanje populacije. Pretpostavimo kako imamo uzorak duljine 10 pri čemu je

$$X_i \sim N\left(0, \left(2\sqrt{2}\right)^2\right) \quad \forall i \in \{1, ..., 50\}.$$

.

Znamo da vrijedi

$$\mathbb{E}[X] = \overline{\mathbf{x}}.$$

Napomenimo kako smo sad u situaciji u kojoj je $F$ potpuno poznata. U slučaju da nam je poznata klasa distribucija (npr. studentova), a ne i parametri, iste bismo morali procijeniti MLE metodom.

Ako nam ni klasa distribucija ne bi bila poznata trebali bismo ju na neki način odrediti iz uzorka. Jedna od metoda bi bila crtanje histograma na temelju kojega bismo dobili neke slutnje te adekvatni statistički testovi kojima bismo svoje teze potvrdili. I u ovom slučaju parametre bismo procjenjivali MLE metodom.

Naš uzorak je dan sa:

$$\mathbf{x} = (-3.543, -3.973, 3.091, -0.607, -1.204, -4.983, -1.655, -3.042, 1.483, -4.322)$$

$$\Rightarrow \overline{\mathbf{x}} = -1.8754.$$

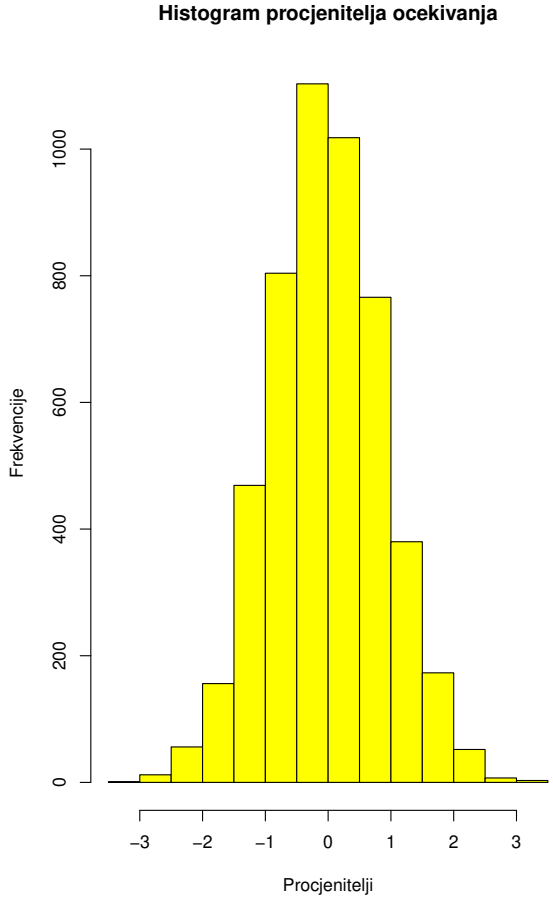
Radimo parametarski bootstrap gdje uzorkujemo $B = 5000$ puta uzorke duljine $n = 10$ iz $N\left(0, \left(2\sqrt{2}\right)^2\right)$ kako bismo dobili traženu procjenu za očekivanje početnog uzorka. Koristeći gornji algoritam dobijemo da je bootstrap procjena očekivanja populacije i standardna greška procjene $\hat{se}$ jednaka:

$$\hat{\theta} = -0.02845307, \quad \hat{se}_B = 0.9.$$

Na slici 1.5 vidimo histogram procjenitelja očekivanja parametarskog bootstrapa iz $N\left(0, \left(2\sqrt{2}\right)^2\right)$ dobivenih uzorkovanjem 5000 puta iz dane razdiobe.

Vidimo da naš uzorak dolazi iz normalne distribucije, točnije $N(0, \left(2\sqrt{2}\right)^2)$. U slučaju neparametarske procjene vidimo kako je naša procjena za očekivanje populacije $\hat{\theta}_{\text{neparam}} \approx -1.87$ dok je u parametarskom slučaju ta procjena $\hat{\theta}_{\text{param}} \approx -0.03$ iz čega vidimo kako je parametarska procjena bliža stvarnoj, koja je 0.

Ovaj rezultat ne iznenađuje s obzirom na to da je uzorkovanje iz originalne distribucije bolja generalizacija nego uzorkovanje s ponavljanjem iz unaprijed definiranog uzorka. Ova procjena potkrijepljena je R-kodom u [6.2].




Slika 1.5: Histogram procjena očekivanja parametarske bootstrap metode

# Poglavlje 2

# Bootstrap $t$ pouzdani interval

### Motivacija

U prethodnom poglavlju spomenuli smo i ukratko objasnili pouzdane intervale i bootstrap metodu.

Naveli smo i izveli 2 formule za procjenu $(1-\alpha)\cdot 100\%$ pouzdanog intervala za populacijsko očekivanje temeljeno na nekom uzorku pod određenim pretpostavkama.

Dali smo kratak uvid u bootstrap metodu procjene parametra na 2 načina: parametarski i neparametarski te smo uz jednu jednostavnu simulaciju vidjeli koliko se te procjene na istom uzorku mogu razlikovati.

U narednim poglavljima spojit ćemo ta 2 pojma te proći kroz neke načine određivanja pouzdanih intervala bootstrap metodom. Korištenjem bootstrap metoda možemo dobiti pouzdane intervale bez pretpostavki normalnosti. Najprije se bavimo *bootstrap-t* pouzdanim intervalima.

## 2.1 Metoda

Sam princip računanja bootstrap t-pouzdanog intervala za očekivanje populacije zasniva se ponovno na formiranju bootstrap uzorka, no ovaj put računamo $T$ statistiku za svaki uzorak.

Podsjetimo se, ona je dana formulom

$$T = \frac{\overline{X}_n - \mu}{S_n}\sqrt{n}.$$

Prije nego krenemo u sam algoritam trebamo objasniti kako parametar $\mu$ iz gornje formule poistovjećujemo s aritmetičkom sredinom polaznog uzorka $\overline{X}$ dok statistiku $\overline{X}_n$ poistovjećujemo s aritmetičkom sredinom $b$-tog bootstrap uzorka.

Također, $S_n$ poistovjećujemo sa standardnom devijacijom $b$-tog bootstrap uzorka.

Sam princip dan je sljedećim algoritmom:

---

**Algorithm 3** Bootstrap t pouzdani interval

---

1) Generiraj slučajni uzorak $(x_1^*, ...x_n^*)$ duljine $n$ iz početnog (s ponavljanjem).

2) Izračunaj T - statistiku, tj. realizaciju:

$$T = \frac{\overline{X}_n - \overline{X}}{S_n^*}\sqrt{n},$$

gdje su:

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k^*, \quad S_n^* = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k^* - \overline{x^*}_n\right)^2};$$

3) Ponoviti korake 1 i 2 $B$ puta - rezultat je niz $T$ statistika:

$$T_1^*, \ldots, T_B^*.$$

4) Sortirati gornji uzorak i pronaći odgovarajuće $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$ - kvantile.

5) Pouzdani interval dan je formulom:

$$\left[\overline{X} - T_{1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \overline{X} - T_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right],$$

gdje je

$$S = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \overline{X}\right)^2}.$$

---

Ovdje je važno napomenuti kako $T_{\frac{\alpha}{2}}$ i $T_{1-\frac{\alpha}{2}}$ možemo izračunati na jednostavan način. Ako zahtijevamo $(1-\alpha)\cdot 100\%$ pouzdan interval sa $B$ bootstrap uzoraka tada $T_{\frac{\alpha}{2}}$ računamo kao $B \cdot \alpha$-ti broj po redu u sortiranom uzorku. Ako taj umnožak nije prirodan broj, zaokružimo ga na najbliži.

Na sasvim analogan način računamo $T_{1-\frac{\alpha}{2}}$, samo tražimo $\left(1 - \frac{\alpha}{2}\right) \cdot B$-tu vrijednost po redu te u slučaju da se ne radi o prirodnom broju postupamo analogno kao gore.

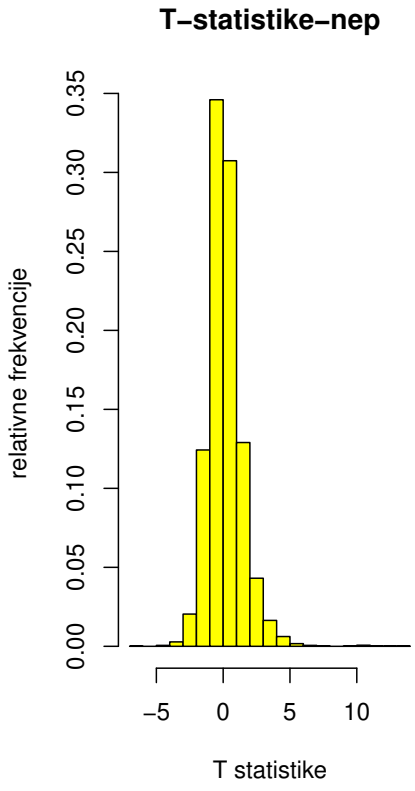
## 2.2 Simulacija

Ovdje ćemo vidjeti simulaciju parametarske i neparametarske procjene pouzdanog intervala. U oba slučaja promatramo populacijsko očekivanje s brojem replikacija 10000.

### Neparametarska procjena

Na iste te podatke studentiziranom neparametarskom bootstrap metodom određujemo pouzdani interval ponovno za populacijsko očekivanje. Simulaciju radimo po gornjem algoritmu generirajući poduzorke duljine 10 čime dolazimo do intervala $[-3.411, -0.292]$.

Na slici vidimo histogram realizacija T-statistika za neparametarsku procjenu. Simulacija je potkrijepljena R-kodom u [6.3].




