UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

# TEORIJA I PRAKSA DOBIJANJA UZORAKA NA OSNOVU RASPOLOŽIVIH PODATAKA

MASTER RAD

SUZANA PRICA

MENTOR:
PROF. DR VESNA JEVREMOVIĆ

BEOGRAD, 2014.

*AUTOR SE ZAHVALJUJE MENTORU PROF. DR VESNI JEVREMOVIĆ, ČLANOVIMA KOMISIJE PROF. DR SLOBODANKI JANKOVIĆ I MR MARKU OBRADOVIĆU I KOLEGINICI MR OLGI MELOVSKI TRPINAC*

*NA SVESRDNOJ POMOĆI PRILIKOM PISANJA OVOG RADA*

# SADRŽAJ

# REZIME

U prvom delu rada su prikazani razni tipovi uzoraka (prost slučajan uzorak, uzorak sa vraćanjem, bez vraćanja, stratifikovani, grupni, dvoetapni, periodični uzorak) i načini dobijanja uzoraka. Prikazan je i najčešći način izbora uzorka ekonomskih jedinica u Republičkom zavodu za statistiku.

U nastavku rada reč je o tome kako se u daljem radu može koristiti reprezentativni uzorak. Jedna od metoda je Bootstrap i u radu je detaljno opisana i ilustrovana primerima. To je metoda koja se razlikuje od klasičnih statističkih metoda, podrazumeva korišćenje računara i omogućava dobijanje rezultata vezanih za tačkaste i intervalne ocene parametara raspodele posmatranog obeležja. Metoda može da se koristi i za testiranje statističkih hipoteza, što je takođe pokazano u radu.

U Dodatku su dati korišćeni programi pisani u R programskom jeziku.

# 1. UZORAK

## 1.1. NAČINI DOBIJANJA UZORAKA I VRSTE UZORAKA

Statistika se bavi proučavanjem skupova sa velikim brojem elemenata koje imaju jedno ili više zajedničkih kvantitativnih ili kvalitativnih svojstava. Oni se posmatraju u velikom broju, u masi, da bi se otkrilo šta je u njima opšte i zakonito.

Skup elemenata koji posmatramo zove se **populacija** (ili generalni skup) a svojstvo definisano za svaki element populacije zove se **obeležje**.

Posebno, **numeričko obeležje** je funkcija X definisana na populaciji $\Omega = \{\omega_1, \omega_2 . .\}$ sa vrednostima u skupu R tj. $X: \Omega \rightarrow R$.

Na istoj populaciji se istovremeno može posmatrati više obeležja a ponekad je od interesa posmatrati njihovu međuzavisnost.

Zadatak statistike je određivanje raspodele obeležja, ili bar nekih parametara te raspodele. Zbog toga se prikupljaju podaci o populaciji.

Podaci se mogu prikupljati na dva načina:

- potpunim ispitivanjem celokupne populacije
- delimičnim ispitivanjem, odnosno ispitivanjem jednog dela populacije

Zbog brojnosti populacije, velikih troškova vezanih sa registrovanjem obeležja kod svakog elementa, velikog gubitka vremena ili pak uništavanja elemenata populacije, nekad nije moguće dobiti kompletnu informaciju o raspodeli obeležja u celoj populaciji, tako da su potpuna ispitivanja retka u praksi. Primer potpunog ispitivanja je popis stanovništva koji se zbog velikih troškova vrši svake desete godine.

Zbog navedenih teškoća pristupa se delimičnom ispitivanju populacije tako što se iz celokupne populacije uzima jedan deo i on se izučava. Taj deo se zove **uzorak**. Broj elemenata u uzorku je konačan i zove se **obim uzorka**. Na izabranom uzorku registruje se obeležje kod svakog elementa a zatim se zaključci o dobijenoj raspodeli obeležja proširuju sa uzorka na ceo skup.

Pri tome se uzorak bira tako da bude **reprezentativan** tj. da se struktura uzorka u što većoj meri poklapa sa strukturom populacije kako bi dobijeni rezultati bili u određenom smislu bliski stvarnim karakteristikama populacije. Jedan od načina postizanja reprezentativnosti je da uzorak izaberemo slučajno.

Metod slučajnog izbora uzorka sastoji se u tome da se slučajno bira element ω iz populacije Ω i registruje njegovo obeležje X = X(ω). Obeležje X je slučajna promenljiva i ako vršimo n takvih biranja elemenata, odnosno registrovanja obeležja X, imamo uzorak obima n, tj. n-dimenzionalnu slučajnu promenljivu $(X_1,X_2,...,X_n)$ gde je $X_i$, (i = 1,2,...,n) obeležje X u i-tom biranju.

Specijalno, **prost slučajan uzorak** je uzorak kod koga su slučajne promenljive $X_i$ (i = 1,2,...,n) nezavisne i imaju istu raspodelu kao obeležje X.

n-dimenzionalni vektor $(x_1, x_2, \dots, x_n)$ gde su $x_i$ realizovane numeričke vrednosti slučajnih promenljivih $X_i$, (i = 1,2,...,n) se naziva **realizovani uzorak.**

Elementi koji se biraju u uzorak mogu posle beleženja osobina da se vraćaju u populaciju, da bi se zatim iz celokupne populacije na slučajan način uzimao sledeći element. Takav izbor se naziva **izbor sa vraćanjem**.

Ukoliko se izabrani element posle beleženja osobina ne vraća u populaciju, a sledeći element uzorka se bira među preostalim elementima populacije reč je o **izboru bez vraćanja**.

Uzorak željenog obima (sa ili bez vraćanja) može se dobiti korišćenjem **tablica slučajnih cifara** ili pomoću **(pseudo)slučajnih brojeva.**

- U **tablicama slučajnih cifara** cifre su grupisane radi lakšeg čitanja, a mogu se čitati sleva na desno ili odozgo prema dole ili po nekom drugom pravilu, počevši od bilo kog mesta u tablici.

Pretpostavimo da populacija ima N elemenata numerisanih brojevima 1,2,...,N i da želimo da formiramo uzorak obima n. Elemente uzorka ćemo dobiti čitanjem (redom) grupa po k cifara iz tablice slučajnih brojeva (gde je k broj cifara broja N). Ukoliko se radi o izboru sa vraćanjem uzorak ćemo formirati od prvih n k-tocifrenih brojeva koji zadovoljavaju uslov da su manji ili jednaki N. Ako se radi o uzorku bez vraćanja pored uslova da su k-tocifreni brojevi manji ili jednaki N mora da bude ispunjen uslov da su izabrani k-tocifreni brojevi različiti.

Nedostatak ove metode dobijanja uzorka je što se za neke vrednosti N odbacuje mnogo k-tocifrenih brojeva. Jedan od načina da se to prevaziđe je da se brojevi „čitaju“ na primer, po modulu 50 ako je $N \leq 50$ i N dvocifren broj.

- **Slučajni brojevi** su nezavisne realizacije slučajne veličine X sa uniformnom U(0,1) raspodelom a **pseudoslučajni brojevi** su brojevi iz intervala (0,1) koji se dobijaju po nekim pravilima. Za razliku od slučajnih brojeva koji se mogu dobiti, na primer, bacanjem kockice ili simetričnog novčića, pseudoslučajni brojevi se, kao što je već rečeno, dobijaju po nekom pravilu tako da je moguće ponoviti generisanje istog niza brojeva.

Niz $Y_1, Y_2, \ldots Y_k, \ldots$pseudoslučajnih brojeva se obično dobija nekom rekurentnom formulom, a jedna od prvih primenjivanih formula je tzv. metoda sredine kvadrata:

$$Y_{m+1} = D[10^{-2k} \cdot C(10^{3k} \cdot Y_m^2)],$$

gde D označava decimalni deo, C ceo deo broja a $Y_m$ ima oblik $0.\alpha_1 \ldots \alpha_{2k}$. Broj $Y_1$ se bira proizvoljno. Problem kod pseudoslučajnih brojeva dobijenih ovom metodom je što obično imaju mali period (dužina niza pre nego što počne da se ponavlja).

Pretpostavimo da populacija ima N elemenata numerisanih brojevima 1,2,...,N i da želimo da formiramo uzorak obima n. Elemente uzorka ćemo dobiti prevođenjem (pseudo)slučajnog broja $Y$ iz intervala [0,1] u prirodan broj m pomoću formule:

$$m = [N \cdot Y] + 1,$$

gde [·] označava ceo deo broja.

Pored prostog slučajnog uzorka postoje i druge vrste uzoraka, koje se takođe dobijaju po principu slučajnosti, a neki od njih su:
**stratifikovani, grupni, dvoetapni, periodični uzorci**.

Svaki od njih biće razmotren detaljnije.

**Stratifikovani uzorak**

U mnogim situacijama se izučavana populacija može podeliti na podgrupe koje se mogu posebno izučavati. Svaka takva podela populacije na disjunktne podgrupe naziva se stratifikacija ili raslojavanje a dobijene podgrupe stratumi ili slojevi.

Kako je osnovni cilj statistike ocenjivanje raspodele posmatranog obeležja populacije ili parametara te raspodele, cilj stratifikacije je da se postigne veća tačnost ocene, ekonomičnost ili jednostavnost ispitivanja. U nekim situacijama je ispitivanje moguće sprovoditi jedino po stratumima.

Neka je populacija obima N podeljena na L disjunktnih stratuma obima $N_j$, j = 1,2,...,L, pri čemu je:

$$N_1 + N_2 + \ldots + N_L = N \quad \text{i neka je} \quad W_j = N_j / N$$

„udeo“ j-tog stratuma u veličini populacije. Ako je populacija beskonačna tada $W_j$ predstavlja verovatnoću da se pri slučajnom izboru elementa populacije dobije element j-tog stratuma. Tada važi:

$$W_1 + W_2 + \ldots + W_L = 1.$$

Iz populacije se izvlači uzorak obima n tako što se iz j-tog stratuma izvlači deo uzorka

$$n_j \leq N_j \quad \text{tako da je} \quad n_1 + n_2 + ... + n_L = n.$$

Izbor elemenata po stratumima može biti sa ili bez vraćanja i izbori elemenata iz različitih stratuma su međusobno nezavisni.

