UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Tadej Šnajder

Statistične metode samovzorčenja

Delo diplomskega seminarja

Mentorica: izr. prof. dr. Maja Pohar Perme

Ljubljana, 2015

Kazalo

Statistične metode samovzorčenja

Povzetek

V delu diplomskega seminarja smo pokazali ključen primer 2.1, kjer s standardnimi analitičnimi postopki ne moremo oceniti lastnosti cenilke $\hat{\theta} = \overline{X}/\overline{Y}$. Obravnavali smo parametrično in neparametrično metodo samovzorčenja, ki omogočata ocenjevanje lastnosti cenilk kot je $\hat{\theta}$. Delovanje metode smo pokazali na primeru 3.1. Navedli smo izrek 4.11, da za vzorčno povprečje metoda smovzorčenja konvergira proti pravi vrednosti. Za to smo potrebovali metrični prostor porazdelitvenih zakonov z Mallowsovo metriko. Z lemama 4.7 in 4.10 ter trditvijo 4.9 smo dokazali izrek 4.11. Navedli in dokazali smo tudi trditev 4.13, ki nam pove za katere cenilke lahko uporabimo metodo samovzorčenja. Na primeru 4.15 smo pokazali, da za cenilko $\hat{\theta} = max(X_1, X_2, ...X_n)$ neparametrična metoda samovzorčenja ne deluje. Prav tako smo obravnavali, kako lahko z metodo samovzorčenja ocenjujemo pristranskost in disperzijo cenilke. Na primeru 5.6 smo pokazali, da lahko z metodo samovzorčenja ocenjujemo cenilko mediano. Obravnavali smo, kako lahko s percentilno metodo samovzorčenja ocenjujemo intervale zaupanja. Pokazali smo dva primera 6.6 in 6.7, kjer percentilna metoda odpove. Na koncu smo z neparametrično metodo samovzorčenja rešili primer 2.1 in povzeli rezutate.

Statistical bootstrap methods

Abstract

In the thesis we showed key example 2.1, where use of standard analytical procedures is not sufficient to estimate properties of estimator $\hat{\theta} = \overline{X}/\overline{Y}$. We used parametric bootstrap method and nonparametric bootstrap method, that allow us esimating properties of estimators, such as $\hat{\theta}$. We have shown how bootstrap works on example 3.1. We have proved that the sample average bootstrap method converges to the mean of population in theorem 4.11. We used Mallows metric and metric spaces to prove the theorem. In addition, we used lemmas 4.7, 4.10 and theorem 4.9. We also proved theorem 4.13, which allows us to use bootstrap method on some other esimators. In the example 4.15, we have shown that for the $\hat{\theta} = max(X_1, X_2, ...X_n)$ estimator the nonparametric bootstrap does not correctly estimate θ. We have shown how to use bootstrap method to estimate bias and variance of the estimator $\hat{\theta}$. In example 5.6 we have shown that bootstrap may be used to etimate the esitmator median. We have also shown, how to use bootstrap method to estimate confidence intervals with the percentile method. We showed two examples 6.6 and 6.7 in which the percentile method does not return the correct confidence interval. In the end we solved the example 2.1 and summarized its results.

Math. Subj. Class. (2015): Statistics

Ključne besede: Metoda samozovčenja, percentilna metoda samovzorčenja, pristranskost, disperzija, parametrična metoda samovzorčenja, neparametrična metoda samovzorčenja, kvocient pričakovanih vrednosti, cenilka maksimum, mediana

Keywords: Bootstrap, percentile bootstrap, bootstrap bias, bootstrap variance, parametric bootstrap, nonparametric boostrap, quotient of expected values, maximum estimator, median

ZAHVALA

Za vse koristne nasvete, dragocene pripombe in pomoč pri pisanju dela diplomskega seminarja se zahvaljujem svoji mentorici izr. prof. dr. Maji Pohar Perme.

Najlepše se zahvaljujem tudi svoji družini in svoji punci Veroniki za vso podporo in razumevanje, ki ste mi jo nudili pri pisanju dela diplomskega seminarja.

Zahvala gre tudi vsem, ki ste me spremljali na študijski poti.

1. Uvod

V statistiki se problemi običajno rešujejo z uporabo standardnih analitičnih postopkov. Pri določenih problemih pa s standardnimi formulami ne moremo oceniti iskanih količin. V okviru dela diplomskega seminarja smo raziskali metode samovzorčenja kot alternativen pristop za reševanje takšnih problemov. Predstavili bomo primer, na katerem ocenjujemo kvocient pričakovanih vrednosti. Obravanavali bomo parametrično in neparametrično metodo samovzročenja, ki omogočata ocenjevanje iskanih količin z ustvarjanjem novih vzorcev. S pomočjo neparametrične metode samovozrčenja bomo ocenili kvocient pričakovanih vrednosti. Ocenili bomo tudi pristranskost, disperzijo in interval zaupanja stopnje zaupanja 0.95 za izbrano cenilko, s katero bomo ocenili kvocient pričakovanih vrednosti. Kako ocenimo iskano količino, pristranskost izbrane cenilke, disperzijo izbrane cenilke in interval zaupanja s pomočjo metod samovzorčenja, bomo pokazali na preprostih primerih. Predstavili bomo tudi primere, kjer uporaba ali parametrične ali neparametrične metode samovzorčenja ne da pravilnih rezutatov.

Cilj naše diplomske naloge je podati teoretične osnove metode samovzorčenja in pokazati na primerih, kdaj in zakaj metoda samovzorčenja deluje in kdaj uporaba metode samovzorčenja ni ustrezna.

2. Osnovni koncept metode samovzorčenja

V delu diplomskega seminarja se bomo seznanili z metodama samovzorčenja.
Za slučajni vzorec predpostavimo, da je sestavljen iz neodvisnih enako porazdeljenih slučajnih spremenljivk.
Na slučajnem vzorcu S_n ocenjujemo neznano količino θ, ki jo lahko izračunamo kot $\theta = g(F)$, to je kot funkcional kumulativne porazdelitvene funkcije F.
Vsi izračuni bodo narejeni s pomočjo programa R studio. Algoritmi, ki jih bomo uporabili bodo priloženi v prilogi. Rezultate bomo zaokrožili na tri decimalke.

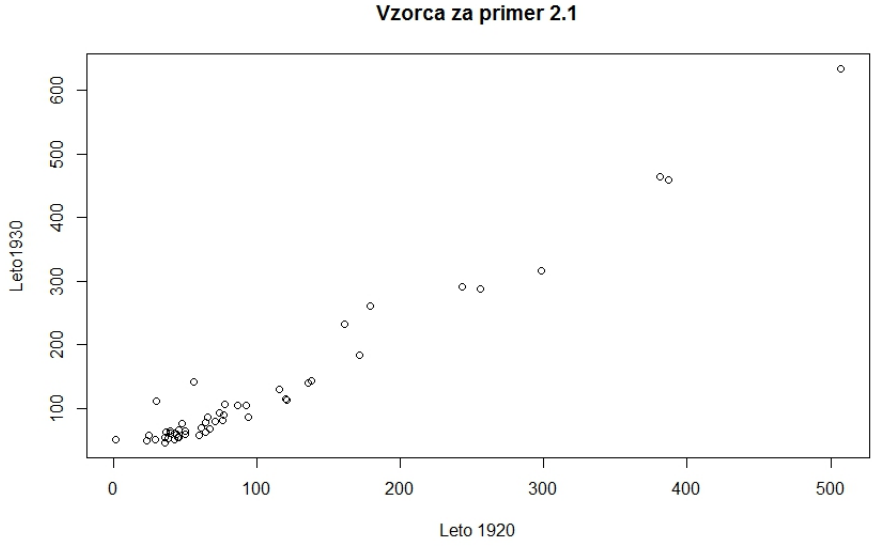
2.1. **Predstavitev ključnega primera.**

Predstavili bomo primer 2.1, pri katerem je s standardnimi analitičnimi postopki za iskano količino θ, kot je na primer kvocient pričakovanih vrednosti, zahtevno izračunati lasnosti cenilk kot sta pristranskost in disperzija. Primer smo vzeli iz literature [1].

Primer 2.1. Podanih imamo $n = 49$ parov, kjer en par $(x_j, y_j), j = 1, 2, ..49$, ustreza številu prebivalcev v letih 1930 in 1920 v j-tem ameriškem mestu.Vzorec označimo z S_n. Zanima nas pričakovana rast prebivalstva v mestih Združenih držav Amerike med letoma 1920 in 1930. Ocena pričakovane rasti bi nam omogočila, da ocenimo število prebivalcev v letu 1930, če poznamo število prebivalcev v letu 1920.
Privzamemo, da je število prebivalcev v j-tem mestu neodvisno od števila prebivalcev ostalih mest za $j = 1, 2, ...49$.

Z X označimo slučajno spremenljivko, ki nam generira vrednosti po mestih v letu 1930, z Y pa slučajno spremenljivko, ki nam generira vrednosti v letu 1920. Iskano količino izrazimo kot $\theta = \frac{E(X)}{E(Y)}$. Ob oceni vrednosti θ nas bodo tudi zanimale pristranskost, disperzija izbrane cenilke in ocena intervala zaupanja stopnje zaupanja 0.95 za θ.

Za lažjo predstavo si oglejmo pare točk na grafu:



V skladu z literaturo [1] trdimo, da za primer 2.1 ni očitnega parametrčnega modela za skupno porazdelitev (X, Y) kot je na primer logaritemsko normalna porazdelitev. Poskusimo θ oceniti s cenilko $\hat{\theta} = \frac{\overline{X}}{\overline{Y}}$. Zanima nas pristranskost izbrane cenilke $\hat{\theta}$.

Za pristranskost potrebujemo pričakovano vrednost cenilke:

$$E(\hat{\theta}) = E(\frac{\overline{X}}{\overline{Y}}) = E(\frac{\frac{1}{n}\sum_{j=1}^{n} X_j}{\frac{1}{n}\sum_{k=1}^{n} Y_k}) = E(\frac{\sum_{j=1}^{n} X_j}{\sum_{k=1}^{n} Y_k}) =$$

$$= \sum_{j=1}^{n} E(\frac{X_j}{\sum_{k=1}^{n} Y_k})$$

Ob izračunu pristranskosti opazimo naslednje težave:

- Slučajni spremenjivki X in Y sta lahko odvisni, posledično X_k in Y_k, $k = 1, 2, ...n$.
- Ne znamo izračunati $E(\overline{X}/\overline{Y})$.
- Posledično ne znamo določiti pristranskosti.
- $E(\hat{\theta})$ ne znamo zapisati kot funkcijo θ, torej ne znamo izraziti pristranskosti.

