Matematički fakultet

Univerzitet u Beogradu

Master rad

# Butstrap metoda i njena primena

*Student:*
**Milica Bošković**

*Mentor:*
**Slobodanka Janković**

Oktobar, 2015

# Sadržaj

# 1 Rezime

Tema ovog rada jeste Butstrap metoda i njena primena. Rad je podeljen u 8 poglavlja.
U poglavlju **2** je opisana kratka istorija i značaj Butstrap metode, kao i osnovni pojmovi uzorkovanja i reuzorkovanja.
U poglavlju **3** je dat princip butstrap metodologije. U poglavlju **4** je opisana primena butstapa na ocenjivanje sredine, varijanse i na smanjenje pristrasnosti ocenjivača, dok u **5** i **6** je opisano korišćenenje butstrapa za ocenjivanje intervala poverenja i testiranje statističkih hipoteza.
U Dodatku, na kraju rada, se nalaze kodovi koji su korišćeni za primere i koji su napisani u programskom softveru R.
Pored imena teorema i definicija se nalaze dva broja. Prvi broj je redni broj poglavlja, a drugi broj je redni broj teoreme ili definicije u tom poglavlju.

# 2 Uvod

Butstrap je jedna od metoda reuzorkovanja podataka iz originalnog skupa podataka. Princip butstrap metodologije uveo je američki statističar Bredli Efron[1] 1979. godine, u radu "*Bootstrap methods: Another look at the Jackknife*[1] " koji je objavljen u časopisu "*The Annals of Statistics*". Izraz "**to pull yourself up by your own bootstraps**[2]" poslužio mu je kao inspiracija da odredi ime za novu statističku tehniku. Izraz potiče od nemačkog pisca Raspea[3], koji je na taj način opisao barona Minhauzena, junaka njegovog dela, koji je jednom prilikom, sam sebe izvukao iz močvare zatezanjem kaiševa na sopstvenim čizmama. To bi značilo da je butstrap široko primenljiv, koristan alat, tj. "kaiš" koji nam omogućava da se izvučemo iz statističke močvare, odnosno problema.

Butstrap je jednostavan, ali veoma moćan statistički metod koji je Efron prvi put predstavio na Stenford Univerzitetu 1977. godine. Međutim, i nakon objavljivanja rada u časopisu "*The Annals of Statistics*", 1979. godine, naučnici ga nisu nisu često koristili jer u to vreme računari nisu bili dovoljno razvijeni tako da nije bilo moguće brzo izvođenje zahtevnih i velikih

[1] B.Efron
[2] Prevod engleske reči *bootstraps* na srpskom znači *čizme*
[3] R.E.Raspe

računa. Članak [2] objavljen u "*Scientific American*", 1983. godine bio je pokušaj da se butstrap popularizuje naučničkom društvu pri čemu je dato laičko objašnjenje metode i njene široke primene. Nažalost, pokušaj da se sve objasni što jednostavnije doveo je do gubljenja tehničkih detalja što je rezultovalo povećanjem skepticizma kod naučnika. Efron je i kasnije pokušavao da ispravi utisak, ali je "*Scientific American*" imao daleko veći uticaj na naučnike i istraživače koji su odbili da koriste ovu tehniku jer u članku nije pisalo jasno objašnjenje zašto je moguće zameniti potrebne prave podatke simuliranim podacima. Ključna asimptotska svojstva koja se pojavljuju u butstrapu bilo je teško dokazati, a matematički dokazi su postali poznati i dostupni tek nakon par godina zahvaljujući istraživanjima drugih naučnika. Dakle, u početku butstrap je izgledao isuviše jednostavno i nije shvaćen kao deo revolucije u statističkom razmišljanju i analizi padataka, međutim, danas je prihvaćen kao regularna metoda reuzorkovanja podataka.

## 2.1 Uzorkovanje

**Statistika** (grčki: *statos* - uređen, fiksiran) je matematička disciplina koja se bavi prikupljanjem, prikazivanjem, analizom podataka i zaključivanjem na osnovu podataka. Danas je prisutna i u svakodnevnom životu, a izvesnim aspektima života bavila se i u svojim počecima. Naime statistika je prvobitno proučavala tzv.masovne pojave u ljudskom društvu kroz prikupljanje, upoređivanje i tumačenje podataka o stanovništvu, imovini, vojnoj snazi itd. U principu se izučava neki skup objekata koji se naziva **populacija** (osnovni skup ili generalna kolekcija) u odnosu na izvesnu varijabilnu kvantitativnu ili kvalitativnu osobinu koja se naziva **obeležje**. Obeležje koje se posmatra može biti jednodimenzionalno, dvodimenzionalno ili višedimenzionalno.

Ako se na slučajan način izabere jedan element populacije, ne zna se unapred vrednost obeležja koju taj element ima. To znači da se vrednost obeležja na slučajno izabranom elementu populacije može shvatiti kao vrednost slučajne veličine. Raspodela verovatnoća te slučajne veličine se zove **raspodela obeležja**. Podaci se mogu prikupljati iz populacije i tako izučavati populacija u celini. Međutim, ako populacija sadrži veliki broj elemenata, tada izučavanje cele populacije može trajati dugo ili prouzrokovati veće materijalne troškove. U većini slučajeva je nemoguće ispitivati celu već samo deo populacije. Taj deo zovemo **uzorak**, a broj elemenata u uzorku je **obim**

**uzorka**.

U zavisnosti od toga šta se ispituje pravi se plan za prikupljanje podataka. Podatke možemo dobiti merenjem i brojanjem. Kriterijum za utvrđivanje da li izabrani element ima određeno svojstvo, način prikupljanja podataka, metoda za utvrđivanje postojanja netačnih podataka i cilj istraživanja su važni za dobijanje merodavnih podataka. Dobijeni uzorak i raspodela verovatnoće obeležja na uzorku se razlikuju od populacije koja se proučava i raspodele verovatnoće obeležja na populaciji. Tako da je od najvećeg interesa analizirati podatke i doneti zaključak o populaciji iz koje su podaci uzeti.

Kriterijumi na osnovu kojih određujemo koji metod uzorkovanja ćemo primeniti su: vrsta, veličina i struktura populacije. Uzorak najčešće dobijamo na jedan od sledećih načina:

(1) **izbor sa vraćanjem** - izabrani element se posle beleženja osobina vraća u populaciju, da bi se zatim iz celokupne populacije na slučajan način uzimao sledeći element,

(2) **izbor bez vraćanja** - izbrani element se posle beleženja osobina ne vraća u populaciju, a sledeći element uzorka se bira među preostalim elementima populacije.

Da bismo odredili raspodelu obeležja na populaciji na osnovu uzorka veoma je važno da je taj uzorak reprezentativan. Metode izbora uzorka treba da isključuju sistematske graške. Dopustive su samo slučajne greške, čiji se uticaj može proceniti na osnovu teorije verovatnoće. Teorijski, reprezentativnost uzorka se obezbeđuje na sledeći način:

**Definicija 2.1 (Prost slučajan uzorak)**
*Neka se u populaciji posmatra obeležje $X$. Prost slučajan uzorak obima $n$ za posmatrano obeležje je $n$-dimenziona slučajna veličina $(X_1, X_2, ..., X_n)$ pri čemu su slučajne veličine $X_1, X_2, ..., X_n$ nezavisne i sve imaju istu raspodelu kao posmatrano obeležje $X$.*

**Realizovani uzorak** predstavlja konkretan niz vrednosti obeležja dobijenih na elementima populacije koji su izabrani u uzorak. Dakle, ukoliko prost slučajan uzorak označimo sa

$$\mathbf{X} = (X_1, X_2, ..., X_n),$$

onda je realizovani uzorak

$$\mathbf{x} = (x_1, x_2, ..., x_n).$$

Pomoću **tablica (pseudo)slučajnih brojeva** možemo dobiti uzorak obima n. Ukoliko populacija ima $N$ elemenata, numerisanih brojevima 1,2,...,$N$ i ako broj $N$ ima $k$ cifara tada iz tablice (pseudo)slučajnih brojeva redom čitamo grupe po $k$ cifara i dobijamo $k$-tocifrene brojeve. Ukoliko je taj $k$-tocifreni broj manji ili jednak od $N$ onda iz populacije biramo element sa tim rednim brojem i stavljamo ga u uzorak, a ukoliko je veći od $N$ onda se ne uzima u obzir, a iz tablice se čita sledeći $k$-tocifreni broj. Ukoliko je u pitanju izbor sa vraćanjem onda nakon beleženja osobina izvučenog elementa taj element se opet vraća u osnovni skup, a ako je u pitanju izbor bez vraćanja onda osim $k$-tocifrenih brojeva većih od $N$, ne uzimaju se u obzir ni oni brojevi koji su se već javili u postupku izbora.

Ako je broj elemenata u populaciji veoma veliki ili beskonačan, onda umesto tablice slučajnih brojeva se bira neki drugi postupak. U zavisnosti od cilja proučavanja i veličine populacije bira se neka od metoda za dobijanja uzorka kao što su: stratifikovani uzorak, grupni uzorak, dvoetapni uzorak, periodični uzorak itd.

**Statistika** je slučajna promenljiva (veličina)

$$Y = f(\mathbf{X}) = f(X_1, X_2, ..., X_n)$$

pri čemu je $f: \mathbf{R^n} \to \mathbf{R^s}$ Borelova funkcija i nema nepoznatih parametara [4].

Za primenjivanje bilo kog statističkog postupka analize podataka (analize varijanse, linearne regresije, linearne korelacije itd.) potrebno je prethodno proveriti da li su ispunjene polazne pretpostavke, a jedna od ključnih pretpostavki je raspodela verovatnoće obeležja na populaciji. U datim okolnostima često se ne možemo uzdati u ispunjenost tih pretpostavki i tada nastaje problem jer procena parametra tada nije adekvatna.

Sada ćemo ilustrovati problem na primeru aritmetičke sredine.
Neka je $\mathbf{X} = (X_1, X_2, ..., X_n)$ prost slučajan uzorak na osnovu kojeg treba da

[4]Radi jednostavnosti, pretpostavićemo da su sve slučajne promenljive u ovom radu jednodimenzione

ocenimo sredinu populacije iz koje je izvučen. Aritmetička sredina

$$\overline{X} \;=\; \frac{X_1 + ... + X_n}{n} \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

je statistika koja se najčešće koristi za ocenjivanje sredine populacije. Ukoliko su slučajne veličine $X_1, X_2, ..., X_n$ normalno raspodeljene,tj.

$$X_i : \; N(\mu, \sigma^2) \;\; i = 1, ..., n,$$

na osnovu teorije verovatnoće aritmetička sredina $\overline{X}$ takođe će imati normalnu raspodelu

$$\overline{X} : \; N(\mu, \frac{\sigma^2}{n}).$$

Problem nastaje kada nemamo nikakve pretpostavke o slučajnom uzorku $X_i, \;\; i = 1, ..., n$.
U slučaju kada je obim uzorka $n$ veliki na osnovu asimptotske teorije možemo doći do zaključka o raspodeli aritmetičke sredine koja će i u ovom slučaju imati normalnu raspodelu sa matematičkim očekivanjem $\mu$ i disperzijom $\frac{\sigma^2}{n}$. Do ovog zaključka dolazimo na osnovu centralne granične teoreme kao jedne od najznačajnijih teorema asimptotske teorije.