Postoje dva osnovna pristupa izbora elemenata po stratumima:

- ravnomerni, kad je obim uzorka $n_j$ iz svakog stratuma isti i jednak:

$$n_j = n / L, \quad j = 1,2,...,L,$$

- proporcionalni, kad je obim uzorka $n_j$ iz svakog stratuma proporcionalan obimu stratuma. Tada je:

$$n_j = n \cdot W_j, \quad j = 1,2,...,L.$$

**Grupni uzorak**

Neka je razmatrana populacija podeljena po nekom principu na više disjunktnih grupa. Za stratifikovani uzorak smo iz svake grupe birali određen broj elemenata. Umesto toga od svih grupa u populaciji možemo da odaberemo određen broj grupa i kao uzorak razmotrimo sve elemente iz odabranih grupa. Takav uzorak se naziva **grupni uzorak**.

Stratifikovani uzorak daje bolju sliku o populaciji nego grupni ali se grupni uzorak jednostavnije i brže dobija. Grupni uzorak je po strukturi jednostavan ali kada je obim grupa veliki može biti nepraktičan ili davati manju tačnost.

**Dvoetapni uzorak**

Dvoetapni uzorak predstavlja kombinaciju grupnog i stratifikovanog uzorka. U prvoj etapi od svih grupa na koje je populacija podeljena biramo određen broj grupa a zatim u drugoj etapi iz svake grupe izabrane u prvoj etapi biramo određen broj elemenata.

**Periodični uzorak**

Pretpostavimo da su elementi populacije poređani u određenom redosledu (na primer proizvodi na traci...). Periodični uzorak formiramo tako što elemente u uzorak izdvajamo periodično (u jednakim razmacima), na primer svaki peti. Preciznije, ako je interval (period ili korak) izbora K, od prvih K elemenata biramo slučajnim izborom jedan kao početni a zatim svaki K-ti. Korak izbora zavisi od obima populacije N i od obima uzorka n tako da je $K = N/n$.

Mogućnost da se na osnovu uzorka $(X_1,X_2,...,X_n)$ odredi raspodela F posmatranog obeležja X daje Centralna teorema matematičke statistike. Pre njenog navođenja biće uveden pojam empirijske funkcije raspodele.

**Definicija 1.1.1. (Empirijska funkcija raspodele)**
Neka je $(X_1,X_2,...,X_n)$ prost slučajan uzorak obima n za posmatrano obeležje. Funkcija:

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} I\{X_j \le x\},$$

gde $I\{X_j \le x\}$ predstavlja indikator događaja $\{X_j \le x\}$, jeste empirijska funkcija raspodele.

Označimo sa $n_x$ broj elemenata uzorka za koje je vrednost obeležja X manja 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 $\{X \le x\}$. To je jedna stepenasta funkcija koja uzima vrednosti iz intervala [0,1] , neopadajuća za svako $x$ i neprekidna sa desne strane.

Iz Bernulijevog zakona velikih brojeva sledi da empirijska funkcija raspodele konvergira u verovatnoći ka funkciji raspodele obeležja:

$$F_n(x) \xrightarrow{P} F(x), \qquad n \to \infty$$

Centralna teorema matematičke statistike tvrdi da je ta konvergencija uniformna po $x$.

**Teorema 1.1.1. (Centralna teorema matematičke statistike)** Ako je $F(x)$ teorijska funkcija raspodele obeležja X, a $F_n(x)$ empirijska funkcija raspodele dobijena na osnovu prostog slučajnog uzorka obima n, tada, uniformno po $x$, funkcija $F_n(x)$ teži ka $F(x)$ sa verovatnoćom 1, tj.

$$\mathrm{P}\{\sup_{x\in R}|F_n(x) - F(x)| \to 0,\ n \to \infty\} = 1.$$

Ova teorema (poznata i pod nazivom Glivenko-Kantelijeva teorema) opravdava aproksimaciju funkcije raspodele $F(x)$ obeležja X njenom empirijskom funkcijom raspodele, a samim tim i prenošenje zaključaka koji se donose o raspodeli obeležja na uzorku na celu populaciju, pod uslovom da je obim uzorka dovoljno veliki.

**Primer 1.1.1.** Na primeru kvartalnog strukturnog istraživanja o poslovanju privrednih društava (SBS03) ilustrovaćemo najčešći način izbora uzorka ekonomskih jedinica u Republičkom zavodu za statistiku. Uzorak se odnosi na 2012. godinu.

Cilj istraživanja SBS03 je da se dobiju podaci o kvartalnoj dinamici rezultata finansijskog poslovanja privrednih društava kao i o promenama u strukturi ekonomskih aktivnosti u oblasti nefinansijske poslovne ekonomije.

Izveštajne jedinice su sva ekonomski aktivna privredna društva koja posluju u oblasti nefinansijske poslovne ekonomije kao i druga pravna lica koja proizvode i pružaju usluge za tržište tj. ako su im prihodi od prodaje proizvoda i usluga veći od 50% poslovnih prihoda.

Osnovna obeležja za koja se prikupljaju podaci su:

**Poslovni prihodi** - prihodi od prodaje roba, proizvoda i usluga, prihodi od premija, subvencija itd.

**Poslovni rashodi** - nabavna vrednost prodate robe, troškovi utrošenog materijala, troškovi zarada i naknada zarada i ostali lični rashodi, troškovi proizvodnih usluga, nematerijalni troškovi osim troškova poreza i nematerijalnih troškova

**Zalihe** materijala, nedovršene proizvodnje, gotovih proizvoda, robe

**Broj zaposlenih** - kvartalni prosek

**Investicije u materijalna osnovna sredstva** (nova i polovna) - gradjevinski radovi, oprema s montažom (domaća i uvozna), dugogodišnji zasadi, zemljište

Istraživanje se sprovodi na uzorku izabranih jedinica. Okvir za izbor uzorka formira se na osnovu skupa poslovnih subjekata Statističkog poslovnog registra koji ispunjavaju sledeće uslove:

- imaju pretežnu delatnost iz svih sektora KD(2010)[1] osim iz sektora Finansijske delatnosti i delatnost osiguranja i sektora Državne uprave i obaveznog socijalnog osiguranja.
- nisu na teritoriji AP Kosovo i Metohija
- odgovarajuća pravna lica (privredna društva i druga pravna lica) predala su završne račune za 2010. godinu

[1] Klasifikacija delatnosti prema Uredbi o klasifikaciji delatnosti iz 2010.godine („Službeni glasnik RS“, broj 54/09)

- odgovarajuća pravna lica nisu ugašena niti su u likvidaciji, prema podacima koje vodi Agencija za privredne registre

- poslovni subjekti čija su odgovarajuća pravna lica po formi druga pravna lica zadržavaju se u okviru za izbor uzorka samo ako im prihod od prodaje čini 50% i više od ukupnog poslovnog prihoda. Uz ovaj uslov treba da važi da je prihod od prodaje veći od 100 miliona dinara, ili da je broj zaposlenih veći od 50.

Od ovako izdvojenog skupa poslovnih subjekata konačni okvir za izbor uzorka formira se tako što se isključuje najmanje jedinice prema prometu iz svake oblasti delatnosti. Na taj način je okvir za izbor uzorka sveden sa 80 hiljada na 28 hiljada jedinica.

Zatim se vrši stratifikacija jedinica okvira za izbor uzorka prema delatnosti, broju zaposlenih (manje od 50 zaposlenih, 50 i više zaposlenih) i prema veličini prometa (oni sa većim prometom koji se popisuju i oni sa manjim prometom od kojih se bira slučajan uzorak).

Po preporuci Radne grupe za rad na integrisanoj bazi regionalnih podataka (REGIO), Naftna industrija Srbije (NIS) je podeljena na delove različite ekonomske aktivnosti. Delovi NIS–a su razvrstani u posebne popisne stratume, prema oblasti delatnosti i broju zaposlenih.

Definisano je 270 stratuma, od kojih je 136 popisnih (koji obavezno ulaze u uzorak) kojima je u fazi pripreme uzorka dodato još 8 stratuma koji sadrže delove NIS-a.

Uzorak je alociran (tačno je određen broj jedinica koje ulaze u uzorak po stratumima) prema *Bethel* [2] algoritmu sa unapred zadanim planiranim greškama za ocene totala za promet i broj zaposlenih po oblastima i sektorima delatnosti. Alociran je uzorak obima 2874 jedinice pri čemu je 1485 popisnih.

Unutar stratuma biran je prost slučajan uzorak sekvencijalnom šemom izbora uz pomoć permanentnih slučajnih brojeva iz intervala (0,1) koji su pridruženi jedinicama okvira za izbor uzorka. Slučajni početak za izbor uzorka bio je 0.25

## 1.2. O MOGUĆNOSTIMA KORIŠĆENJA JEDNOG UZORKA ZA DOBIJANJE NOVIH UZORAKA

U matematičkoj statistici se od prostog slučajnog uzorka $(X_1,X_2,...,X_n)$ formiraju slučajne veličine $Y = f(X_1,X_2,...,X_n)$ koje se nazivaju **statistike**. Funkcija f treba da bude merljiva u odnosu na $\sigma$ – algebru generisanu slučajnim veličinama $X_1,X_2,...,X_n$ da bi se mogle računati verovatnoće i određivati raspodela za Y. Nepoznati parametri raspodele posmatranog obeležja se ocenjuju na osnovu uzorka kao realizovane vrednosti pogodno odabranih statistika.

Na osnovu prostog slučajnog uzorka može se odrediti i interval koji sa unapred zadatom verovatnoćom sadrži nepoznati parametar. Takav interval naziva se interval poverenja i on predstavlja intervalnu ocenu nepoznatog parametra raspodele.

[2] Bethel, J."Sample allocation in multivariate surveys". *Survey methodology*, 15 (1989): 47-57

Kako su vrednosti posmatrane statistike za različite uzorke izvučene iz iste populacije različiti, možemo definisati raspodelu verovatnoće dobijanja određene vrednosti statistika za uzorak izvučen iz date populacije. Ova raspodela se naziva uzoračka raspodela i njena standardna devijacija predstavlja standardnu grešku date statistike.