Slika 2.1: T-statistike - neparametarska procjena

Uočimo kako nam stvarno populacijsko očekivanje, 0 ne upada u pouzdani interval što svakako nismo htjeli pa zaključujemo da u ovom slučaju baš studentizirani bootstrap pouzdani interval nije najoptimalnija procjena.

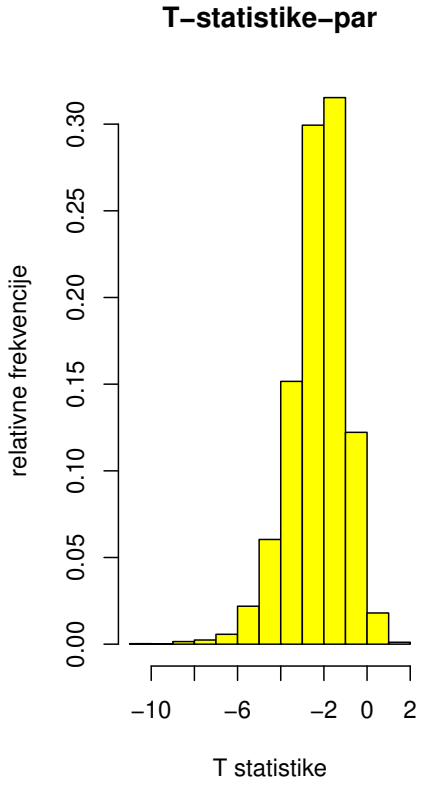
Pogledajmo sada rezultate za parametarsku procjenu.

## Parametarska procjena

Procjenu smo dobili na analogan način kao gore. Jedina razlika je u simulaciji koju nismo radili iz uzorka već iz distribucije kojoj pripada naš uzorak, $N(0.2\ \sqrt{2})$.
parametarski. Primjenjujući gornji algoritam na naše podatke dolazimo do procjene parametarskog studentiziranog pouzdanog intervala za populacijsko očekivanje koji glasi:

$$[-1.772917, 2.48077].$$

Na slici ponovno vidimo histogram T-statistika. Simulacija je potkrijepljena R-kodom u [6.4].




Slika 2.2: T-statistike - parametarska procjena

Uočimo kako nam ovdje stvarno očekivanje upada u gornji interval što nas navodi na zaključak da je parametarska procjena ovoga puta točnija nego neparametarska.

Razlozi za takav zaključak jasni su sami od sebe. Naime, uvijek je bolja generalizacija uzorkovanjem iz veće klase interesnih uzoraka nego iz jednog koji nam je dostupan. Nadalje, ako gledamo histograme T-statistika za parametarsku i neparametarsku procjenu vidimo kako ni jedna nije simetrična. Parametarska je nagnuta desno, dok je neparametarska lijevo. To nas ne iznenađuje budući da naš polazni uzorak ima više negativnih nego pozitivnih vrijednosti stoga se kod neparametarskog uzorkovanja više sredina nalazi lijevo od 0. S druge strane, kod parametarske simulacije vidimo obrnutu situaciju.

Budući da naši histogrami nisu simetrični drugačije metode daju poprilično drugačije procjene stoga nas gornji rezultat i ne iznenađuje toliko. Stoga, ako histogrami naših statistika nisu simetrični neparametarska procjena bootstrap pouzdanih intervala možda i neće dati najbolje rješenje.

# Poglavlje 3

# Bootstrap percentilni pouzdani intervali

### Motivacija

U ovom poglavlju opisujemo drugi pristup za određivanje bootstrap pouzdanog intervala zasnovanog na percentilima distribucije bootstrap uzorka dane statistike.

## 3.1 Metoda

Diskusija u prošlom poglavlju dala je jednostavan algoritam za određivanje pouzdanog intervala. Metoda koju obrađujemo u ovom poglavlju još je jednostavnija.

Ponovno koristimo bootstrap metodu ponovnog uzorkovanja te na temelju nje dolazimo do kumulativne funkcije distribucije bootstrap vrijednosti statistike od interesa. Na temelju te distribucije određujemo granice pouzdanog intervala preko kvantila prethodno navedene distribucije.

Drugim riječima za $(1-\alpha)\cdot 100\%$ pouzdan interval, vrijedi:

$$[L_n, D_n] = \left[\hat{F}^{-1}\left(\frac{\alpha}{2}\right), \hat{F}^{-1}\left(1-\frac{\alpha}{2}\right)\right].$$

gdje je $\hat{F}$ kumulativna funkcija distribucije bootstrap vrijednosti statistike od interesa.

Algoritam je ukratko opisan na sljedećoj stranici:

**Algorithm 4** Bootstrap percentilni intervali

1) Generiraj slučajni uzorak $(x_1^*, ...x_n^*)$ duljine $n$ iz početnog (s ponavljanjem).
2) Izračunaj statistiku od interesa, npr. aritmetičku sredinu:

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k^*.$$

3) Ponoviti korake 1 i 2 $B$ puta - rezultat je niz statistika:

$$\overline{X}_1^*, \ldots, \overline{X}_B^*.$$

4) Sortirati gornji uzorak i pronaći odgovarajuće kvantile $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$.
5) Pouzdani interval dan je formulom:

$$\left[\hat{F}^{-1}\left(\frac{\alpha}{2}\right), \hat{F}^{-1}\left(1 - \frac{\alpha}{2}\right)\right],$$

gdje je $\hat{F}(x)$ kumulativna funkcija distribucije statistike od interesa,

Kvantile gornje distribucije tražimo na način opisan u drugom poglavlju.

## 3.2 Simulacija

Radimo 2 simulacije na našem početnom uzorku procjene pouzdanog intervala bootstrap percentilnom metodom. U parametarskom i neparametarskom slučaju promatramo populacijsko očekivanje s brojem replikacija 10000.

### Parametarska procjena

Kako bismo došli do parametarske procjene primijenit ćemo gornji algoritam. Najprije uzorkujemo 10000 puta uzorke duljine 10 iz naše distribucije $N(0, 2\sqrt{2})$ te za svaki novi uzorak računamo aritmetičku sredinu.
Rezultat je niz aritmetičkih sredina koje na kraju sortiramo.
U koraku (5) uveli smo funkciju $\hat{F}$ koja je definirana na uzorku od 10000 aritmetičkih sredina sljedećom formulom: $\hat{F} : \mathbb{R} \to [0, 1]$ :

$$\hat{F}(x) = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}_{\{X_i \leq x\}}.$$

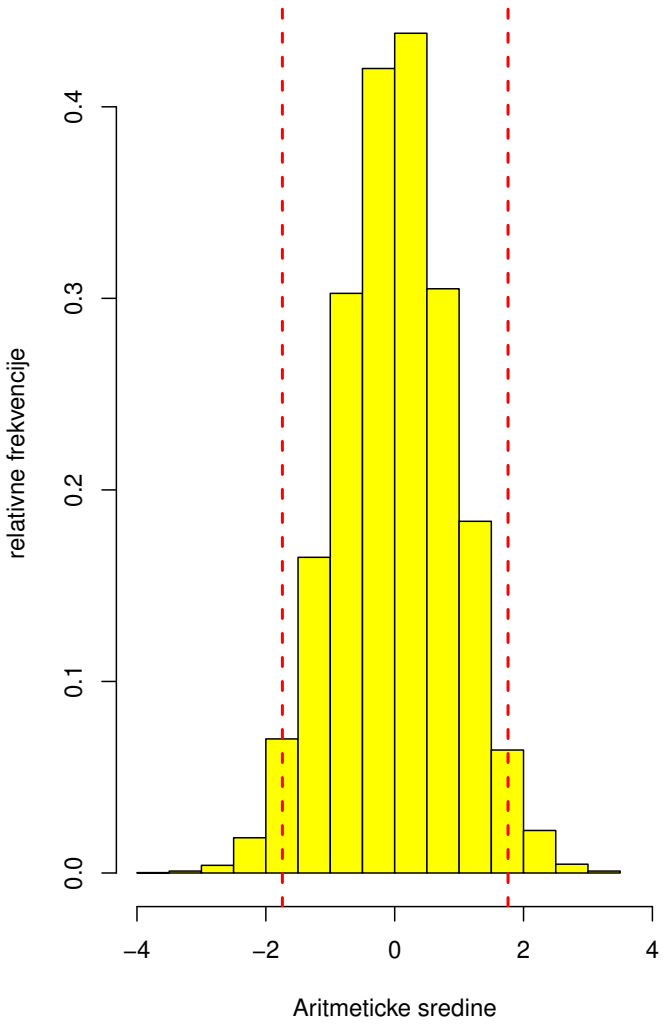
Budući da imamo 10000 aritmetičkih sredina u našem slučaju $B = 10000$.
Ako tražimo 95% pouzdan interval moramo uzeti $\alpha = 0.05$ čime su granice našeg intervala dane sa:

$$\hat{F}^{-1}(0.025) = -1.757, \quad \hat{F}^{-1}(0.975) = 1.758.$$

U konačnici dobivamo 95% pouzdan interval za populacijsko očekivanje koji iznosi

$$[-1.757, 1.758].$$

Na slici 3.1 vidimo histogram naših aritmetičkih sredina, a crvenom bojom su iscrtane granice našeg intervala. Simulacija je potkrijepljena R-kodom u [6.5].




