UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Greta Gašparac

Ocenjevanje standardne napake in intervalov zaupanja z metodo bootstrap

DIPLOMSKO DELO

UNIVERZITETNI ŠTUDIJSKI PROGRAM
PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: izr. prof. dr. Erik Štrumbelj

Ljubljana, 2018

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Glavni cilj naloge je proučiti metodo bootstrap in njeno uporabo pri ocenjevanju intervalov zaupanja. Kandidatka naj povzame glavne ideje, prednosti in slabosti metode bootstrap, opiše nekaj najpogostejših načinov ocenjevanja intervalov zaupanja in empirično ovrednoti njihovo uspešnost v praksi. Poudarek naj bo na neparametrični metodi bootstrap.

Rada bi se zahvalila svoji družini za spodbudo, potrpljenje in oporo skozi celoten proces mojega izobraževanja.

Hvala sošolcem in prijateljem za vse lepe trenutke, ki smo jih skupaj preživeli tekom dodiplomskega študija.

Hvala Tei za lektoriranje in Tilnu za najbolj »life-changing« vožnjo v Ljubljano.

Za konec pa bi se rada zahvalila še svojemu mentorju, izr. prof. dr. Eriku Štrumblju. Najlepša hvala za ves čas in trud, za znanje, ki sem ga pridobila skozi skupno sodelovanje, za razumevanje in spodbudne besede, ko je bilo to potrebno, in še posebej za (slabe) šale, ki so me vedno nasmejale.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
CA	classification accuracy	klasifikacijska točnost
CI	confidence interval	interval zaupanja
LOOCV	leave-one-out cross-validation	prečno preverjanje z metodo *izloči enega*
LS	log-score	log-ocena
MSE	mean square error	srednja kvadratna napaka
RF	random forests	naključni gozdovi
RMSE	root mean square error	koren srednje kvadratne napake
SE	standard error	standardna napaka

Seznam uporabljenih simbolov

simbol	**pomen**
F	populacija
$\mathbf{x}$	vektor
I	indikatorska funkcija
P	verjetnost
θ^*, $\mathbf{x}^*$	spremenljivke bootstrap
$\hat{F}$, $\hat{\theta}$	ocenjene vrednosti
$\xrightarrow{a.s.}$	skoraj gotova konvergenca
$\xrightarrow{\mathcal{D}}$	konvergenca v porazdelitvi

Povzetek

Naslov: Ocenjevanje standardne napake in intervalov zaupanja z metodo bootstrap

Avtor: Greta Gašparac

Povzetek: V diplomskem delu predstavimo metodo bootstrap, ki jo uvrščamo v družino metod samovzorčenja ter je preprostejša in bolj intuitivna alternativa tradicionalnim statističnim metodam za ocenjevanje negotovosti. Osredotočimo se na neparametrično različico metode. Opišemo njene lastnosti in jih predstavimo s praktičnimi primeri s področja strojnega učenja – ocenjevanje in primerjava različnih modelov. Izpostavimo tudi šibke točke metode. Predstavimo in primerjamo tri intervale zaupanja bootstrap: standardnega normalnega z uporabo standardne napake bootstrap in dva klasična intervala bootstrap, centilnega in BCa. Pričakovano se v večini primerov najbolje obnese BCa.

Ključne besede: metoda bootstrap, standardna napaka, intervali zaupanja.

Abstract

Title: Boostraping standard errors and confidence intervals

Author: Greta Gašparac

Abstract: We introduce the reader to the bootstrap, a simple and flexible resampling-based alternative for quantifying uncertainty. We describe the basic characteristics of the non-parametric bootstrap and illustrate its practical behaviour with simulations in the context of a typical task in machine learning - estimating and comparing the performance of different prediction models. We also present some of the method's weaknesses. We introduce and compare three standard intervals: the standard normal using bootstrap standard error and two more typical bootstrap confidence intervals, the percentile and the BCa interval. As theory suggests, the BCa performs the best over a wide range of situations.