Primenom statističkih metoda, na osnovu standardne greške, moguće je odrediti intervale poverenja za nepoznati parametar. Pri tome su naši zaključci bazirani na određenim polaznim pretpostavkama (na primer da je raspodela obeležja na populaciji normalna). Međutim, kada neka od pretpostavki nije ispunjena ili sumnjamo u njenu ispunjenost bolje je upotrebiti neku drugu metodu za procenu parametra u raspodeli obeležja.

Jednu mogućnost pružaju **resampling metode.**

Ove metode tretiraju realizovani uzorak $(x_1, x_2, \dots, x_n)$ kao populaciju iz koje se na određen način generišu novi uzorci, što se naziva reuzorkovanje. Uzorci dobijeni reuzorkovanjem treba da simuliraju višestruko dobijanje uzoraka (tzv. uzorkovanje) iz osnovnog skupa.

Jedina pretpostavka koja se pravi prilikom korišćenja resampling metoda je da uzorak u razumnoj meri predstavlja populaciju iz koje je uzet tj. da je uzorak reprezentativan. To je veoma važno jer se podaci dobijeni na uzorku multiplikuju, pa ako se polazni uzorak po svojim karakteristikama značajno razlikuje od populacije to može dovesti do pogrešnih zaključaka o populaciji.

Postoje četiri osnovna tipa **resampling metoda**, a to su:

- **Permutacioni testovi** (R.A. Fisher, 1935)

Permutacioni testovi se koriste za testiranje hipoteza i sprovode se na sledeći način:

Od raspoloživog uzorka $(x_1, x_2, \dots, x_n)$ se kreiraju novi uzorci permutovanjem njegovih elemenata. Na taj način dobijamo $n!$ permutacija, odnosno novih uzoraka. Za svaki od njih računamo test statistiku. Pod uslovom da je nulta hipoteza tačna test statistika ima istu raspodelu verovatnoća za svih $n!$ permutacija.

Označimo sa T test statistiku izračunatu iz polaznog uzorka a sa $T_i$ , i = 1,2,..., $n!$ test statistike izračunate za nove uzorke.

Na primer, u slučaju levostranog testa, $p$-vrednost računamo na sledeći način:

$$p = \frac{\sum_{i=1}^{n!} I\{T_i \leq \mathrm{T}\}}{n!}.$$

Test je statistički značajan ako je $p < \alpha$, gde je $\alpha$ unapred zadat nivo značajnosti.

Ako je obim uzorka (n) veliki umesto korišćenja $n!$ novih uzoraka, radi smanjenja obima računanja, možemo formirati slučajan uzorak od R permutacija i $p$ vrednost izračunati na sledeći način:

$$p = \frac{\sum_{i=1}^{R} I\{T_i \leq \mathrm{T}\}}{R}.$$

- **Cross validation** (A.K. Kurtz, 1948)

Cross-validation (unakrsna validacija) se koristi za ocenu tačnosti modela.

Jedan od tipova unakrsne validacije je *K cross-validation*:

Raspoloživi uzorak se podeli na k međusobno različitih particija iste veličine. Model generišemo koristeći (k-1) particiju a na preostaloj jednoj particiji se model testira. Postupak se ponavlja k puta tako da je svaka particija po jednom u ulozi particije na kojoj se model testira.

Greška unakrsne validacije se računa po formuli:

$$\mathrm{E} = \frac{1}{k}\sum_{i=1}^{k} E_i \,,$$

gde su $E_i$ greške ocenjene koristeći elemente particija na kojima se model testira.

Ukoliko je broj particija k jednak obimu raspoloživog uzorka radi se o *Leaving-one-out cross validation*:

Model se generiše korišćenjem (n-1) elemenata uzorka $(x_1, x_2, \dots, x_n)$ a testira se na jednom preostalom elementu. Postupak se ponavlja n puta tako da se svaki element originalnog uzorka jednom koristi kao element za testiranje. Greška unakrsne validacije se dobija po formuli:

$$\mathrm{E} = \frac{1}{n}\sum_{i=1}^{n} E_i \,,$$

gde su $E_i$ greške ocenjene koristeći elemente za testiranje.

- **Jackknife** (M. Quenouille, 1949)

Jackknife metoda se koristi za računanje ocene pristrasnosti i standardne greške ocene nepoznatog parametra raspodele.

Raspoloživi uzorak se podeli na određen broj grupa čija veličina može varirati od polovine uzorka do jednog elementa uzorka. U praksi je najčešći slučaj da se uzorak deli na onoliko grupa koliki je obim raspoloživog uzorka, tako da je veličina grupe jednaka jedinici.

Novi uzorci se prave tako što sve jedinice osim jedne grupe ulaze u uzorak tj. izbacivanjem po jedne grupe. U slučaju kad je veličina grupa jednaka jedinici novi uzorci se prave izbacivanjem po jednog elementa iz uzorka. Na taj način dobićemo onoliko novih uzoraka koliko ima elemenata u originalnom uzorku, pri čemu je njihov obim za jedan manji od obima polaznog uzorka.

- **Bootstrap** (B. Efron, 1979)

U nastavku rada biće detaljno objašnjena Bootstrap metoda.

# 2. BOOTSTRAP METODA

Tvorac ove metode je američki statističar Bradley Efron. Efron je 1979. godine predložio Bootstrap metodu kao novu kompjutersku statističku tehniku i predstavio je u radu: „Bootstrap methods: Another look at the Jackknife“. Naziv Bootstrap potiče od engleskog izraza „ to pull oneself by one's bootstraps“, što, u prenesenom smislu, znači postići uspeh bez oslanjanja na pomoć spolja.

Bootstrap se može definisati kao metoda kojom se na osnovu raspoloživih podataka iz nekog uzorka kreira veliki broj novih uzoraka, istog obima kao i izvorni uzorak, slučajnim biranjem sa vraćanjem iz skupa raspoloživih podataka.

Ovo znači da svaki element ima jednaku verovatnoću da uđe u uzorak i da neki element može da se pojavi više puta a neki nijednom.

Osnovni cilj ove metode je procena parametara populacije. Za svaki od novih uzoraka računa se statistika koja nas interesuje, a zatim se formira njena uzoračka raspodela. Ova raspodela je empirijska i ne oslanja se ni na kakve polazne pretpostavke. Na osnovu njenih dobijenih vrednosti računaju se, na primer, intervali poverenja za parametar koji ocenjujemo, a može se, kako će kasnije biti reči, koristiti i za testiranje statističkih hipoteza.

Da bismo dobili ocenu uzoračke raspodele statistike $\hat{\theta}$ koristimo Monte Karlo metodu.

Monte Karlo metode se mogu okarakterisati kao numeričke metode za rešavanje matematičkih problema pomoću modeliranja slučajnih veličina i statističkog ocenjivanja karakteristika tih veličina.

U ovom slučaju Monte Karlo metoda se sastoji u slučajnom izvlačenju velikog broja (B) uzoraka obima n iz osnovnog uzorka (isto obima n) i računanja statistike $\hat{\theta}$ za svaki od tih uzoraka. Raspodela verovatnoća dobijenih vrednosti $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ je ocena uzoračke raspodele statistike $\hat{\theta}$.

Dakle, osnovni problem je oceniti uzoračku raspodelu i to se može rešiti sledećim postupkom:

Pretpostavimo da je **X** = ($X_1$,$X_2$,...,$X_n$) uzorak za obeležje X sa nepoznatom funkcijom raspodele F i da smo nepoznati parametar Θ = Θ(F) ocenili na osnovu uzorka **X** statistikom $\widehat{\Theta}$. Cilj nam je da odredimo uzoračku raspodelu statistike $\widehat{\Theta}$.

Ideja Bootstrap metode je da se nepoznata funkcija raspodele F oceni empirijskom i da se generiše B prostih slučajnih uzoraka iz empirijske funkcije raspodele:

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} I\{X_j \le x\}, \quad x \in \mathrm{R}$$

formirane na osnovu raspoloživih podataka, tj. realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$.

To znači da se B Bootstrap uzoraka dobijaju putem izvlačenja sa vraćanjem iz originalnog skupa $\{x_1, x_2, \dots, x_n\}$ tako da svaka od vrednosti $x_i$ pri svakom izvlačenju ima jednaku verovatnoću (1/n) da bude izabrana.

Drugim rečima, ako sa $X^* = (X_1^*, X_2^*, \dots, X_n^*)$ obeležimo Bootstrap prost slučajan uzorak, slučajne promenljive $X_i^*$ , i = 1,2,...,n su međusobno nezavisne i raspodeljene po pravilu:

$$X_i^*: \begin{pmatrix} x_1 & \dots & x_n \\ 1/n & \dots & 1/n \end{pmatrix}.$$

Pri tome važi: $E(X_i^*) = \bar{x}$

$$\mathrm{Var}(X_i^*) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}.$$

Za svaki Bootstrap uzorak:

$$x_1^* = (x_{11}^*, x_{21}^*, \dots, x_{n1}^*)$$
...

$$x_B^* = (x_{1B}^*, x_{2B}^*, \dots, x_{nB}^*),$$

gde su $x_j^*$ realizacije uzorka $X^*$,

računamo odgovarajuću Bootstrap ocenu $\widehat{\Theta}_j^*$, j = 1,2,...,B

$\hat{\Theta}_1^* = \hat{\Theta}(x_1^*)$

....

$\hat{\Theta}_B^* = \hat{\Theta}(x_B^*)$.

Zatim konstruišemo histogram relativnih frekvencija od dobijenih B Bootstrap ocena $\hat{\Theta}_1^*, \hat{\Theta}_2^*, \dots, \hat{\Theta}_B^*$ dodeljujući verovatnoću 1/B svakoj od tih vrednosti.

Raspodela koju smo dobili je Bootstrap ocena uzoračke raspodele statistike $\hat{\Theta}$. Ona se sad može upotrebiti za donošenje zaključaka o parametru Θ.