Slika 3.1: Aritmetičke sredine parametarske bootstrap metode sa kvantilima

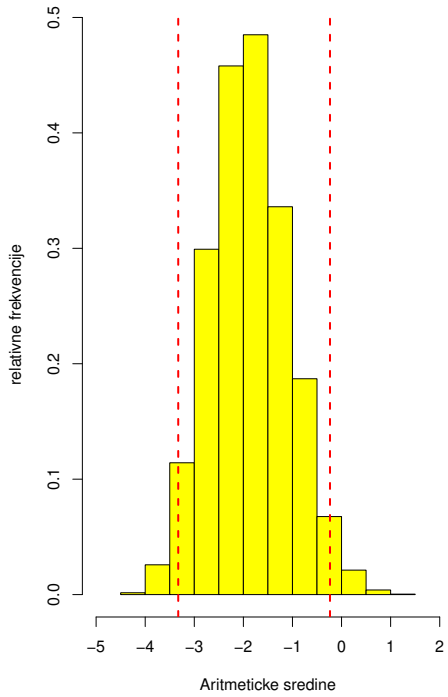
## Neparametarska procjena

Na iste te podatke percentilnom neparametarskom bootstrap metodom određujemo pouzdani interval ponovno za populacijsko očekivanje. Princip određivanja ostaje isti kao i funkcija $\hat{F}$ jedino što pri generiranju našeg uzorka svaki put s ponavljanjem uzorkujemo

iz početnog uzorka. Analogno određujemo granice pomoću $\hat{F}$ uz $\alpha = 0.05$
te neparametarskom metodom dolazimo do rezultata

$$[-3.411, -0.292].$$

Na slici vidimo histogram naših statistika, a crvenom bojom su iscrtane granice našeg intervala. Simulacija je potkrijepljena R-kodom u [6.6].




Slika 3.2: Aritmetičke sredine neparametarske bootstrap metode sa kvantilima

Vidimo kako je neparametarska procjena puno lošija od parametarske, bez obzira na velik broj simuliranja uzorka. U tom slučaju vidimo da nam stvarno populacijsko očekivanje, tj. 0 ne upada u 95% pouzdan interval.

U parametarskom slučaju stvarno očekivanje nam upada u 95% pouzdan interval i sredina intervala je jako blizu 0. No, nekako imamo osjećaj da je ta procjena mogla biti oštrija. I to je upravo ono o čemu pišemo u sljedećem poglavlju.

# Poglavlje 4

# Bolji bootstrap pouzdani intervali

### Motivacija

Jedan od glavnih ciljeva bootstrapa je dati interval koji je blizu pravog pouzdanog intervala za populacijski parametar u slučaju kada nam teorija omogućava da to izračunamo. Minimalan uvjet koji očekujemo je da nam stvarna vrijednost parametra upadne u dobiveni interval. No, u gornjim primjerima smo vidjeli kako nam to i nije uvijek ispunjeno. Posebno u poglavlju 2 s bootstrap t-metodom.

Malo bolji rezultati su bili kod percentilne metode gdje je u parametarskoj procjeni stvarna vrijednost očekivanja bila unutar procijenjenog pouzdanog intervala. No, ipak te rezultate možemo još poboljšati što je i tema ovog poglavlja.

U ovom poglavlju opisat ćemo poboljšanu verziju percentilne metoda, zvane $\mathrm{BC}_a$ čija skraćenica dolazi od engleskog termina *bias-corrected and accelerated*.

## 4.1 $\mathrm{BC}_a$ metoda

$\mathrm{BC}_a$ metoda zapravo predstavlja poboljšanje percentilne metode. Sam proces računanja ovih pouzdanih intervala je kompleksniji.

Podsjetimo se, pri računanju percentilnog bootstrap pouzdanog intervala definirali smo niz statistika dobivenim ponovnim uzorkovanjem, bilo iz uzorka bilo iz distribucije kojoj on pripada. Rezultat je bio niz statistika:

$$\overline{X}_1^*, \ldots, \overline{X}_B^*.$$

Percentilnom metodom vrlo jednostavno smo računali pouzdani interval pomoću $\alpha$-kvantila, točnije pouzdani interval bio je definiran sa

$$(\overline{X^*}^{(\alpha)}, \overline{X^*}^{(1-\alpha)}).$$

$\mathrm{BC}_a$ pouzdani intervali također su određeni kvantilima bootstrap uzorka, ali ne istim kao percentilna metoda. Ova metoda koristi kvantile koji ovise o parametrima $\hat{a}$ i $\hat{z}_0$. Te parametre u engleskom jeziku nazivamo *acceleration* i *bias-correction*.

$\mathrm{BC}_a$ pouzdani interval površine $1 - 2\alpha$ dan je sljedećim izrazom:

$$\mathrm{BC}_a : \left(\widehat{\theta}_{\mathrm{lo}}, \widehat{\theta}_{\mathrm{up}}\right) = \left(\widehat{\theta}^{*(\alpha_1)}, \widehat{\theta}^{*(\alpha_2)}\right), \tag{4.1}$$

gdje su

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right) \tag{4.2}$$

i

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right). \tag{4.3}$$

Funkcija $\Phi(x)$ nam označava funkciju distribucije standardne normalne slučajne varijable, a $z^{(\alpha)}$ je $\alpha$-ti kvantil standardne normalne distribucije.

Uočimo da ako su $\hat{a}$ i $\hat{z}_0$ jednaki 0 tada je

$$\alpha_1 = \Phi(z^{(\alpha)}) = \alpha,\ \Phi(z^{(1-\alpha)}) = 1 - \alpha,$$

pa je se zapravo radi o percentilnoj metodi. Ono što nam zapravo mijenja kvantile zapravo su $\hat{a}$ i $\hat{z}_0$.

Kako računamo $\hat{a}$ i $\hat{z}_0$?
$\hat{z}_0$ je dobiven iz postotka bootstrap replikacija koje su manje od originalne procjene na početnom uzorku. Na taj omjer primjenjujući inverz standardne normalne funkcije distribucije dolazimo do traženog $\hat{z}_0$.
Točnije, izraz je dan sljedećom formulom:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\widehat{\theta}^*(b) < \hat{\theta}\right\}}{B}\right). \tag{4.4}$$

Parametar $\hat{a}$ dolazi od tzv. *jackknife* procjene. Budući da to nije tema ovog rada zainteresirane čitatelje upućujemo da pogledaju u [2], poglavlje 11. U uvodu smo dali kratak opis metode. Sukladno tamošnjim oznakama parametar $\hat{a}$ dan je sa:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\widehat{\theta}_{(jack)} - \widehat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\widehat{\theta}_{(jack)} - \widehat{\theta}_{(i)}\right)^2\right\}^{3/2}}. \tag{4.5}$$

**Algorithm 5** Bootstrap $BC_a$ pouzdani intervali

1) Generiraj slučajni uzorak $(x_1^*, ...x_n^*)$ duljine $n$ iz početnog (s ponavljanjem).
2) Izračunaj statistiku od interesa, npr. aritmetičku sredinu:

$$\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k^*.$$

3) Ponoviti korake 1 i 2 $B$ puta - rezultat je niz statistika:

$$(\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*) = (\overline{X}_1^*, \ldots, \overline{X}_B^*).$$

4) Izračunaj $\hat{a}$ i $\hat{z_0}$ prema formulama (4.5) i (4.4).
5) Izračunaj tražene $\alpha_1$ i $\alpha_2$ kvantile formulama (4.2) i (4.3).
6) Pouzdani interval dan je formulom (4.1).

Ukratko ćemo dati algoritam za računanje $BC_a$ pouzdanih intervala kako bismo cijelu gornju diskusiju objasnili sistematičnije:

Pogledajmo sada simulaciju na našem uzorku. Ponovimo kako je procjena pouzdanog intervala dana sljedećom formulom:

$$\left(\widehat{\theta}^{*(\alpha_1)}, \widehat{\theta}^{*(\alpha_2)}\right) = \left(\widehat{\theta}^{*\left(\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right)\right)}, \widehat{\theta}^{*\left(\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right)\right)}\right).$$

Promotrimo sada neparametarsku i parametarsku procjenu $BC_a$ bootstrap pouzdanog intervala za populacijsko očekivanje na temelju našeg uzorka.

Sam algoritam simulacije ponekad je vrlo teško i skupo provesti budući da statistika od interesa može biti kompleksna, a broj simulacija koji je potreban za ovu metodu poprilično je velik.

U tom slučaju $BC_a$ rubove intervala moguće je analitički aproksimirati. Tu metodu nazivamo ABC što je skraćenica od *approximate bias confidence*.

Analitička aproksimacija rubova intervala je suviše kompleksna da bismo ju obradili u ovom radu stoga zainteresiranog čitatelja upućujemo na [2], poglavlje 14.3.

Spomenimo samo kako nju ne možemo koristiti za svaku statistiku, već samo za određeni tip dok je brzina algoritma zadovoljavajuća u većini slučajeva.

## 4.2 Simulacija

### Neparametarska procjena

Primjenjujući gornji algoritam na uzorak iz poglavlja 1 neparametarski bootstrap nam daje sljedeći 95% pouzdan interval za populacijsko očekivanje:

$$[-3.098, -0.112].$$

Rezultat je potkrijepljen kodom u [6.7].
Usporedimo li ovo s rezultatima neparametarske percentilne procjene vidimo kako nam i dalje 0 ne ulazi u interval, no ova procjena daje uži interval te uočavamo da je desni rub u ovom slučaju bliži stvarnoj vrijednosti nego u percentilnom slučaju.

### Parametarska procjena

Budući da R nema gotovu funkciju za računanje parametarskog $BC_a$ pouzdanog intervala istu ćemo napisati ručno prema algoritmu 5. Kao i uvijek do sada, tu procjenu radimo na našem početnom uzorku iz $N(0, 2\sqrt{2})$.

Na kraju, dolazimo do sljedeće procjene 95% pouzdanog intervala:

$$[-3.312578, -3.312558]\,.$$

6ezultat je potkrijepljen kodom u [5.8].
Vidimo kako nam je širina intervala $\approx 2 \cdot 10^{-5}$ što je dosta uže u odnosu na neparametarsku procjenu, ali ponovno nam stvarna vrijednost populacijskog očekivanja, odnosno 0, ne upada u dobiveni procijenjeni interval.