Keywords: bootstrap, standard error, confidence intervals.

1

Uvod

Empirične raziskave so ključni del znanosti. Rezultati takšnih raziskav imajo vedno neko stopnjo negotovosti in pomembno je, da jo ovrednotimo in navedemo skupaj s končnim rezultatom ter se tako izognemo napačnim sklepom. Običajno to storimo z intervali zaupanja, ki temeljijo na predpostavki o normalnosti porazdelitve podatkov, ali pa s statističnimi preizkusi domnev.

Primer 1.1. Zanima nas pričakovana vrednost cene študentskih bonov v različnih restavracijah po Sloveniji. Recimo, da podatki niso dosegljivi na spletu in ceno izvemo s telefonskim klicem. Ne želimo klicati vseh 490 restavracij[1], zato jih naključno izberemo le 15.

V programskem jeziku `R` trikrat vzorčimo in za vsak vzorec izračunamo pričakovano vrednost. Podatki so v evrih. Spomnimo, da so cene študentskih bonov navzgor omejene s $4,37$ evra.

```
> set.seed(0)
> vzorec1 <- sample(cene, 15)
[1] 3.87 3.20 3.27 2.37 3.37 2.87 2.57 2.60 3.17 0.87 3.57 2.97 4.37 3.90 3.37
> vzorec2 <- sample(cene, 15)
[1] 2.99 3.37 3.00 2.70 3.10 0.60 1.37 2.17 3.90 0.00 2.98 4.00 3.90 3.37 2.50
> vzorec3 <- sample(cene, 15)
[1] 3.27 1.37 4.00 2.00 3.00 3.37 4.37 4.00 2.40 3.00 1.00 2.07 2.20 2.99 3.17

# Pricakovane vrednosti vzorcev
> sapply(list(vzorec1, vzorec2, vzorec3), mean)
[1] 3.089333 2.663333 2.814000

# Standardni interval zaupanja
> std_interval <- function(vzorec, n, alpha) {
    c(mean(vzorec) - qnorm(1-alpha)*(sd(vzorec)/sqrt(n)),
      mean(vzorec) + qnorm(1-alpha)*(sd(vzorec)/sqrt(n)))
 }
> sapply(list(vzorec1, vzorec2, vzorec3), std_interval, 15, 0.025)
         [,1]     [,2]     [,3]
[1,] 2.675523 2.063792 2.324129
[2,] 3.503144 3.262874 3.303871
```

Dobljeni vzorci imajo zelo različne pričakovane vrednosti in negotovost okoli 1 evro pokriva pravzaprav približno 25 % možnih vrednosti, kar je veliko. Velikost vzorca je premajhna in ne moremo se zanesti na centralni limitni izrek.

[1]Podatki so dostopni na spletni strani `https://www.studentska-prehrana.si/sl/restaurant`.

Podatki pa so redko porazdeljeni normalno in uporaba takšnih intervalov zaupanja vodi do napačnih rezultatov (razen če nas zanima pričakovana vrednost in naš vzorec ni premajhen). Tudi druge metode imajo svoje omejitve - nekatere temeljijo na normalnosti podatkov, druge pa so uporabne

samo takrat, kadar nas zanimata pričakovana vrednost ali mediana. Kako pa izmeriti negotovost, če nas zanima katera druga cenilka, recimo korelacijski koeficient ali klasifikacijska točnost? Efron je leta 1979 predstavil rešitev za to težavo, metodo, ki jo je poimenoval *bootstrap* [9].

Metoda bootstrap je osrednja tema tega dela. V slovenščini jo poimenujemo tudi *metoda stremena* ali *metoda samovzorčenja*, vendar bomo v tem delu uporabljali angleško poimenovanje. Z izrazom *metoda bootstrap* bomo poimenovali najbolj osnovno, neparametrično različico metode.