**Teorema 2.1 (Centralna granična teorema)**
*Neka su $X_1, X_2, ...$ nezavisne slučajne promenljive sa istom raspodelom, konačnim disperzijama $\sigma^2$ i matematičkim očekivanjima $\mu$. Tada slučajna promenljiva*

$$Z_n = \frac{X_1 + ... + X_n \;-\; n\mu}{\sigma\sqrt{n}}$$

*konvergira u raspodeli ka $Z \;\sim\; N(0, 1), \; tj.$*

$$\lim_{n\to\infty} P(Z_n \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

*za svako $x \epsilon \mathbf{R}$.*

Ukoliko slučajnu promenljivu $Z_n$ iz prethodne teoreme pomnožimo sa $\sigma/\sqrt{n}$ dobijamo

$$\frac{X_1 + ... + X_n}{n} \;=\; \mu + \frac{\sigma}{\sqrt{n}} Z_n,$$

što znači da aritmetička sredina $\frac{X_1+...+X_n}{n}$ ima približno $N(\mu, \frac{\sigma^2}{n})$ raspodelu.

Dakle, problem smo rešili kada je obim uzorka veliki, ali kako doći do rešenja ako je obim mali?
U praktičnim uslovima često se srećemo sa uzorcima malog obima pri čemu zaključke ne možemo doneti na osnovu asimptotske teorije. Metode reuzorkovanja nude se kao rešenje za prevazilaženje ovog problema.

## 2.2 Reuzorkovanje

Primenjivanje metoda parametarske statistike podrazumeva da su ispunjeni određeni uslovi. Te pretpostavke najčešće se odnose na raspodelu promenljive, ali da bi adekvatno primenili kompleksniju analizu treba da bude ispunjen veći broj uslova. Ukoliko sumnjamo da neka od pretpostavki za primenjivanje željenog postupka analize nije ispunjena, umesto uzorkovnja adekvatnije je primeniti reuzorkovanje (resampling).

**Resampling** je engleska reč koja kada se prevede na srpski jezik znači reuzorkovanje. Reuzoračke metode imaju isti princip koji se zasniva na recikliranju informacija iz samo jednog uzorka populacije. Realizovani uzorak $\mathbf{x} = (x_1, x_2, ..., x_n)$ dobijen iz osnovne populacije, posmatramo kao novu populaciju iz koje pravimo veliki broj novih uzoraka (reuzoraka):

$$\begin{aligned}
\mathbf{x}_1^* &= (x_{11}^*, x_{21}^*, ..., x_{k1}^*),\\
\mathbf{x}_2^* &= (x_{12}^*, x_{22}^*, ..., x_{k2}^*),\\
&\vdots\\
\mathbf{x}_B^* &= (x_{1B}^*, x_{2B}^*, ..., x_{kB}^*),
\end{aligned}$$

pri čemu zvezdica $(*)$ označava da je u pitanju element dobijen reuzorkovanjem.
Kako polazni uzorak $\mathbf{x}$ reprezentuje osnovnu populaciju, dobijeni reuzoraci koje ćemo označiti sa $\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_B^*)$ treba da simuliraju višestruko uzorkovanje iz osnovnog skupa.
Na osnovu svih $B$ reuzoraka potrebno je oceniti parametar, tj. naći uzoračku raspodelu verovatnoće statistike $Y = f(x)$ kojom ocenjujemo parametar.

Ideja je da uzoračku raspodelu statistike $Y$ aproksimiramo reuzoračkom raspodelom računajući

$$y_i^* = f(x_i^*) \;\; i = 1, ..., B.$$

Dakle, na osnovu reuzoraka računamo ocenu parametra, pri čemu nije neophodno praviti bilo kakve pretpostavke o raspodeli statistike na osnovnom uzorku. Suštinska i jedina pretpostavka koju pravimo jeste da polazni uzorak $\mathbf{x}$ u razumnoj meri predstavlja populaciju iz koje je uzet, jer taj uzorak prilikom korišćenja metode reuzorkovanja, predstavlja novu populaciju i ukoliko u njemu postoji greška onda se ta greška u daljem radu multiplikuje i može dovesti do pogrešnih rezultata.
Vidimo da reuzorci ne moraju biti istog obima kao i polazni uzorak, tj. $n$ ne mora biti jednako $k$. Obim reuzorka zavisi od metode koju primenjujemo. Postoje četiri osnovna tipa reuzoračkih metoda: cross-validation, permutacioni testovi, jackknife i butstrap. U daljem radu mi ćemo se baviti butstrap metodom.
Za izradu poglavlja **2** korišćena je literatura [1], [2], [5], [6], [8] i [10].

# 3 Butstrap

## 3.1 Uvod

Butstrap metoda je neparametarska metoda reuzorkovanja koja omogućava brzu i jednostavnu procenu bez pretpostavki na tip raspodele (ukoliko je raspodela nepoznata ili kompleksna) i ne oslanja se na asimptotske rezultate.

Sada ćemo objasniti način funkcionisanja butstrap metode.
Neka je je $\mathbf{X} = (X_1, X_2, ..., X_n)$ prost slučajan uzorak za obeležje populacije $X$ sa funkcijom raspodele $F$ i neka je $\mathbf{x} = (x_1, x_2, ..., x_n)$ realizovani uzorak generisan funkcijom $F$. $Z = f(\mathbf{X})$ je neka statistika uzorka $\mathbf{X}$. Naš zadatak je da odredimo funkciju raspodele $H$ statitike $Z$ na osnovu uzorka $\mathbf{X}$, tj. da nađemo uzoračku raspodelu.

U zavisnosti od toga da li je oblik funkcije $F$ poznat razlikujemo dva slučaja:

(1) **parametarski butstrap** - oblik funkcije raspodele $F$ je poznat, ali tu figuriše nepoznati parametar $\theta = \theta(F)$ (npr. znamo da $F$ ima normalnu raspodelu, ali ne znamo parametre raspodele),

(2) **neparametarski butstrap** - nemamo nikakve pretpostavke o funkciji raspodele $F$.

U oba slučaja od raspoloživih podataka $\mathbf{x} = (x_1, x_2, ..., x_n)$ pravimo B butstrap uzoraka

$$\mathbf{x}_1^* = (x_{11}^*, x_{21}^*, ..., x_{n1}^*),$$
$$\mathbf{x}_2^* = (x_{12}^*, x_{22}^*, ..., x_{n2}^*),$$
$$\vdots$$
$$\mathbf{x}_B^* = (x_{1B}^*, x_{2B}^*, ..., x_{nB}^*),$$

koji su istog obima kao i polazni uzorak $\mathbf{x}$.
Zatim na osnovu dobijenih butstrap uzoraka $\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_B^*)$ računamo vrednosti statistike $Z$,tj.

$$z_i = f(\mathbf{x}_i^*) \;\; i = 1, ..., n, \quad \mathbf{z} = (z_1, z_2, ..., z_B)$$

Sada, umesto da računamo raspodelu $H$ statistike $Z$ računaćemo njenu empirijsku funkciju raspodele.
Ovom prilikom ćemo se prisetiti definicije empirijske funkcije raspodele.

**Definicija 3.1 (Empirijska funkcija raspodele)**
*Neka je $(X_1, X_2, ..., X_n)$ prost slučajan uzorak obima n za posmatrano obeležje. Funkcija*

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I\{X_k \leq x\}$$

*je empirijska funkcija raspodele.*

Pri tome je $I\{X_k \leq x\}$ indikator događaja koji broji elemente uzorka koji imaju manju ili jednaku vrednost od $x$. Neka je $n_x$ broj elemenata uzorka za koje je vrednost obeležja $X$ manja ili jednaka od realnog broja $x$. Tada se realizovana vrednost empirijske funkcije raspodele u tački $x$ dobija po formuli: $F_n(x) = \frac{n_x}{n}$.

Empirijska funkcija raspodele je jednaka relativnoj učestalosti događaja $\{X_k \leq k\}$. To je stepenasta funkcija koja uzima vrednosti iz segmenta $[0, 1]$, neopadajuća je za svako $x$ i neprekidna sa desne strane.

**Empirijsku funkciju raspodele** statistike $Z$ izračunata za kolekciju $\mathbf{z} = (z_1, z_2, ..., z_B)$ definišemo na sledeći način:

$$H(z|\mathbf{z}) = \frac{1}{B}\sum_{i=1}^{B} I(Z_i \leq z), \quad z\epsilon\mathbf{R}$$

S obzirom da funkciju raspodele menjamo empirijskom funkcijom raspodele navešćemo i teoreme koje opravdavaju taj postupak.

**Teorema 3.1 (Slabi zakon velikih brojeva)**
*Neka su $X_1, X_2, ...$ nezavisne slučajne promenljive sa $E(X_k) = \mu$ i sa konačnim varijansama $Var(X_k) \leq V$ za svako $k = 1, 2, ...$ gde je $V$ pozitivna konstanta. Tada niz aritmetičkih sredina $(X_1 + ... + X_n)/n$ konvergira u verovatnoći ka $\mu$, tj.*

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| \geq \varepsilon\right\} = 0, \text{ za svako } \varepsilon > 0.$$

**Dokaz.** Neka je $Y_n = \frac{X_1+...+X_n}{n}$. Tada je $EY_n = \mu$, dok za varijansu važi relacija

$$VarY_n = \frac{1}{n^2}(VarX_1 + VarX_2 + ... + VarX_n) \leq \frac{nV}{n^2} = \frac{V}{n}.$$

Primenom Čebišovljeve nejednakosti dobijamo

$$P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| \geq \varepsilon\right\} = P\{|Y_n - EY_n| \geq \varepsilon\} \leq \frac{VarY_n}{\varepsilon^2} \leq \frac{V}{n\varepsilon^2} \to 0 \; kad \; n \to \infty.$$

U posebnom slučaju, kada su $X_n$ Bernulijeve slučajne promenljive sa verovatnoćom uspeha $p$, dobija se tzv. Bernulijev zakon velikih brojeva:

$$\lim_{n\to\infty} P\left\{\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right\} = 0,$$

gde je sa $S_n = X_1 + ... + X_n$ označen broj uspeha u $n$ eksperimenata.