Generalno ova metoda smatra se preciznijom, ali neovisno o tome ponovno ne moramo imati najoptimalniji rezultat.

# Poglavlje 5

# Dodatni primjeri simulacija

Pogledajmo sada neke simulacije parametara iz uzoraka koji dolaze iz različitih distribucija kako bismo usporedili rezultate s dosadašnjim.
Podsjetimo se još jednom kako smo sve gornje simulacije proveli na fiksnom uzorku koji dolazi iz simetrične distribucije, točnije $N(0, 2\sqrt{2})$.

Spomenimo još jednom sve dosad dobivene procjene 95% pouzdanih intervala za očekivanje populacije našeg početnog uzorka. Svi rezultati nalaze se u donjoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[-1.773, 2.88]$ | $[-3.41, -0.3]$ |
| Percentilna metoda | $[-1.76, 1.758]$ | $[-3.411, -0.292]$ |
| $BC_a$ metoda | $[-3.313, -3.3126]$ | $[-3.098, -0.112]$ |

Tablica 5.1: Tablica svih pouzdanih intervala za očekivanje populacije

Uočimo kako u gornjoj tablici u preko 50% slučajeva nam se dogodilo kako stvarna vrijednost populacijskog očekivanja nije upala u traženi pouzdani interval. Razlog tome vjerojatno se krije u samom uzorku i činjenici kako je velika većina vrijednosti lijevo od 0.

Pogledajmo sada još neke primjere simulacija za parametar očekivanja primjenjujući sve gornje metode. Cijeli R-kod nam se nalazi u [6.10].

**Primjer 1:** Prvo ćemo pogledati simulaciju pouzdanih intervala za parametar očekivanja iz eksponencijalne razdiobe s parametrom 1. Podsjetimo se,

$$X \sim Exp(\lambda) \Rightarrow \mathbb{E}[X] = \frac{1}{\lambda}.$$

Naš uzorak je duljine 10 i glasi:

$$\mathbf{x} = (0.14, 1.859, 2.818, 1.746, 1.063, 0.422, 0.536, 0.337, 0.053, 0.491).$$

Računamo 95% pouzdane intervale. Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[-0.1011, 1.44599]$ | $[-0.0955, 1.47558]$ |
| Percentilna metoda | $[0.484, 1.722]$ | $[0.4563, 1.512]$ |
| $BC_a$ metoda | $[-0.142, 0.509]$ | $[0.506, 1.626]$ |

Tablica 5.2: Tablica svih pouzdanih intervala za očekivanje populacije

Ponovno vidimo kako stvarno očekivanje, tj. 1, u velikom broju slučajeva upada u traženi interval čime smo zadovoljni. Ponovno nam $BC_a$ daje najuže intervale, dok bootstrap-t i percentilna metoda daju najšire intervale.

**Primjer 2:** Pogledajmo sada simulaciju za parametar očekivanja ponovno iz eksponencijalne razdiobe, samo s parametrom 10. Ponovno radimo procjene 95% pouzdanih intervala. Cijeli R-kod ostaje isti kao u [6.10] osim generiranja uzorka što je trivijalno stoga ga ovdje nećemo navoditi. Naš uzorak ponovno je duljine 10 i ovog puta glasi:

$$\mathbf{x} = (0.17, 0.16, 0.24, 0.017, 0.02, 0.19, 0.013, 0.04, 0.2978, 0.293).$$

Pogledajmo rezultate u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0.04399, 0.105]$ | $[-0.008, 0.092]$ |
| Percentilna metoda | $[0.0480, 0.1692]$ | $[0.0352, 0.0923]$ |
| $BC_a$ metoda | $[0.0190, 0.0244]$ | $[0.0378, 0.0963]$ |

Tablica 5.3: Tablica svih pouzdanih intervala za očekivanje populacije

Vidimo kako stvarno očekivanje, tj $1/10$ ne upada u traženi interval kod $BC_a$ metode iako imamo najuži pouzdani interval. Kod ostalih metoda rezultat je ovisio o vrsti simulacije. Opet vidimo kako bootstrap-t i percentilna metoda daju poprilično široke intervale smo mogli i očekivati.

**Primjer 3:** Pogledajmo sada još jednu simulaciju za parametar očekivanja. Ovoga puta promatramo uzorak iz gamma distribucije s parametrima 1 i 2. Podsjetimo se kako vrijedi

$$X \sim \Gamma(a, b) \Rightarrow \mathbb{E}[X] = ab.$$

Stoga naše populacijsko očekivanje iznosi 2. Kao i do sada računamo 95% pouzdane intervale. Ponovno, R-kod ostaje isti kao u [6.10] osim simulacijskog dijela. Naš uzorak i dalje je duljine 10 i glasi:

$$\mathbf{x} = (0.18, 1.715, 0.05, 0.02, 0.347, 0.086, 0.071, 0.713, 0.359, 0.268).$$

Pogledajmo rezultate simulacije u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0.2309, 0.5608]$ | $[0.2505, 0.4173]$ |
| Percentilna metoda | $[0.2366, 0.8489]$ | $[0.1146, 0.4599]$ |
| $BC_a$ metoda | $[0.0609, 0.0068]$ | $[0.1345, 0.5249]$ |

Tablica 5.4: Tablica svih pouzdanih intervala za očekivanje populacije

Ovdje već možemo uočiti kako u nijednom slučaju nam populacijski parametar ne upada u traženi interval. Razlog tome vrlo vjerojatno je asimetrična distribucija. Stoga zaključujemo kako u ovom slučaju nemamo optimalnu procjenu pouzdanog intervala za parametar očekivanja.

**Primjer 4:** Pogledajmo sada simulacije za medijan. Svi pouzdani intervali koje konstruiramo bit će 95% pouzdani.
Pogledajmo prvo eksponencijalnu distribuciju s parametrom 1. Podsjetimo se, ako je $\mathbf{x} = (x_1, ...x_n)$ sortirani početni uzorak medijan je statistika definirana sa

$$m = x_{(\frac{n+1}{2})}.$$

Ako naš uzorak dolazi iz eksponencijalne razdiobe s parametrom 1, vrijednost medijana dobivamo koristeći sljedeću formulu pri čemu je $f(x)$ funkcija gustoće naše razdiobe:

$$0.5 = \int_0^m f(x)dx.$$

To nas dovodi do sljedeće vrijednosti populacijskog medijana:

$$\begin{aligned} 0.5 &= \int_0^m e^{-x}\,dx \\ 0.5 &= 1 - e^{-m} \\ 0.5 &= e^{-m} \quad \Big/ \ln \\ \ln 0.5 &= -m \\ -m &= -0.6931 \Rightarrow m = 0.6931. \end{aligned}$$

Pogledajmo procjene pouzdanih intervala na jednom takvom uzorku. Cijeli R-kod nalazi se u [5.9]. Ponovno je naš uzorak duljine 10 i glasi:

$$\mathbf{x} = (0.1407, 1.859, 2.8183, 1.746, 1.063, 0.4228, 0.536, 0.337, 0.053, 0.491).$$

Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[-0.5533, 1.0174]$ | $[-0.423, 1.043]$ |
| Percentilna metoda | $[-2.0855, 2.0830]$ | $[0.3158, 1.7460]$ |
| $BC_a$ metoda | $[-1.0313527, -0.3461549]$ | $[0.2388, 1.7460]$ |

Tablica 5.5: Tablica svih pouzdanih intervala za medijan populacije

Vidimo da nam je u većini slučajeva medijan populacije upao u procijenjeni interval što i je cilj procjene pouzdanih intervala. Ponovno, bootstrap-t metoda dala je najdulje intervale dok je $BC_a$ dala nešto uže intervale pouzdanosti kako smo i očekivali.

**Primjer 5:** Pogledajmo sada rezultat na uzorku iz eksponencijalne razdiobe s parametrom 10. R-kod nam ostaje isti kao u [6.9] osim simulacije uzorka u prvom redu što je trivijalno pa nećemo navoditi. Ovoga puta medijan populacije iznosi 6.93. Uzorak je duljine 10 i glasi:

$$\mathbf{x} = (0.182, 0.04, 0.006, 0.0073, 0.0478, 0.0369, 0.1382, 0.025, 0.1106, 0.219).$$

Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[-0.232, 0.129]$ | $[-0.168, 0.157]$ |
| Percentilna metoda | $[0.2724, 0.174]$ | $[0.0356, 0.1523]$ |
| $BC_a$ metoda | $[-0.0252, 0.1415]$ | $[0.041, 0.057]$ |

Tablica 5.6: Tablica svih pouzdanih intervala za medijan populacije

Vidimo kako nam u ovom slučaju nijednom populacijski medijan nije upao u 95% pouzdani interval. Stoga zaključujemo kako niti u ovom slučaju nismo dobili optimalnu procjenu pouzdanog intervala. Ponovno $BC_a$ metoda je dala najuže intervale.

**Primjer 6:** Pogledajmo sada rezultate za medijan na uzorku iz normalne distribucije s parametrima $\mu = 0$ i $\sigma = 2\sqrt{2}$. R-kod nam ostaje isti kao u [6.9] osim simulacijskog dijela pa ga nećemo posebno navoditi. Podsjetimo se, kod normalne distribucije vrijednost medijana je zbog simetričnosti jednaka vrijednosti očekivanja. Stoga u našem slučaju medijan populacije iznosi 0. Simulaciju ovoga puta radimo na uzorku iz poglavlja 1 stoga ga nećemo posebno navoditi. Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[-1.184, 1.887]$ | $[-0.698, 2.75]$ |
| Percentilna metoda | $[-2.0726, 1.1116]$ | $[-1.0791, 1.8275]$ |
| $BC_a$ metoda | $[-0.735, -0.353]$ | $[-1.8145, 1.4810]$ |

Tablica 5.7: Tablica svih pouzdanih intervala za medijan populacije

Ponovno vidimo kako nam u velikoj većini slučajeva medijan populacije upada u 95% pouzdani interval što je zbog simetričnosti distribucije bilo očekivano. Ponovno, $BC_a$ metoda je dala najuže intervale.