Opazimo še naslednji težavi:

- Ne znamo izračunati točne disperzije $\hat{\theta}$, ker ne znamo izračunati $E(\hat{\theta}^2)$ in $E(\hat{\theta})^2$:

$$D(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2.$$

- Ne vemo, kako je porazdeljena cenilka $\hat{\theta}$ (za $\hat{\theta}$ imamo eno samo vrednost), zato ne znamo oceniti aproksimativnega intervala zaupanja stopnje zaupanja 0.95 za θ.

2.2. **Izračun θ iz ključnega primera ob predpostavki, da poznamo skupno porazdelitev vzorca.**

Če bi poznali skupno porazdelitev, bi nam ta omogočila empirično oceno lastnosti cenilke $\hat{\theta}$. To bomo pokazali v naslednjih dveh točkah.
V točki 1. bomo predpostavili, da sta slučajni spremenljivki X in Y neodvisni, porazdeljeni logaritemsko normalno. V točki 2. bomo predpostavili da poznamo skupno porazdelitev (X,Y), ki naj bo porazdeljena bivariatno logaritemsko normalno.

1.) Predpostavimo, da sta slučajni spremenljivki X in Y neodvisni, porazdeljeni logaritemsko normalno.
Pričakujemo, da je cenilka pristranska, saj

$$E(\hat{\theta}) = E(\frac{\sum_{j=1}^{n} X_j}{\sum_{j=1}^{n} Y_j}) = nE(X)E(\frac{1}{\sum_{j=1}^{n} Y_j}) \geq nE(X)\frac{1}{E(\sum_{j=1}^{n} Y_j)} = \frac{E(X)}{E(Y)} = \theta.$$

Pri prvem enačaju smo uporabili neodvisnost, pri neenačaju pa Jensenovo neenakost.

Ob trenutnih predpostavkah dodatno predpostavimo, da poznamo točne vrednosti parametrov (vzorca smo logaritmirali; logaritmirana vzorca sta po predpostavki porazdeljena normalno. Za točne vrednosti smo vzeli ocene, ki smo jih izračunali s standardnimi cenilkami na logaritmiranih vzorcih):

-$\mu_X = \overline{X}$=4.574

-$\mu_Y = \overline{Y}$=4.257

-$\sigma_X^2 = \frac{1}{48}\sum_{j=1}^{49}(X_i - \overline{X})^2 = 0.451$

-$\sigma_Y^2 = \frac{1}{48}\sum_{j=1}^{49}(Y_i - \overline{Y})^2 = 0.83$

Po predpostavki o neodvisnosti je korelacijski koeficient enak 0.

Točna vrednost

$$\theta = \frac{E(X)}{E(Y)} = \frac{exp(4.574 + 0.451/2)}{exp(4.257 + 0.83/2)} = 1.135$$

Poskusimo pristranskost oceniti empirično. To bomo naredili na naslednji način:

-Generiramo 100 vzorcev za X in 100 vzorcev za Y velikosti $n = 49$.

-Ocenimo za prvi vzorec iz X in prvi vzorec iz Y kvocient povprečij, nato za druga dva, do zadnjih dveh vzorcev.

-Ocene povprečimo in povprečje primerjamo z točno vrednostjo.

Dobimo vrednost 1.104. Vidimo, da je cenilka pri teh predpostavkah pristranska, razlika empirične ocene in točne vrednosti θ je $1.135 - 1.104 \neq 0$.

2.) Predpostavimo, da poznamo skupno porazdelitev (X,Y), ki naj bo porazdeljena bivariatno logaritemsko normalno.
Vrednosti vzorca iz primera 2.1 logaritmiramo. Za parametre porazdelitve bomo vzeli
naslednje vrednosti (spet uporabimo standarne cenilke za izračun parametrov normalne porazdelitve):

- μ_X =4.574

-μ_Y =4.257

-$\sigma_X^2 = 0.451$

-$\sigma_Y^2 = 0.83$

-$\rho = 0.842$

Tudi pri teh predpostavkah je $\theta = \frac{E(X)}{E(Y)} = 1.136$

Ocenimo empirično pristranskost tako, da generiramo 100 vzorcev velikosti $n = 49$ in na vsakem ocenimo vrednost cenilke.
Dobimo vrednost 1.071. Tudi v tem primeru je cenilka pristranska.

Ker skupne porazdelitve (X, Y) ne poznamo, se bomo problema lotili z metodo samovzorčenja. Za boljše razumevanje bomo posamezno metodo samovzorčenja predstavili na primeru in jo nato uporabili na primeru 2.1.

Na primerih bomo obravnavali parametrično in neparametrično metodo samovzorčenja. Za boljše razumevanje delovanja metode samovzorčenja smo izbrali preprost primer, ki ga je možno rešiti tudi brez uporabe metode samovzorčenja. Primer najdemo v literaturi [1].

Primer 3.1. Podana je tabela podatkov:

j	1	2	3	4	5	6	7	8	9	10	11	12
x_j	3	5	7	18	43	85	91	98	100	130	230	497

Tabela prikazuje povezavo med številom okvar klimatskih naprav (označene z j) vgrajenih v letala Boeing 720 in številom njihovih obratovalnih ur (označene z x_j). Oceniti hočemo pričakovano vrednost in standardni odklon vzorčnega povprečja ($\overline{x}_j$).

Primer bomo rešili z metodo samovzorčenja. Ločili bomo dva pristopa reševanja:

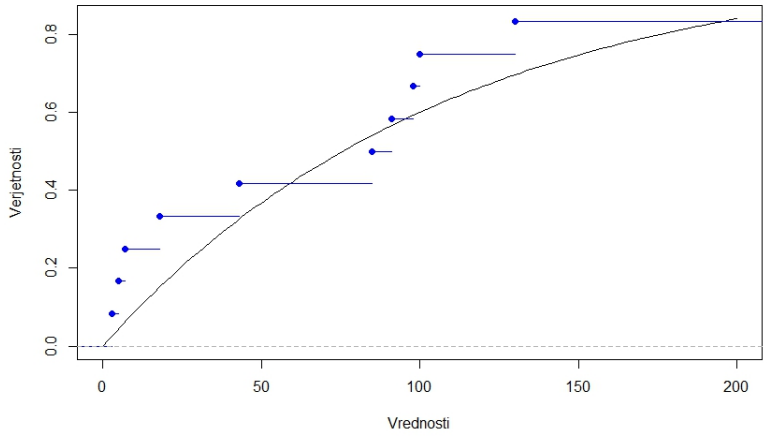
a) Označimo dani vzorec z S_n.
Iz S_n sestavimo empirično porazdelitveno funkcijo

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i<x\}}.$$

Z F_n generiramo nove vzorce velikosti n = 12 (torej enake velikosti kot originalni vzorec), kjer i-ti generirani vzorec označimo z $S^*_{n,i}$. Vzorec sestavimo tako, da izbiramo nove neodvisne realizacije iste slučajne spremenljivke, ki ima kumulativno porazdelitveno funkcijo F_n. Izbiramo s ponavljanjem, kar pomeni, da lahko imamo v vzorcu tudi več enakih realizacij.
Na vsakem na novo dobljenem vzorcu ocenimo iskano količino (v tem primeru ali pričakovano vrednost ali standardni odklon) in ocenjene vrednosti povprečimo. Če je iskana količina pričakovana vrednost vzorčnega povprečja, označimo njeno oceno $\overline{x}^*_R$, če je iskana količina standardni odklon pa označimo s^*_R. Z indeksom R povemo, kolkio novih vzorcev smo generirali.

b) Na podlagi podatkov sklepamo, da je vzorec porazdeljen eksponentno.
Naredili bomo grafično primerjavo s kumulativno porazdelitveno funkcijo eksponentno porazdeljene slučajne spremenljivke. Testa χ^2 ne uporabimo, saj je vzorec premajhen, torej test ni primeren. Parameter λ ocenimo z momentno cenilko $\hat{\lambda} = \frac{n}{\sum_{j=1}^{n} X_j}$.

Primerjava empirične porazdelitve vzorca (modra) in kumulativne porazdelitvene funkcije eksponentno $\hat{\lambda}$ porazdeljene slučajne spremenljivke (črna) je prikazana na naslednjem grafu:



Na podlagi grafične primerjave sklepamo, da je model sprejemljiv, čeprav imamo opravka s premajhnim vzorcem, da bi sklepanje lahko ovrednotili.
Označimo kumulativno porazdelitveno funkcijo eksponentne slučajne spremenljivke z F_{S_n}, kjer smo parameter λ ocenili z $\hat{\lambda}$. S tem želimo poudariti, da smo kumulativno porazdelitveno funkcijo določili na podlagi vzorca S_n.
Na enak način kot v točki a) iz F_{S_n} generiramo nove vzorce. Ocenimo iskano količino na vsakem novem vzorcu ter ocene povprečimo.

Metodo iz točke a) imenujemo *neparametrična metoda samovzorčenja*, metodo samovzorčenja iz točke b) imenujemo *parametrična metoda samovzorčenja*.

S pomočjo algoritmov iz priloge ocenimo pričakovano vrednost in standardni odklon z neparametrično metodo in ocenimo povprečje in standardni odklon s parametrično metodo.

Ocenjene vrednosti so:
Povprečje (neparametrično): 108.819
Standardni odklon (neparametrično):123.539
Povprečje (parametrično): 110.623
Standardni odklon (parametrično):104.922

Standardni odklon po parametrični metodi odstopa od neparametrične, kar lahko pripišemo izbiri modela.

S ponovno simulacijo ne bomo dobimo istih $S^*_{n,i}$, $i = 1, ...R$. Posledično ne dobimo enakih vrednosti za x^*_R ali s^*_R.
Z neparametrično metodo samovzorčenja naredimo deset novih simulacij in ocenimo standardni odklon:

```
121.813
121.913
129.233
124.569
117.445
124.866
```

126.077
130.161
122.934
118.806

Metoda je intuitivna; če bo porazdelitev vzorca F_n" blizu" dejanski porazdelitveni funkciji F ter smo za ocenjevanje θ izbrali ustrezno cenilko $\hat{\theta}$, potem pričakujemo, da bo tako dobljena ocena θ_R^* konvergirala proti iskani vrednosti θ.
Prav tako potrebujemo sestavljanje vzorcev s ponavljanjem; v nasprotnem primeru je realizacija novo generiranega vzorca odvisna od predhodnih realizacij.