**Teorema 3.2 (Pomoćna teorema)** *Neka je $\{X_n\}$ niz slučajnih promenljivih takvih da za svaki prirodan broj $m$ važi*

$$\sum_{n=1}^{\infty} P\left\{|X_n| \geq \frac{1}{m}\right\} < \infty.$$

*Tada je*

$$P\left\{\lim_{n\to\infty} X_n = 0\right\} = 1.$$

**Teorema 3.3 (Borelov strogi zakon velikih brojeva)** *Neka je $S_n$ broj uspeha u $n$ Bernulijevih eksperimenata, sa verovatnoćom uspeha $p$. Tada je*

$$P\left\{\lim_{n\to\infty} \frac{S_n}{n} = p\right\} = 1.$$

**Dokaz.** Neka je $Y_n = S_n/n$. Imamo da je $EY_n = p$, $VarY_n = p(1-p)/n$, pa primenom Čebišovljeve nejednakosti dobijamo, za svako $m \in \mathbf{N}$,

$$\sum_{k=1}^{\infty} P\left\{|Y_{k^2} - p| \geq \frac{1}{m}\right\} \leq m^2 p(1-p) \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty.$$

Na osnovu Pomoćne teoreme sa $X_n = Y_{n^2} - p$, imamo da je

$$P\{\lim_{k\to\infty} Y_{k^2} = p\} = 1.$$

Primetimo da Pomoćnu teoremu ne možemo primeniti direktno na niz $Y_n - p$ jer red $\sum 1/n$ divergira. Ali, za svako $n \in \mathbf{N}$ postoji $k = k(n)$ takvo da je $k^2 \leq n < (k+1)^2$; za takvo $n$ i $k$ imamo da je

$$|Y_n - Y_{k^2}| = \left|\frac{S_n}{n} - \frac{S_{k^2}}{k^2}\right| = \left|\left(\frac{1}{n} - \frac{1}{k^2}\right) S_{k^2} + \frac{1}{n}(S_n - S_{k^2})\right| \leq \frac{(n-k^2)k^2}{nk^2} + \frac{n-k^2}{n},$$

jer je $S_{k^2} < k^2$, kao broj uspeha u $k^2$ eksperimenata.
Takođe je $S_n - S_{k^2} \leq n - k^2$.
Dalje, iz $k^2 \leq n < (k+1)^2$ nalazimo da je $0 \leq n - k^2 \leq 2k$, pa je

$$\frac{(n-k^2)k^2}{nk^2} + \frac{n-k^2}{n} = \frac{2(n-k^2)}{n} \leq \frac{4k}{n} = \frac{4k^2}{nk} \leq \frac{4}{k},$$

odakle sledi da je $|Y_n - Y_{k^2}| \leq 4/k$. Prema tome,

$$|Y_n - p| \leq |Y_n - Y_{k^2}| + |Y_{k^2} - p| \leq \frac{4}{k} + |Y_{k^2} - p|,$$

pa je događaj

$$\lim_{n\to\infty} Y_n = p$$

ekvivalentan događaju

$$\lim_{k\to\infty} Y_{k^2} = p,$$

a dokazali smo da je verovatnoća ovog događaja jednaka jedinici. Time je dokaz završen.

Empirijska funkcija raspodele nije deterministička funkcija. U svakom eksperimentu se, iz uzorka obima n, dobija drugačija empirijska funkcija raspodele.
Prema tome, empirijska funkcija raspodele postiže, u fiksiranoj tački $x$, vrednosti $k/n$ sa nekom verovatnoćom. Definišimo $Y_i = 1$ ako je $X_i \leq x$ i $Y_i = 0$ ako je $X_i > x$. Tada zbir $S_n = Y_1 + ... + Y_n$ predstavlja broj onih slučajnih promenljivih iz uzorka čije su vrednosti $\leq x$, pa je

$$F_n(x) = \frac{Y_1 + ... + Y_n}{n} = \frac{S_n}{n}, \tag{1}$$

Imamo da je $EY_k = P\{X_k \leq x\} = F(x)$, gde je $F$ funkcija raspodele iz koje je uzet uzorak.
Prema Bernulijevom zakonu velikih brojeva, empirijska funkcija raspodele $F_n$ konvergira u verovatnoći ka funkciji raspodele $F$. Na osnovu Borelovog zakona velikih brojeva, $F_n$, tj. aritmetička sredina (1) konvergira ka $F(x)$ skoro svuda. Tačnije, za svako $x\epsilon\mathbf{R}$ važi da je

$$P\left\{\lim_{n\to\infty} F_n(x) = F(x)\right\} = 1.$$

Ovaj rezultat opravdava aproksimaciju funkcije raspodele njenom empirijskom raspodelom dobijenom iz uzorka. Sledeća teorema, poznata i pod nazivom "Centralna teorema statistike" tvrdi da je ta aproksimacija uniformna po x:

**Teorema 3.4 (Glivenko-Kantelijeva teorema)** *Neka je $F_n$ empirijska funkcija raspodele dobijena iz prostog slučajnog uzorka obima n iz raspodele sa funkcijom raspodele $F$. Tada je, sa verovatnoćom 1,*

$$\lim_{n\to\infty} \sup_{x\in\mathbf{R}} |F_n(x) - F(x)| = 0.$$

Osnovni cilj butstrap metode, kao što je već i pomenuto, jeste procena parametara raspodele obeležja populacije. Zato ćemo nadalje statistiku $Z = f(X)$ čija nas raspodela zanima, posmatrati kao ocenu nepoznatog parametra $\theta = \theta(F)$ raspodele $F$ na osnovu prostog slučajnog uzorka $\mathbf{X}$ i umesto $Z = f(X)$ koristićemo oznaku $\hat{\theta} = \hat{\theta}(\mathbf{X})$.

Od polaznog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ obima $n$, gde su svi elementi međusobno različiti, u slučaju neparametarskog butstrap metoda, moguće je napraviti $n^n$ butstrap uzoraka oblika $\mathbf{x}_j^*$, što je veliki broj.
Zahvaljuju razvoju računara moguće je izračunati sve te uzorke, ali krajnji cilj je uštedeti na vremenu, a to znači da za ispitivanje nećemo računati sve moguće reuzorke već ćemo se ograničiti na određeni broj.
Upravo iz tog razloga u butstrapu pojavljuju se dva izvora greške:

(1) zamena $F$ sa $\hat{F}$,

(2) procena raspodele od $\hat{\theta}$ simulacijama iz $\hat{F}$.

Tačnije, greška može nastati ako polazni uzorak $\mathbf{x} = (x_1, x_2, ..., x_n)$ ne predstavlja baš najbolje populaciju iz koje je izabran, ili, pak, greška može nastati usled ne uzimanja svih butstrap uzoraka, ali ova greška se može smanjiti odabirom većeg broja butstrap uzoraka.

## 3.2 Parametarski butstrap

U slučaju parametarskog butstrapa pretpostavljamo da je poznat oblik funkcije raspodele $F$ prostog slučajnog uzorka $\mathbf{X}$, ali da ta funkcija zavisi od nepoznatog parametra $\theta$ za koji važi $\theta \epsilon \Theta$. Kao što je ranije rečeno, potrebno je naći funkciju raspodele neke ocene $\hat{\theta} = \hat{\theta}(\mathbf{X})$ parametra $\theta$.
Postupak je sledeći:

(1) na osnovu originalnog realizovanog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ računamo vrednost ocene $\hat{\theta}(\mathbf{x})$ parametra $\theta$,

(2) funkciju $F_\theta$ aproksimiramo funkcijom[5] $F_{\hat{\theta}}$,

[5]Oznake $\theta$ i $\hat{\theta}$ u donjem indeksu ukazuju na vrednosti koje figurišu u funkciji $F$

(3) funkcijom $F_{\hat{\theta}}$ generišemo B nezavisnih uzoraka

$$\mathbf{x}_i^* = (x_{1i}^*, x_{2i}^*, ..., x_{ni}^*), \quad i = 1, ..., B,$$

koje nazivamo butstrap uzorcima,

(4) za svaki od butstrap uzoraka računamo vrednost ocene

$$\hat{\theta}_i^* = \hat{\theta}(\mathbf{x}_i^*), \quad i = 1, ..., B.$$

Te vrednosti se zovu butstrap ocene ili butstrap replike posmatrane statistike,

(5) konstruišemo empirijsku funkciju raspodele,tj. butstrap raspodelu ocene $\hat{\theta}$

$$H(\cdot|\hat{\theta}^*), \ \hat{\theta}^* = (\hat{\theta_1^*}, \hat{\theta_2^*}, ..., \hat{\theta_B^*})$$

kao butstrap ocenu nepoznate funkcije $H$.

## 3.3 Neparametarski butstrap

U neparametarskom butstrapu polazimo od pretpostavke da funkcija raspodele $F$ obeležja $X$ nije poznata.
Problem rešavamo na sledeći način:

(1) na osnovu originalnog realizovanog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ računamo empirijsku funkciju raspodele

$$F_n(x|\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \quad x\epsilon\mathbf{R}$$

(2) na osnovu dobijene empirijske funkcije pravimo B novih uzoraka.Praktično to se svodi na dobijanje uzoraka

$$\mathbf{x}_i^* = (x_{1i}^*, x_{2i}^*, ..., x_{ni}^*), \quad i = 1, ..., B,$$

putem izvlačenje sa vraćanjem iz osnovnog skupa podataka $\{x_1, x_2, ..., x_n\}$. Svaki element ima jednaku verovatnoću $\frac{1}{n}$, da bude izvučen i nakon beleženja njegovih svojstava on se opet vraća u populaciju iz koje je izabran. To znači da jedan element može više puta da se pojavi u uzorku jer se verovatnoća njegovog biranja ne menja tokom procesa pravljenja uzorka.

Naredni koraci su isti kao kod parametarskog butstrapa:

(3) za svaki od butstrap uzoraka računamo vrednost ocene

$$\hat{\theta}_i^* = \hat{\theta}(\mathbf{x}_i^*), \;\; i = 1, ..., B.$$

(4) konstruišemo empirijsku funkciju raspodele

$$H(\cdot|\hat{\theta}^*),$$

na osnovu $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*)$, kao butstrap ocenu nepoznate funkcije $H$.

Za izradu poglavlja **3** korišćena je literatura [5], [6], [8] i [9].

# 4 Ocenjivanje sredine, varijanse i redukcija pristrasnosti

## 4.1 Uvod

Butstrap metod funkcioniše na sledeći način:
Neka nam je dato obeležje populacije $X$ sa funkcijom raspodele $F$[6] u kojoj figuriše nepoznati parametar $\theta$. Krajnji cilj je odraditi funkciju raspodele $H$ neke ocene $\hat{\theta}$ parametra $\theta$.
Nepoznatu funkciju raspodele $H$ ocene $\hat{\theta}$ ocenjujemo empirijskom funkcijom raspodele butstrap uzoraka $H(\cdot|\hat{\theta}^*)$, pri čemu je $H(\cdot|\hat{\theta}^*)$ aproksimacija idealne butstrap raspodele $H^*$.
Na osnovu jedne realizacije prostog slučajnog uzorka, $\mathbf{x} = (x_1, x_2, ..., x_n)$, potrebno je funkciju $F$ zameniti funkcijom $F_{\hat{\theta}}$ ili funkcijom $F(x|\mathbf{x})$, a zatim jednom od ove dve funkcije simulirati butstrap uzorke $\mathbf{x}_i^* = (x_{1i}^*, x_{2i}^*, ..., x_{ni}^*)$, $i = 1, ..., B$ i izračunati vrednosti $\hat{\theta}_i^* = \hat{\theta}(\mathbf{x}_i^*), \;\; i = 1, ..., B$.
Dobijamo kolekciju $\hat{\theta}^* = (\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$, čija empirijska funkcija raspodele

$$H(\rho|\hat{\theta}^*) = \frac{1}{B}\sum_{i=1}^{B} I(\{\hat{\theta}_i^* \leq \rho\}), \;\; \rho \epsilon \mathbf{R}$$

za $B \to \infty$ konvergira ka funkciji raspodele $H^*$ idealne butstrap ocene

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

[6] Oblik funkcije $F$ ne mora biti poznat

## 4.2 Ocenjivanje sredine i varijanse

U ovom odeljku ćemo pokazati kako se ocenjuju sredina i varijansa ocenjivača.
Pretpostavimo da važe sve pretpostavke navedene u prethodnom odeljku i da treba da nadjemo sredinu $E(\hat{\theta})$ i varijansu $D(\hat{\theta})$ ocene $\hat{\theta}$ nepoznatog parametra $\theta$, u odnosu na originalnu raspodelu $F$.

Na osnovu statističke teorije poznato je da za ocenu sredine $\mu$, obeležja X, koristimo uzoračku sredinu $\hat{\mu} = \overline{X_n}$.