Opisana metoda predstavlja **neparametarsku** Bootstrap metodu.

Pored nje postoji i **parametarska** Bootstrap metoda koja se primenjuje kada je poznat oblik funkcije raspodele $F_\Theta$ iz koje je uzet posmatrani uzorak $x = (x_1, x_2, \dots, x_n)$ ali je nepoznat parametar raspodele – Θ. Ideja parametarskog Bootstrap - a je da se generiše B nezavisnih uzoraka iz

$$F_{\hat{\Theta}(x)},$$

gde $\hat{\Theta}(x)$ predstavlja ocenjenu vrednost parametra Θ na osnovu datog uzorka x. Dalje se primenjuje isti postupak kao kod neparametarske Bootstrap metode.

Pretpostavimo da imamo uzorak $\mathbf{X} = (X_1,\dots,X_n)$ za obeležje X i da je nepoznati parametar Θ ocenjen na osnovu uzorka **X** statistikom $\hat{\Theta}$. U najvećem broju slučajeva uzoračka raspodela statistike $\hat{\Theta}$ (za $n \geq 30$) ima oblik normalne raspodele sa očekivanjem Θ i standardnom devijacijom $a/\sqrt{n}$, gde je a pozitivan broj koji zavisi od tipa statistike $\hat{\Theta}$ (Centralna granična teorema).

Neka je $\hat{\Theta}^*$ ista statistika kao $\hat{\Theta}$ samo dobijena Bootstrap metodom.

Kad $n \to \infty$ uzoračka raspodela statistike $\hat{\Theta}^*$ je takođe slična normalnoj raspodeli sa očekivanjem $\hat{\Theta}$ i standardnom devijacijom $a/\sqrt{n}$.

Stoga Bootstrap raspodela za $\hat{\Theta}^* - \hat{\Theta}$ aproksimira uzoračku raspodelu za $\hat{\Theta} - \Theta$. (prema [5])

**Primer 2.1.** Razmotrimo uzorak koji se sastoji od 200 elemenata generisanih iz N(0,1) raspodele. To je originalni uzorak. U ovom slučaju uzoračka raspodela aritmetičke sredine je aproksimativno normalna sa srednjom vrednošću 0 i standardnom devijacijom $1/\sqrt{200}$.

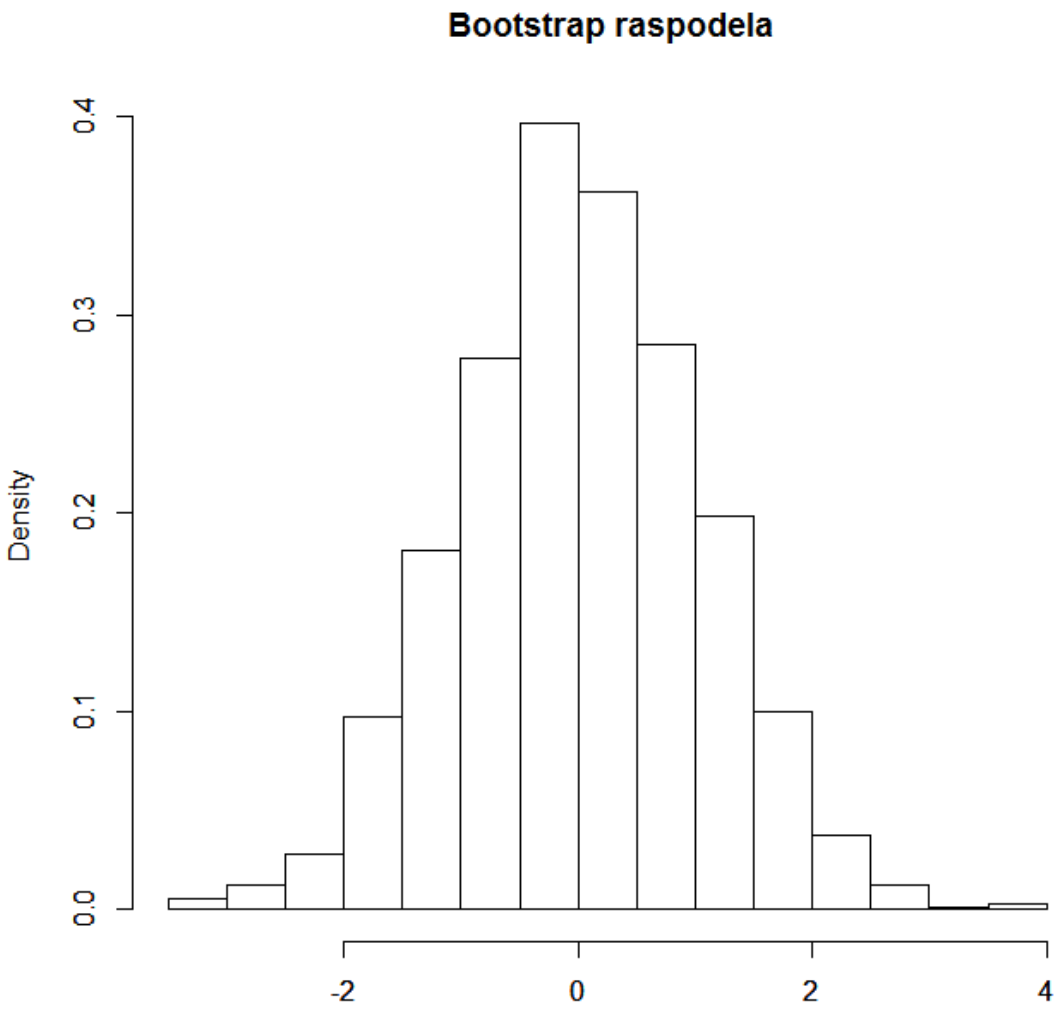
Primenićemo Bootstrap metodu da uporedimo rezultate. Izvučeno je 1500 uzoraka iz originalnog uzorka i izračunata je aritmetička sredina za svaki novi uzorak.

Program koji postoji u programskom jeziku R:

Korak 1. Generišemo uzorak od 200 elemenata iz standardne normalne populacije
gauss < −rnorm(200, 0, 1)

Korak 2. Formiramo 1500 bootstrap uzoraka pomoću neparametarskog Bootstrap-a
bootmean < −1 : 1500
for(i in 1 : 1500) bootmean[i] < −mean(sample(gauss, replace = T))

Korak 3. Crtanje grafika
bootdistribution < −sqrt(200) * (bootmean − mean(gauss))
hist(bootdistribution, freq = FALSE,main = ”Bootstrap raspodela”, xlab = ””)




Dobijen je grafik uzoračke raspodele aritmetičke sredine koji ima približno normalnu raspodelu što pokazuje da Bootstrap metoda daje rezultate bliske rezultatima koje daju klasične statističke metode.

Koliko je dobra ocena dobijena Bootstrap metodom Efron je ilustrovao u već pomenutom radu: „Bootstrap methods: Another look at the Jackknife“ na sledećim primerima:

**Primer 2.2.** Pretpostavimo da je **X** = ($X_1$,$X_2$,...,$X_n$) uzorak za obeležje X i da X$\epsilon${0,1}. Neka je F funkcija raspodele obeležja i $\Theta$(F) = $P_F${X = 1}.

Posmatramo slučajnu veličinu:

R(**X**,F) = $\hat{\Theta}$(**X**) - $\Theta$(F),

gde je $\hat{\Theta}$(**X**) ocena parametra $\Theta$ na osnovu uzorka **X**.

Pri tome važi:

R(**X**,F) = $\bar{X}$- $\Theta$(F),

gde je $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Na osnovu realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$, Bootstrap metodom dobijamo nove uzorke $X^* = (X_1^*, X_2^*, \dots, X_n^*)$, gde su $X_i^*$ raspodeljene po pravilu:

$$X_i^*: \begin{pmatrix} 1 & 0 \\ \bar{x} & 1-\bar{x} \end{pmatrix}.$$

Raspodelu slučajne veličine R(**X**,F) aproksimiraćemo Bootstrap raspodelom slučajne veličine:

$R^*$= R($X^*$, $\hat{F}$) = $\hat{\Theta}$($X^*$) - $\Theta$($\hat{F}$) = $\bar{X}^*$- $\bar{x}$,

gde je $\hat{F}$ empirijska funkcija raspodele.

Na osnovu svojstava binomne raspodele dobijamo da su očekivna vrednost i varijansa:

$E_*(\bar{X}^* - \bar{x}) = 0$

$Var_*(\bar{X}^* - \bar{x}) = \bar{x}(1-\bar{x})$ / n,

gde oznake $E_*$ i $Var_*$ označavaju izračunavanja koja se odnose na Bootstrap raspodelu za $X^*$ sa fiksnim x i $\hat{F}$.

Na osnovu dobijenih rezultata zaključujemo da je $\bar{X}$ nepristrasna ocena za $\Theta$ sa varijansom aproksimativno jednakom $\bar{x}(1-\bar{x})$ / n što je tačno.

**Primer 2.3.** Pretpostavimo da je **X** = ($X_1$,$X_2$,...,$X_n$) uzorak za obeležje X sa funkcijom raspodele F i da ocenjujemo parametar Θ(F) = $Var_F$X ocenom:

$\widehat{\Theta}(\mathbf{X}) = \sum_{i=1}^{n}(X_i - \bar{X})^2 \,/\, (n-1)$ .

Posmatramo slučajnu veličinu:

R(**X**,F) = $\widehat{\Theta}(\mathbf{X})$ - Θ(F).

Označimo sa $\mu_k$ k-ti centralni momenat:

$\mu_k(F) = E_F(X - E_F X)^k$ i $\hat{\mu}_k = \mu_k(\hat{F})$,

gde je $\hat{F}$ empirijska funkcija raspodele.

Bootstrap metodom, na osnovu realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$, dobijamo nove uzorke $X^* = (X_1^*, X_2^*, \dots, X_n^*)$.

Raspodelu slučajne veličine R(**X**,F) aproksimiraćemo Bootstrap raspodelom slučajne veličine:

$R^* = R(X^*, \hat{F}) = \widehat{\Theta}(X^*) - \Theta(\hat{F})$.