Metodo bootstrap se uporablja za ocenjevanje standardne napake, pristranskosti, računanje intervalov zaupanja in tudi za statistično preverjanje domnev (t. i. testiranje hipotez). Spada v družino metod samovzorčenja. Te so v teoriji prisotne že dolgo, vendar jih zaradi računske zahtevnosti nismo mogli uporabljati v praksi. To nas danes več ne omejuje, vendar z izjemo razširjene uporabe prečnega preverjanja pri strojnem učenju potenciala metod samovzorčenja ne izkoriščamo dovolj. Podobno v svojem članku trdi tudi Hesterberg [13] in zagovarja umestitev takšnih metod v učni načrt. Ta argument je glavna motivacija za to diplomsko delo – metodo želimo predstaviti kot bolj preprosto in intuitivno alternativo tradicionalnim statističnim pristopom. Poleg tega pa je tudi bolj splošno uporabna, saj je ne omejujejo številne predpostavke.

V 2. poglavju pojasnimo glavno idejo metode bootstrap, opišemo njen algoritem in ocenjevanje standardne napake. Poglavje 3 namenimo opisu treh intervalov zaupanja bootstrap: standardnemu, centilnemu in BCa. V 4. poglavju se osredotočimo na matematično ozadje metode in pojasnimo, zakaj deluje. V 5. poglavju navedeno podpremo še s simulacijami. Opazujemo, kolikšno je primerno število vzorcev bootstrap in kako se opisani intervali odrežejo na praktičnih primerih pri različnih velikostih začetnega vzorca. Predstavimo nekaj primerov, v katerih metoda bootstrap odpove.

2

Glavna ideja metode bootstrap

Naj bo F populacija, $\mathbf{x} = (x_1, x_2, ..., x_n)$ slučajni vzorec velikosti n iz te populacije in $\theta = t(\cdot, F)$ neka lastnost populacije, ki nas zanima. Če je θ pričakovana vrednost vzorca, se lahko zanesemo na centralni limitni izrek in predpostavimo, da so naši podatki normalno porazdeljeni, ter za izračun intervalov zaupanja uporabimo oceno $\hat{\theta} = t(\mathbf{x}, F)$, tabelo standardne normalne porazdelitve in formulo za standardno napako pričakovane vrednosti. Če pa nas zanima interval zaupanja mediane ali katere druge statistike, postopek ni tako preprost. Za izračun potrebujemo vzorčno porazdelitev, ki jo načeloma dobimo iz F, če imamo na voljo poljubno mnogo vzorcev in za vsakega posebej izračunamo statistiko $\hat{\theta}$. To pa je v praksi običajno neizvedljivo (ponavljanje eksperimenta je drago, nimamo dovolj časa ...). Tukaj nastopi metoda bootstrap.

Če je $\mathbf{x}$ dovolj velik preprost slučajni vzorec, naj bi dobro odražal lastno-

sti F, in je torej približek naši populaciji. Namesto da bi vzorčili iz populacije F, bomo vzorčili iz ocene populacije $\hat{F}$. Ta se razlikuje glede na različico metode bootstrap, v primeru neparametričnega bootstrapa pa je enaka empirični porazdelitveni funkciji