**Definicija 4.1 (Uzoračka sredina)**
*Neka je $(X_1, X_2, ..., X_n)$ prost slučajan uzorak obima n za posmatrano obeležje $X$. Uzoračka sredina je statistika*

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

*Ako je realizovani uzorak dat sa $(x_1, x_2, ..., x_n)$ onda je realizovana vrednost uzoračke sredine jednaka*

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Prirodno je očekivati da se **butstrap ocena srednje vrednosti** ocene $\hat{\theta}$ parametra $\theta$ dobija računanjem sredine kolekcije butstrap ocena, $\hat{\theta}^* = (\theta_1^*, ..., \theta_B^*)$,

$$\overline{\theta}^* = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i^*. \tag{2}$$

Takođe, na osnovu statističke teorije znamo da za ocenu varijanse $\sigma^2$, obeležja $X$, koristimo uzoračku disperziju, $\overline{S_n^2}$. Ukoliko želimo nepristrasnu ocenu varijanse onda koristimo popravljenu uzoračku disperziju, $\widetilde{S_n^2}$.

**Definicija 4.2 (Uzoračka disperzija)**
*Neka je $(X_1, X_2, ..., X_n)$ prost slučajan uzorak obima n za posmatrano obeležje $X$. Ukoliko matematičko očekivanje obeležja nije poznato i $\overline{X}_n$ je uzoračka sredina, tada je statistika*

$$\overline{S_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

*uzoračka disperzija , a popravljena uzoračka disperzija je*

$$\widetilde{S_n^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

*Ako je realizovani uzorak dat sa* $(x_1, x_2, ..., x_n)$ *onda je realizovana vrednost uzoračke disperzije jednaka*

$$\overline{s_n^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2,$$

*odnosno, realizovana vrednost popravljene uzoračke disperzije je*

$$\widetilde{s_n^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2.$$

Analogno, **butstrap ocena varijanse** ocene $\hat{\theta}$ parametra $\theta$ je

$$\hat{\sigma}^{2*} = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \overline{\theta}^*)^2.$$

## 4.3 Ocenjivanje pristrasnosti

Neka $E(X)$ označava očekivanu vrednost slučajne promenljive $X$ i neka je $\hat{\theta}$ ocena nepoznatog parametra $\theta$. Kažemo da je $\hat{\theta}$ nepristrasna ili centrirana ocena nepoznatog parametra $\theta$ ukoliko važi

$$E(\hat{\theta}) = \theta.$$

Ukoliko ne važi prethodna jednakost onda je $\hat{\theta}$ pristrasna ocena nepoznatog parametra $\theta$ i postoji neko $b(\theta) \neq 0$ tako da važi $E(\hat{\theta}) = \theta + b(\theta)$. Sa $b$ je označena **pristrasnost** ili **bias** koja se računa po formuli:

$$b(\theta) = E(\hat{\theta}) - \theta.$$

**Ocena sa redukovanom pristrašnošću** je

$$\hat{\theta}_{red} = \hat{\theta} - b(\theta). \tag{3}$$

Svrha ocenjivanja pristrasnosti jeste da se poboljša pristrasnost ocene i da

se dobije ako je moguće nepristrasna ocena.
Ocenu sa redukovanom pristrasnošću lako je izračunati ukoliko je funkcionalni oblik pristrasnosti poznat. Međutim, u praksi to često nije slučaj i u takvim situacijama problem se jednostavno rešava primenom butstrap metode.
Računamo butstrap ocenu sa redukovanom pristrasnošću, $\hat{\theta}^*_{red}$, nepoznatog parametra $\theta$ koja se za veliko B zanemarljivao razlikuje od $\hat{\theta}_{red}$.

Neka je $\mathbf{x} = (x_1, x_2, ..., x_n)$ realizovani prost uzork na osnovu kojeg odredjujemo vrednost $\hat{\theta}(\mathbf{x})$ ocene parametra $\theta$ i funkciju $F_{\hat{\theta}(\mathbf{x})}$, a zatim po pravilu $F_{\hat{\theta}(\mathbf{x})}$ pravimo B butstrap uzoraka $\mathbf{x}^*_i$, $i = 1, ..., B$ i računamo butstrap ocene $\hat{\theta}^*_i = \hat{\theta}(\mathbf{x}^*_i)$, $i = 1, ..., B$.
Na osnovu zakona velikih brojeva, aritmetička sredina ovih ocena[7] zadovoljava

$$\overline{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^*_i \ \rightarrow E_{\hat{\theta}}(\hat{\theta}(\mathbf{X}^*)) = \hat{\theta} + b(\hat{\theta}), \quad B \rightarrow \infty. \tag{4}$$

Na osnovu prethodnog izraza dobijamo butstrap ocenu pristrasnosti:

$$b(\hat{\theta}) = \overline{\theta}^* - \hat{\theta} \tag{5}$$

odnosno

$$b(\hat{\theta}) = \sum_{i=1}^{B}(\hat{\theta}^*_i - \hat{\theta})/B. \tag{6}$$

Na osnovu formula (3) i (5) dobijamo butstrap ocenu sa redukovanom pristrasnošću:

$$\hat{\theta}^*_{red} = \hat{\theta} - (\overline{\theta}^* - \hat{\theta}) = 2\hat{\theta} - \overline{\theta}^*$$

**Primer**:

(a) Ako znamo da raspodela obeležja $X$ pripada dopustivoj familiji $N(\mu, \sigma^2)$, $\mu \epsilon \mathbf{R}$, $\sigma > 0$, primenom parametarskog butstrapa naći ocenu sredine i varijanse ocene $\hat{\theta}$ nepoznatog parametra $\theta = \mu$.
Ovo ćemo da ilustrujemo na uzorku iz populacije sa normalnom $N(4, 1)$ raspodelom

[7]Oznaka $\hat{\theta}$ u donjem indeksu znači da matematičko očekivanje računamo u odnosu na butstrap uzorke

(b) Ukoliko o funkciji raspodele $F$ obeležja $X$ nemamo nikakve podatke, odrediti ocenu sredine i varijanse ocene $\hat{\theta}$ nepoznatog parametra $\theta = \mu$ primenom neparametarskog butstrapa.

Oba dela primera, i pod (a) i pod (b), uraditi za različite obime uzorka ($n$=5, 150) i za različite brojeve butstrap replika ($B$=15,150, 1500).

Rezultati:

(a): Potrebne rezultate dobijamo primenom statističkog programa R (algoritam za rešavanje ovog zadatka nalazi se u Dodatku na kraju rada, pod nazivom `parametarskiButstrap`). Zbog prostorne ograničenosti, detaljniji prikaz rezultata biće dat samo u slučaju $n$=5 i $B$=15, a ostale rezultate ćemo prikazati u tabeli.

```
> parametarskiButstrap(5,15,4,1) (n=5,B=15,mi=4,sigma=1)
  polazni uzorak x
   3.400108
   2.631839
   3.046638
   1.439148
   4.414398

   Sredina od x:                 2.986426
   Disperzija od x:              1.183416
   Standardna devijacija od x:   1.087849

Butstrap uzorci koje smo generisali parametarskim butstrapom:
> matrica
     noviuzorak  noviuzorak noviuzorak noviuzorak noviuzorak
[1,] 4.420604   1.637118   3.646361   1.557288   2.444789
[2,] 3.660552   2.957044   3.129816   2.906601   3.918517
[3,] 3.394259   2.108705   3.158341   2.030331   1.328363
[4,] 2.475016   3.372885   2.495452   3.965002   2.437528
[5,] 3.502346   1.911738   2.604270   4.806921   4.697447

     noviuzorak noviuzorak noviuzorak noviuzorak noviuzorak
[1,] 4.232598   2.671716   1.187819   5.067180   1.909780
[2,] 2.684551   2.825651   3.272882   4.676875   3.013504
[3,] 3.358274   2.865083   2.100242   3.398291   2.749024
[4,] 3.128077   2.410790   3.229086   3.799654   2.673970
[5,] 2.593667   4.028596   3.960240   2.939385   4.402566

     noviuzorak noviuzorak  noviuzorak noviuzorak noviuzorak
[1,] 2.2840041   2.903891   1.332319  3.053234  3.3548120
[2,] 3.4431076   2.273510   3.287282  2.095170  3.4751861
[3,] 3.3019343   4.980251   4.449991  4.465226  3.1324722
[4,] 4.7717693   4.895439   2.233496  3.310019  3.6196344
[5,] 0.8731765   2.760660   4.070601  4.680168  0.4216661
```

```
Butstrap replike ocene teta kapa dobijene
parametarskim butstrapom:

     0.000000    2.397498   3.006848  3.053229  2.965329
     3.199433    2.960367   2.750054  3.976277  2.949769
     2.934798    3.562750   3.074738  3.520764  2.800754

Butstrap ocena sredine  teta kapa: 2.876841
Butstrap ocena varijanse teta kapa: 0.7740316
```

Stvarna raspodela $F$ našeg uzorka $\mathbf{x}$ bila je $N(4,1)$, ali mi se "pravimo" kao da poznajemo oblik raspodele ali ne i vrednosti njenih parametara $\theta = \mu = 4$ i $\rho = \sigma^2 = 1$. Nas interesuje da pronađemo sredinu i varijansu ocene $\hat{\theta} = \overline{X}_n$ nepoznatog parametra $\theta = \mu$.
Na osnovu statističke teorije znamo da ako slučajan uzorak $\mathbf{X}$, obima $n$, ima $N(\mu, \sigma^2)$ raspodelu, onda za veliko $n$ i aritmetička sredina uzorka $\overline{X}_n$ ima normalnu raspodelu $N(\mu, \frac{\sigma^2}{n})$, tj.

$$\hat{\theta} \, \epsilon \, N(4, 0.2).$$

Pretpostavili smo da nam ovo teorijsko tvrđenje nije poznato i zato smo generisali B=15 butstrap uzoraka $\mathbf{x}_i^*, \ i = 1, ..., 15$, parametrarskim butstrapom po pravilu

$$F_{(\hat{\theta}(x), \hat{\rho}(x))} = N(\overline{x}_n, \overline{s}_n^2),$$

gde su $\overline{x}_n$ i $\overline{s}_n^2$ sredina i varijansa polaznog uzorka $\mathbf{x}$.
Zatim smo za svaki uzorak izračunali butstrap replike $\hat{\theta}_i^*, \ i = 1, ..., 15$ i pomoću njih našli butstrap ocenu sredine i varijanse ocene $\hat{\theta}$.
Istu funkciju primenićemo i na uzorak obima n=150 čime dobijamo polazni uzorak sa sledećim osobinama:

```
> parametarskiButstrap(150,15,4,1) (n=150,B=15,mi=4,sigma=1)

   Sredina od x:                3.960713
   Disperzija od x:             1.052489
   Standardna devijacija od x:  1.025909
```

U tabeli prikazujemo dobijene butstrap ocene sredine i varijanse od $\hat{\theta}$ za podatke veličina $n$=5,150 i $B$=15,150,1500.

| n↓ B→ | 15 |
|---|---|
| 5 | $\overline{\theta}^* = 2.8768, \hat{\sigma}^{2*} = 0.77403$ |
| 15 | $\overline{\theta}^* = 3.9421, \hat{\sigma}^{2*} = 0.01572$ |
| n↓ B→ | 150 |
| 5 | $\overline{\theta}^* = 2.9932, \hat{\sigma}^{2*} = 0.3728$ |
| 15 | $\overline{\theta}^* = 3.9687, \hat{\sigma}^{2*} = 0.00782$ |
| n↓ B→ | 1500 |
| 5 | $\overline{\theta}^* = 2.9573, \hat{\sigma}^{2*} = 0.27084$ |
| 15 | $\overline{\theta}^* = 3.9602, \hat{\sigma}^{2*} = 0.0075$ |

**Sredina**:
Za originalan uzorak izvučen iz normalne raspodele sa sredinom $\mu = 4$, na osnovu statističke teorije znamo da i matematičko očekivanje njene ocene treba da bude identično. Međutim, iz tabele iznad, vidimo da su butstrap ocene sredine $\hat{\theta}$ približnije odgovarajućim uzoračkim vrednostima, nego teorijskim, tj. $\overline{x}_n = 2.986426$ za $n$=5 i $\overline{x}_n = 3.960713$ za $n$=150. Time smo dobili praktičnu potvrdu tvrdnje da jedan deo greške pri butstrap zaključivanju potiče iz toga što nismo izabrali reprezentativan polazni uzorak. Kako $B$ raste tako je i ocena tačnija i za veće $n$ imamo bolje rezultate.