Na osnovu statističke teorije važi:

$E_*(R^*) = 0$

$Var_*(R^*) = \frac{\hat{\mu}_4 - ((n-3)/(n-1))\hat{\mu}_2{}^2}{n}$ .

Na osnovu toga zaključujemo da je $\widehat{\Theta}(\mathbf{X}) = \sum_{i=1}^{n}(X_i - \bar{X})^2 \,/\, (n-1)$ nepristrasna ocena za $Var_F$X što je tačno i da je:

$Var_F\widehat{\Theta}(\mathbf{X}) \approx Var_*(R^*)$,

što je ujedno i Jackknife ocena za $Var_F\widehat{\Theta}$.

Bootstrap metoda omogućava dobijanje rezultata vezanih za tačkaste i intervalne ocene parametara raspodele posmatranog obeležja, a može se koristiti i za testiranje statističkih hipoteza.

Svaka od ovih oblasti biće u nastavku rada posebno razmatrana.

## 2.1. BOOTSTRAP I TAČKASTE OCENE PARAMETARA

Neka je dat prost slučajan uzorak $\mathbf{X} = (X_1,X_2,...,X_n)$ obima n za posmatrano obeležje X. Ako se statistikom $Y = f(X_1,X_2,...,X_n)$ ocenjuje parametar $\Theta$, tada se statistika Y naziva ocena parametra $\Theta$ i označava se sa $\widehat{\Theta}$. Pri tome je poželjno da ona ima određena svojstva kao što su nepristrasnost, postojanost i efikasnost. Ovakve ocene parametara se nazivaju tačkaste ocene.

Jedne od mera tačnosti ocene su njena standardna greška i pristrasnost (Bias) i Bootstrap metoda se može primeniti u njihovom ocenjivanju.

### 2.1.1. BOOTSTRAP OCENA STANDARDNE GREŠKE

Pretpostavimo da je $\mathbf{X} = (X_1,X_2,...,X_n)$ uzorak za obeležje X sa nepoznatom funkcijom raspodele F i da smo nepoznati parametar $\Theta = \Theta(F)$ ocenili na osnovu uzorka **X** ocenom $\widehat{\Theta}$.

Bootstrap procedura za ocenu standardne greške je primenljiva bez obzira koliko je komplikovana ocena $\widehat{\Theta}$ i sastoji se od sledećih koraka:

1. Generišemo B nezavisnih Bootstrap uzoraka $x_1^*, x_2^*, \dots, x_B^*$ iz empirijske funkcije raspodele $F_n$ formirane na osnovu realizovanog uzorka $x = (x_1, x_2, \dots, x_n)$
2. Za svaki Bootstrap uzorak izračunamo Bootstrap ocenu $\widehat{\Theta}_j^* = \widehat{\Theta}(x_j^*)$, j = 1,2,...,B
3. Ocenimo standardnu grešku $se_F(\widehat{\Theta})$ sa:

$$\widehat{se}_B = \left\{\sum_{j=1}^{B}\left[\widehat{\Theta}_j^* - \widehat{\Theta}^*(.)\right]^2/(B-1)\right\}^{1/2},$$

gde je $\widehat{\Theta}^*(.) = \sum_{j=1}^{B}\widehat{\Theta}_j^*/B$.

Za računanje standardne greške dovoljno je 25-200 Bootstrap uzoraka.

### 2.1.2. BOOTSTRAP OCENA ZA PRISTRASNOST

Još jedna mera tačnosti ocene, pored standardne greške, je pristrasnost (Bias).

Pretpostavimo da je **X** = $(X_1,X_2,...,X_n)$ uzorak za obeležje X sa nepoznatom funkcijom raspodele F i da smo nepoznati parametar $\Theta = \Theta(F)$ ocenili na osnovu uzorka **X** ocenom $\hat{\theta}$.

Pristrasnost ocene $\hat{\theta}$ je razlika između očekivane vrednosti ocene $\hat{\theta}$ i parametra $\Theta$:

$$Bias_F(\hat{\theta}, \Theta) = E_F(\hat{\theta}) - \Theta.$$

Bootstrap metodom možemo oceniti pristrasnost tako što ćemo:

1. Generisati B Bootstrap uzoraka $x_1^*, x_2^*, \dots, x_B^*$ iz empirijske funkcije raspodele $F_n$ formirane na osnovu realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$
2. Za svaki Bootstrap uzorak izračunamo Bootstrap ocenu $\hat{\theta}_j^* = \hat{\theta}(x_j^*)$, j = 1,2,...,B
3. Matematičko očekivanje ćemo oceniti sa:

$$\hat{\theta}^*(.) = \sum_{j=1}^{B} \hat{\theta}_j^* / B.$$

4. Pristrasnost (Bias) ocenjujemo sa:

$$\widehat{Bias}_B = \hat{\theta}^*(.) - \hat{\theta}. \qquad (1)$$

Uobičajeni razlog zašto ocenjujemo pristrasnost jeste da bismo mogli da korigujemo ocenu da bude manje pristrasna. Tako korigovana ocena je oblika:

$$\overline{\theta} = \hat{\theta} - \widehat{Bias}_B.$$

Zamenjujući $\widehat{Bias}_B$ formulom (1) dobijamo:

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

Korigovanje ocene sa $\widehat{Bias}_B$ može, međutim, uzrokovati povećanje standardne greške ocene.

# 2.2. BOOTSTRAP I INTERVALI POVERENJA

## 2.2.1. INTERVALI POVERENJA

Realizovana vrednost tačkaste ocene parametra može dosta odstupati od stvarne vrednosti parametra. Na osnovu prostog slučajnog uzorka može se odrediti interval koji sa unapred zadatom verovatnoćom sadrži nepoznati parametar.

**Definicija 2.2.1.1.** Neka je $(X_1,X_2,...,X_n)$ uzorak za obeležje X čija je raspodela $F(x, \Theta)$ i neka su $\hat{\theta}_1 = \hat{\theta}_1(X_1,X_2,...,X_n)$ i $\hat{\theta}_2 = \hat{\theta}_2(X_1,X_2,...,X_n)$ dve statistike koje zavise od nepoznatog parametra $\Theta$ takve da je:

$$P(\hat{\theta}_1 \leq \hat{\theta}_2) = 1,$$
$$P(\hat{\theta}_1 \leq \Theta \leq \hat{\theta}_2) = \beta = 1 - \alpha,$$

gde je $\beta$ unapred zadata verovatnoća. Tada se slučajni interval $[\hat{\theta}_1, \hat{\theta}_2]$ koji zavisi od uzorka $(X_1,X_2,...,X_n)$ zove interval poverenja za parametar $\Theta$, a verovatnoća $\beta$ nivo poverenja.

Vrednost nivoa poverenja $\beta$ je najčešće 0.9 ili 0.95 .

Bootstrap metoda se može primeniti i za određivanje intervala poverenja.

## 2.2.2. BOOTSTRAP INTERVAL POVERENJA

Pretpostavimo da je $\mathbf{X} = (X_1,X_2,...,X_n)$ uzorak za obeležje X i da smo nepoznati parametar $\Theta$ raspodele obeležja ocenili na osnovu uzorka **X** ocenom $\hat{\theta}$ .

Razmotrimo slučaj kad $\hat{\theta}$ ima $N(\Theta, \sigma_{\hat{\theta}}^2)$ raspodelu.

Tada je $(1- \alpha)100$ % interval poverenja za $\Theta$ jednak $(\hat{\theta}_L, \hat{\theta}_U)$ , gde je:

$$\hat{\theta}_L = \hat{\theta} - Z^{(1-\alpha/2)}\sigma_{\hat{\theta}} \quad \text{i} \quad \hat{\theta}_U = \hat{\theta} - Z^{(\alpha/2)}\sigma_{\hat{\theta}} , \qquad (2)$$

a sa $Z^{(\Upsilon)}$ je označen $\Upsilon$ 100 percentil od standardne normalne raspodele.

$Z^{(\Upsilon)} = F^{-1}(\Upsilon)$ , gde je F funkcija raspodele slučajne veličine koja ima N(0,1) raspodelu.

Pretpostavimo da smo na osnovu realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$ ocenili $\widehat{\Theta}$ i $\sigma_{\widehat{\Theta}}$ i izračunali $\widehat{\Theta}_L$ i $\widehat{\Theta}_U$.

Ako je $\widehat{\Theta}^*$ slučajna veličina sa N($\widehat{\Theta}, {\sigma_{\widehat{\Theta}}}^2$) raspodelom tada važi:

$$P(\widehat{\Theta}^* \leq \widehat{\Theta}_L) = P\left(\frac{\widehat{\Theta}^* - \widehat{\Theta}}{\sigma_{\widehat{\Theta}}} \leq -Z^{(1-\alpha/2)}\right) = \alpha/2$$

i $P(\widehat{\Theta}^* \leq \widehat{\Theta}_U) = 1 - \alpha/2,$

što znači da su $\widehat{\Theta}_L$ i $\widehat{\Theta}_U$: $(\alpha/2)100$ i $(1-\alpha/2)100$ percentili raspodele slučajne veličine $\widehat{\Theta}^*$ i oni formiraju $(1-\alpha)100$ % interval poverenja za Θ.

Međutim, često $\widehat{\Theta}$ nema normalnu raspodelu ali postoji transformacija ϕ = m(Θ) takva da raspodela od $\widehat{\phi}$ = m($\widehat{\Theta}$) bude normalna N(ϕ, ${\sigma_c}^2$). Na primer m(Θ) može biti logΘ.

Tada važi kao u (2), da

$$(\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c\ ,\ \widehat{\phi} - Z^{(\alpha/2)}\sigma_c)$$

predstavlja $(1-\alpha)100$ % interval poverenja za ϕ.