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=0}^{n} I(x_i = x),$$

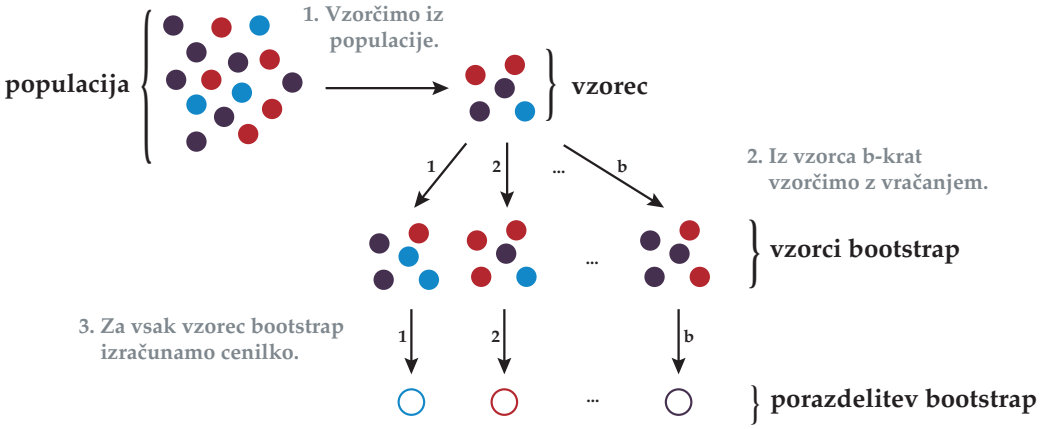
ki v vzorcu $\mathbf{x}$, velikosti n, vsaki točki x_i priredi verjetnost $\frac{1}{n}$. Z drugimi besedami, iz $\mathbf{x}$ naredimo nov vzorec velikosti n s pomočjo vzorčenja z vračanjem.

Dobimo n^n različnih vzorcev $\mathbf{x}^{*1}, \mathbf{x}^{*2}, ..., \mathbf{x}^{*n^n}$, ki jih poimenujemo *vzorci bootstrap*. Za vsakega izračunamo vrednost statistike $\hat{\theta}^*(i) = t(\mathbf{x}^{*i}, \hat{F})$ in tako dobimo *porazdelitev bootstrap*:

$$H^*(x) = P(\hat{\theta}^* \leq x).$$

Porazdelitev bootstrap H^* je pravzaprav vzorčna porazdelitev ocene naše populacije $\hat{F}$, ki ocenjuje neko lastnost prave vzorčne porazdelitve $H(x) = P(\hat{\theta} \leq x)$ iz prave populacije F. Glavni namen metode bootstrap je ovrednotiti negotovost prvotne ocene statistike $\hat{\theta}$.

Če upoštevamo vseh n^n vzorcev, govorimo o *teoretični metodi bootstrap*. Ker pa je to v praksi računsko zahtevno, si izberemo neko število vzorcev b (o primerni velikosti tega števila razpravljamo v podpoglavju 5.1) in vzorčimo b-krat, čemur pravimo *implementacija vzorčenja Monte Carlo* (celoten postopek je grafično predstavljen na sliki 2.1). Tako ustvarimo dve ravni napake. Sprva z vzorcem aproksimiramo populacijo, nato pa ovrednotimo le del vseh možnih vzorcev bootstrap. Napake, ki nastane z vzorčenjem Monte Carlo, sicer ne obravnavamo, saj jo z velikostjo b lahko poljubno zmanjšamo.



Slika 2.1: Shematični prikaz delovanja metode bootstrap.

Definicija 2.1 (standardna napaka metode bootstrap)**.** *Naj bo $\hat{\theta}$ vrednost statistike, ki nas zanima, izračunana iz naših podatkov, b število vzorcev bootstrap in $\hat{\theta}^*(i)$ izračunana statistika iz i-tega vzorca. Potem je standardna napaka bootstrap*

$$\hat{se}_b = \sqrt{\frac{1}{b-1}\sum_{i=0}^{b}\left(\hat{\theta}^*(i) - \hat{\theta}^*(\cdot)\right)^2},$$

pri čemer

$$\hat{\theta}^*(\cdot) = \frac{1}{b}\sum_{i=0}^{b}\hat{\theta}^*(i),$$

kar je dejansko standardni odklon vzorca porazdelitve bootstrap.