**Disperzija**:
Za originalan uzorak izvučen iz normalne raspodele sa varijansom $\sigma^2 = 1$, na osnovu statističke teorije znamo da varijansa ocene $\hat{\theta}(\mathbf{X}) = \overline{X}_n$ treba da bude $\sigma^2/5 = 0.2$ za $n$=5, odonosno $\sigma^2/150 = 0.00667$ za $n$=150. Da bismo uporedili rezultate dobijene na osnovu originalnih uzoraka sa butstrap ocenama potrebno je da uskladimo formule koje smo koristili, tj. da od uzoračke disperzije dobijemo popravljenu uzoračku disperziju:

$$n = 5,\ \overline{s}_n^2 = 1.183416,\ \widetilde{s}_n^2 = (5/(5-1))\overline{s}_n^2 = 1.47927,\ \frac{\widetilde{s}_n^2}{5} = 0.295854$$

$$n = 150,\ \overline{s}_n^2 = 1.052489,\ \widetilde{s}_n^2 = (150/(150-1))\overline{s}_n^2 = 1.05955,\ \frac{\widetilde{s}_n^2}{150} = 0.007064$$

Iz tabele vidimo da su varijanse i ovog puta približnije uzoračkim nego teorijskim vrednostima. Kako $B$ raste tako i ocena postaje približnija.

(b): Za deo pod (b) koristićemo algoritam `neparametarskiButstrap` (nalazi se u Dodatku na kraju rada), s tim što smo konkretno za ovaj deo, (b), iskoristili iste polazne uzorke kao iz dela (a), za $n$=5 i za $n$=150. Zatim smo izvlačenjem sa vraćanjem iz tih uzoraka napravili butstrap uzorke veličine $B$=15,150 i 1500. Navešćemo krajnje rezultate bez analize.

| n↓ B→ | 15 |
|---|---|
| 5 | $\overline{\theta}^* = 2.8892, \hat{\sigma}^{2*} = 0.76253$ |
| 15 | $\overline{\theta}^* = 3.9875, \hat{\sigma}^{2*} = 0.00439$ |
| n↓ B→ | 150 |
| 5 | $\overline{\theta}^* = 2.9851, \hat{\sigma}^{2*} = 0.2718$ |
| 15 | $\overline{\theta}^* = 3.9576, \hat{\sigma}^{2*} = 0.00625$ |
| n↓ B→ | 1500 |
| 5 | $\overline{\theta}^* = 2.9829, \hat{\sigma}^{2*} = 0.19959$ |
| 15 | $\overline{\theta}^* = 3.9604, \hat{\sigma}^{2*} = 0.00746$ |

Za izradu poglavlja **4** korišćena je literatura [3], [5] i [8].

# 5 Intervali poverenja

## 5.1 Uvod

Prilikom proučavanja obeležja na osnovu uzorka moguća su dva slučaja:

(1) zna se oblik raspodele obeležja, ali su nepoznati parametri,

(2) ne zna se raspodela obeležja.

Neka je dat prost slučajan uzorak $\mathbf{X} = (X_1, X_2, ..., X_n)$ obima $n$ za posmatrano obeležje $X$. Statistika $Y = f(X_1, X_2, ..., X_n)$ je slučajna veličina koja implicitno zavisi od parametra $\theta$ u raspodeli $F$ obeležja $X$ . Ako statistikom $Y$ ocenjujemo parametar $\theta$, tada se statistika $Y$ naziva ocena parametra $\theta$ i označava se sa $\hat{\theta}$.
Nepoznati parametar $\theta$ raspodele $F$ ocenjuje se na osnovu realizovanog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$. Taj postupak naziva se ocenjivanje parametra. Pošto je

realizovana vrednost statistike neki realni broj, tj. neka tačka na realnoj pravoj, ovakva ocena parametra se naziva **tačkasta ocene**.

Statistike koje se koriste treba da imaju određene osobine kao što su: nepristrasnost, postojanost i efikasnost. Tim osobinama se opravdava njihova primena u ocenjivanju parametara.

Realizovana vrednost tačkaste ocene parametra može dosta odstupati od stvarne vrednosti parametra, a da je pri tome nepoznato koliko je to odstupanje. Verovatnoća da statistika $Y$ uzme vrednost jednaku stvarnoj vrednosti može biti nula, što i jeste slučaj kod neprekidnih raspodela. Stoga se, na osnovu prostog slučajnog uzorka određuje interval koji, sa unapred zadanom pouzdanošću, sadrži nepoznati parametar. Tada se govori o **intervalnoj oceni parametra** ili o **intervalu poverenja**.

Na osnovu uzorka definišu se statistike $L(X_1, X_2, ..., X_n)$ i $S(X_1, X_2, ..., X_n)$ tako da važe uslovi:

$$P\{L \leq S\} = 1,$$

$$P\{L \leq \theta \leq S\} = \alpha, \quad \alpha\epsilon[0, 1].$$

Tada se $[L, S]$ naziva **interval poverenja za nepoznati parametar $\theta$ sa nivoom poverenja** $\alpha$. Obično se kaže da je to $100\alpha\%$ interval poverenja za nepoznati parametar $\theta$. Statistike $L$ i $S$ nazivaju se donja i gornja granica intervala poverenja. U primenama se uzima da je vrednost nivoa poverenja $\alpha$ blisko jedinici, najčešće je $\alpha$=0.9 ili $\alpha$=0.95.

Isto ime koristimo i za realizovano $[l, s]$. Razlika je u tome što neslučajni interval $[l, s]$ sadrži ili ne sadrži $\theta$, dok za slučajni interval $[L, S]$ vezujemo tzv. verovatnoću pokrivanja nepoznatog parametra.

Poznato je više butstrap načina za ocenjivanje intervala poverenja, a neki od njih su:

(1) Efronov percentilni butstrap,

(2) Efronov percentilni butstrap sa korekcijom pristrasnosti,

(3) percentilni t-butstrap,

(4) normalni t-butstrap.

## 5.2 Efronov percentilni butstrap

Neka je dat prost slučajan uzorak $\mathbf{X} = (X_1, X_2, ..., X_n)$ obeležja $X$ sa raspodelom $F$ u kojoj figuriše nepoznati parametar $\theta$. Potrebno je konstruisati interval poverenja za parametar $\theta$. Primenjujemo uobičajenu butstrap proceduru tj. na osnovu realizovanog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ računamo vrednost ocene $\hat{\theta} = \hat{\theta}(\mathbf{X})$, a zatim generišemo butstrap replike $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*$ po pravilu $F_{\hat{\theta}(\mathbf{x})}$. Ove replike treba u određenoj meri da reflektuju slučajnost originalne ocene $\hat{\theta}$. To nam intuitivno sugeriše da je potrebno poređati butstrap ocene u varijacioni niz

$$\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq ... \leq \hat{\theta}_{(B)}^*,$$

a onda da za gornju i donju granicu intervala poverenja adekvatno odaberemo veću i manju vrednost.
Efron je predložio da ukoliko hoćemo da napravimo npr. 90% interval poverenja, izaberemo 90% $\hat{\theta}_i^*$-ova koji se nalaze u sredini varijacionog niza, a da odbacimo 5% najmanjih i 5% najvećih vrednosti.
Već smo u prethodnim odeljcima uveli oznaku $H^*$ koja je označavala funkciju raspodele idealne butstrap ocene $\hat{\hat{\theta}}^* = \hat{\theta}(\mathbf{X}^*)$. Sada ćemo je označiti sa $H_{\hat{\theta}}(h)$ da bismo naglasili da je reč o parametarskom butstrapu.

$$H^*(h) = H_{\hat{\theta}}(h) = P_{\hat{\theta}}\{\hat{\theta}(\mathbf{X}^*) \leq h\}, \quad h\epsilon\mathbf{R}.$$

**Gornju granicu nominalnog 100$\alpha$% intervala poverenja** za $\theta$ nalazimo rešavanjem jednačine

$$H_{\hat{\theta}}(h) = (1 + \alpha)/2$$

po h. Dobijamo[8]:

$$\hat{\theta}^g(\alpha) = H_{\hat{\theta}}^{-1}((1 + \alpha)/2). \tag{7}$$

Podsetimo se da kada kada $B \to \infty$, empirijska funkcija raspodele

$$H_B(\rho|\hat{\theta}^*) = \frac{1}{B}\sum_{i=1}^{B} I(\{\hat{\theta}_i^* \leq \rho\}), \ \rho\epsilon\mathbf{R}$$

[8]Indeks $g$ u gornjem uglu označava da je u pitanju gornja granica

kolekcije $\hat{\theta}^* = (\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$ konvergira ka funkciji raspodele $H_{\hat{\theta}}$ idealne butstrap ocene

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

Kada je B veliko, rešenje jednačine (7) možemo dobiti primenom empirijske funkcije raspodele $H_B(\cdot|\hat{\theta}^*)$, odnosno zamenom

$$H_{\hat{\theta}}^{-1}((1+\alpha)/2) \to H_B^{-1}((1+\alpha)/2|\hat{\theta}^*), B \to \infty.$$

U praksi ovu vrednost dobijamo određivanjem broja

$$g = \frac{1+\alpha}{2}B,$$

a zatim g-tu vrednost po redu, $\hat{\theta}_{(g)}^*$, iz niza

$$\hat{\theta}_{(1)}^* \leq \hat{\theta}_{(2)}^* \leq ... \leq \hat{\theta}_{(B)}^*$$

proglašavamo za gornju granicu intervala poverenja. Ukoliko $g$ nije ceo broj onda nalazimo najveći broj $n$ za koji važi $n \leq g$ , $n\epsilon\mathbf{Z}$, a zatim interpolacijom dobijamo gornju granicu koja je oblika:

$$\hat{\theta}_{(g)}^* = \hat{\theta}_{(n)}^* + (g-n)(\hat{\theta}_{(n+1)}^* - \hat{\theta}_{(n)}^*).$$

Analogno računamo i **donju granicu 100$\alpha$% intervala poverenja** parametra $\theta$:

$$\hat{\theta}^d(\alpha) = H_{\hat{\theta}}^{-1}((1-\alpha)/2) \to H_B^{-1}((1-\alpha)/2|\hat{\theta}^*), B \to \infty. \tag{8}$$

U praksi računamo broj $d$ tako da važi:

$$d = \frac{1-\alpha}{2}B,$$

a onda statistiku poretka, $\hat{\theta}_{(d)}^*$, proglašavamo donjom granicom. Ukoliko $d$ nije ceo broj, onda već opisanim postupkom interpolacije nalazimo granicu. Granice (7) i (8) određuju **nominalni 100$\alpha$% intervala poverenja sa jednakim repovima** za parametar $\theta$ dobijen Efronovim percentilnim metodom:

$$I = (\hat{\theta}^d(\alpha), \hat{\theta}^g(\alpha)).$$

## 5.3 Efronov percentilni butstrap sa korekcijom pristrasnosti

Kao i u prethodnom odeljku, potrebno je naći interval poverenja za parametar $\theta$.