**Primjer 7:** Sada pogledajmo simulaciju medijana za gamma distribuciju s parametrima 1 i 2. Budući da ne postoji zatvorena forma kod računanja medijana gamma distribucije istu ćemo aproksimirati koristeći naredbu *qgamma* u R-u. Ovime dolazimo do aproksimacije medijana gornje distribucije $\approx 0.346$. R-kod ostaje isti kao u [6.9] osim simulacijskog dijela stoga ga nećemo posebno navoditi. Uzorak je duljine 10 i glasi:

$$\mathbf{x} = (0.816, 0.16, 0.038, 0.711, 0.12, 0.021, 0.645, 0.593, 0.358, 1.0354).$$

Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0.113, 0.676]$ | $[0.209, 0.792]$ |
| Percentilna metoda | $[0.137, 0.7403]$ | $[0.2271, 0.903]$ |
| $BC_a$ metoda | $[0.364, 0.4739]$ | $[0.1968, 0.794]$ |

Tablica 5.8: Tablica svih pouzdanih intervala za medijan populacije

Vidimo kako nam populacijski medijan upada u pouzdani interval u svim slučajevima osim parametarske $BC_a$ metode. Ovoga puta ne vidimo preveliku razliku obzirom na širinu intervala kod procjena kao što je to dosad bio slučaj.

**Primjer 8:** Pogledajmo sada simulaciju za varijancu populacije. Svi pouzdani intervali koje računamo bit će 95% pouzdani.

Pogledajmo prvo rezultat za uzorak iz normalne distribucije. Radimo procjenu na uzorku s početka rada, iz poglavlja 1. Varijanca populacije iznosi $(2\sqrt{2})^2 = 8$. R-kod se nalazi u [6.11]. Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | [7.0239, 11.3159] | [4.2978, 8.5178] |
| Percentilna metoda | [2.469, 17] | [1.73, 10.979] |
| $BC_a$ metoda | [2.480, 3.376] | [3.457, 14.378] |

Tablica 5.9: Tablica svih pouzdanih intervala za varijancu populacije

Vidimo kako nam se u svim procjenama osim $BC_a$ varijanca populacije nalazi u procijenjenom intervalu. Jedino gdje to nije slučaj je parametarska $BC_a$ metoda čija procjena je ujedno i najuža.

**Primjer 9:** Pogledajmo sada procijenjene pouzdane intervale za varijancu populacije na uzorku iz eksponencijalne distribucije s parametrom 1. Podsjetimo se,

$$X \sim Exp(\lambda) \Rightarrow Var[X] = \frac{1}{\lambda^2}.$$

Cijeli R kod ostaje isti kao u [6.11]. Simulacije smo radili na sljedećem uzorku duljine 10:

$$\mathbf{x} = (2.81, 1.42, 0.395, 1.807, 0.0592, 0.217, 0.202, 0.005, 1.282, 2.2783).$$

Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0, 0.6574]$ | $[0, 0.755]$ |
| Percentilna metoda | $[0.133, 3.3866]$ | $[0.0871, 0.6586]$ |
| $BC_a$ metoda | $[0.0365, 0.006]$ | $[0.1706, 0.9134]$ |

Tablica 5.10: Tablica svih pouzdanih intervala za varijancu populacije

Vidimo da nam se varijanca populacije nalazi u samo jednom pouzdanom intervalu, a to je percentilna parametarska procjena. Ne vidimo veliku razliku u odnosu na duljine intervala stoga zaključujemo da u ovom slučajnu nismo dobili najoptimalniju procjenu pouzdanog intervala.

**Primjer 10:** Pogledajmo sada procjene varijance populacije na uzorku iz eksponencijalne distribucije s parametrom 10. R kod ostaje isti kao u [6.11]. Uzorak je duljine 10 i glasi:

$$\mathbf{x} = (0.07, 0.263, 0.0143, 0.026, 0.191, 0.102, 0.0753, 0.054, 0.047, 0.007).$$

Rezultati su dani u sljedećoj tablici:

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0, 0.049]$ | $[0, 0.0732]$ |
| Percentilna metoda | $[0.0013, 0.0330]$ | $[0.0039, 0.017]$ |
| $BC_a$ metoda | $[0.0021, 0.0037]$ | $[0.0058, 0.0133]$ |

Tablica 5.11: Tablica svih pouzdanih intervala za varijancu populacije

Vidimo da u velikoj većini slučajeva varijanca populacije ulazi u procijenjeni pouzdani interval. Ponovno nam $BC_a$ metoda daje najuže intervale.

**Primjer 11:** Pogledajmo sada pouzdane intervale za varijancu populacije na uzorku iz gamma distribucije s parametrima 1 i 2. Podsjetimo se,

$$X \sim \Gamma(1, 2) \Rightarrow VarX = ab^2.$$

. R kod je isti kao u [6.11]. Uzorak je duljine 10 i glasi:

$$\mathbf{x} = (0.16, 0.014, 0.308, 0.387, 0.3683, 0.94883, 0.0083, 0.353, 0.951, 0.0143).$$

Rezultati su dani u sljedećoj tablici: Vidimo kako nam varijanca populacije, 0.25 upada

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | $[0, 0.5702]$ | $[0, 0.888]$ |
| Percentilna metoda | $[0.0346, 0.848]$ | $[0.0543, 1.1194]$ |
| $BC_a$ metoda | $[0.34, 0.547]$ | $[0.0875, 1.421]$ |

Tablica 5.12: Tablica svih pouzdanih intervala za varijancu populacije

u pouzdani interval osim u slučaju parametarske $BC_a$ metode. Stoga zaključujemo kako smo u ovom slučaju dobili optimalnu procjenu 95% pouzdanog intervala.

**Primjer 12:** Budući da su svi uzorci do sada bili duljine 10 sada ćemo pogledati jednu simulaciju na uzorku duljine 30. Ponovit ćemo simulaciju za medijan eksponencijalne distribucije s parametrom 10 te ćemo vidjeti hoćemo li dobiti bolje rezultate. Podsjetimo se, populacijski medijan iznosi 6.93. R kod ostaje isti kao i u [6.9]. Naš uzorak glasi:

$$\mathbf{x} = (0.073, 0.1587, 0.093, 0.068, 0.162, 0.3363, 0.141, 0.134, 0.285, 0.178, 0.249, 0.0376, 0.184, 0.136, 0.104, 0.12, 0.078, 0.05, 0.0384, 0.011, 0.0261, 0.065, 0.2382, 0.0017, 0.2198, 0.0832, 0.01368, 0.0685, 0.119, 0.124).$$

Ponovno radimo 95% pouzdane intervale. Rezultati su dani u sljedećoj tablici: Vidimo

| Metoda | Parametarski | Neparametarski |
|---|---|---|
| Bootstrap-t | [0.099, 0.1828] | [0.026, 0.246] |
| Percentilna metoda | [0.0349, 0.115] | [0.0707, 0.1385] |
| $BC_a$ metoda | [0.1114, 0.1285] | [0.0685, 0.1385] |

Tablica 5.13: Tablica svih pouzdanih intervala za medijan populacije

kako nam ni u ovom slučaju medijan populacije ni približno ne upada u dobivene procjene pouzdanih intervala. Ovim primjerom smo pokazali da i veći uzorak ne mora dati bolje rezultate za pouzdane intervale. Već to uvelike ovisi o samoj statistici i distribuciji.

Na temelju svih dosadašnjih simulacija i primjera zaključujemo kako naš pouzdani interval uvelike ovisi o uzorku, distribuciji iz koje dolazi uzorak te samoj statistici. Vidjeli smo da u velikoj većini slučajeva $\mathrm{BC}_a$ pouzdani intervali daju najuže intervalne procjene, no vrlo često populacijski parametar ne upada u procijenjeni pouzdani interval. Također, vidjeli smo kako su nam procjene puno optimalnije kod uzoraka iz simetrične distribucije. Nadalje, procjene pouzdanih intervala za očekivanje i varijancu dale su bolje rezultate nego procjene za medijan pogotovo kod asimetrične distribucije. Također, u zadnjem primjeru vidjeli smo kako ni duljina uzorka ne mora biti ključ bolje procjene. Stoga vidimo kako s ovakvim pristupom procjene pouzdanih intervala moramo biti oprezni. Ako imamo statistiku poput medijana dobra praksa je napraviti što više procjena pa vidjeti koliko se rubovi razlikuju. Ako je ta razlika velika sama procjena vjerojatno nije optimalna.