Oznaka 3.2. *Definiramo* samovzorčno porazdelitveno funkcijo *in jo označimo z* F_n^* *kot skupno ime za empirično porazdelitveno funkcijo* F_n *in za model* F_{S_n}*, izbran na podlagi vzorca* S_n*.*

Oznaka 3.3. *Naj bo dan vzorec* S_n*.*
Označimo z $S_{n,1}^*, S_{n,2}^*, ..., S_{n,R}^*$ *nove neodvisno generirane vzorce iz porazdelitvene funkcije* F_n^**.*
Označimo z $S^* = (S_{n,1}^*, S_{n,2}^*, ..., S_{n,R}^*)$ *vektor vzorcev, v katerem imamo zbrane na novo generirane vzorce.*

Zapišimo postopek ocenjevanja iskane količine θ z metodo samovzorčenja. Postopek je povzet po [8].

Postopek 3.4.
- Iz danih podatkov sestavimo empirično porazdelitveno funkcijo F_n ali model F_{S_n}.
- Sestavimo vektor R-tih vzorcev velikosti n s ponavljanjem in ga zapišemo kot $S^* = (S_{n,1}^*, S_{n,1}^*, ..., S_{n,R}^*)$.
- Na vsakem vzorcu $S_{n,i}^*$ s cenilko $\hat{\theta}$ izračunamo za iskano vrednost θ, izračunano vrednost označimo s θ_i.
- Izračunane vrednosti θ_i povprečimo; dobljeno povprečje θ_R^* je ocena za iskano količino θ.

Komentar 3.5.

- *Funkcijo* $g(.)$ *bomo uporabljali kot oznako za algoritem, s katerim bomo iz vzorca* S_n *izračunali vrednost cenilke* $\hat{\theta}$ *za* θ*.*
- θ_R^* *je odvisna od nabora vzorcev, ki ga dobimo z metodo samovzorčenja, torej je slučajna spremenljivka.*
- *Z enim samim vzorcem* S_n *navadno lahko izračunamo samo eno vrednost cenilke* $\hat{\theta}$*. Metoda samovzorčenja nam da več vzorcev, na katerih lahko izračunamo* θ_i*, to je ocenjena vrednost cenilke na* i*-tem generiranem vzorcu* $S_{n,i}^*$*. Iz vrednosti* θ_i *lahko sestavimo empрično porazdelitveno funkcijo, s katero aproksimiramo kumulativno porazdelitveno funkcijo cenilke* $\hat{\theta}$*.*

3.1. Ocenitev θ iz primera 2.1 z neparametrično metodo samovzorčenja.

Na primeru 2.1 bomo s pomočjo neparametrične metode samovzorčenja ocenili θ. Ocenjevanja se bomo lotili na naslednji način:

Sestavimo empirično porazdelitveno funkcijo vzorca S_n

$$F_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i<x, y_i<y\}}.$$

Nadaljujemo po postopku 3.4. Vrednost θ_i na vzorcu $S^*_{n,i}$ izračunamo s cenilko $\hat{\theta} = \overline{X}/\overline{Y}$, za $i = 1, 2, ...R$. Dobimo oceno $\theta^*_R = 1.236$.

Opomba 3.6. Vrednost θ bi lahko ocenjevali tudi s parametrično metodo samovzorčenja, vendar je omenjena metoda za izračun vrednosti θ^*_R manj primerna. Težave se pojavijo pri določitvi ustreznega modela za slučajni vektor (X, Y) kot smo povedali v primeru 2.1.

3.2. **Primerjava ocen po parametrični in neparametrični metodi samovzorčenja.**

Primer 3.7. Poglejmo na primeru ali je parametrična metoda boljša kot neparametrična metoda, če poznamo porazdelitev vzorca.
Naj bo dan slučajni vzorec velikosti $n = 50$, sestavljen iz realizacij x_i neodvisno enako porazdeljenih slučajnih spremenljivk X_i (reprezentativna slučajna spremenljivka je X), ki so porazdeljene enakomerno zvezno na intervalu $[0, k], k \in \mathbb{R}, k > 0$. Ocenjujemo $\theta = D(X)$. θ bomo ocenjevali s cenilko
$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{50}(X_i - \overline{X})^2$.

1.) Predpostavimo, da je prava vrednost $k = 5$.
Ocenjujemo k z nepristransko momentno cenilko $\hat{k} = 2\overline{X}$. Ker vemo, da je vzorec porazdeljen enakomerno zvezno, lahko uporabimo parametrično metodo samovzorčenja (če porazdelitve vzorca ne bi poznali, bi lahko poskusili preveriti domnevo, da je vzorec porazdeljen po znani porazdelitvi z uporabo χ^2-testa).

S parametrično metodo samovzorčenja po postopku 3.4 dobimo oceno za θ $\theta^*_{200} = 2.461$. Ocenimo tudi vzorčno disperzijo cenilke $\hat{\theta}$ po formuli

$$\frac{1}{199}\sum_{j=1}^{200}(\theta_j - \theta^*_{200})^2.$$

Dobimo vrednost 0.091.

2.) Uporabimo neparametrično metodo samovzorčenja in dobimo oceno $\theta^*_{200} = 2.476$ in vrednost vzorčne disperzije cenilke $\hat{\theta}$, ki jo izračunamo po formuli zgoraj, 0.076.

Spomnimo se: s ponovno simulacijo ne dobimo istih rezultatov. Zanima nas hitrost konvergence, ko spreminjamo število generiranih vzorcev.

R	5	10	50	200	1000
Param.	2.532	2.380	2.545	2.443	2.439
Neparam.	2.534	2.370	2.374	2.429	2.445

Izračunana vrednost cenilke $\hat{\theta} = 2.493$. V naslednji tabeli imamo podana absolutna odstopanja θ_R^* od izračunane vrednosti cenilke $\hat{\theta}$:

R	5	10	50	200	1000
Param.	0.039	0.113	0.052	0.05	0
Neparam.	0.041	0.123	0.119	0.064	0.048

Iz tabele je razvidno, da v tem primeru parametrična metoda hitreje konvergira.

4. Konvergenca metode samovzorčenja

V tem razdelku bomo definirali in dokazali pogoje za konvergenco $\theta_R^* \to \theta$. Pri definiranju in dokazovanju si bomo pomagali z metričnimi prostori. Teorija je povzeta po literaturi [2], vendar je poenostavljena.

4.1. **Mallowsova metrika in pripadajoči metrični prostori.**

Omejili se bomo na enorazsežne slučajne spremenljivke.

Definicija 4.1. Naj bo Γ_p množica vseh porazdelitvenih zakonov definiranih na Borelovi σ-algebri $\mathscr{B}(\mathbb{R})$ realne osi $\mathbb{R}$, tako da je

$$\Gamma_p := \{f; \int |x|^p f(dx) < \infty\}, p > 0.$$

Na Γ_p vpeljemo tako imenovano *Mallowsovo metriko*:

$$d_p(f_X, f_Y) := \inf_{f_{X,Y}} E(|X - Y|^p)^{\frac{1}{p}},$$

kjer infimum teče po skupnih porazdelitvah $f_{X,Y}$, kjer sta X in Y slučajni spremenljivki z vrednostmi v $\mathbb{R}$, katere robna porazdelitvena zakona sta f_X in $f_Y \in \Gamma_p$.

Komentar 4.2. *Osredotočili se bomo na Γ_1, Γ_2, ker želimo vpeljati naslednjo obliko konvergence za $f_n, f \in \Gamma_2$:*
f_n konvergira proti f, če in samo če $f_n \to f$ v porazdelitvi in $\int x^2 f_n(dx) \to \int x^2 f(dx)$.
Videli bomo, da nam to obliko konvergence omogoča d_2-metrika.

Za metodo samovzorčenja potrebujemo (Γ_1, d_1), (Γ_2, d_2).

Trditev 4.3. *Naj bosta dana Γ_p in d_p definirana zgoraj, $p \in \{1, 2\}$.*
Tedaj veljajo naslednje trditve:

- *Infimum Mallowsove metrike je vedno dosežen.*
- *(Γ_1, d_1) je metrični prostor.*
- *(Γ_2, d_2) je metrični prostor.*

Dokaz trditve bomo izpustili, ker je del dokaza prezahteven za dodiplomsko stopnjo. Dokaz najdemo v literaturi [2].

4.2. **Konvergenca povprečja metode samovzorčenja k pričakovani vrednosti populacije.**

Z naslednjo lemo pokažemo, da d_2 zadošča pogojem v komentarju 4.2.

Lema 4.4. *Naj bosta f in $f_n \in \Gamma_p$, $p \in \{1, 2\}$. Tedaj je trditev*
$d_p(f_n, f) \to 0$, ko gre $n \to \infty$ ekvivalentna trditvi $f_n \to f$ v porazdelitvi in $\int |x|^p f_n(dx) \to \int |x|^p f(dx)$.

Dokaz. Dokazali bomo samo $\Rightarrow$. Dokaz $\Leftarrow$ je prezahteven za dodiplomsko stopnjo. Tudi ta dokaz najdemo v literaturi [2].
Predpostavimo $d_p(f_n, f) \to 0$, $p \in \{1, 2\}$. Naj bodo slučajna spremenljivka X in členi zaporedja slučajnih spremenljivk $(X_n)_{n\in\mathbb{N}}$ takšni, da za njihove porazdelitvene zakone f in $(f_n)_{n\in\mathbb{N}}$ velja

$$d_p(f, f_n) = E(|X - X_n|^p)^{\frac{1}{p}}, \forall n \in \mathbb{N}.$$

Takšne slučajne spremenljivke obstajajo po trditvi 4.3. Poglejmo si razliko

$$\Big(\int |x|^p f_n(dx)\Big)^{\frac{1}{p}} - \Big(\int |x|^p f(dx)\Big)^{\frac{1}{p}} =$$

$$= E(|X_n|^p)^{\frac{1}{p}} - E(|X|^p)^{\frac{1}{p}}$$

uporabimo neenakost Minkowskega

$$\leq E(|X_n - X|^p)^{\frac{1}{p}} = d_p(f_n, f) \to 0,$$

po predpostavki.

Iz L^p konvergence sledi konvergenca v porazdelitvi.

□

Oznaka 4.5. *Naj bo dan slučajni vzorec S_n velikosti n.*
Z μ_n bomo označili vzorčno povprečje S_n. Potem z μ_R^* označimo oceno za vzorčno povprečje, ki smo ga dobili z metodo samovzorčenja.