Kažemo da je ocena $\hat{\theta}$ nepristrasna u odnosu na medijanu ukoliko ispunjava uslov

$$H_\theta(\theta) = P_\theta\{\hat{\theta} \leq \theta\} = 0.5. \tag{9}$$

Ukoliko govorimo u terminima butstrap raspodele onda bi važilo:

$$H_{\hat{\theta}}(\hat{\theta}) = 0.5.$$

Ako sa $N$ obeležimo funkciju normalne raspodele sa matematičkim očekivanjem 0 i disperzijom 1, onda je

$$k = N^{-1}(H_{\hat{\theta}}(\hat{\theta}))$$

vrednost koja predstavlja Efronov predlog za korekciju u slučaju kada se vrši korigovanje pristrasnosti ocena koje nisu nepristrasne u odnosu na medijanu. Ako je $k = 0$ onda $\hat{\theta}$ ispunjava uslov (9). Efron za gornju granicu nominalnog $100\alpha\%$ intervala poverenja sa jednakim repovima predlaže:

$$\hat{\theta}^g(\alpha) = H^{-1}_{\hat{\theta}}(N(2k + n_{(1+\alpha)/2})), \tag{10}$$

gde je $n_t = N^{-1}(t)$ kvantil reda $t$ raspodele $N(0, 1)$. Analogno dobijamo i donju granicu nominalnog $100\alpha\%$ intervala poverenja:

$$\hat{\theta}^g(\alpha) = H^{-1}_{\hat{\theta}}(N(2k + n_{(1-\alpha)/2})).$$

U praksi ovaj metod bismo primenili tako što generišemo butstrap replike $\hat{\theta}^*_1, \hat{\theta}^*_2, ..., \hat{\theta}^*_B$ prema pravilu $F_{\hat{\theta}}$, zatim ih poređamo u varijacioni niz

$$\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq ... \leq \hat{\theta}^*_{(B)},$$

i odredimo vrednost broja $\hat{q}$, tj. broj $\hat{\theta}^*_{(i)}$-ova koje su manje ili jednake od $\hat{\theta}(\mathbf{x})$.
Za aproksimaciju korelacionog faktora $k$ uzimamo $N^{-1}(\hat{q})$.
Računamo

$$q_{(1+\alpha)/2} = N(2k + n_{(1+\alpha)/2}),$$

$$g = Bq_{(1+\alpha)/2}.$$

Zatim, $g$-ti element, $\hat{\theta}^*_{(g)}$, varijacionog niza ocena $\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq ... \leq \hat{\theta}^*_{(B)}$ proglašavamo za gornju granicu intervala poverenja parametra $\theta$. Ako $g$ nije ceo broj onda postupkom interpolacije, koji je objašnjen u prethodnom odeljku, nalazimo gornju granicu.
Analogno nalazimo i donju granicu, tj. računamo

$$d = Bq_{(1-\alpha)/2},$$

a zatim $d$-ti element po redu proglašavamo za donju granicu intervala poverenja za parametar $\theta$.

## 5.4 Percentilni t-butstrap

Percentilni t-butstrap lakši je za računanje u odnosu na Efronov percentilni butstrap.
Polazeći od prostog slučajnog uzorka $\mathbf{X} = (X_1, X_2, ..., X_n)$ obeležja $X$ generisanog funkcijom raspodele $F = F(\theta)$, i zadate ocene $\hat{\theta} = \hat{\theta}(\mathbf{X})$ od $\theta$, konstruisaćemo interval poverenja za ovaj nepoznati parametar.
Pretpostavimo da je $\hat{\sigma}$ ocena standardne devijacije ocene $\hat{\theta}$ i da je dobijena od polaznog uzorka.
Konstruišemo butstrap kolekciju replika $\hat{\theta}^*_1, \hat{\theta}^*_2, ..., \hat{\theta}^*_B$ i za svaku od njih računamo butstrap ocenu standardne devijacije $\hat{\sigma}^*_i$ ($\hat{\sigma}^*_i$ je ocena standardne devijacije koja odgovara $\hat{\theta}^*_i$). Definišimo:

$$T^*_i = (\hat{\theta}^*_i - \hat{\theta})/\hat{\sigma}^*_i; \quad i = 1, ..., B,$$

$$T^* = (T^*_1, T^*_2, ..., T^*_B), \ \hat{\sigma}^* = (\hat{\sigma}^*_1, \hat{\sigma}^*_2, ..., \hat{\sigma}^*_B).$$

Odnosno

$$T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$$

kao butstrap analog za

$$T = (\hat{\theta} - \theta)/\hat{\sigma}.$$

Ukoliko bi $\theta$ bila srednja vrednost obeležja populacije i $\hat{\theta}$ njena ocena dobijena na osnovu uzorka onda bi vrednost $T$ pripadala familiji t-statistika, a ukoliko još i uzorak potiče iz normalne raspodele onda je $T$ pivot vrednost. To znači da je raspodela verovatnoće za $T$ nezavisna od parametara modela. U našem slučaju,iznad, pod pretpostavkom normalnosti, statistika $T$ ima

Studentovu raspodelu sa $n-1$ stepeni slobode. Prisetimo se da Studentova raspodela zavisi samo od broja stepeni slobode koji je definisan poznatim brojem veličine uzorka ($n$) i ne zavisi od srednje vrednosti obeležja populacije $\theta$ i standardne devijacije $\hat{\sigma}$ ocene $\hat{\theta}$. Ovakve vrednosti zovemo "pivoti", zato što za njih možemo napraviti tačan izraz za verovatnoću, a onda procesom "pivotiranja" taj izraz možemo transformisati u interval poverenja za parametar.
Ukoliko je $\theta$ srednja vrednost obeležja populacije,$\hat{\theta}$ njena ocena dobijena na osnovu uzorka izvučenog iz normalne raspodele, butstrap statistika $T^*$ je asimptotski pivot, što znači da njena raspodela asimptotski teži raspodeli nezavisnoj od parametara i njeni percentili konvergiraju ka percentilima statistike $T$.
U većini slučajeva parametrar $\theta$ je znatno komplikovaniji od srednje vrednosti obeležja populacije, pa je i butstrap ocena $\hat{\theta}^*$ nepoznata. Zato primenjujemo butstrap postupak kako bi dobili replike $\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*$ i izračunali vrednosti $\hat{T}_i^*, \quad i = 1, ..., B$. Zatim pravimo neopadajući niz $\hat{T}_{(1)}^* \leq \hat{T}_{(2)}^* \leq ... \leq \hat{T}_{(B)}^*$, i na njemu primenjujemo obični percentilni butstrap koji podrazumeva da se interval poverenja gradi od $100\alpha\%$ $\hat{T}_{(i)}^*$-iova koji se nalaze u sredini niza (tj. odbacuje se (1-$\alpha$)/2 najmanjih i (1+$\alpha$)/2 najvećih vrednosti).

Drugim rečima, Efron je za nominalni $100\alpha\%$ interval poverenja predložio

$$[\hat{\theta} - T^*_{\frac{1+\alpha}{2}}\hat{\sigma}, \hat{\theta} + T^*_{\frac{1-\alpha}{2}}\hat{\sigma}],$$

gde je $T^*_{\frac{1+\alpha}{2}}$ $100\frac{1+\alpha}{2}$ percentil od $T^*$, a $T^*_{\frac{1-\alpha}{2}}$ $100\frac{1-\alpha}{2}$ percentil. Da bi dobili $T^*_{\frac{1+\alpha}{2}}$ i $T^*_{\frac{1-\alpha}{2}}$ iz butstrap histograma $T_i^*$-ova, mora biti ispunjen uslov

$$B\frac{1-\alpha}{2} \, \epsilon \, \mathbf{Z}.$$

Ukoliko to nije slučaj, Efron je predložio da se izaberu najbliži percentili, tj. da se izračuna broj $k = [(B+1)\frac{1-\alpha}{2}]$ (zagrade [ ] označavaju ceo deo broja) koji je najveći broj manji od $(B+1)\frac{1-\alpha}{2}$, a onda da $T_k^*$ bude zamena za $T^*_{\frac{1-\alpha}{2}}$, a $T^*_{B+1-k}$ zamena za $T^*_{\frac{1+\alpha}{2}}$.

Veliko ograničenje za primenu percentilnog t-butstrapa jeste potreba za ocenom standardne devijacije ocene $\hat{\theta}$ i butstrap ocena standardne devijacije $\hat{\sigma}_i^*, \quad i = 1, ..., B$, koje nam nije uvek poznato.

Postoji još jedan butstrap metod za ocenjivanje intervala poverenja, koji

Hesterberg[9] naziva "butstrap-t metod". Zbog sličnog imena sa percentilnim t-butstrapom, često se u literaturi može naći i pod imenom "normalni t-butstrap". Taj metod koristi se samo za ocenjivanje standardne devijacije, i to u slučaju kada je butstrap raspodela ocene približno Gausova, sa malom pristrasnošću. Hesterberg je za nominalni $100\alpha\%$ interval poverenja predložio

$$[\hat{\theta} - t^*\hat{\sigma}^*, \hat{\theta} + t^*\hat{\sigma}^*],$$

gde je $t^*$ $100\frac{1+\alpha}{2}$ percentil Studentove t-raspodele sa $n-1$ stepeni slobode ($n$ je veličina polaznog uzorka).
Uočimo razliku između percentilnog t-butstrapa i normalnog t-butstrapa:
Za percentilni t-butstrap, percentili su uzeti iz butstrap raspodele za $T^*$, a standardna devijacija ocene $\hat{\theta}$ iz originalnog uzorka. U normalnom t-butstrapu percentili su uzeti is Studentove raspodele, ali je standardna devijacija izračunata na osnovu butstrap uzoraka.

**Primer**:
Neka je $X$ obeležje sa raspodelom koja pripada dopustivoj familiji $\{N(\mu, \sigma^2),\ \mu\epsilon\mathbf{R}, \sigma > 0\}$. Treba konstruisati dvostrani interval poverenja sa jednakim repovima na nominalnom nivou $\alpha = 0.95$ za nepoznatu varijansu obeležja, $\theta = \sigma^2$, primenom Efronovog percentilnog butstrapa.
Ovo ćemo da ilustrujemo na uzorku obima $n = 10$ iz populacije $N(3, 1)$.