Pošto važi:

$$(1-\alpha) = P\left[\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c < \phi < \widehat{\phi} - Z^{(\alpha/2)}\sigma_c\right] =$$

$$= P\left[m^{-1}(\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c) < \Theta < m^{-1}(\widehat{\phi} - Z^{(\alpha/2)}\sigma_c)\right], \qquad (3)$$

znači da $(m^{-1}(\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c),\ m^{-1}(\widehat{\phi} - Z^{(\alpha/2)}\sigma_c))$ predstavlja $(1-\alpha)100$ % interval poverenja za Θ.

Označimo sa H funkciju raspodele ocene $\widehat{\Theta}$ i H zavisi od Θ. Neka je $\widehat{H}$ funkcija raspodele sa realizacijom $\widehat{\Theta}$ umesto Θ i neka je $\widehat{\Theta}^*$ slučajna veličina sa funkcijom raspodele $\widehat{H}$ a $\widehat{\phi}$ = m($\widehat{\Theta}$) i $\widehat{\phi}^*$ = m($\widehat{\Theta}^*$). Tada važi:

$$P\left[\widehat{\Theta}^* \leq m^{-1}(\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c)\right] = P\left[\widehat{\phi}^* \leq (\widehat{\phi} - Z^{(1-\alpha/2)}\sigma_c)\right] =$$

$$= P\left(\frac{\widehat{\phi}^* - \widehat{\phi}}{\sigma_c} \leq -Z^{(1-\alpha/2)}\right) = \alpha/2$$

tj. $m^{-1}(\hat{\Phi} - Z^{(1-\alpha/2)}\sigma_c)$ je $(\alpha/2)100$ percentil raspodele $\hat{H}$ i

$m^{-1}(\hat{\Phi} - Z^{(\alpha/2)}\sigma_c)$ je $(1-\alpha/2)100$ percentil raspodele $\hat{H}$.

Kada bismo znali $\hat{H}$ mogli bismo da odredimo i interval poverenja za Θ. Međutim mi obično ne znamo $\hat{H}$.

Ideja Bootstrap procedure je da se od polaznog uzorka formira veliki broj uzoraka $x_1^*, x_2^*, \dots, x_B^*$ putem izvlačenja sa vraćanjem i za svaki oceni $\hat{\Theta}_j^* = \hat{\Theta}(x_j^*)$, j = 1,2,...,B.

Zatim formiramo histogram relativnih frekvencija od ocena $\hat{\Theta}_j^*$, j = 1,2,...,B. Tada će $(\alpha/2)100$ i $(1-\alpha/2)100$ percentili od tog histograma formirati interval (3), tj. $(1-\alpha)100\,\%$ interval poverenja za parametar Θ, što se može predstaviti sledećim koracima:

1) Od polaznog uzorka x = $(x_1, x_2, \dots, x_n)$ formiramo B Bootstrap uzoraka $x_j^*$, j = 1,2,...,B putem izvlačenja sa vraćanjem

2) Za svaki od uzoraka izračunamo ocenu $\hat{\Theta}_j^* = \hat{\Theta}(x_j^*)$, j = 1,2,...,B

3) Sortiramo ih po veličini

$$\hat{\Theta}_{(1)}^* \le \hat{\Theta}_{(2)}^* \le \dots \le \hat{\Theta}_{(B)}^*$$

i označimo sa m veličinu $[(\alpha/2)B]$, gde je [.] ceo deo broja.

Tada interval $(\hat{\Theta}_{(m)}^*, \hat{\Theta}_{(B+1-m)}^*)$ predstavlja **Bootstrap interval poverenja** za parametar Θ.

**Primer 2.2.2.1.** Konstruisaćemo $100\beta$ % interval poverenja dobijen primenom klasičnih statističkih metoda na konkretnom uzorku i uporediti sa intervalom poverenja koji se dobija primenom Bootstrap metode.

Pretpostavimo da nam je dat slučajan uzorak **X** = ($X_1$,$X_2$,...,$X_n$) za obeležje X sa normalnom raspodelom: N($\mu$,$\sigma^2$), $\mu \in R$, $\sigma>0$.

Na osnovu uzorka **X** = ($X_1$,$X_2$,...,$X_n$) nepoznato matematičko očekivanje ocenjujemo uzoračkom sredinom $\bar{X}_n$ i pretpostavimo da je disperzija $\sigma^2$ nepoznata. Statističkim metodama interval poverenja za matematičko očekivanje $\mu$ obeležja X sa normalnom raspodelom i nepoznatom disperzijom određujemo polazeći od uslova:

$P\{ \bar{X}_n - \varepsilon \le \mu \le \bar{X}_n + \varepsilon \} = \beta,$

gde je β nivo poverenja.

Odnosno iz uslova $P\{|\bar{X}_n - \mu| \leq \mathcal{E}\} = \beta$.

Pošto statistika $\frac{\bar{X}_n - \mu}{\bar{S}_n}\sqrt{n-1}$ ima Studentovu raspodelu sa n-1 stepeni slobode, iz tablica za Studentovu raspodelu nalazimo vrednost $t_{n-1;\beta}$ tako da za dato β važi:

$$P\left\{\left|\frac{\bar{X}_n - \mu}{\bar{S}_n}\sqrt{n-1}\right| \leq t_{n-1;\beta}\right\} = \beta.$$

Traženi interval poverenja je oblika:

$$\left[\bar{X}_n - \frac{t_{n-1;1-\beta}}{\sqrt{n-1}}\,\bar{S}_n\,,\ \bar{X}_n + \frac{t_{n-1;1-\beta}}{\sqrt{n-1}}\,\bar{S}_n\right]. \qquad (4)$$

Na konkretnom primeru odredićemo tačne vrednosti intervala poverenja.

Pretpostavimo da nam je dat uzorak obima 10 iz $N(100,3^2)$ raspodele koji smo generisali, koristeći programski paket R, primenom naredbe:

```
x<-rnorm(10,100,3)
```

```
x = (102.13920  105.83821  100.86308  104.59968  97.46180
      99.14003  100.50945   99.17493   96.86652  95.73759)
```

A sada, pretpostavljajući da su nam nepoznati parametri $\mu$ i $\sigma^2$, na osnovu formule (4) odredićemo 95% interval poverenja za nepoznato matematičko očekivanje.

$$\left[\bar{X}_n - \frac{t_{n-1;1-\beta}}{\sqrt{n-1}}\,\bar{S}_n\,,\ \bar{X}_n + \frac{t_{n-1;1-\beta}}{\sqrt{n-1}}\,\bar{S}_n\right] =$$

$$= \left[100.23305 - \frac{2.262}{\sqrt{9}} \cdot 3.10047, 100.23305 + \frac{2.262}{\sqrt{9}} \cdot 3.10047\right] =$$

$$= [97.89529,\ 102.57080].$$

Pozivajući funkciju **percentciboot** (koja je data u Dodatku) naredbom:

Percentciboot(x,1000,0.05),

gde je 1000 broj Bootstrap uzoraka a 0.05 nivo značajnosti testa, odredićemo donju i gornju granicu 95% intervala poverenja dobijenog Bootstrap metodom. Dobijeni su sledeći rezultati:

Donja granica: 98.32317
Gornja granica: 102.0147

a primenom naredbe:

Percentciboot(x,3000,0.05)

dobijamo sledeće rezultate:

Donja granica: 98.38048
Gornja granica: 102.1796

Zaključujemo da postoji velika saglasnost između rezultata dobijenih klasičnim statističkim metodama i rezultata dobijenih Bootstrap metodom.

## 2.3. BOOTSTRAP I TESTIRANJE HIPOTEZA

### 2.3.1. O TESTIRANJU STATISTIČKIH HIPOTEZA

Postupkom testiranja hipoteza proveravamo neku teoriju ili uverenje o parametru osnovnog skupa (populacije) ili o raspodeli obeležja uopšte. Koristeći podatke iz uzorka, primenom statističkih metoda utvrđujemo da li se i sa kojom verovatnoćom može prihvatiti pretpostavka o konkretnoj brojčanoj vrednosti nekog parametra ili hipoteza o konkretnoj raspodeli obeležja.

Svaka pretpostavka tj. hipoteza o nepoznatom parametru raspodele obeležja se naziva parametarska hipoteza a postupak njenog potvrđivanja ili odbacivanja na osnovu podataka iz uzorka je parametarski test. Statistika koja se koristi u tom postupku se naziva test statistika.

Hipoteza koja se testira se zove nulta hipoteza i označava se sa $H_0$. Alternativna hipoteza je tvrđenje o nekom parametru koje će biti istinito ako je nulta hipoteza neistinita.

Statistička hipoteza je prosta ako je njom potpuno određena raspodela obeležja ($\theta = \theta_0$). U suprotnom je statistička hipoteza složena (na primer $\theta \in \{\theta_1, \theta_2, \theta_3\}$).

Prilikom testiranja hipoteza mogu se pojaviti dve greške: greška prve i druge vrste.

Greška prve vrste se javlja kada se istinita nulta hipoteza odbaci. Verovatnoća javljanja greške prve vrste se obično obeležava sa $\alpha$ i naziva se nivo značajnosti testa:

$\alpha$ = P( $H_0$ se odbacuje| $H_0$ je istinita).

Greška druge vrste se javlja kada se neistinita nulta hipoteza ne odbaci. Verovatnoća javljanja greške druge vrste se obično obeležava sa β , odnosno:

β = P( $H_0$ se ne odbacuje| $H_0$ je neistinita).

Vrednost (1 - β) se naziva jačina testa (moć testa) i ona predstavlja verovatnoću da se greška druge vrste ne javi tj. verovatnoću da se odbaci netačna nulta hipoteza.

Postoje dva postupka testiranja hipoteza:

- **pristup zasnovan na kritičnoj vrednosti**

U ovom pristupu nalazimo kritičnu vrednost (ili vrednosti) i izračunavamo realizovanu vrednost test statistike za realizovanu vrednost statistike uzorka i poredimo je sa kritičnom vrednošću.

Nulta hipoteza će biti proglašena neprihvatljivom ako registrovana vrednost test statistike pripadne kritičnoj oblasti.

Oblik kritične oblasti određuje alternativna hipoteza, a veličinu kritične oblasti i njene granice određuje nivo značajnosti $\alpha$.