Želimo pokazati, da slučajna spremenljivka $\sqrt{R}(\mu_R^* - \mu_n)$ konvergira k normalni porazdelitvi s pričakovano vrednostjo 0. To bomo pokazati s pomočjo (Γ_1, d_1) in (Γ_2, d_2).
Za dokaz konvergence bomo potrebovali še naslednje porazdelitvene zakone:

Oznaka 4.6. *Naj bodo $X_1, X_2, ...$ zaporedje neodvisnih enako porazdeljenih slučajnih spremenljivk s porazdelitvenim zakonom f in končno disperzijo. Označimo $f^{(R)}$ kot porazdelitveni zakon slučajne spremenljivke*

$$S_R = \sqrt{R}\sum_{i=1}^{R}(X_i - E(X_i))$$

in $f^{(R)*}$ kot porazdelitveni zakon slučajne spremenljivke

$$S_R^* = \frac{1}{R}\sum_{i=1}^{R} X_i.$$

Definiramo še slučajno spremenljivko

$$s_n^{*2} := \frac{1}{n}\sum_{i=1}^{n}(X_i^* - \mu_n^*)^2.$$

Če je $f \in \Gamma_2$, potem je tudi $f^{(R)} \in \Gamma_2$, ker ima tedaj tudi S_R končno disperzijo, posledično tudi končen drugi moment.

Velja tudi naslednja ocena:

Lema 4.7. *Naj bodo $X_1, X_2, ...X_R$ neodvisne in enako porazdeljene s porazdelitvenim zakonom $g \in \Gamma_p, p \in \{1, 2\}$. Naj bodo $Y_1, Y_2, ..Y_R$ neodvisne in enako porazdeljene in neodvisne od $X_i, i = 1, 2, ..R$ s porazdelitvenim zakonom $f \in \Gamma_p p \in \{1, 2\}$. Tedaj velja naslednja ocena:*

$$d_p(h^{(R)}, g^{(R)}) \leq d_p(h, g), p \in \{1, 2\}$$

$$d_p(h^{(R)*}, g^{(R)*}) \leq d_p(h, g), p \in \{1, 2\}$$

Dokaz. Pari (X_i, Y_i) so neodvisni, porazdelitveni zakon X_1 je g, porazdelitveni zakon Y_1 je h. Brez škode za splošnost predpostavimo, da je dosežen infimum Mallowsove metrike. Označimo $X_i' = X_i - E(X)$,$Y_i' = Y_i - E(Y)$.

$$d_p(h^{(R)}, g^{(R)}) = E(|\frac{1}{\sqrt{R}} \sum_{i=1}^{R} X_i' - \frac{1}{\sqrt{R}} \sum_{i=1}^{R} Y_i'|^p)^{\frac{1}{p}} = \frac{1}{\sqrt{R}} E(|\sum_{i=1}^{R} X_i' - \sum_{i=1}^{R} Y_i'|^p)^{\frac{1}{p}} \leq$$

uporabimo neenakost Minkowskega

$$\leq \frac{1}{R} \sum_{j=1}^{R} E(|X_i' - Y_i'|^p)^{\frac{1}{p}} = E(|X_1' - Y_1'|^p)^{\frac{1}{p}} =$$

$$E(|X_1 - Y_1|^p)^{\frac{1}{p}} - |E(X) - E(Y)|^2 \leq d_p(h, g)$$

Dokaz

$$d_p(h^{(R)*}, g^{(R)*}) \leq d_p(h, g), p \in \{1, 2\}$$

se pokaže enako. □

Nasleddnjo definicijo smo vzeli iz vira [7].

Definicija 4.8. Naj bo $(\Omega, \mathscr{F}, P)$ pozitiven prostor z mero. Množica $A \subset L^1(\mu)$ je *enakomerno integrabilna*, če za vsak $\epsilon > 0$ obstaja tak $\delta > 0$, da je

$$|\int_E f d\mu| < \epsilon,$$

ko je $f \in A$ in $\mu(E) < \delta$.
Za razred C slučajnih spremenljivk rečemo,da so *enakomerno integrabilne*, če:

- obstaja $K > 0$ tako, da za vsak $X \in C$ velja $E(|X|) < K$ in

- za vsak $\epsilon > 0$ obstaja $\delta > 0$ tako, da za vsako merljivo množico A, za katero je $P(A) < \delta$ in za vsak $X \in C$, $E(|X|\mathbb{1}_A) \leq \epsilon$.

Naslednjo trditev potrebujemo za dokaz leme 4.10.Trditev najdemo v viru [**?**]

Trditev 4.9. *Naj bo $X_1, X_2, ...X_n$ nabor neodvisnih enako porazdeljenih slučajnih spremeljivk s kumulativno porazdelitveno funkcijo F in naj bo F_n empirična porazdelitvena funkcija slučajnega vzorca S_n, ki ga dobimo kot realizacijo slučajnih spremenljivk X_i, $i = 1, 2...n$. Tedaj velja*

$$P[\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0] = 1.$$

Dokaz. Naj bo $\epsilon > 0$. Naj bo $k \in \mathbb{Z}, k > 2$ in $k > \frac{1}{\epsilon}$, fiksen. Sestavimo delitev

$$-\infty < d_1 \leq d_2, ... \leq d_k < \infty$$

realne osi $\mathbb{R}$ tako, da velja

$$F(d_i^-) \leq \frac{i}{k} \leq F(d_i), i = 1, 2, ...k-1,$$

kjer je za vsak i

$$F(d_i^-) = P(X < d_i) = F(d_i) - P(X = d_i).$$

Potem za $d_{j-1}^- < d_j$ velja

$$F(d_j^-) - F(d_j) = \frac{j}{k} - \frac{j-1}{k} = \frac{1}{k} < \epsilon.$$

Spomnimo se definicije empirične porazdelitvene funkcije za slučajni vzorec S_n:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}}.$$

Po krepkem zakonu velikih števil, ko $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i < x\}} \to E(\mathbb{1}_{\{X_i < x\}}) = P(X_i < x) = F(x),$$

skoraj gotovo, torej $F_n(x) \to F(x)$ po točkah.
Ko $n \to \infty$ velja
$|F_n(d_i^-) \to F(d_i^-)| \to 0$ in $|F_n(d_i) \to F(d_i)| \to 0$. Posledično tudi

$$\Delta_n = \max_{i=1,2...k-1} \{|F_n(d_i^-) - F(d_i^-)|, |F_n(d_i) - F(d_i)|\} \to 0,$$

ko $n \to \infty$. Za poljuben $x \in \mathbb{R}$ naj bo i tak, da $d_{i-1} \leq x < d_i$. Od tod sledi

$$F_n(x) - F(x) \leq F_n(d_i^-) - F(d_{i-1}) \leq F_n(d_i^-) - F(d_i^-) + \epsilon$$

$$F_n(x) - F(x) \geq F_n(d_{i-1}) - F(d_i^-) \geq F_n(d_{i-1}) - F(d_{i-1}) - \epsilon.$$

Posledično za poljuben $x \in \mathbb{R}$ velja

$$F_n(d_{i-1}) - F(d_{i-1}) - \epsilon \leq F_n(x) - F(x) \leq F_n(d_i) - F(d_i) + \epsilon$$

in

$$|F_n(x) - F(x)| < \Delta_n + \epsilon \to \epsilon, n \to \infty$$

skoraj gotovo. Ker konvergenca velja za poljuben x, velja tudi

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to \epsilon, n \to \infty$$

skoraj gotovo. Zgornje lahko zapišemo kot:

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

Označimo z A_ϵ množico $\omega \in \Omega$, za katere drži zgornja konvergenca. Potem je $P(A_\epsilon) = 1$.
Konvergenca velja za vsak $\epsilon > 0$.
Definiramo $A = \bigcap_{\epsilon > 0} A_\epsilon$, tedaj za A prav tako velja

$$P(A) = P(\bigcap_{\epsilon > 0} A_\epsilon) = \lim_{\epsilon \to 0} P(A_\epsilon) = 1.$$

Od tod sledi

$$P[\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|=0]=1$$

□

Lema 4.10. *Naj bodo $(X_i)_{i\geq 1}$ zaporedje neodvisnih enako porazdeljenih slučajnih spremenljivk s porazdelitvenim zakonom $f\in\Gamma_p$, $p\in\{1,2\}$. Naj bo f_n porazdelitveni zakon, ki ga določa empirična porazdelitvena funkcija slučajnega vzorca $(X_1,...X_n)$. Potem je $d_p(f_n,f)\to 0$, $p\in\{1,2\}$.*

Dokaz. Pri trditvi 4.9 smo že pokazali, da $F_n(x)\to F(x), n\to\infty$ po točkah, torej v porazdelitvi. Ker kumulativna porazdelitvena funkcija enolično določa porazdelitveni zakon, velja enako za porazdelitvena zakona.
Ker imajo X in X_i končen 1. in 2. moment, so njihovi porazdelitveni zakoni iz Γ_p. Za uporabo leme 4.4 moramo pokazati še

$$\int|x|^p f_n(dx)\to\int|x|^p f(dx).$$

Naj bo $p=2$. Potem je $\int|x|^p f(dx)<\infty$ po predpostavki. Z B označimo množico $\omega\in\Omega$ za katere $\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n(x)-F(x)|\neq 0$. Po trditvi 4.9 je $P(B)=0$.

$$\lim_{n\to\infty}(\int_\Omega|x|^p f_n(dx)-\int_\Omega|x|^p f(dx))=$$

$$=\lim_{n\to\infty}(\int_B|x|^p f_n(dx)-\int_B|x|^p f(dx)+\int_{B^c}|x|^p f_n(dx)-\int_{B^c}|x|^p f(dx))=^*$$

$$=\lim_{n\to\infty}(0-0+\int_{B^c}|x|^p f_n(dx)-\int_{B^c}|x|^p f(dx))=^{**}$$

$$=0.$$

* integral po množici z mero 0 je enak 0.
**upoštevamo trditev 4.9 in enakomerno integrabilnost $g(x)=|x|^p$. □

Sedaj lahko dokažemo naslednji izrek:

Izrek 4.11. *Naj bo $X_1,X_2,...$ zaporedje neodvisnih enako porazdeljenih slučajnih spremenljivk s končnima μ in σ^2. Naj bo $S_n=(X_1,X_2,...X_n)$. Potem pri danem vzorcu S_n, ko $n\to\infty$ in $R\to\infty$ velja:*

a) Pogojna porazdelitev $\sqrt{R}(\mu_R^*-\mu_n)|S_n$ konvergira v porazdelitvi k $N(0,\sigma^2)$,

b) $s_n^*\to\sigma$ v pogojni verjetnosti, to je za vsak $\epsilon>0$ velja

$$P(|\ s_n^*-\sigma\ |>\epsilon|X_1,X_2,..X_n)\to 0$$

.