**Rezultati**:
Ako je dat prost slučajan uzorak $\mathbf{X} = (X_1, X_2, ..., X_n)$, nepoznate parametre $\mu$ i $\sigma^2$ ocenjujemo uobičajenim ocenama $\overline{X}_n$ i $\overline{S}_n^2$, respektivno.
Klasičan dvostrani $100\alpha\%$-ni interval poverenja za $\sigma^2$ obeležja $X$ ima oblik

$$\left(\frac{n\overline{S}_n^2}{c_2}, \frac{n\overline{S}_n^2}{c_1}\right), \tag{11}$$

gde su $c_1$ i $c_2$ kvantili reda $(1-\alpha)/2$ i $(1+\alpha)/2$ hi-kvadrat raspodele sa $n-1$ stepeni slobode, $\chi^2_{n-1}$.
Statistika od koje polazimo je

$$\frac{n\overline{S}_n^2}{\sigma^2}.$$

Ona ima hi-kvadrat raspodelu sa $n-1$ stepeni slobode, pa $c_1$ i $c_2$ određujemo

[9] T.Hesterberg

pomoću izraza

$$P\{c_1 \leq \frac{n\overline{S}_n^2}{\sigma^2} \leq c_2\} = P\{c_1 \leq \chi^2_{n-1} \leq c_2\} = \alpha.$$

Butstrap postupak započinjemo izvlačenjem originalnog uzorka,
$\mathbf{x} = (x_1, x_2, ..., x_n)$, na osnovu kojeg računamo vrednosti ocena nepoznate sredine i varijanse, $\overline{x}_n$ i $\overline{s}_n^2$. Zatim, generišemo $B$ butstrap uzoraka $\mathbf{x}_i^*,\ i = 1, ..., B$, iz raspodele $N(\overline{x}_n, \overline{s}_n^2)$ i određujemo kolekciju butstrap replika, $\hat{\theta}_i^* = \hat{\theta}(\mathbf{x}_i^*) = (\overline{s}_n^2)_i^*,\ i = 1, ..., B$.
Odgovarajuća empirijska funkcija raspodele $H_B(\cdot|\hat{\theta}^*)$ dobro aproksimira funkciju raspodele $H_{\hat{\theta}}(\cdot)$ idealne butstrap ocene $\hat{\hat{\theta}}^* = \hat{\theta}(\mathbf{X}^*)$. Kada je $B$ veliko možemo odrediti njen tačan oblik[10]:

$$H_{\hat{\theta}}(x) = P_{\hat{\theta}}\left\{\hat{\theta}(\mathbf{X}^*) \leq x\right\} = P_{\hat{\theta}}\left\{\frac{n\overline{S}_n^{2*}}{\overline{s}_n^2} \leq \frac{nx}{\overline{s}_n^2}\right\} = \chi^2_{n-1}\left(\frac{nx}{\overline{s}_n^2}\right),\ x\epsilon\mathbf{R}. \quad (12)$$

Ovaj izraz sledi iz činjenice da slučajna promenljiva $\frac{n\overline{S}_n^{2*}}{\overline{s}_n^2}$ ima hi-kvadrat raspodelu sa $n-1$ stepeni slobode.
Da bismo odredili gornju granicu intervala poverenja koristimo izraz (7) iz poglavlja **5.2**:

$$\hat{\theta}^g(\alpha) = H_{\hat{\theta}}^{-1}((1+\alpha)/2) = \frac{\overline{s}_n^2}{n}(\chi^2_{n-1})^{-1}((1+\alpha)/2). \quad (13)$$

Potpuno analogno, donja granica $100\alpha\%$-nog intervala poverenja za nepoznatu varijansu obeležja $X$ je:

$$\hat{\theta}^d(\alpha) = H_{\hat{\theta}}^{-1}((1-\alpha)/2) = \frac{\overline{s}_n^2}{n}(\chi^2_{n-1})^{-1}((1-\alpha)/2). \quad (14)$$

Sada ćemo da generišemo jedan uzorak, obima $n$=10, iz $N(3,1)$ raspodele pozivom funkcije `rnorm`. Dobijamo:

```
x<-rnorm(10,3,1)
Polazni uzorak x:
 2.4001077 1.6318391 2.0466381 0.4391476 3.4143977
 3.7923220 3.8099106 4.1328964 3.7423154 2.7227997
```

[10]Funkcija hi-kvadrat raspodele sa $n-1$ stepeni slobode

Pretvarajući se da nemamo informaciju o $\mu$ i $\sigma^2$, putem simulacija konstruisaćemo Efronov interval poverenja za nepoznatu varijansu obeležja na nominalnom nivou $\alpha = 0.95$, pozivom funkcije `intPovPercentilni`:

```
intPovPercentilni(x,0.95,10000)  alfa=0.95, B=10000
Interval poverenja dobijen Efronovim percentilnim butstrapom:

donja granica:  0.396242
gornja granica: 2.326952
```

Prema statističkoj teoriji, klasičan 95%-ni interval poverenja (izraz (11)) je:

$$\left(\frac{n\overline{s}_n^2}{c_2}, \frac{n\overline{s}_n^2}{c_1}\right) = \left(\frac{10 \cdot 1.412953}{19.0228}, \frac{10 \cdot 1.412953}{2.7004}\right) = (0.7427682, 5.232384)$$

gde su odgovarajući kvantili hi-kvadrat raspodele

$$c_1 = \chi^2_{9,0.025} = 2.7004,\ c_2 = \chi^2_{9,0.975} = 19.0228,$$

a varijansa uzorka, $\overline{s}_n^2$, je izračunata naredbom `var(x)`:

$$\hat{\theta}(\mathbf{x}) = \overline{s}_{10}^2 = 1.412953.$$

Kada je broj butstrap uzoraka $B$ veliki, onda možemo i da odredimo analitički oblik Efronovog percentilnog intervala poverenja, tj. formule (13) i (14) daju gornju i donju granicu:

$$\hat{\theta}^g(\alpha) = \frac{\overline{s}_{10}^2}{10}\chi^2_{9,0.975} = 2.687832,$$

$$\hat{\theta}^d(\alpha) = \frac{\overline{s}_{10}^2}{10}\chi^2_{9,0.025} = 0.3815538.$$

Zaključujemo da postoji visok stepen saglasnosti između analitičkih rezultata i rezultata dobijenih upotrebom simulacija, a to je postignuto simuliranjem velikog broja butstrap replikacija.
Za izradu poglavalja **5** korišćena je literatura [3], [4] i [8].

# 6 Testiranje hipoteza

## 6.1 Uvod

Statističkom hipotezom nazivamo tvrđenje, pretpostavku u vezi sa svojstvom populacije koje nas interesuje, tj. o nekom njenom obeležju. Hipoteza koja se testira se zove nulta hipoteza i označava se sa $H_0$. U paru sa nultom hipotezom uvek se javlja hipoteza $H_1$ ili alternativna hipoteza koja na neki način protivreči hipotezi $H_0$. Obe hipoteze mogu biti proste ili složene. Za hipotezu se kaže da je prosta, ako se odnosi na jednu vrednost parametra kojom je raspodela obeležja potpuno određena, npr. $H_0(\theta = \theta_0)$. Ako hipoteza nije prosta, onda je složena kao što su hipoteze:

$$H_0(\theta < \theta_0),\ H_0(\theta > \theta_0).$$

Statističkim testom nazivamo postupak (ne)odbacivanja nulte hipoteze na osnovu realizovanog uzorka. Baš kao sto ne možemo naći 100%-ni interval poverenja, ni statistički testovi ne garantuju 100%-nu sigurnost. Može se desiti da dva različita uzorka dovedu do suprotnih odluka, pa je pravilno zaključiti da *na osnovu datog uzorka* odbacujemo ili nemamo razlog da odbacimo $H_0$.
Ako se odbacuje nulta hipoteza kada je ona tačna, čini se greška prvog tipa, koju označavamo sa $\alpha$. Ako se prihvata nulta hipoteza kada je tačna alternativna hipoteza, čini se greška drugog tipa, $\beta$. Broj $\alpha$ se još naziva prag značajnosti ili nivo značajnosti. U praksi se najčešće uzima da je $\alpha$ jednako 0.01 ili 0.05

Statistički test je u potpunosti određen kritičnom oblašću $W$. Oblik kritične oblasti određuje alternativna hipoteza, a veličinu kritične oblasti i njene granice određuje prag značajnosti. Ako realizovani uzorak pripada kritičnoj oblasti onda se nulta hipoteza odbacuje, u suprotnom se prihvata.
Da li je razlika između onoga što očekujemo na osnovu $H_0$ i informacije koje nam pruža izvučeni uzorak realna ili je posledica slučajnosti, testira se upotrebom odgovarajuće test statistike $T$, tj. slučajne promenljive koja ima poznatu raspodelu verovatnoće ukoliko je nulta hipoteza tačka, dok je pod alternativnom raspodeljena drugačije.
Postupak je sledeći:
∘ Kada je registrovana vrednost test statistike, $T$, takva da ju je pod $H_0$ sa velikom verovatnoćom moguće dobiti na slučajan način, zaključujemo da test nije obezbedio adekvatan dokaz protiv ove hipoteze.

∘ Ako se vrednost $T$ može okarakterisati kao "ekstrem", u smislu neuobičajnosti njene slučajne realizacije pod $H_0$, to predstavlja argument koji ne ide u prilog polaznoj pretpostavci pa možemo doneti odluku o njenom odbacivanju.

Odluka o tome da li odbaciti nultu hipotezu ili ne, može biti doneta na osnovu tzv. **p-vrednosti** pridružene registrovanom $\hat{T}$, gde je sa $\hat{T}$ označena vrednost statistike $T$ izračunata na osnovu realizovanog uzorka. p-vrednost testa je verovatnoća da test statistika $T$ uzme još "ekstemniju"vrednost od ove realizacije, pod pretpostavkom da je nulta hipoteza tačna. Ako je u pitanju desnostrani test, tj.

$$\alpha = P_{H_0}\left\{T > c\right\},$$

gde smo sa $c$ obeležili granicu kritične oblasti, odgovarajuća p-vrednost se definiše kao

$$p = P_{H_0}\left\{T > \hat{T}\right\}.$$

Tada nultu hipotezu odbacujemo za $p \leq \alpha$, dok za $p > \alpha$ nema osnova da se ona odbaci.

## 6.2 Postupak testiranja

Pretpostavimo da vršimo testiranje **desnostrane alternativne hipoteze** pomoću test statistike $T$, tj. kritična oblast se nalazi u gornjem repu raspodele test statistike, a nultu hipotezu odbacujemo za dovoljno velike realizacije test statistike. Testiranje vršimo na nivou značajnosti $\alpha$.

Na osnovu raspoloživog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ računamo realizovanu vrednost test statistike

$$\hat{T} = T(x_1, x_2, ..., x_n).$$

Sa $G(T)$ označavamo funkciju raspodele od $T$, pri tačnoj nultoj hipotezi. Tada odgovarajuću p-vrednost dobijamo po formuli

$$p(\hat{T}) = 1 - G(\hat{T}). \tag{15}$$

Nultu hipotezu odbacujemo ukoliko je $p(\hat{T}) < \alpha$.
Alternativni pristup bi dobili putem računanja kritične oblasti $c(\alpha)$. U tom slučaju nultu hipotezu odbacujemo ako i samo ako važi:

$$\hat{T} > c(\alpha),$$

$$P_{H_0}\{\hat{T} > c(\alpha)\} = \alpha.$$

Bez obzira na način koji izaberemo da izvršimo testiranje, krajnji zaključci moraju da budu isti.

Najčešći problem sa kojim se srećemo jeste taj što funkcija raspodele $G(T)$ test statistike $T$ nije poznata. Rešenje nalazimo u aproksimaciji funkcije $G(T)$ nekom drugom funkcijom. Zbog jednostavnosti primene i usled stalnog rasta računarske moći, primenićemo butstrap testiranje, koje u odnosu na druge postupke često bolje aproksimira nepoznato $G(T)$.