Porazdelitev bootstrap je preozka za faktor $\sqrt{(n-1)/n}$, kadar govorimo o pričakovani vrednosti, in za podoben faktor tudi pri drugih statistikah. Preveč optimistične ocene so posledica obravnavanja vzorca kot populacijo. Našo populacijo predstavlja empirična porazdelitvena funkcija, katere varianca je $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=0}^{n}(x_i - \bar{x})^2$. Standardna napaka teoretične metode bootstrap je torej $se_b = \hat{\sigma}/\sqrt{n}$. Ker pa je ta »populacija« v resnici le vzorec, bi varianco običajno računali z izrazom $s^2 = \frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2$ in

dobili standardno napako $se = s/\sqrt{n}$, iz česar sledi

$$\frac{se_b}{se} = \frac{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2}}{\sqrt{n}\sqrt{n}} : \frac{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2}}{\sqrt{n-1}\sqrt{n}} = \sqrt{\frac{n-1}{n}}.$$

Primer 1.1 (nadaljevanje)**.** Na našem primeru uporabimo metodo bootstrap.

```
> vzorec
 [1] 3.87 3.20 3.27 2.37 3.37 2.87 2.57 2.60 3.17 0.87 3.57 2.97 4.37 3.90 3.37

# generiraj b vzorcev bootstrap in izracunaj porazdelitev bootstrap
> b <- 10000
> vzorci_b <- replicate(b, sample(vzorec, replace = T))
> porazdelitev_b <- apply(vzorci_b, 2, mean)

# izracunaj standardno napako bootstrap
> se_b <- sd(porazdelitev_b)
[1] 0.2039123

# izracunaj standardno napako z enacbo za pricakovano vrednost
> se <- sd(vzorec)/sqrt(15)
[1] 0.2111318
```

Standardno napako smo izračunali na dva načina: z metodo bootstrap in z uporabo enačbe za standardno napako pričakovane vrednosti

$$se = \frac{s^2}{\sqrt{n}},$$

pri kateri vzorčni standardni odklon delimo s korenom velikosti vzorca. Standardna napaka bootstrap je res manjša za faktor $\sqrt{14/15}$, s čimer potrdimo prej omenjeno.

Iz primera je razvidno, kako preprosta je implementacija metode bootstrap. Ena izmed prednosti metode je tudi to, da postopek ostaja enak ne glede na statistiko. Če nas zanima kaj drugega, zgolj spremenimo tretji argument funkcije `apply` v želeno funkcijo. Seveda pa ima programski jezik `R` na voljo tudi kar nekaj knjižnic, s katerimi si lahko pomagamo pri eksperimentih.

3

Intervali zaupanja bootstrap

Intervali zaupanja so postali skoraj nepogrešljivi spremljevalci rezultatov v empirični znanosti, saj tako le z dvema dodatnima vrednostma podamo veliko bolj informativno oceno. Od prve objave metode bootstrap leta 1979 je bilo predlaganih kar nekaj različnih metod za izračun mej intervalov zaupanja, ki izkoriščajo prednosti metode bootstrap. Z njimi lahko preprosto, ne da bi nas omejevale številne predpostavke, dobimo meje intervalov tudi za netrivialne statistike. V tem poglavju spoznamo tri intervale zaupanja: standardni interval z uporabo standardne napake bootstrap, ki temelji na centralnem limitnem izreku, in pa dva bolj klasična intervala bootstrap: centilni in njegovo izboljšano različico, interval zaupanja BCa.

Najprej naj navedemo nekaj definicij, ki bodo pripomogle k razumevanju lastnosti opisanih intervalov.

Definicija 3.1. *Pravimo, da so intervali zaupanja enakostranski, če sta verjetnosti, da je θ nad zgornjo mejo $\hat{\theta}_{zg}$ ali pod spodnjo mejo $\hat{\theta}_{sp}$, enaki:*

$$P(\theta < \hat{\theta}_{sp}) = \alpha \quad \text{in} \quad P(\theta > \hat{\theta}_{zg}) = \alpha. \tag{3.1}$$

Definicija 3.2 (natančnost)**.** *Zgornja meja intervala $\hat{\theta}[\alpha]$ je natančna prvega reda, če se verjetnost nepokritega območja od α razlikuje le za neki člen reda velikosti $n^{-1/2}$, kjer je n velikost vzorca:*

$$P(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1/2}).$$

O natančnosti drugega reda govorimo, če je ta razlika le reda velikosti n^{-1}:

$$P(\theta \leq \hat{\theta}[\alpha]) = \alpha + O(n^{-1}).$$

Podobno velja za spodnjo mejo.