Dokaz. Dokažimo točko a).

Naj bo f porazdelitveni zakon za slučajno spremenljivko X, ki je reprezentativna slučajna spremenljivka za zaporedje $(X_i)_{i\geq 1}$. Naj bo f_n porazdelitveni zakon, ki ga določa empirična porazdelitvena funkcija vzorca S_n.
V oznakah 4.6 je $f_n^{(R)}$ porazdelitveni zakon slučajne spremenljivke $\sqrt{R}(\mu_R^*-\mu_n)|S_n$. Vemo:

- $f_n^{(R)}$ je porazdelitveni zakon $\sqrt{R}(\mu_R^* - \mu_n)$ pogojno na S_n,
- po lemi 4.10 velja $d_p(f_n, f) \to 0$,
- potem je po lemi 4.7: $d_2(f_n^{(R)}, f^{(R)}) \leq d_2(f_n, f) \to 0$.

Od tod sledi:

- $f^{(R)}$ konvergira proti $N(0, \sigma^2)$ v porazdelitvi po centralnem limitnem izreku.
- $f_n^{(R)} \to f^{(R)}$ v porazdelitvi in v 1. in 2. momentu.
- $\sqrt{R}(\mu_R^* - \mu_n) \to N(0, \sigma^2)$ v porazdelitvi, po prejšnjh dveh točkah.

Pokažimo še točko b).
Označimo z $f_n^{(n)*}$ porazdelitveni zakon za $\frac{1}{n}\sum_{i=1}^n X_i^*$ in z $f^{(n)*}$ porazdelitveni zakon za $\frac{1}{n}\sum_{i=1}^n X_i$.

- Po lemi 4.7 velja $d_2(f^{(n)*}, f_n^{(n)*}) \leq d_2(f, f_n)$,
- po lemi 4.10 pa velja $d_2(f, f_n) \to 0$.

Torej $f^{(n)*}$ in $f_n^{(n)*}$ konvergirata v 2. momentu po lemi 4.4.
Naj bo $\epsilon > 0$, tedaj je

$$P(\mid s_n^* - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n) = P(\mid s_n^* + s_n - s_n - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n) \leq$$

$$P(\mid s_n^* - s_n \mid > \epsilon \mid X_1, X_2, ...X_n) + P(\mid s_n - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n).$$

Posebej obravnavamo:

- $P(\mid s_n^* - s_n \mid > \epsilon \mid X_1, X_2, ...X_n) \to 0$,
  ker $\frac{1}{n}\sum_{i=1}^n X_i^*$ konvergira proti $\frac{1}{n}\sum_{i=1}^n X_i$ v Mallowsovi metriki, torej tudi v drugem momentu, posledično v verjetnosti. Funkcija $g(x) = \sqrt{x}$ je zvezna, od tod sledi $s_n^* \to s_n$ verjetnostno.

- $P(\mid s_n - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n) \to 0$, saj

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

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

  Po krepkem zakonu velikih števil $\overline{X} \to \mu$ in po krepkem zakonu velikih števil $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \to E((X_i - \mu)^2) = \sigma^2$. Kot prej $g(x) = \sqrt{x}$ je zvezna, od tod sledi $s_n \to \sigma$ verjetnostno.

$P(\mid s_n^* - s_n \mid > \epsilon \mid X_1, X_2, ...X_n) \to 0$ in $P(\mid s_n - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n) \to 0$, ko $n \to 0$ in $R \to 0$, posledično tudi $P(\mid s_n^* - \sigma \mid > \epsilon \mid X_1, X_2, ...X_n)$ konvergira proti 0. □

Posledica 4.12. *Naj bo f porazdelitveni zakon normalno porazdeljene slučajne spremenljivke.*
Na podlagi dokaza izrek 4.11 pridemo do naslednje ugotovitve: za vsako $\tilde{f}_n \in \Gamma_2$ iz okolice f_n, ki konvergira k f v d_p-metriki, izrek 4.11 še vedno drži. To pa pomeni, da lahko namesto empirične porazdelitvene funkcije predpostavimo parametrični model, če le-ta zadošča predpostavkam izreka in konvergira proti f. Po lemi 4.4 je dovolj, če velja $\tilde{f}_n \to f$ v porazdelitvi in $\int |x|^2 \tilde{f}_n(dx) \to \int |x|^2 f(dx)$.

4.3. **Konvergenca metode samovzorčenja za zvezne preslikave.**

Z izrekom 4.11 smo pokazali, da μ_R^* konvergira proti μ populacije. Z metodo samovzorčenja želimo ocenjevati parametre porazdelitev in funkcije parametrov. Zanimajo nas pogoji, ki jim moramo zadostiti, da bo metoda samovzorčenja ocenjevala iskano vrednost.

Naj bodo X_i neodvisne enako porazdeljene slučajne spremenljivke. Vemo, da $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ konvergira proti $N(0,1)$ po centralnem limitnem izreku. Konvergenco lahko razširimo na naslednji način:
a) Naj bo $h : \mathbb{R} \to \mathbb{R}$, $\overline{h}(X) = \frac{1}{n}\sum_{i=1}^n h(X_i)$ in $\mu' = E(h(X_1)) < \infty$.
b) Naj bo $g : \mathbb{R} \to \mathbb{R}$, odvedljiva.
c) Naj bo $E(h(X_1)^2) < \infty$.
d) Naj bo $\dot{g}(\mu') \neq 0$.
Potem velja naslednja konvergenca v porazdelitvi:

$$\sqrt{n}(g(\overline{h}(X)) - g(\mu')) \to N(0, D^2).$$

Dokaz. Razvijemo $\sqrt{n}(g(\overline{h}(X)) - g(\mu')) = \sqrt{n}\dot{g}(\mu')[\overline{h}(X) - \mu'] + \Delta$.

Slučajna spremenljivka $\overline{h}(X) = \frac{1}{n}\sum_{i=1}^n h(X_i)$ zadošča predpostavkam centralnega limitnega izreka; $h(X_i)$ so neodvisne, enako porazdeljene in imajo končni 1. ter 2. moment, torej je $\overline{h}(X)$ asimptotično normalno porazdeljena.
Predpostavimo, da $\Delta \to 0$ verjetnostno.
Skličemo se na izrek Slutskyja; če $X_n \to X$ v porazdelitvi in $Y_n \to c \in \mathbb{R}$ verjetnostno, potem $X_n + Y_n \to X + c$ in $\frac{X_n}{Y_n} \to \frac{X}{c}$ v porazdelitvi.
Torej $T_n = \sqrt{n}(g(\overline{h}(X)) - g(\mu')) = \sqrt{n}\dot{g}(\mu')[\overline{h}(X) - \mu']$ konvergira k normalno porazdeljeni slučajni spremenljivki s pričakovano vrednostjo 0.

□

Podobno lahko pokažemo za metodo samovzorčenja. Trditev označimo:
-$\tilde{h}(X^*) = \frac{1}{n}\sum_{i=1}^n h(X_i^*)$ in
- $\tilde{r}(X^*) = \frac{1}{n}\sum_{i=1}^n h(X_i^*)$.
Tedaj je

$$\sqrt{n}(g(\tilde{h}(X^*)) - g(\overline{h}(X))) = \sqrt{n}\dot{g}(\mu')(\tilde{h}(X^*) - \overline{h}(X)) + \Delta_2$$

Če zahtevamo, da je r zvezna in $\Delta_2 \to 0$ v pogojni verjetnosti (pogojno na vzorec S_n), potem je $\tilde{h}(X^*) - \overline{h}(X)$ porazdeljena normalno po izreku 4.11 in izreku Slutskyja.
Ugovitev lahko povzamemo v naslednjo trditev, ki temelji na [1]:

Trditev 4.13. *Naj bo dan slučajni vzorec S_n, sestavljen iz realizacij neodvisnih enako porazdeljenih slučajnih spremeljivk s končno disperzijo. Naj bodo funkcije h, g definirane kot prej. Naj velja:*
1) $\Delta_2 \to 0$ v pogojni verjetnosti (pogojno na dan vzorec S_n),
2) naj bo $E(h(X_1)^2) < \infty$.
3)Naj bodo $X_1^*, X_2,^* ...X_R^*$ neodvisno enako porazdeljene pogojno na S_n.
Potem

$$\sqrt{R}(g(\tilde{h}(X^*)) - g(\overline{h}(X))) \to N(0, C^2),$$

kjer je $C^2 = s_n^2\big(\dot{g}(\mu')\big)^2$.

Komentar 4.14. *Naj bo θ vrednost, ki jo ocenjujemo s cenilko $\hat{\theta}$. $\hat{\theta}$ je slučajna spremenljivka, odvisna od vzorca S_n. Na danem vzorcu S_n lahko izračunamo samo eno vrednost $\hat{\theta}$. Za aproksimativno kumulativno porazdelitveno funkcijo cenilke $\hat{\theta}$ bi potrebovali več vzorcev, za katere bi izračunali vrednost cenilke.*
Tu si lahko pomagamo z metodami samovzorčenja; z njo dobimo več vzorcev, na katerih lahko ocenjujemo $\hat{\theta}$.
S pomočjo novih vrednosti sestavimo empirično porazdelitveno funkcijo za $\hat{\theta}$. Kumulativne porazdelitvene funkcije za cenilke, ki zadoščajo trditvi 4.13, lahko ocenjujemo s pomočjo metode samovzorčenja.
S tem dejstvom si bomo pomagali pri ocenjevanju disperzije cenilk in intervalov zaupanja.

4.4. **Ustreznost uporabe metode samovzorčenja.**

Za boljše razumevanje si bomo na naslednjih primerih ogledali, kdaj metoda samovzočenja ne deluje pravilno. Tako bomo pokazali, kdaj moramo biti pri uporabi metode smovzorčenja previdni. Primer 4.15 smo povzeli po viru [6].

Primer 4.15. Dan je vzorec S_n velikosti $n = 10$ neodvisno enako porazdeljenih slučajnih spremenljivk, porazdeljenih enakomerno zvezno na intervalu $[0, c]$. Oceniti želimo c, njegovo cenilko označimo s $\hat{\theta}$. Primer bomo poskusili rešiti z uporabo neparametrične metode samovzorčenja.