U slučaju da testiramo hipotezu $H_{0:}(\Theta = \Theta_0)$ protiv alternativne $H_{1:}(\Theta \neq \Theta_0)$ kritična oblast je dvostrana i sačinjena je od unije intervala $(-\infty,f)$ i $(g,+\infty)$ .

U slučaju da je alternativna hipoteza $H_{1:}(\Theta < \Theta_0)$ ili $H_{1:}(\Theta > \Theta_0)$ kritična oblast je jednostrana i predstavlja intervale $(-\infty,f)$ i $(g,+\infty)$ respektivno.

Pri tome se statistike f i g određuju u funkciji statistike kojom se ocenjuje parametar Θ.

- **pristup zasnovan na *p* vrednosti**

Ovaj pristup se obično primenjuje u statističkim paketima. Sastoji se u izračunavanju tzv. *p* vrednosti za realizovanu vrednost statistike uzorka.

**Definicija 2.3.1.1.** Pod pretpostavkom da je nulta hipoteza istinita, *p* vrednost može da se definiše kao verovatnoća da statistika uzorka odstupa od hipotetičke vrednosti parametra u smeru alternativne hipoteze, barem toliko koliko i realizovana vrednost statistike uzorka u izabranom uzorku.

Na primer, u slučaju desnostranog testa, ako smo prethodno odredili nivo značajnosti $\alpha$, nultu hipotezu odbacujemo ako je:

$p$ vrednost $< \alpha$,

a ne odbacujemo ako je:

$p$ vrednost $\geq \alpha$.

Bootstrap metoda se može koristiti i za testiranje statističkih hipoteza.
Biće razmotrena dva slučaja (sa dva i sa jednim uzorkom).

### 2.3.2. SLUČAJ SA DVA UZORKA

Pretpostavimo da je $\mathbf{X} = (X_1, X_2, ..., X_n)$ uzorak za obeležje X sa funkcijom raspodele $F(x)$ i $\mathbf{Y} = (Y_1, Y_2, ..., Y_m)$ uzorak za obeležje Y sa funkcijom raspodele $F(x - \Delta)$ gde je $\Delta \in R$. Ako sa $\mu_X$ i $\mu_Y$ označimo matematička očekivanja obeležja tada važi $\Delta = \mu_Y - \mu_X$ .

Hipoteza koju hoćemo da testiramo je:

$H_0: \Delta = 0$ protiv alternativne $H_1: \Delta > 0$.

Za test statistiku uzećemo:

$$V = \bar{Y} - \bar{X},$$

gde su $\bar{X}$ i $\bar{Y}$ aritmetičke sredine uzoraka **X** i **Y**.

Naša odluka o prihvatanju ili odbacivanju nulte hipoteze zasnovana je na *p* vrednosti koju ćemo oceniti sa:

$$\hat{p} = P_{H_0}[V \geq \bar{y} - \bar{x}], \qquad (5)$$

gde su $\bar{x}$ i $\bar{y}$ realizovane vrednosti aritmetičkih sredina uzoraka **X** i **Y**.

Bootstrap metodom određujemo raspodelu statistike V pod pretpostavkom da je hipoteza $H_0: \Delta = 0$ tačna.

Jednostavan način da to uradimo je da spojimo uzorke **X** i **Y** u jedan veliki uzorak i onda putem izvlačenja sa vraćanjem kreiramo nove uzorke obima n i m iz empirijske funkcije raspodele formirane na osnovu uzorka $z = (x_1, x_2, \dots x_n, y_1, y_2, \dots y_m)$.

Svaka od tih vrednosti ima istu verovatnoću (1 / (n+m)) da bude izvučena. Time je zadovoljen uslov da su uzorci iz iste raspodele.

Postupak se može opisati sledećim koracima:

1. Od realizovanih uzoraka $x = (x_1, x_2, \dots, x_n)$ i $y = (y_1, y_2, \dots, y_m)$ formiramo jedan $z = (x_1, x_2, \dots x_n, y_1, y_2, \dots y_m)$
2. Putem izvlačenja sa vraćanjem iz skupa $\{x_1, x_2, \dots x_n, y_1, y_2, \dots y_m\}$ formiramo uzorak obima n , $x^* = (x_1^*, x_2^*, \dots, x_n^*)$ i izračunamo:

$$\bar{x}_j^*$$

3. Putem izvlačenja sa vraćanjem iz skupa $\{x_1, x_2, \dots x_n, y_1, y_2, \dots y_m\}$ formiramo uzorak obima m , $y^* = (y_1{}^*, y_2{}^*, \dots, y_m{}^*)$ i izračunamo:

$$\bar{y}_j{}^*$$

4. Izračunamo vrednost:

$$v_j{}^* = \bar{y}_j{}^* - \bar{x}_j{}^*$$

5. Ponavljamo korake 2-4 B puta i dobijemo vrednosti $v_1{}^*, v_2{}^*, \dots, v_B{}^*$

Histogram relativnih frekvencija kreiran na osnovu vrednosti $v_1{}^*, v_2{}^*, \dots, v_B{}^*$ dodeljujući svakoj vrednosti verovatnoću 1/B predstavlja raspodelu statistike V pod uslovom da je nulta hipoteza tačna i na osnovu te raspodele i formule (5) ocenimo *p* vrednost sa:

$$\hat{p}^* = \frac{\sum_{j=1}^{B} I\{v_j{}^* \geq \bar{y} - \bar{x}\}}{B}.$$

Dobijenu ocenu poredimo sa nivoom značajnosti $\alpha$ , pa na osnovu toga donosimo odluku o prihvatanju ili odbacivanju nulte hipoteze.

**Primer 2.3.2.1.** Testiraćemo hipotezu o jednakosti matematičkih očekivanja za dva nezavisna obeležja sa normalnom raspodelom i poznatim i jednakom disperzijama primenom klasičnih statističkih metoda i primenom Bootstrap metode.

Pretpostavimo da je dat uzorak **X** = ($X_1$,$X_2$,...,$X_n$) obima n za obeležje X sa N($\mu_1$, $\sigma^2$) i uzorak **Y** = ($Y_1$,$Y_2$,...,$Y_m$) obima m za obeležje Y sa N($\mu_2$, $\sigma^2$) raspodelom.

Hipoteza koju hoćemo da testiramo je:

$H_0$: $\mu_1 = \mu_2$ protiv alternativne $H_1$: $\mu_2 > \mu_1$.

Pod uslovom da je nulta hipoteza tačna statistika:

$Z = \frac{\bar{Y}_m - \bar{X}_n}{S}$ ima N(0,1) raspodelu,

gde je $S = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}$.

*p* vrednost nalazimo iz uslova: $P\left\{\frac{\bar{Y}_m-\bar{X}_n}{s} \geq \frac{\bar{y}_m-\bar{x}_n}{s}\right\} = p$

Na konkretnom primeru odredićemo *p* vrednosti.

Pretpostavimo da nam je dat uzorak obima 10 iz $N(90,6^2)$ raspodele koji smo generisali, koristeći statistički paket R, primenom naredbe:

x<-rnorm(10,90,6)

x = (91.09427 98.37224 85.12329 94.83729 87.07225
86.36860 88.10942 94.25709 77.90737 88.28919)

i uzorak obima 15 iz $N(95,6^2)$ raspodele koji smo generisali primenom naredbe:

y<-rnorm(15,95,6)

y = (99.00327 97.85534 99.64330 87.88428 85.94843 89.78400 85.88152
91.63612 91.57083 100.87483 99.99031 83.80554 94.04961
94.91306 80.06482)

Primenom klasične statistike dobijamo sledeći rezultat:

$$p = P\left\{\frac{\bar{Y}_m-\bar{X}_n}{s} \geq \frac{\bar{y}_m-\bar{x}_n}{s}\right\} = P\left\{\frac{\bar{Y}_m-\bar{X}_n}{s} \geq \frac{92.86035-89.1431}{\sqrt{\frac{36}{10}+\frac{36}{15}}}\right\} =$$

$$= P\left\{\frac{\bar{Y}_m-\bar{X}_n}{s} \geq 1.51756\right\} \approx 0.066 .$$

Primenom Bootstrap metode, pozivajući funkciju **boottesttwo** (koja je data u Dodatku) naredbom:

boottesttwo (x,y,1000),

gde je 1000 broj Bootstrap uzoraka dobijena je *p* vrednost:

$p = 0.071$ ,

a primenom naredbe:

boottesttwo (x,y,3000)

dobijena je p vrednost:

$p = 0.061$ .

Zaključujemo da na nivou značajnosti od 5% i primenom klasične statistike i primenom Bootstrap metode nultu hipotezu prihvatamo i da postoji velika saglasnost između dobijenih rezultata.

### 2.3.3. SLUČAJ SA JEDNIM UZORKOM

Pretpostavimo da je **X** = ($X_1$,$X_2$,...,$X_n$) uzorak za obeležje X sa funkcijom raspodele F i označimo sa $\mu$ matematičko očekivanje obeležja.

Hoćemo da testiramo hipotezu:

$H_0$: $\mu = \mu_0$ protiv alternativne $H_1$: $\mu > \mu_0$.

Za test statistiku ćemo uzeti $\bar{X}$ – aritmetičku sredinu uzorka **X**.

Naša odluka o prihvatanju ili odbacivanju nulte hipoteze zasnovana je na *p* vrednosti koju ćemo oceniti sa

$$\hat{p} = P_{H_0}[\bar{X} \geq \bar{x}], \qquad (6)$$

gde je $\bar{x}$ realizovana vrednost aritmetičke sredine uzorka **X**.

Bootstrap metodom ćemo oceniti raspodelu statistike $\bar{X}$ pod uslovom da je hipoteza $H_0$: $\mu = \mu_0$ tačna. Zbog uslova da bude zadovoljena nulta hipoteza Bootstrap metodu ćemo umesto na realizovan uzorak $(x_1, x_2, \dots, x_n)$ primeniti na skup z = $\{z_1, z_2, \dots, z_n\}$ gde su:

$$z_i = x_i - \bar{x} + \mu_0 , \qquad i = 1,2,...,n$$

i putem izvlačenja sa vraćanjem kreirati Bootstrap uzorke. Tada će za opserviranu vrednost $Z^*$ važiti $E(Z^*) = \mu_0$.