Definicija 3.3 (točnost)**.** *Naj bo $\hat{\theta}_t[\alpha]$ prava meja intervala, za katero velja $P(\theta \leq \hat{\theta}_t[\alpha]) = \alpha$. Meja intervala $\hat{\theta}[\alpha]$ je točna prvega reda, če se od prave meje intervala razlikuje le za neki člen reda velikosti $n^{-1/2}$, kjer je n velikost vzorca:*

$$\hat{\theta}[\alpha] = \hat{\theta}_t[\alpha] + O(n^{-1/2}).$$

O točnosti drugega reda govorimo, če je ta razlika le reda velikosti $n^{-3/2}$:

$$\hat{\theta}[\alpha] = \hat{\theta}_t[\alpha] + O(n^{-3/2}).$$

Trditev 3.1. *Točnost i-tega reda implicira natančnost tega reda [10, 12].*

Izraza definicij 3.2 in 3.3 sta definirana enostransko (vendar navedeno velja tudi za posamezne meje enakostranskega dvostranskega intervala). Ko v nadaljevanju govorimo o natančnosti in točnosti, se pojma vedno navezujeta na posamezne meje intervala. Stremimo k uporabi intervalov zaupanja, ki so točni drugega reda, saj z njimi v limiti, ko $n \to \infty$, dobimo boljše rezultate.

Definicija 3.4 (invariantnost)**.** *Interval zaupanja je invarianten, če za vsako monotono transformacijo parametra $\phi = m(\theta)$ velja*

$$[\hat{\phi}_{sp}, \hat{\phi}_{zg}] = [m(\hat{\theta}_{sp}), m(\hat{\theta}_{zg})],$$

pri čemer sta $\hat{\theta}_{sp}$ in $\hat{\theta}_{zg}$ spodnja in zgornja meja intervala.

Invariantnost nam omogoča preprosto izračunati meje intervalov zaupanja za različne transformacije podatkov, ne da bi postopek vzorčenja znova ponavljali.

Definicija 3.5 (ohranjanje razpona)**.** *Če je parameter omejen na interval $[a, b]$ in interval zaupanja to upošteva (meje intervala so v zalogi vrednosti parametra), pravimo, da tak interval ohranja razpon.*

Pri nekaterih statistikah je pomembno, da interval zaupanja ohranja razpon, saj se tako izognemo nesmiselnim mejam intervalov.

Ob opisu vsakega intervala zaupanja dodamo lastno implementacijo intervala v programskem jeziku R. Lastnosti intervalov povzamemo v tabeli 3.1 na koncu poglavja.

3.1 Standardni interval zaupanja

Naj bo $\hat{\theta}$ ocena parametra θ, ki nas zanima, $\hat{se}_b$ ocenjena standardna napaka metode bootstrap in $100(1-2\alpha)$ % želena stopnja zaupanja. Z $z^{(\alpha)}$ označimo 100α. centil standardne normalne porazdelitve:

$$P(z^{(\alpha)} \leq \frac{\hat{\theta} - \theta}{\hat{se}_b} \leq z^{(1-\alpha)}) = 1 - 2\alpha.$$

Meje intervala so torej naslednje:

$$[\hat{\theta} - z^{(1-\alpha)} \cdot \hat{se}_b, \hat{\theta} - z^{(\alpha)} \cdot \hat{se}_b].$$

Ker pa je normalna porazdelitev simetrična, velja $z^{(\alpha)} = -z^{(1-\alpha)}$ in zapis lahko poenostavimo v

$$\hat{\theta} \pm z^{(1-\alpha)} \cdot \hat{se}_b.$$

Interval je konstruiran s predpostavko, da so naši podatki porazdeljeni normalno in je simetričen. Zaradi te lastnosti so meje standardnega intervala lahko nenatančne, kadar imamo opraviti z nagnjenimi porazdelitvami. Lahko si pomagamo s transformacijo podatkov, ki porazdelitev $\hat{\theta}^*$ normalizira, izračunamo meji intervala in jih potem transformiramo nazaj, vendar se hitro pojavi težava, saj nam taka transformacija ni vedno znana.