# Poglavlje 6

# Dodatak R-kod

Listing 6.1: Histogram procjenjitelja očekivanja neparametarskim bootstrapom

```
uzorak <- c( -3.5427851, -3.9725894,  3.0908902,
-0.6072817, -1.2040964, -4.9828332,
-1.6554067, -3.0415893,
1.4828888, -4.3215476)
library('boot')
procjenitelji2 <- numeric(5000)
for (i in 1:5000)
{
  uz2 <- sample(uzorak, size = 10 , replace = T,
  prob = NULL)
  procjenitelj2 <- mean(uz2)
  procjenitelji2[i] <- procjenitelj2
}
hist(procjenitelji2,xlab='Procjenitelji',ylab='Frekvencije',
main='Histogram procjenitelja ocekivanja',c='yellow')
```

Listing 6.2: Histogram procjenjitelja očekivanja parametarskim bootstrapom

```
#parametarski Boostrap. Uzorkujemo 5000 puta
#i svaki put racunamo MLE procjenitelj
procjenitelj <- mean(uzorak)
procjenitelji1 <- numeric(5000)
for (i in 1:5000)
{
  uz1 <- rnorm(10,0,sd= 2 * sqrt(2))
```

```
  procjenitelj1 <- mean(uz1)
  procjenitelji1[i] <- procjenitelj1
}
procjenitelji1
mean(procjenitelji1)
hist(procjenitelji1,xlab='Procjenitelji',ylab='Frekvencije',
main='Histogram  procjenitelja ocekivanja',c='yellow')
```

Listing 6.3: Histogram T- statistika i pouzdani intervali neparametarske bootsrtrap t metode

```
#prvo generiramo uzorak 10 000 puta
#neparametarska procjena
sredina <- mean(uzorak)
vektor_t_statistika_nep <- numeric(10000)
for (i in 1 : 10000)
{
  generirani <- sample(uzorak, size = 10 , replace = T,
  prob = NULL)
  vektor_t_statistika_nep[i] <- (sredina - mean(generirani))
  / sd(generirani) * sqrt(10)
}
vektor_t_statistika_nep <- sort(vektor_t_statistika_nep,
decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_nep, probs = c(0.025, 0.975))
donji <- -1.987612
gornji <- 3.050887

#granice pouzdanog intervala
neparametarski_t_donja_granica <-
sredina - gornji *sd(uzorak) / sqrt(10)

neparametarski_t_gornja_granica <-
sredina - donji *sd(uzorak)/ sqrt(10)
neparametarski_t_donja_granica
neparametarski_t_gornja_granica

hist(vektor_t_statistika_nep,probability = T,c='yellow',
xlab = 'T statistike',
     ylab='relativne frekvencije',
```

```
      main = 'T-statistike-nep')
```

Listing 6.4: Histogram T- statistika i pouzdani intervali parametarske bootsrtrap t metode

```
#bootstrap t pouzdani interval parametarski
vektor_t_statistika_par <- numeric(10000)
for (i in 1:10000)
{
  generirani <- rnorm(10,0,sd= 2 * sqrt(2))
  vektor_t_statistika_par[i] <-
  (sredina - mean(generirani))  / sd(generirani)
  * sqrt(10)

}

vektor_t_statistika_par <- sort(vektor_t_statistika_par,
decreasing = F)
#sada iz gornjeg uzorka trazimo kvantile
quantile(vektor_t_statistika_par, probs = c(0.025, 0.975))
donji_par <- -5.2388811
gornji_par <- -0.1232905

#sada tra imo granice intervala
parametarski_t_donja_granica <-
sredina - gornji_par*sd(uzorak)/sqrt(10)

parametarski_t_gornja_granica <-
sredina - donji_par*sd(uzorak)/sqrt(10)
parametarski_t_donja_granica
parametarski_t_gornja_granica

hist(vektor_t_statistika_par,probability = T,c='yellow',
xlab = 'T statistike',
      ylab='relativne frekvencije',
      main = 'T-statistike-par')
```

Listing 6.5: Histogram očekivanja i pouzdani intervali percentilne metode - parametarski

```
#prvo gledamo statistiku koju racunamo
S <- function(d,i)
{
  d1 <- d[i]
  return(mean(d1))

}

#prvo radimo parametarski Boostrap
#ova funkcija nam svaki put generira uzorak za pramaterasku
bootstrap procjenu

funkcija <- function(data,mle)
{
  generirani <- rnorm(10,0,sd= 2 * sqrt(2))
  return(generirani)
}


boot_uzorak <- boot(uzorak, statistic = S, sim='parametric',
R=10000, ran.gen = funkcija, mle = mean(uzorak))
boot_uzorak
boot_uzorak$t#dane statistike
dim(boot_uzorak$t)

#bootstrap percentilni pouzdani intervali - parametarski


boot_percentilni <- boot(uzorak, statistic = S,
sim='parametric', R=10000, ran.gen = funkcija,
mle = mean(uzorak))
boot_percentilni



interval_pouzdanosti_perc_param <- boot.ci(boot_percentilni,
conf = 0.95, type = 'perc')
interval_pouzdanosti_perc_param #(-1.757,1.758)
```

```
h <- hist(boot_percentilni$t, probability = T, c='yellow',
xlab = 'Aritmeti ke sredine', ylab='relativne frekvencije',
main = '', xlim=c(-4,4))
donji <- quantile(boot_percentilni$t,0.025)
donji
gornji <- quantile(boot_percentilni$t,0.975)
gornji
abline(v=donji, col='red',lwd=2,lty=2)
abline(v=gornji, col='red',lwd=2,lty=2)
```

Listing 6.6: Histogram očekivanja i pouzdani intervali percentilne metode - neparametarski

```
#neparametarska procjena
boot_percentilni_nep <- boot(uzorak,S,R=10000)
boot_percentilni_nep

intervali_pouzdanosti_perc_nep <-
boot.ci(boot_percentilni_nep,conf = 0.95,type = 'perc')
intervali_pouzdanosti_perc_nep #(-3.411, -0.292 )

#sada crtamo histogram
h1 <- hist(boot_percentilni_nep$t, probability = T, c='yellow',
xlab = 'Aritmeti ke sredine',
          ylab='relativne frekvencije',main='',xlim=c(-5,2))
donji2 <- quantile(boot_percentilni_nep$t,0.025)
donji2
gornji2 <- quantile(boot_percentilni_nep$t,0.975)
gornji2
abline(v=donji2, col='red',lwd=2,lty=2)
abline(v=gornji2, col='red',lwd=2,lty=2)
```

Listing 6.7: Percentilni pouzdani intervali BCa metode - neparametarski

```
boot_BC_nep <- boot(uzorak,S,R=1000)
boot_BC_nep

intervali_pouzdanosti_BC_nep <- boot.ci(boot_BC_nep,
conf = 0.95,type = 'bca')
intervali_pouzdanosti_BC_nep
```

Listing 6.8: Percentilni pouzdani intervali BCa metode - parametarski

```
boot_BC_par <- boot(uzorak, statistic = S, sim='parametric'
,R=10000 ,ran.gen = funkcija, mle = mean(uzorak))
boot_BC_par

boot_BC_par$t #procjene statistika

#prvo ra unamo z_O
#treba prebojiti u gornjem uzorku vrijednosti koje
su manje od izvorne statistike



broj_manjih <- length(boot_BC_par$t
[boot_BC_par$t < sredina])
broj_manjih #5149

z_0 <- qnorm(p = (broj_manjih / 10000))
z_0


#sada ra unamo a za  to  nam treba jackknife procjena
jknife <- function(x, fxn) {
  theta <- fxn(x)
  n <- length(x)
  partials <- rep(0,n)

  #hackknife procjene parametra
  for (i in 1:n){
    partials[i] <- fxn(x[-i])
  }
```

```
  #sada dolazimo do jackknife procjene na eg parametra
  jack_procjena <- mean(partials)

  #sada ra unamo parametar a
  brojnik <- sum((jack_procjena - partials)^3)
  nazivnik <- 6 * (sum((jack_procjena-partials)^2))^(1.5)

  a <- brojnik / nazivnik


  return (a)
}

#sada dolazimo do a
procjena_a <- jknife(uzorak,S)
procjena_a #0.03484736

z_alpa1 <- qnorm(0.025)
z_alpa2 <- qnorm(0.0975)

#sada tra imo kvantile
alpha1 <- pnorm(z_0 + (z_0 + z_alpa1) /
(1 - procjena_a * (z_0 + z_alpa1)))
alpha1

alpha2 <- pnorm(z_0 + (z_0 + z_alpa2) /
(1 - procjena_a * (z_0 + z_alpa2)))
alpha2

#sada dolazimo do pouzdanog intervala
boot_BC_par$t <- sort(boot_BC_par$t, decreasing = F)
quantile(boot_BC_par$t, probs = c(alpha1, alpha2))
donja_granica_BC_par <- -3.523984
gornja_granica_BC_par <- -3.523984
```