Postupak se može opisati sledećim koracima:

1) Od realizovanog uzorka x = $(x_1, x_2, \dots, x_n)$ formiramo vektor z = $(z_1, z_2, \dots, z_n)$, gde su:

$$z_i = x_i - \bar{x} + \mu_0$$

2) Formiramo B uzoraka obima n dobijenih putem izvlačenja sa vraćanjem iz skupa $\{z_1, z_2, \dots, z_n\}$ i za svaki izračunamo:

$\bar{z}_j^*$, j = 1,2,...,B

3) Histogram relativnih frekvencija kreiran na osnovu vrednosti $\bar{z}_j^*$, j = 1,2,...,B dodeljujući svakoj vrednosti verovatnoću 1/B predstavlja raspodelu statistike $\bar{X}$ pod uslovom da je nulta hipoteza tačna.

Na osnovu te raspodele i formule (6) ocenimo *p* vrednost sa:

$$\hat{p}^* = \frac{\sum_{j=1}^{B} I\{\bar{z}_j^* \geq \bar{x}\}}{B}$$

i dobijenu ocenu poredimo sa nivoom značajnosti $\alpha$, pa na osnovu toga donosimo odluku o prihvatanju ili odbacivanju nulte hipoteze.

**Primer 2.3.3.1.** Pretpostavimo da je dat uzorak **X** = ($X_1$,$X_2$,...,$X_n$) obima n za obeležje X sa N($\mu$, $\sigma^2$) raspodelom i da su $\mu$ i $\sigma^2$ nepoznati parametri.

Testiraćemo hipotezu:

$H_0$: $\mu = \mu_0$ protiv alternativne $H_1$: $\mu > \mu_0$

primenom klasičnih statističkih metoda i primenom Bootstrap metode.

Ako je hipoteza $H_0$ tačna statistika:

T = $\frac{\bar{X}_n - \mu_0}{\bar{S}_n}\sqrt{n-1}$ ima Studentovu raspodelu sa n-1 stepeni slobode.

*p* vrednost nalazimo iz uslova: $P\left\{\frac{\bar{X}_n - \mu_0}{\bar{S}_n}\sqrt{n-1} \geq \frac{\bar{x}_n - \mu_0}{\bar{S}_n}\sqrt{n-1}\right\} = p$

Pretpostavimo da nam je dat uzorak obima 10 iz N($90,6^2$) raspodele koji smo generisali, pomoću statističkog paketa R, primenom naredbe:

x<-rnorm(10,90,6)

x = (87.85529 82.51220 85.48065 94.15581 82.82237
91.77458 84.14647 81.97834 82.31771 87.50563)

i neka je $\mu_0 = 85$.

T = $\frac{\bar{X}_n - \mu_0}{\bar{S}_n}\sqrt{n-1} = \frac{86.05491-85}{4.01550} \cdot \sqrt{9} = 0.788125$

$$p = \mathrm{P}\left\{\frac{\bar{X}_n - \mu_0}{\bar{S}_n}\sqrt{n-1} \geq 0.788125\right\}$$

Iz tablica vidimo da je $p \in (0.2,\ 0.25)$.

Pozivajući funkciju **boottestonemean** (koja je data u Dodatku) naredbom:

boottestonemean (x,85,1000),

gde je 1000 broj Bootstrap uzoraka dobijena je *p* vrednost:

$p = 0.212$ ,

a primenom naredbe:

boottestonemean (x,85,3000)

dobijena je *p* vrednost:

$p = 0.2087$ .

I u ovom primeru se zaključuje da postoji velika saglasnost između rezultata dobijenih primenom klasičnih statističkih metoda i rezultata dobijenih primenom Bootstrap metode. Na nivou značajnosti od 5% i primenom klasične statistike i primenom Bootstrap metode nultu hipotezu prihvatamo.

Što se tiče testiranja neparametarskih hipoteza (hipoteza o raspodeli obeležja) oblika:

$H_0: F = F_0$ protiv alternativne $H_1: F \neq F_0$,

Bootstrap metoda se ne primenjuje u njihovom testiranju. Razlog je taj što Bootstrap uzorke generišemo iz empirijske funkcije raspodele (kojom ocenjujemo raspodelu obeležja F), a ona može ali i ne mora da zadovoljava nultu hipotezu. Dakle, Bootstrap uzorci se ne generišu iz raspodele koja zadovoljava nultu hipotezu pa je Bootstrap metoda beskorisna u ovom slučaju.

# ZAKLJUČAK

Istraživači koji su se bavili resampling metodama navode više razloga zbog kojih smatraju upotrebu ovih metoda opravdanim. Jedan od njih je što one predstavljaju pogodnu alternativu za klasične statističke metode koje su bazirane na teorijskim raspodelama, što podrazumeva određene pretpostavke o populaciji i uzorku. Za upotrebu resampling metoda potrebno je napraviti samo jednu pretpostavku – da podaci u razumnoj meri predstavljaju populaciju iz koje su izvučeni tj. da je uzorak reprezentativan.

Neki kritičari smatraju da se resampling ne može smatrati formom statističkog zaključivanja zato što vrši generalizaciju na osnovu samo jednog uzorka. Još jedna od zamerki je da kad uzorak loše predstavlja populaciju iz koje je izvučen resampling metodama se može samo ponavljati ili čak i povećati greška.

Međutim, ove zamerke se odnose na situacije kada bismo i primenom klasičnih statističkih metoda pravili greške u zaključivanju o parametrima populacije ili odluci u statističkom testu.

# DODATAK - Korišćene R funkcije

1) BOOTTESTONEMEAN

```
boottestonemean<-function (x , theta0 , b) {
#
# x   je vektor koji sadrži originalni uzorak
# theta0   je pretpostavljena vrednost matematičkog očekivanja u nultoj hipotezi
# b  je broj bootstrap uzoraka
#
# origtest je vrednost test statistike za originalni uzorak
# pvalue je bootstrap  p-vrednost
# teststatall  je vektor b bootstrap vrednosti test statistike
#
n<-length (x)
v<-mean(x)
z<-x-mean(x) +theta0
counter<-0
teststatall<-rep (0 , b)
for ( i in 1 : b ) {xstar<-sample (z , n , replace=T)
vstar<-mean (xstar)
if (vstar>= v) {counter<-counter+1}
teststatall [i] <-vstar}
pvalue<-counter/b
list ( origtest=v , pvalue=pvalue)
}
```

2) BOOTTESTTWO

```
boottesttwo<-function (x , y , b ) {
#
# x   je vektor koji sadrži prvi uzorak
# y  je vektor koji sadrži drugi uzorak
# b  je broj bootstrap uzoraka
#
# origtest je vrednost test statistike za originalni uzorak
# pvalue je bootstrap  p-vrednost
# teststatall  je vektor b bootstrap vrednosti test statistike
#
n1<-length(x)
```

```
n2<-length(y)
v<-mean(y) - mean(x)
z<-c(x , y)
counter<-0
teststatall<-rep(0 , b)
for ( i in 1 : b) {xstar<-sample (z , n1 , replace=T)
ystar<-sample(z , n2 , replace=T)
vstar<-mean(ystar) - mean(xstar)
if (vstar>= v) {counter<-counter+1}
teststatall [i] <-vstar}
pvalue<-counter/b
list ( origtest=v , pvalue=pvalue)
}
```

3) PERCENTCIBOOT

```
percentciboot<-function (x , b , alpha) {
# x   je vektor koji sadrži originalni uzorak
# b  je broj bootstrap uzoraka
# alpha : (1 - alpha) je nivo poverenja
#
# theta je tačkasta ocena
# lower  je donja granica intervala poverenja
# upper  je gornja granica intervala poverenja
# thetastar  je vektor b bootstrap tačkastih ocena
#
theta<-mean (x)
thetastar<-rep (0 , b)
n<-length (x)
for ( i in 1 : b) {xstar<-sample (x , n , replace=T)
thetastar [i] <-mean (xstar)
}
thetastar<-sort (thetastar)
pick<-round ( ( alpha/2 ) * (b+ 1 ) )
lower<-thetastar [pick]
upper<-thetastar [b-pick+1]
list (theta=theta , lower=lower , upper=upper)
}
```

# LITERATURA

[1] Bradley Efron, R.J.Tibshirani: *An_Introduction_to the Bootstrap* (1993)

[2] Bradley Efron: *Bootstrap methods: Another look at the jackknife* (The Annals of Statistics 1979, Vol. 7, No. 1, 1-26)

[3] Danka Purić, Goran Opačić: *Poduzorkovanje, samouzorkovanje, postupak univerzalnog noža i njihova upotreba u postupcima za statističku analizu multivarijacionih podataka* (Primenjena psihologija, 2013, Vol 6(3), str. 249-266)

[4] Ivana Malić: *Mali uzorci i primena Bootstrap metoda u ekonometriji* (Master rad, Novi Sad, 2012)

[5] K.Singh, M.Xie: *Bootstrap: A Statistical Method* (Rutgers University)

[6] Prem S.Mann: *Uvod u statistiku* (2009)

[7] Robert V. Hogg, Joseph W. McKean, Allen T. Craig: *Introduction to Mathematical Statistics* (2005)

[8] Republički zavod za statistiku: *Kvartalno poslovanje privrednih društava, 2012 – radni dokument br. 84* (2013)

[9] Vesna Jevremović: *Verovatnoća i statistika* (2009)

[10] Z.Ivković, D.Banjević, P.Peruničić, Z.Glišić: *Statistika* (1980)

[11] http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/notebootstrap.pdf

[12] http://starisajt.elfak.ni.ac.rs/phptest/new/html/Studije/predavanja-literatura/matematika-odabrana-poglavlja/statistika.pdf

[13] http://scholar.lib.vt.edu/theses/available/etd-61697-14555/unrestricted/Ch4.pdf

[14] http://dms.irb.hr/tutorial/hr_tut_mod_eval_4.php