Recimo, da uporabimo cenilko $\hat{\theta} = max(X_1, ...X_n)$, to je makismalno vrednost na danem vzorcu. Cenilka ne ustreza predpostavkam trditve 4.13.

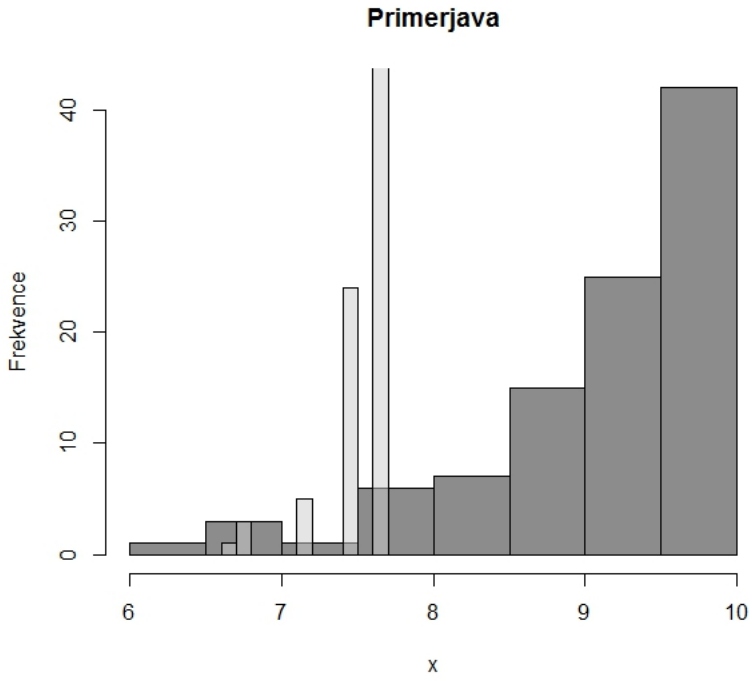
Pogledali si bomo empirično primerjavo porazdelitve cenilke $\hat{\theta}$. Naj bo slučajna spremenljivka X, iz katere smo dobili vzorec S_n, porazdeljena enakomerno zvezno na intervalu $[0, 10]$. Simuliramo 100 novih vzorcev velikosti $n = 10$ in na vsakem uporabimo cenilko $\hat{\theta}$. Vektor ocen nam bo dal aproksimacijo za porazdelitev cenilke $\hat{\theta}$.

Iz originalnega vzorca S_n z neparametrično metodo samovzorčenja tudi sestavimo 100 novih vzorcev in na vsakem novem vzorcu uporabimo cenilko $\hat{\theta}$.

Hitro vidimo, da bo samovzorčna ocena zmeraj manjša ali enaka kot ocena na originalnem vzorcu.

Težava se pojavi, ker je največja vrednost, ki jo lahko dosežemo z empirično porazdelitveno funkcijo, že dosežena na vzorcu S_n. Ustvarjanje novih vzorcev nam ne bo dalo večje vrednosti.

Primerjamo histograma; s svetlosivo barvo je prikazan histogram ocen θ^*_{100}, ki smo ga dobili z neparametrično metodo samovzorčenja, s temnosivo je prikazan histogram ocen dobljenih s pomočjo slučajne spremenljivke X in cenilke $\hat{\theta}$.



Na podlagi primerjave histogramov sklepamo, da neparametriča metoda ne deluje, ker porazdelitveni zakon $max\{X_1^*, X_2^*, ...X_n^*\}$ ne konvergira v porazdelitvi k porazdelitvenemu zakonu $max\{X_1, X_2, ...X_n\}$.
Težavi se lahko izognemo z uporabo druge cenilke; če izberemo cenilko $\hat{\theta} = 2\overline{X}$, metoda deluje pravilno (cenilka ustreza predpostavkam trditve 4.13; vzamemo $g(x) = 2x$).

Največji problem parametrične metode je določitev ustrezne porazdelitve, kot v primeru 2.1. Problem se pojavi predvsem, ko je velikost vzorca premajhna (ne moremo ovrednotiti ustreznosti modela). V tem primeru priporočamo uporabo neparametrične metode. Vendar v skladu za literaturo [3] ne priporočamo uporabe za vzorce velikosti < 10.

5. Pristranskost in disperzija samovzorčne cenilke

Teorija temelji na viru[4].

Definicija 5.1. Naj bo dan slučajni vzorec S_n iz neke populacije in naj bo $\theta \in \mathbb{R}$ iskana količina na tej populaciji. Pravimo, da je $\hat{\theta}$ *nepristranska* cenilka za θ, če velja:

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

Če je $\hat{\theta}$ pristranska, definiramo *pristranskost cenilke* kot:

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

Prav tako definiramo *disperzijo cenilke* za θ kot

$$\nu(\hat{\theta}) = D(\hat{\theta}).$$

Če s $\hat{\theta}$ označimo poljubno cenilko za θ, lahko $\hat{\theta}$ zapišemo kot funkcijo vzorca, torej $\hat{\theta} = f(S_n)$. V splošnem je takšna cenilka lahko pristranska

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

V primeru, ko poznamo "obliko pristranskosti" $b(\hat{\theta})$, tj. predpis za $b(.)$, lahko vrednost cenilke $\hat{\theta}$ zmanjšamo s cenilko za pristranskost. Pristranskost lahko odstranimo v celoti, če je cenilka za pristranskost nepristranska, saj takrat z oceno ne dobimo nobene dodatne napake. Težave nastopijo takrat, ko ne poznamo oblike funkcije $b(.)$ kot v primeru 2.1. Z metodo samovzorčenja se težavi izognemo, saj zanjo ne potrebujemo predpisa $b(.)$.
Označimo s θ_R^* samovzorčno cenilko za θ, ki ustreza pogojem trditve 4.13. Ko $R \to \infty$ velja

$$\theta_R^* \to E(\theta_R^*) = \hat{\theta} = \theta + b(\hat{\theta}),$$

kjer je $\hat{\theta}$ cenilka, s katero smo ocenili θ_i na posameznem novem vzorcu (dobljenem z metodo samovzorčenja). θ_R^* je nepristranska za $\hat{\theta}$, saj jo dobimo kot vzorčno povprečje θ_i.
Naslednja postopka najdemo tudi v [8].

Definicija 5.2. Pri metodi samovzorčenja definiramo *samovzorčno cenilko za pristranskost* kot

$$b(\theta_R^*) := \theta_R^* - \hat{\theta}.$$

Pristranskost cenilke z metodo samovzorčenja ocenjujemo po naslednjem postopku.

Postopek 5.3.
-Sestavimo samovzorčno porazdelitveno funkcijo F_{S_n}.
-Iz samovzorčne porazdelitvene funkcije s ponavljanjem ustvarimo nove vzorce S_i^* za $i = 1, 2, ...R$.
-Na vsakem novem vzorcu ocenimo iskano vrednost θ.
-Ocene povprečimo, dobimo θ_R^* in od θ_R^* odštejemo s $\hat{\theta}$ ocenjeno vrednost originalnega vzorca $\hat{\theta}$.
-Dobljena vrednost je ocena za pristranskost cenilke $b(\theta_R^*)$.

Oznaka 5.4. *Naj bo dan vzorec S_n, na katerem ocenjujemo $\nu(\hat{\theta})$. Tedaj bomo z*

$$\nu(\theta_R^*) = \frac{1}{R-1}\sum_{i=1}^{R}(\theta_i - \theta_R^*)^2$$

označili disperzijo cenilke $\hat{\theta}$ ocenjene z metodo samovzorčenja. Pri tem je θ_i vrednost cenilke na vzorcu $S_{n,i}^*$, $i = 1, 2, ...R$.

Zapišimo še postopek za ocenjevanje disperzije.

Postopek 5.5.
-Sestavimo samovzorčno porazdelitveno funkcijo F_n.
-Ustvarimo nove vzorce iz samovzorčne porazdelitvene funkcije s ponavljanjem.
-Na vsakem generiranem vzorcu izračunamo vrednost cenilke $\hat{\theta}$.
-Izračunamo samovzorčno cenilko θ_R^*.
-Izračunamo disperzijo cenilke $\nu(\theta_R^*)$.

5.1. Izračun pristranskosti in disperzije cenilke θ iz ključnega primera.

Vrnimo se na primer 2.1.
Zanima nas pristranskost izbrane cenilke $\overline{X}/\overline{Y}$.
Ko smo pristranskost pokusili izračunati direktno preko $E(\hat{\theta})$, smo naleteli na težave. Težave se pojavijo tudi pri izračunu disperzije. Pristranskost in disperzijo zato ocenimo z metodo samovzorčenja.
Pristranskost in disperzijo cenilke bomo ocenili s postopkoma 5.3 in 5.5.

Oceni pristranskosti in disperzije sta 0.003 in 0.001.

5.2. Ocenjevanje mediane z neparametrično metodo samovzorčenja.

Primer 5.6. Generiramo slučajni vzorec velikosti $n = 1001$ iz Cauchyjeve porazdelitve, katere gostota je

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2};$$

Zanima nas vrednost mediane na vzorcu in pristranskost ter disperzija cenilke za mediano.

Z neparametrično metodo samovzorčenja $R = 100$ dobimo:
Vrednost: 0.020
Pristranskost: -0.390
Disperzija: 0.002

Iz gostote porazdelitve vidimo, da je prava vrednost mediane enaka 0 (simetričnost gostote z vrhom v 0). Če ocenimo mediano z neparametrično metodo samovzorčenja, dobimo $\theta_{100}^* = 0.020$. Vidimo, da je izračunana vrednost blizu pričakovane vrednosti mediane, ki je enaka 0.
Empirični oceni, dobljeni iz 100 vzorcev velikosti $n = 1001$ iz Cauchyjeve porazdelitve:

Mediana: -0.001
Disperzija mediane: 0.002

Mediana je primer cenilke, ki ne zadošča pogojem trditve 4.13, vendar neparametrična metoda za mediano deluje pravilno.
Primer dokazuje, da je trditev 4.13 zadostna, ni pa potrebna za pravilno delovanje metode.
Izpeljave eksaktnih formul za pristranskost in disperzijo cenilk so lahko v določenih primerih zahtevne (primer: mediana). Prav tako so lahko zahtevni izračuni vrednosti za pristranskost (ko točne vrednosti mediane ne poznamo) in disperzijo. Z metodo samovzorčenja pa relativno hitro in idejno preprosto pridemo do dobrih ocen za iskane vrednosti.