Od zadatog uzorka $\mathbf{x} = (x_1, x_2, ..., x_n)$ pravimo B butstrap uzoraka $\mathbf{X}^* = (X_1^*, X_2^*, ..., X_B^*)$ (parametarskim ili neparametarskim butstrap metodom). Za svaki butstrap uzorak računamo vrednost butstrap test statistike:

$$\hat{T}_i^* = T(x_i^*), \quad i = 1, ..., B,$$

najčešće na isti način na koji je i $\hat{T}$ izračunato iz $\mathbf{x}$. Preporuka je da upotrebljeni butstrap treba da zadovoljava nultu hipotezu, ali to nije uvek moguće, pa se u takvim situacijam $\hat{T}_i^*$ ne mogu računati potpuno isto kao $\hat{T}$ već je potrebno izvršiti transformaciju polaznog uzorka kako bi butstrap zadovoljio nultu hipotezu.

Konstruišemo empirijsku funkciju raspodele dobijene kolekcije,

$$G_B(x|\mathbf{T}) = \frac{1}{B}\sum_{i=1}^{B} I(\hat{T}_i^* \leq x), \quad x\epsilon\mathbf{R}, \quad \mathbf{T} = (\hat{T}_1^*, \hat{T}_2^*, ..., \hat{T}_B^*)$$

i računamo butstrap p-vrednost, kao ocenu stvarne p-vrednosti, (15):

$$\hat{p}^*(\hat{T}) = 1 - G_B(\hat{T}|\mathbf{T}) = \frac{1}{B}\sum_{i=1}^{B} I(\hat{T}_i^* > \hat{T}). \tag{16}$$

Butstrap p-vrednost je u opštem slučaju udeo butstrap test statistika $\hat{T}_i^*$, čije su vrednosti ekstremnije u odnosu na originalnu realizaciju $\hat{T}$. Izraz (16) je empirijski analog za (15), zato nultu hipotezu odbacujemo kada je $\hat{p}^*(\hat{T}) < \alpha$. Ako pustimo da broj generisanih butstrap uzoraka B teži beskonačnosti, empirijska funkcija raspodele $G_B(x|\mathbf{T})$ će biti dobra aproksimacija za stvarnu funkciju raspodele vrednosti $\hat{T}_i^*$, u oznaci $G^*(T)$, pa će time i butstrap p-vrednost (16) biti bliska idealnoj butstrap p-vrednosti, označenoj sa $p^*(\hat{T}) = 1 - G^*(\hat{T})$.

Ukoliko želimo da izvršimo **levostrani test** tj. da nultu hipotezu $H_0$ odbacimo za vrednosti $\hat{T}$ koje se nalaze u donjem repu raspodele test statistike, onda je $\alpha = P_{H_0}\{\hat{T} \leq c(\alpha)\}$, a odgovarajuća p-vrednost je oblika

$p(\hat{T}) = G(\hat{T})$. Time, u formuli (16), dobijamo nejednakost suprotnog smera, tj.

$$\hat{p}^*(\hat{T}) = G_B(\hat{T}|\mathbf{T}) = \frac{1}{B}\sum_{i=1}^{B} I(\hat{T}_i^* \leq \hat{T}).$$

Ukoliko želimo da izvršimo **dvostrani test**, moramo formirati butstrap p-vrednost na odgovarajući način. Pretpostavimo da je statistika $T$ simetrično raspodeljena oko nule. Tada koristimo izraz

$$\hat{p}^*_{sim}(\hat{T}) = \frac{1}{B}\sum_{i=1}^{B} I(|\hat{T}_i^*| > |\hat{T}|), \tag{17}$$

kako bi označili **simetričnu butstrap p-vrednost**. Ovaj izraz praktično konvertuje dvostrani test u jednostrani. U suprotnom, (17) moguće je zameniti **butstrap p-vrednošću za jednake repove**, koja ima oblik

$$\hat{p}^*_{jr}(\hat{T}) = 2min\left\{\frac{1}{B}\sum_{i=1}^{B} I(\hat{T}_i^* \leq \hat{T}), \frac{1}{B}\sum_{i=1}^{B} I(\hat{T}_i^* > \hat{T})\right\}. \tag{18}$$

Ovde, zapravo, izvodimo dva testa, koja odbacuju nultu hipotezu $H_0$ za one vrednosti test statistike koje se nalaze u donjem i gornjem repu raspodele. Izrazi (17) i (18) ne moraju dati slične rezultate. Uglavnom su testovi bazirani na simetričnoj butstrap p-vrednosti, pod nultom hipotezom, pouzdaniji od onih koji računaju butstrap p-vrednost za jednake repove.

Samim tim, izraz (18) se koristi ukoliko želimo da odbacimo nultu hipotezu i za suviše male i za suviše velike vrednosti $\hat{T}$.
Izraz (16) koristimo kada test statistika uzima uvek pozitivne vrednosti, kao $\chi^2$, dok izraz (17) je primenljiv samo na test statistike čije vrednosti mogu imati ma koji predznak, poput t-statistike.
Za izradu poglavla **6** korišćena je literatura [3], [4], [5] i [8].

# 7 Zaključak

Od 1977.godine, kada je Efron predstavio osnove butstrap metodologije, pa do danas, razvijen je veliki broj butstrap postupaka koji se koriste u različite svrhe. Neke od oblasti u kojima se primenjuje butstrap smo naveli u radu: ocenjivanje tačkastih i intervalnih ocena, testiranje hipoteza, pravljenje intervala poverenja. Međutim, butstrap se koristi i u regresionoj analizi,

u prognoziranju vremenskih serija i dr.
Za primenjivanje ove metode potrebno je napraviti samo jednu pretpostavku, da uzorak, na osnovu kojeg vršimo reuzorkovanje, dobro predstavlja populaciju iz koje je izvučen. Zatim,računanjem dolazi se do butstrap rezultata. Uglavnom je dovoljno da broj butstrap uzoraka bude oko 1000.
Neki od naučnika su veliki protivnici ove metode. Oni smatraju da reuzorkovanjem iz samo jednog uzorka nije moguće dobiti validne rezultate jer se može desiti da dobijeni uzorak ne predstavlja najbolje populaciju iz koje je izvučen.
U svakom slučaju, ukoliko postoje dodatne pretpostavke o populaciji, bolje je koristiti uobičajene statističke metode, a ukoliko sumnjamo u neku pretpostavku ili ih uopšte nemamo onda je bolje primeniti butstrap metod.

# 8 Dodatak

Za izradu ovog poglavlja korišćena je literatura [7] i [11].

```
parametarskiButstrap<-function(n,B,mi,sigma){
#generisemo polazni uzorak x sa normalnom(mi,sigma)raspodelom
x<-rnorm(n,mi,sigma)
#sredina, disperzija i standardna devijacija polaznog uzorka
mean(x)
var(x)
sd(x)
#vektor u koji smestamo butstrap ocene sredine za svaki
#reuzorak
bootsredina<- rep(0,B)
 bootdisperzija<- rep(0,B)
noviuzorak<-rep(0,B)
 for(i in 1 : B){
 noviuzorak<-rnorm(n,mean(x),var(x))
 bootsredina[i]<-mean(noviuzorak)
 bootdisperzija[i]<-var(noviuzorak)
 }
#racunamo butstrap ocenu sredine
boot_ocena_sredine<-(1/B)*sum(bootsredina)
boot_ocena_sredine
#racunamo butstrap ocenu disperzije
boot_ocena_disperzije<-(1/(B-1))*
*sum((bootsredina-boot_ocena_sredine)^2)
boot_ocena_disperzije
}
```

```
neparametarskiButstrap<-function(n,B,mi,sigma){
#generisemo polazni uzorak x sa normalnom(mi,sigma)raspodelom
x<-rnorm(n,mi,sigma)
#sredina, disperzija i standardna devijacija polaznog uzorka
mean(x)
var(x)
sd(x)
#vektor u koji smestamo butstrap ocene sredine za
# svaki reuzorak
bootsredina<- rep(0,B)
 bootdisperzija<- rep(0,B)
noviuzorak<-rep(0,B)
 for(i in 1 : B){
 noviuzorak<-sample(x, replace = T)
 bootsredina[i]<-mean(noviuzorak)
 bootdisperzija[i]<-var(noviuzorak)
 }
#racunamo butstrap ocenu sredine
boot_ocena_sredine<-(1/B)*sum(bootsredina)
boot_ocena_sredine
#racunamo butstrap ocenu disperzije
boot_ocena_disperzije<-(1/(B-1))*
*sum((bootsredina-boot_ocena_sredine)^2)
boot_ocena_disperzije
}
```

```
intPovPercentilni<-function(x,alfa,B){
sredina<-mean(x)
varijacija<-var(x)
stdev<-sd(x)
novi_uzorak<-rep(0,length(x)) #ovde cemo stavljati novi
# uzorak
niz_teta<-rep(0,B) #ovde cemo stavljati ocene
# varijanse novih uzoraka
for(i in 1:B){#pravimo novi uzorak
novi_uzorak<-rnorm(length(x),sredina,varijacija)
niz_teta[i]<-var(novi_uzorak) #varijansa novog uzorka
}
niz_teta<-sort(niz_teta)#sortiramo niz u rastuci poredak
#pravimo gornju granicu intervala poverenja
gornja_granica<-c(0)
g<-((1+alfa)/2)*B
n<-round(g)
if( g==n) {
#ako je ceo broj onda uzmemo g-ti element iz
# sortiranog niza
gornja_granica<-niz_teta[g]} else{
gornja_granica<-niz_teta[n]+
+(g-n)*(niz_teta[n+1]-niz_teta[n])}
#pravimo donju granicu
donja_granica<-c(0)
d<-((1-alfa)/2)*B
t<-round(d)
if( d==t) { #ako je ceo broj onda uzmemo d-ti element
# iz sortiranog niza
donja_granica<-niz_teta[d]} else{
donja_granica<-niz_teta[t]+
+(d-t)*(niz_teta[t+1]-niz_teta[t])
}
k<-c(donja_granica,gornja_granica)
k}
```

# Literatura


[1] B.Efron, Bootstrap methods: Another look at the Jackknife, The Annals of Statistics,1979, Vol.7, No.1, 1-26

[2] B.Efron, P.Diaconis, Computer-Intesive Methods in Statistics, Scientific American,1983, Vol.248, No.5, 116-132

[3] Michael R. Chernick, Bootstrap Methods: A Guide for Practitioners and Researchers, John Wiley & Sons, Inc., Hoboken, New Jersey, 2008

[4] Phillip Good, Permutation, Parametric and Bootstrap Tests of Hypotheses, Springer Science+Business Media, Inc., USA, 2005

[5] Prof. dr Vesna Jevremović, Verovatnoća i statistika, Matematički fakultet, Beograd, 2009

[6] Milan Merkle, Verovatnoća i statistika, Akademska misao, Beograd, 2010

[7] http://cran.r-project.org/doc/contrib/Kasum+Legovic-UvodUr.pdf

[8] http://www.dmi.uns.ac.rs/site/dmi/download/master/primenjena_matematika/IvanaMalic.pdf

[9] http://web.math.pmf.unizg.hr/ wagner/userfiles/nastava/sp2_vjezbe10.pdf

[10] http://psihologija.ff.uns.ac.rs/primenjena/clanci/20133249.pdf

[11] http://www.ievbras.ru/ecostat/Kiril/R/Biblio/R_eng/Chernick2011.pdf