Listing 6.9: Bootstrap pouzdani intervali medijan - sve simulacije

```
uz <- c(0.1407, 1.859, 2.8183, 1.746, 1.063, 0.4228,
0.536, 0.337, 0.053, 0.491)
or_med <- median(uz)
or_med

library('boot')
#radimo statistiku od interesa
S <- function(d,i)
{
  d1 <- d[i]
  return(median(d1))

}

funkcija <- function(data,mle)
{
  generirani <- rexp(10,1)
  return(generirani)
}

#parametarski_percentilna metoda
boot_uz <- boot(uz, statistic = S, sim='parametric',
R=10000,ran.gen = funkcija, mle = median(uz))
boot_uz


#simulacija
intervali_pouzdanosti_uz_par <- boot.ci(boot_uz,
conf = 0.95,type = 'perc')
intervali_pouzdanosti_uz_par


#neparametarski_percentilna metoda
interval_percentilni_nep <- boot(uz,S,R=10000)
interval_percentilni_nep

intervali_pouzdanosti_perc_nep_uz <-
```

```
boot.ci(interval_percentilni_nep, conf = 0.95,
type = 'perc')
intervali_pouzdanosti_perc_nep_uz


#bootstrap-t-neparametarski
or_med <- median(uz)
sredina_exp <- mean(uz)
vektor_t_statistika_nep_exp <- numeric(10000)
for (i in 1 : 10000)
{
  generirani <- sample(uz, size = 10 , replace = T,
  prob = NULL)
  vektor_t_statistika_nep_exp[i] <-
  (mean(generirani)-sredina_exp)/sd(generirani)*sqrt(10)
}
vektor_t_statistika_nep_exp <-
sort(vektor_t_statistika_nep_exp, decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_nep_exp, probs=c(0.025,0.975))
donji_exp_t_nep <- -1.838430
gornji_exp_t_nep <- 3.248278

#granice pouzdanog intervala
neparametarski_t_donja_granica_exp <-
or_med-gornji_exp_t_nep * sd(uz)/sqrt(10)

neparametarski_t_gornja_granica_exp <-
or_med - donji_exp_t_nep * sd(uz)/sqrt(10)
neparametarski_t_donja_granica_exp
neparametarski_t_gornja_granica_exp


#parametarska procjena
vektor_t_statistika_par_exp <- numeric(10000)
for (i in 1:10000)
{
  generirani_uz <- rexp(10,1)
  vektor_t_statistika_par_exp[i] <- (sredina_exp -
```

```
  mean(generirani_uz)) / sd(generirani_uz) * sqrt(10)

}

vektor_t_statistika_par_exp <-
sort(vektor_t_statistika_par_exp, decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_par_exp,
probs = c(0.025,0.975))
donji_par_exp <- -1.750411
gornji_par_exp <- 3.705965

#sada tra imo granice intervala
parametarski_t_donja_granica_exp <-
or_med - gornji_par_exp* sd(uz) /sqrt(10)

parametarski_t_gornja_granica_exp <-
or_med - donji_par_exp* sd(uz) /sqrt(10)
parametarski_t_donja_granica_exp
parametarski_t_gornja_granica_exp

#BCa procjena
#neparamtarski
boot_BC_nep_exp <- boot(uz,S,R=10000)
boot_BC_nep_exp

intervali_pouzdanosti_BC_nep_exp <- boot.ci(boot_BC_nep_exp,
conf = 0.95,type = 'bca')
intervali_pouzdanosti_BC_nep_exp

#parametarski BCa
boot_BC_par_exp <- boot(uz, statistic = S, sim='parametric',
R=10000, ran.gen = funkcija, mle = median(uz))
boot_BC_par_exp

boot_BC_par_exp$t
#prvo ra unamo z_O
#treba prebojiti u gornjem uzorku vrijednosti koje su manje
#od izvorne statistike
```

```
broj_manjih_exp <- length(boot_BC_par_exp$t
[boot_BC_par_exp$t < or_med])
broj_manjih_exp

z_0_exp <- qnorm(p = (broj_manjih_exp / 10000))
z_0_exp

#sada ra unamo a za  to  nam treba jackknife procjena
jknife_exp <- function(x, fxn) {
  theta <- fxn(x)
  n <- length(x)
  partials <- rep(0,n)

  #hackknife procjene parametra
  for (i in 1:n){
    partials[i] <- fxn(x[-i])
  }

  #sada dolazimo do jackknife procjene na eg parametra
  jack_procjena <- mean(partials)

  #sada ra unamo parametar a
  brojnik <- sum((jack_procjena - partials)^3)
  nazivnik <- 6*(sum((jack_procjena-partials)^2))^(1.5)

  a <- brojnik / nazivnik


  return (a)
}

#sada dolazimo do a
procjena_a_exp <- jknife_exp(uz,S)
procjena_a_exp #-3.94249e-16

z_alpa1_exp <- qnorm(0.025)
```

```
z_alpa2_exp <- qnorm(0.0975)

alpha1_exp <- pnorm(z_0_exp + (z_0_exp + z_alpa1_exp)/
(1 - procjena_a_exp * (z_0_exp + z_alpa1_exp)))
alpha1_exp

alpha2_exp <- pnorm(z_0_exp + (z_0_exp + z_alpa2_exp)/
(1 - procjena_a_exp * (z_0_exp + z_alpa2_exp)))
alpha2_exp

boot_BC_par_exp$t <- sort(boot_BC_par_exp$t,
decreasing = F)
quantile(boot_BC_par_exp$t, probs=c(alpha1_exp, alpha2_exp))
```

Listing 6.10: Bootstrap pouzdani intervali očekivanje - sve simulacije

```
#simulirajmo sada uzorak duljine 10 iz exponencijalne
#razdiobe s parametrom 1
#uz <- rexp(10, rate = 1)
#print(uz)
uz <- c(0.1407, 1.859, 2.8183, 1.746, 1.063,
0.4228, 0.536, 0.337, 0.053, 0.491)
or_oc <- mean(uz)
or_oc

library('boot')
#radimo statistiku od interesa
S <- function(d,i)
{
  d1 <- d[i]
  return(mean(d1))

}

funkcija <- function(data,mle)
{
  generirani <- rexp(10,1)
  return(generirani)
}
```

```
#parametarski_percentilna metoda
boot_uz <- boot(uz, statistic = S, sim='parametric',
R=10000,ran.gen = funkcija, mle = mean(uz))
boot_uz


#simulacija
intervali_pouzdanosti_uz_par <- boot.ci(boot_uz,
conf = 0.95,type = 'perc')
intervali_pouzdanosti_uz_par


#neparametarski_percentilna metoda
interval_percentilni_nep <- boot(uz,S,R=10000)
interval_percentilni_nep

intervali_pouzdanosti_perc_nep_uz <- boot.ci(
interval_percentilni_nep,conf = 0.95,type = 'perc')
intervali_pouzdanosti_perc_nep_uz


#bootstrap-t-neparametarski
or_oc <- mean(uz)
sredina_exp <- mean(uz)
vektor_t_statistika_nep_exp <- numeric(10000)
for (i in 1 : 10000)
{
  generirani <- sample(uz, size = 10 , replace = T,
  prob=NULL)
  vektor_t_statistika_nep_exp[i] <- (mean(generirani)-
  sredina_exp) / sd(generirani) * sqrt(10)
}
vektor_t_statistika_nep_exp <-
sort(vektor_t_statistika_nep_exp, decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_nep_exp,probs=c(0.025,0.975))
donji_exp_t_nep <- -1.837276
gornji_exp_t_nep <- 3.620536
```

```
#granice pouzdanog intervala
neparametarski_t_donja_granica_exp
<- or_oc-gornji_exp_t_nep*sd(uz)/sqrt(10)

neparametarski_t_gornja_granica_exp
<- or_oc-donji_exp_t_nep* sd(uz)/sqrt(10)
neparametarski_t_donja_granica_exp
neparametarski_t_gornja_granica_exp


#parametarska procjena
vektor_t_statistika_par_exp <- numeric(10000)
for (i in 1:10000)
{
  generirani_uz <- rexp(10,1)
  vektor_t_statistika_par_exp[i] <- (mean(generirani_uz)-
  sredina_exp) / sd(generirani_uz) * sqrt(10)

}

vektor_t_statistika_par_exp
<- sort(vektor_t_statistika_par_exp, decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_par_exp,probs=c(0.025,0.975))
donji_par_exp <- -1.734495
gornji_par_exp <- 3.639854

#sada tra imo granice intervala
parametarski_t_donja_granica_exp <-
or_oc - gornji_par_exp* sd(uz) /sqrt(10)

parametarski_t_gornja_granica_exp <-
or_oc - donji_par_exp* sd(uz)/sqrt(10)
parametarski_t_donja_granica_exp
parametarski_t_gornja_granica_exp

#BCa procjena
#neparamtarski
boot_BC_nep_exp <- boot(uz,S,R=10000)
```

```
boot_BC_nep_exp

intervali_pouzdanosti_BC_nep_exp <-
boot.ci(boot_BC_nep_exp,conf = 0.95, type = 'bca')
intervali_pouzdanosti_BC_nep_exp

#parametarski BCa
boot_BC_par_exp <- boot(uz, statistic = S, sim='parametric'
,R=10000, ran.gen = funkcija, mle = mean(uz))
boot_BC_par_exp

boot_BC_par_exp$t
#prvo ra unamo z_O
#treba prebojiti u gornjem uzorku vrijednosti koje su
#manje od izvorne statistike



broj_manjih_exp <- length(
boot_BC_par_exp$t[boot_BC_par_exp$t < or_oc])
broj_manjih_exp

z_0_exp <- qnorm(p = (broj_manjih_exp / 10000))
z_0_exp #0.489624

#sada ra unamo a za  to  nam treba jackknife procjena
jknife_exp <- function(x, fxn) {
  theta <- fxn(x)
  n <- length(x)
  partials <- rep(0,n)

  #hackknife procjene parametra
  for (i in 1:n){
    partials[i] <- fxn(x[-i])
  }

  #sada dolazimo do jackknife procjene na eg parametra
  jack_procjena <- mean(partials)
```

```
  #sada ra unamo parametar a
  brojnik <- sum((jack_procjena - partials)^3)
  nazivnik <- 6*(sum((jack_procjena-partials)^2))^(1.5)

  a <- brojnik / nazivnik


  return (a)
}

#sada dolazimo do a
procjena_a_exp <- jknife_exp(uz,S)
procjena_a_exp

z_alpa1_exp <- qnorm(0.025)
z_alpa2_exp <- qnorm(0.0975)

alpha1_exp <- pnorm(z_0_exp+(z_0_exp+z_alpa1_exp)/
(1 - procjena_a_exp * (z_0_exp + z_alpa1_exp)))
alpha1_exp #0.1633664

alpha2_exp <- pnorm(z_0_exp + (z_0_exp + z_alpa2_exp)/
(1 - procjena_a_exp * (z_0_exp + z_alpa2_exp)))
alpha2_exp # 0.3757429

boot_BC_par_exp$t <- sort(boot_BC_par_exp$t,
decreasing = F)
quantile(boot_BC_par_exp$t,
probs=c(alpha1_exp, alpha2_exp))
```