Do sedaj smo se ukvarjali s točkovno oceno za θ. Sedaj pa se bomo ukvarjali z intervalsko oceno za θ, torej z intervali zaupanja. Privzamemo, da so intervali zaupanja simetrični.

Definicija 6.1. Naj bo dan slučajni vzorec S_n. Tedaj interval

$$IZ_\beta(\theta) = [L(S_n), U(S_n)],$$

za katerega velja $P(\theta \in [L(S_n), U(S_n)]) = \beta$, imenujemo *interval zaupanja za* θ *stopnje zaupanja* β.

Pri določanju intervalov zaupanja se lahko pojavijo podobne težave kot pri ocenjevanju θ. Največja težava je, da ne poznamo porazdelitve $\hat{\theta}$. Težave se lahko pojavijo tudi, ko je vzorec premajhen za izpolnjevanje konvergenčnih kriterijev. Zlasti se omenjene težave pojavijo, ko statističnih testov ne moremo izvesti ali ne dajo zadovoljivega odgovora. V takšnih primerih je smiselno uporabiti neparametrično metodo samovzorčenja.

6.1. **Percentilna metoda.**

Pri neparametrični metodi samovzorčenja empirična porazdelitvena funkcija F_n pri danem vzorcu S_n nadomesti teoretično porazdelitveno funkcijo F. Najbolj preprost način, kako uporabiti metodo samovzorčenja za iskanje intervalov zaupanja, je naslednji:
z empirično porazdelitveno funkcijo generiramo nove vzorce in izračunamo $\hat{\theta}$ na vsakem novem vzorcu. Izračunane vrednosti θ_i lahko ponovno zapišemo v empirično porazdelitveno funkcijo cenilke $\hat{\theta}$. Slednja nam da aproksimacijo za kumulativno porazdelitveno funkcijo $\hat{\theta}$. Vrednost $\frac{1-\beta}{2}R$ ustrezno zaokrožimo na celo število. Če torej razvrstimo dobljene vrednosti θ_i po velikosti, odstranimo najmanjše $\frac{1-\beta}{2}R$ vrednosti in največje $\frac{1-\beta}{2}R$ vrednosti, nam preostali najmanjša in največja vrednost predstavljata aproksimativni $IZ_{0.95}(\theta)$. Ko odstranjujemo najmanjše vrednosti zaokrožimo $\frac{1-\beta}{2}R$ navzdol, ko odstranjujemo največje vrednosti pa zaokrožimo $\frac{1-\beta}{2}R$ navzgor.
Pri parametrični metodi je razlika samo v tem, da vzorce ustvarimo s pomočjo ustrezne porazdelitve. V skladu z literaturo za iskanje intervala zaupanja priporočamo vsaj $R \geq 1000$ simulacij [1].

Poglejmo uporabo metode na preprostem primeru.

Primer 6.2. Recimo, da smo 20 osebam izmerili IQ. Dobili smo naslednje podatke:

61, 88, 89, 89, 90, 92, 93, 94, 98, 98,

101, 102, 105, 108, 109, 113, 114, 115, 120, 138.

Oceniti želimo intervala zaupanja stopnje zaupanja 0.95 za povprečje in mediano.

Privzamemo lahko normalno porazdeljen vzorec z neznanim parametrom $\sigma^2 > 0$. Interval zaupanja za povprečje znamo izračunati po formuli $[\overline{X} - t_{19}^{-1}(\frac{1+\beta}{2})\frac{s_n}{\sqrt{n}}, \overline{X} + t_{19}^{-1}(\frac{1+\beta}{2})\frac{s_n}{\sqrt{n}}]$, t označuje Studentovo t porazdelitev. V primeru mediane nimamo obrazcev, s katerimi bi lahko izračunali interval zaupanja, zato bomo intervale zaupanja za mediano ocenili s pomočjo metode samovzorčenja.

Ocena Percentilni interval zaupanja za meadiano $R = 2000$:
Neparametrična metoda [92.5, 108.5]
Parametrična metoda [92.5 109.0]
Percentilni interval zaupanja za pričakovano vrednost:[94.1, 107.8]
Eksaktni interval zaupanja za povprečje je $[93.4, 108.3]$.

Naslednji postopek najdemo v viru [8].

Postopek 6.3.
-Iz podatkov sestavimo samovzorčno porazdelitveno funkcijo F_n.
-Ustvarimo nove vzorce, za vsak vzorec izračunamo $\hat{\theta}$ in označimo vrednost na i-tem vzorcu s θ_i.
-Uredimo izračunane vrednosti po velikosti.
-Odrežemo $\frac{1-\beta}{2}R$ (zaokroženo navzdol) najmanjših vrednosti in $\frac{1-\beta}{2}R$ (zaokroženo navzgor) največjih vrednosti.
-Najmanjša preostala vrednost je naša ocena za L, največja preostala vrednost pa ocena za U.

Komentar 6.4. *Če vrednosti iz θ_i sestavimo v empirično porazdelitveno funkcijo, nam ta ocenjuje dejansko porazdelitveno funkcijo parametra $\hat{\theta}$, po komentarju 4.14. Percentilna metoda za cenilke, ki zadoščajo pogojem iz trditve 4.13, deluje pravilno.*

Oznaka 6.5. *Naj bo dan slučajni vzorec S_n. Aproksimativni interval zaupanja za θ, ki ga dobimo s pomočjo metode samovzorčenja, označimo $IZ^*_\beta(\theta)$.*

Porazdelitveno funkcijo za cenilko $\hat{\theta}$, označimo jo z $F_{\hat{\theta}}$, lahko aproksimiramo z metodo samovzorčenja. Označimo empirično porazdelitveno funkcijo θ^*_R s $F_{\theta^*_R}$. S temi oznakami lahko zapišemo

$$IZ^*_\beta(\theta) = [F^{-1}_{\theta^*_R}(\alpha), F^{-1}_{\theta^*_R}(1-\alpha)],$$

$\alpha = \frac{1-\beta}{2}$ in $1-\alpha = \frac{1+\beta}{2}$. Velja $\beta = 1 - 2\alpha$.

6.2. **Kdaj precentilna metoda odpove.**

Obravnavali bomo dva primera.

Primer 6.6. Na primeru 4.15 smo pokazali, da empirična porazdelitvena funkcija, ki jo dobimo z neparametrično metodo samovzorčenja, za cenilko $\hat{\theta} = max(X_1, X_2, ..., X_n)$ ne konvergira k dejanski porazdelitveni funkciji cenilke $\hat{\theta}$. Ocenimo interval zaupanja za $\hat{\theta}$ stopnje zaupanja 0.95.
Dan je vzorec S_n: 6.07, 1.64, 7.11, 5.84, 8.72, 6.87, 3.20, 4.03, 2.64, 5.63.

Maksimalna vrednost vzorca je 8.72.

Oceno intervala zaupanja naredimo s pomočjo algoritmov iz priloge. Za izračun uporabimo neparametrično percentilno metodo.
Dobimo interval zaupanja $[6.87, 8.72]$.
Vidimo, da je robna vrednost intervala zaupanja vrednost cenilke na originalnem vzorcu S_n. Naredimo 100 nadaljnjih simulacij in pogledamo ali je maksimum del intervala zaupanja.

Vidimo, da je vedno del intervala zaupanja, moral pa bi biti samo v približno 95 simulacijah. Metoda v tem primeru ne deluje pravilno.
V trditvi 4.13 smo za funkcijo g zahtevali odvedljivost. Na naslednjem primeru bomo pokazali, da neparametrična percentilna metoda odpove v točki, kjer g ni odvedljiva. Primer temelji na [5]

Primer 6.7. Naj bo dan slučajni vzorec S_n velikosti $n = 50$, porazdeljen $N(0, \sigma^2)$, z neznanim $\sigma^2 > 0$. Pričakovano vrednost μ bomo ocenjevali s cenilko $\hat{\mu} = \sqrt{|\overline{X}|}$. Vemo, da je točna vrednost $\mu = 0$. Zanima nas interval zaupanja stopnje zaupanja 0.95 za μ.

Izračun $IZ^*_{0.95}(\hat{\mu}), R = 2000$ nam da $[0.14, 1.08]$. Vidimo, da točna vrednost sploh ni del intervala zaupanja. V tem primeru je za 100 dodatnih simulacij pokritost 0. Vidimo, da percentilna metoda tudi v tem primeru ne deluje pravilno.

6.3. **Izračun intervala zaupanja ključnega primera s percentilno metodo.**

Ocenimo še iskani interval zaupanja v primeru 2.1.
Dobimo: [1.180, 1.319].

Povzemimo rezultate, dobljene z neparametrično metodo samovzorčenja, za primer 2.1:
-θ smo ocenili z neparametrično metodo samovzorčenja in dobili vrednost 1.236.
-Izračunali smo pristranskost cenilke 0.003; popravljena vrednost za θ je 1.233.
-Disperzija cenilke je 0.001.
-Interval zaupanja stopnje zaupanja 0.95 za θ je $[1.180, 1.319]$, ki ga popravimo za pristranskost : [1.177, 1.316].

7. Zaključek

V delu diplomskega seminarja smo obravnavali statistične metode samovzorčenja. Ugotovili smo, da so metode samovzorčenja idejno preproste in enostavne za uporabo ter omogočajo dobre ocene iskanih količin, v mnogih primerih, kjer standardne metode odpovejo. Kljub temu, da lahko metode samovzorčenja uporabimo za reševanje številnih problemov, moramo biti pri izbiri metode samovzorčenja previdni. Pokazali smo, da parametrična metoda ni ustrezna, ko ne poznamo ustrezne porazdelitve vzorca. Pri uporabi neparametrične metode pa moramo biti previdni pri izbiri cenilk. V prvem razdelku smo obravnavali primer, ki ga s standardnimi analitičnimi postopki brez dodatnih predpostavk ne moremo rešiti. Razložili smo osnovni princip, po katerem delujejo metode samovzorčenja. Predstavili smo dva primera; primer, ko za ponovno vzorčenje uporabimo empirično porazdelitveno funkcijo danega vzorca in primer, ko za ponovno vzorčenje uporabimo ustrezno porazdelitev.

Dokazali smo, kdaj metoda konvergira, če s pomočjo metod samovzorčenja ocenjujemo povprečje in dokazali konvergenco metode, ko cenilka zadošča predpostavkam trditve 4.13.

Pokazali smo, da lahko s pomočjo metod samovzorčenja ocenjujemo pristranskost in disperzijo cenilke. Pokazali smo tudi, kako z metodo samovzorčenja ocenjujemo intervale zaupanja.