Listing 6.11: Bootstrap pouzdani intervali varijancu - sve simulacije- normalna razdioba

```
#simulacija pramaetarkog bootstrapa
#bootstrap t intervali
#budu i da R nema gotovu funkciju ta studentizirani
#pouzdani interval
#radimo ru no
library('boot')

#prvo generiramo uzorak 10 000 puta
```

```
#neparametarska procjena
sredina <- mean(uzorak)
v <- var(uzorak)
vektor_t_statistika_nep <- numeric(10000)
for (i in 1 : 10000)
{
  generirani <- sample(uzorak, size = 10 , replace = T,
  prob = NULL)
  vektor_t_statistika_nep[i] <-
  (mean(generirani)-sredina) / sd(generirani) * sqrt(10)
}
vektor_t_statistika_nep <- sort(vektor_t_statistika_nep,
decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_nep,
probs = c(0.025, 0.975))
donji <- -1.928579
gornji <- 3.146389

#granice pouzdanog intervala
neparametarski_t_donja_granica <- v-gornji * sd(uzorak)/
sqrt(10)

neparametarski_t_gornja_granica <- v-donji * sd(uzorak)/
sqrt(10)
neparametarski_t_donja_granica
neparametarski_t_gornja_granica


#bootstrap t pouzdani interval parametarski
vektor_t_statistika_par <- numeric(10000)
for (i in 1:10000)
{
  generirani <- rnorm(10,0,sd= 2 * sqrt(2))
  vektor_t_statistika_par[i] <-
  ( mean(generirani) - sredina)/sd(generirani) * sqrt(10)

}
```

```
vektor_t_statistika_par <- sort(vektor_t_statistika_par,
decreasing = F)
#sada iz gornjeg uzorka tra imo kvantile
quantile(vektor_t_statistika_par, probs = c(0.025, 0.975))
donji_par <- -5.2937310
gornji_par <- -0.1320278

#sada tra imo granice intervala
parametarski_t_donja_granica <-
v - gornji_par * sd(uzorak)/sqrt(10)

parametarski_t_gornja_granica <-
v - donji_par * sd(uzorak)/sqrt(10)
parametarski_t_donja_granica
parametarski_t_gornja_granica


#prvo gledamo statistiku koju racunamo
S <- function(d,i)
{
  d1 <- d[i]
  return(var(d1))

}

#prvo radimo parametarski Boostrap
#ova funkcija nam svaki put generira uzorak za
#pramaterasku bootstrap procjenu

funkcija <- function(data,mle)
{
  generirani <- rnorm(10,0,sd= 2 * sqrt(2))
  return(generirani)
}


boot_uzorak <- boot(uzorak, statistic = S,
sim='parametric', R=10000 ,
ran.gen = funkcija, mle = var(uzorak))
```

```
boot_uzorak #bootsrap parametarska procjena meana iz dane
#razdiobe
boot_uzorak$t#dane statistike
dim(boot_uzorak$t)

#bootstrap percentilni pouzdani intervali - parametarski
boot_percentilni <- boot(uzorak, statistic = S,
sim='parametric', R=10000, ran.gen = funkcija,
mle = var(uzorak))
boot_percentilni

interval_pouzdanosti_perc_param <- boot.ci(boot_percentilni
,conf = 0.95,type = 'perc')
interval_pouzdanosti_perc_param


#neparametarska procjena
boot_percentilni_nep <- boot(uzorak,S,R=10000)
boot_percentilni_nep

intervali_pouzdanosti_perc_nep <-
boot.ci(boot_percentilni_nep, conf = 0.95,type = 'perc')
intervali_pouzdanosti_perc_nep


#BC_a pouzdani intervali
#neparametarska procjena
boot_BC_nep <- boot(uzorak,S,R=1000)
boot_BC_nep

intervali_pouzdanosti_BC_nep <- boot.ci(boot_BC_nep,
conf = 0.95,type = 'bca')
intervali_pouzdanosti_BC_nep


#BC_a parametarska procjena

#generiranje uzorka
boot_BC_par <- boot(uzorak, statistic = S, sim='parametric'
```

```
,R=10000, ran.gen = funkcija, mle = mean(uzorak))
boot_BC_par

boot_BC_par$t #procjene statistika

#prvo ra unamo z_O
#treba prebojiti u gornjem uzorku vrijednosti koje su
#manje od izvorne statistike

v

broj_manjih <- length(boot_BC_par$t[boot_BC_par$t < v])
broj_manjih #5149

z_0 <- qnorm(p = (broj_manjih / 10000))
z_0 #0.03735745


#sada ra unamo a za  to  nam treba jackknife procjena
jknife <- function(x, fxn) {
  theta <- fxn(x)
  n <- length(x)
  partials <- rep(0,n)

  #hackknife procjene parametra
  for (i in 1:n){
    partials[i] <- fxn(x[-i])
  }

  #sada dolazimo do jackknife procjene na eg parametra
  jack_procjena <- mean(partials)

  #sada ra unamo parametar a
  brojnik <- sum((jack_procjena - partials)^3)
  nazivnik <- 6 * (sum((jack_procjena - partials)^2))^(1.5)

  a <- brojnik / nazivnik
```

```
  return (a)
}

#sada dolazimo do a
procjena_a <- jknife(uzorak,S)
procjena_a #0.03484736

z_alpa1 <- qnorm(0.025)
z_alpa2 <- qnorm(0.0975)

#sada tra imo kvantile
alpha1 <- pnorm(z_0 + (z_0 + z_alpa1) /
(1 - procjena_a * (z_0 + z_alpa1)))
alpha1

alpha2 <- pnorm(z_0 + (z_0 + z_alpa2) /
(1 - procjena_a * (z_0 + z_alpa2)))
alpha2

#sada dolazimo do pouzdanog intervala
boot_BC_par$t <- sort(boot_BC_par$t, decreasing = F)
quantile(boot_BC_par$t, probs = c(alpha1, alpha2))
```

# Bibliografija


[1] C. Davison, D. V. Hinkley, *Bootstrap methods and their application*, Cambridge Series in Statistical and Probabilistic Mathematics, 1997.

[2] B. Efron, R. J. Tibshirani, *An introduction to the Bootstrap*, Springer-Science-Business Media, B.V., 1993.

[3] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, 2002.

[4] Rudi Mrazović, *Mjera i integral-skripta*, 2020.

# Sažetak

U ovom diplomskom radu ukratko smo opisali bootstrap metodu. Napravili smo jednostavan primjer parametarske i neparametarske simulacije te smo promotrili histograme aritmetičkih sredina. Objasnili smo načine na koje se pomoću nje može doći do procjene pouzdanih intervala.

Opisali smo bootstrap t-metodu, percentilnu te $BC_a$ metodu za procjenu pouzdanog intervala. Sve metode potkrijepili smo simulacijama za parametar očekivanja na slučajnom uzorku iz normalne distribucije.

Vidjeli smo kako ovisno o metodi naši intervali mogu biti širi i uži te neovisno o parametarskoj i neparametarskoj procjeni stvarna vrijednost parametra uvijek ne mora upadati u dobiveni procijenjeni interval. $BC_a$ metoda je većinom davala najuži interval dok je u većini slučajeva studentizirana ili bootstrap-t metoda davala najširi interval.

Pogledali smo simulacije za očekivanje, medijan i varijancu iz raznih razdioba. U ovom radu simulaciju smo proveli za normalnu, eksponencijalnu i gamma razdiobu. Zaključili smo kako točnost naše procjene uvelike ovisi o samoj statistici, uzorku i distribuciji. Najbolji rezultati su za simetrične distribucije.

# Summary

In this master thesis we presented bootstrap method. We made simple example of parametric and nonparametric simulation based on which we observed histogram of means. We explained methods where using bootstrap can be useful method to estimate confidence intervals for statistic of interest.

We described bootstrap-t method, percentile and $BC_a$ method for estimation of confidence interval. All methods were supported by simulations for mean confidence intervals on sample from normal distribution.

We saw that depending on method our intervals can be more or less wide and independent from parametric or nonparametric simulation real value of parameter does not always have to be contained in estimated confidence interval. $BC_a$ method mostly gave us the narrowest interval. On the other hand, student or bootstrap-t method in most of the cases gave us the widest interval for parameter of interest.

In the last chapter we saw simulations for mean, median and variance on samples that came from diverse distributions. In this thesis we did simulation for normal, exponential and gamma distribution. We concluded that accuracy of our estimation mostly depends on statistic, sample and distribution. The best results were obtained for symmetrical distributions.

# Životopis

Rođena sam 15.09.1997. u Zadru. Odrasla sam u Obrovcu gdje sam 2012. godine završila osnovnu školu. Iste godine u Zadru upisujem Gimnaziju Jurja Barakovića, smjer prirodoslovno-matematička gimnazija. Tijekom cijelog školovanja pohađala sam dodatnu nastavu iz matematike te sam išla na natjecanja. Najveće postignuće bilo je 2015. godine kada sam u A kategoriji osvojila 3. mjesto na županijskom natjecanju.
2016. godine upisujem preddiplomski sveučilišni studij matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Studij sam završila 2020. nakon čega upisujem diplomski sveučilišni studij Matematička statistika na istom fakultetu.
Također, učlanjujem se u najveću studentsku udrugu eSTUDENT kao članica tima LUMEN Data science te radim na organizaciji najvećeg regionalnog natjecanja u dubinskoj analizi velike količine podataka. Godinu nakon nastavljam svoj put u eSTUDENTu kao voditeljica tima Informacijske tehnologije.
2021. godine zapošljavam se kao Junior R&D Engineer u Robotiq.ai.