Primer iz prvega razdelka smo v celoti rešili z uporabo neparametrične metode samovzorčenja.

Viri in literatura

[1] A. C. Davison, D. V. Hinkley *Bootstrap Methods and Their Application*, Cambridge University Press, New York, 2008.

[2] P. J. Bickel, D. A. Freedman *Some asymptotic theory for the bootstrap*, Annals of Stat. **9** (1981) 1196–1217
D. A. Stephens,*The Glivenko-Cantelli lema*, [20.9. 2015], dostopno na `http://wwwf.imperial.ac.uk/~das01/MyWeb/M3S3/Handouts/GlivenkoCantelli.pdf`

[3] S. Holmes ,[19.9 2015],dostopno na `http://statweb.stanford.edu/~susan/courses/s208/node11.html`

[4] F.W. Scholz, *The Botstrap Small Sample Properties*, [17.9. 2015], dostopno na `http://www.stat.washington.edu/fritz/Reports/bootstrap-report.pdf`

[5] A. Dasgupta, *The bootstrap*,[19.8. 2015], dostopno na `http://www.stat.purdue.edu/~dasgupta/bootstrap.pdf`

[6] S. Holmes,*More about the theoretical underpinnings of the Bootstrap*,[19.9. 2015], dostopno na `http://statweb.stanford.edu/~susan/courses/s208/node15.html`

[7] *Uniform integrability*, [19.9. 2015], dosopno na `https://en.wikipedia.org/wiki/Uniform_integrability`

[8] k. Singh, M. Xie *Bootstrap: A Statistical Method*, [17.9. 2015] dostopno na`http://www.stat.rutgers.edu/home/mxie/rcpapers/bootstrap.pdf`

8. Priloga

Osnovni algoritmi s pomočjo katerih smo izračunali vrednosti primerov:

```
#Pomožna funkcija, ki nam ustvari nove vzorce
generiraj<-function(vzorec, funkcija, R, model=0){
  vektor<-c()
  if (class(model)=="function"){#Parametričen način
    for (i in 1:R){
      nov_vzorec<-model(n)
      vektor<-c(vektor,funkcija(nov_vzorec))
    }
    return(vektor)
  }
  if(model==0){#Neparametričen način
    for (i in 1:R){
      nov_vzorec<-c()
      #Naredi nov vzorec s ponavljanjem
      if (class(vzorec)=="numeric"){
        nov_vzorec<-sample(vzorec,length(vzorec), replace=T)
      }
      if (class(vzorec)=="data.frame" ){
       #Vzorci naj  bodo navpično neodvisni
        slovar<-1:length(vzorec[,1])
        premesaj<-sample(slovar,length(vzorec[,1]), replace=T)
        nov_vzorec<-c()
        for (j in 1:length(premesaj)){nov_vzorec<-rbind(nov_vzorec,vzorec[premesaj[
      }

      #Izračuna funkcijo na novo ustvarjenem vzorcu
      vektor<- c(vektor,funkcija(nov_vzorec))
```

```
    }
    return(vektor)
  }
  else{print("Prosim vnesite primeren model")}

}

#OSNOVNI ALGORITMI
#Neparametrična metoda samovzorčenja
neparametricno<- function(vzorec,funkcija, R){
  return(mean(generiraj(vzorec, funkcija, R)))
}

#Parametrična metoda samovzorčenja
parametricno<- function(vzorec,funkcija,R, model){

  return(mean(generiraj(vzorec,funkcija,R, model)))
}

#Algoritem za disperzijo cenilke
disperzija<-function(vzorec, funkcija, R, model=0){
  return(var(generiraj(vzorec, funkcija, R, model=0)))
}

#Algoritem za pristranskost
pristranskost<-function(vzorec, funkcija, R, model=0){
  return(mean(generiraj(vzorec, funkcija, R,
 model=0))-funkcija(vzorec))
}


#Percentilni interval zaupanja
percentilni<-function(vzorec, funkcija,R,model=0, beta=0.95){
  #Ustvari vektor ocenjenih vrednosti
  vektor<-generiraj(vzorec, funkcija,R,model=0)
  #Uredi vektor po velikosti
  vektor<-sort(vektor)
  alpha<-(1-beta)/2
  #Odreze odvecne vrednosti
  vektor<-vektor[floor(alpha*length(vektor)):
                   floor((1-alpha)*length(vektor))+1]
  return(c(vektor[1], vektor[length(vektor)]))
}
```

P

Primer 2.1:

```
#Vzorec najdem v knjičnici boot pod imenom bigcity
library(boot)
```

```
print(neparametricno(bigcity,function(x)
{mean(x[[2]])/mean(x[[1]])},200))
[1] 1.236237
print(pristranskost(bigcity,function(x)
{mean(x[[2]])/mean(x[[1]])},200))
[1] 0.003015721
print(disperzija(bigcity,function(x)
{mean(x[[2]])/mean(x[[1]])},200))
[1] 0.00134043
percentilni(bigcity,function(x)
{mean(x[[2]])/mean(x[[1]])},1000)
[1] 1.180650 1.319012
```

Primer 3.1:

```
vzorec<-c(3 , 5 , 7 , 18 , 43 , 85 , 91 , 98 , 100 , 130 , 230 , 497)

print(neparametricno(vzorec,mean,200))
[1] 108.8196
print(neparametricno(vzorec,sd,200))
[1] 123.5393
print(parametricno(vzorec, mean, 200, function(n){rexp(12,
1/mean(vzorec))}))
[1] 110.6239
print(parametricno(vzorec, sd, 200, function(n){rexp(12,
 1/mean(vzorec))}))
[1] 104.9222
#Deset novih ocen
for (i in 1:10){print(neparametricno(vzorec,sd,200))}
[1] 121.8136
[1] 121.9138
[1] 129.2339
[1] 124.5694
[1] 117.4456
[1] 124.8667
[1] 126.0778
[1] 130.1615
[1] 122.9347
[1] 118.8067
```

Primer3.7:

```
x<-runif(50,0,5)
#Vzorec
print(x)
 [1] 4.9607190 4.1949396 1.8113593 2.2504256 4.8283413
2.5303796 3.1541280 2.5524335 4.9217531 3.2116672
3.6829500 2.6424569 4.5663401 4.2093093 1.0065682
[16] 1.4111448 4.0777739 3.0555277 4.5273223 4.5946027
0.4659029 4.0012775 4.8607204 0.8278078 1.9748345
```

```
 0.4431025 3.5277704 0.8287817 2.4615573 0.3598061
[31] 4.5599692 1.7509793 4.0485467 0.1995119 2.1864290
 4.6257297 0.4437690 1.5915217 1.6584894 4.7209187
 4.1391382 0.7786232 1.4699316 2.4722785 3.3124818
[46] 0.3984912 0.3596698 0.8895832 3.0269489 4.6201356

#Vzorci generirani s parametricno metode samovzorcenja
theta_j<-generiraj(x,var,200,function(n){runif(50,0,2*mean(x))})
#Ocenjena vrednost theta
print(mean(theta_j))
[1] 2.461768
#Disperzija cenilke, izračunana o formuli iz primera 3.7
print(var(theta_j))
[1] 0.09134741
#Vzorci generirani z neparametricno metode samovzorcenja
theta_i<-generiraj(x,var,200)
#Ocenjena vrednost theta
print(mean(theta_i))
[1] 2.47621
#Disperzija cenilke, izračunana o formuli iz primera 3.7
print(var(theta_i))
[1] 0.07605227
#Podatki za prvo tabelo
#Neparametricno
for (i in c(5,10,50,200,1000)){print(neparametricno(x,var,i))}
[1] 2.53272
[1] 2.37965
[1] 2.545257
[1] 2.443757
[1] 2.439022
#Parametricno
for (i in c(5,10,50,200,1000)){print(parametricno
(x,var,i,function(n){runif(50,0,2*mean(x))}))}
[1] 2.533838
[1] 2.370445
[1] 2.374679
[1] 2.428986
[1] 2.444722
#Podate durge  tabele smo izračunali ročno
```

Primer 4.14:

```
vzorec<-runif(10,0,10)
param<-generiraj(vzorec,function(x){max(x)},100)
emp<-c()
for (i in 1:100){emp<-c(emp,max(runif(10,0,10)))}
hist(emp, main="Primerjava", ylab="Frekvence", xlab="x",
col="darkgray")
hist(param, col="lightgrey",add=T)
```

Primer 5.6:

```
 vzorec_4<-rcauchy(1001,0,1)
nov_vzorec<-generiraj(vzorec_4, median, 200)
print(mean(nov_vzorec))
[1] 0.02066222
print(mean(nov_vzorec)-mean(vzorec_4))
[1] -0.3904924
print(var(nov_vzorec))
[1] 0.00206909
vzorec<-c()
for (i in 1:100){vzorec<-c(vzorec,median(rcauchy(1001,0,1)))}
mean(vzorec)
[1] 0.007715004
var(vzorec)
[1] 0.003099303
```

Primer 6.2:

```
#Dan vzorec
vzorec_5<-c(61,88,89,89,90,92,93,
94,98,98,101,102,105,108
,109,113,114,115,120,138)
print(percentilni(vzorec_5,mean,2000))
[1]  94.1 107.8
#Neparametricen izracun
print(percentilni(vzorec_5,median,2000))
[1]  92.5 108.5
#Parametricen izracun, predpostavimo normalno porazdelitev
print(percentilni(vzorec_5,median,2000,
function(n){rnorm(n,mean(vzorec_5),sd(vzorec_5))}))
[1]  92.5 109.0
```

Primer 6.6:

```
#Vzorec
vzorec_6<-c( 6.07, 1.64, 7.11, 5.84, 8.72,
6.87, 3.20, 4.03, 2.64, 5.63)
print(percentilni(vzorec_6, max, 2000))
[1] 6.07 8.72
maks<-max(vzorec_6)
k<-0
for (i in 1:100){x<-percentilni(vzorec_6, max, 2000)
if (x[[1]]<=maks & maks<=x[[2]]){k<-k+1}}
print(k)
[1] 100
```

Primer 6.7:

```
vzorec<-rnorm(50,0,1)
percentilni(vzorec,function(x){sqrt(abs(x))}, 2000)
k<-0
for (i in 1:100){x<-percentilni(vzorec_6, max, 2000)
```

```
if (x[[1]]<= 0& 0<=x[[2]]){k<-k+1}}
print(k)
[1] 0
```