Standardni interval je točen in natančen prvega reda, ni invarianten in ne ohranja razpona [7, 10]. Kadar je naš vzorec majhen, je lahko preveč optimističen. Če velja $n \leq 20$, namesto tabele standardne normalne porazdelitve običajno uporabimo tabelo t-porazdelitve z $n-1$ prostostnimi stopnjami in dobimo interval:

$$\hat{\theta} \pm t_{n-1}^{(1-\alpha)} \cdot \hat{se}_b,$$

ki je malo širši od standardnega normalnega. Z naraščanjem n Studentova t-porazdelitev konvergira k standardni normalni porazdelitvi (glej prilogo A), zato se intervala posledično vedno manj razlikujeta.

Primer 1.1 (nadaljevanje)**.** Izračunajmo standardni interval za naš primer.

```
# Poiscemo 100(1-alpha). centil normalne porazdelitve, ga pomnozimo s
# standardno napako bootstrap in pristejemo/odstejemo pricakovani vrednosti vzorca.
> alpha <- 0.025
> theta_sp <- mean(vzorec) - qnorm(1 - alpha) * se_b
[1] 2.689549
> theta_zg <- mean(vzorec) + qnorm(1 - alpha) * se_b
[1] 3.489117
```

3.2 Centilni interval zaupanja

Definicija 3.6. *[14] Naj bo F porazdelitvena funkcija, zvezna z desne, u in x pa naj bosta realni števili, za kateri velja* $u \in [0,1]$ *in* $x \in (-\infty, \infty)$*. Potem je*

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\},$$

inverzna funkcija F oz. t. i. kvantilna funkcija, če velja:

$$Q(u) \leq x.$$

Naj bo $\hat{G}$ kumulativna porazdelitvena funkcija parametra $\hat{\theta}^*$. Po definiciji 3.6 velja $\hat{G}^{-1}(\alpha) = \hat{\theta}^{*(\alpha)}$, kar je 100α. centil porazdelitve bootstrap. Centilni interval s stopnjo zaupanja $100(1-2\alpha)$ % je torej

$$[\hat{G}^{-1}(\alpha),\, \hat{G}^{-1}(1-\alpha)] = [\hat{\theta}^{*(\alpha)},\, \hat{\theta}^{*(1-\alpha)}].$$

Pri centilnem intervalu nimamo predpostavk o porazdelitvi podatkov. Je točen in natančen prvega reda, invarianten ter ohranja razpon [10, 12]. Če ima $\hat{G}$ normalno porazdelitev, potem standardni in centilni interval vračata enake rezultate.

Lema 3.1. *Naj transformacija* $\hat{\phi} = m(\hat{\theta})$ *vrne normalno porazdelitev*

$$\hat{\phi} \sim N(\hat{\phi},\, c^2)$$

za neki standardni odklon c. Potem je centilni interval parametra $\hat{\theta}$ *enak*

$$[m^{-1}(\hat{\phi} - z^{(1-\alpha)}c),\, m^{-1}(\hat{\phi} - z^{(\alpha)}c)].$$

Če obstaja transformacija, ki našim podatkom da normalno porazdelitev s stabilno varianco, potem centilni interval deluje pravilno (dokaz najdemo v [18]). V nasprotnem primeru tega zagotovila nimamo. Centilni interval odpravlja pomanjkljivosti standardnega intervala, pri katerem moramo poznati ustrezno transformacijo, kadar podatki niso porazdeljeni normalno, še vedno pa ne more odpravljati napak, ki nastanejo, če je naša

ocena $\hat{\theta}$ pristranska. To pa upošteva naslednji opisani interval.

Primer 1.1 (nadaljevanje)**.** Izračunajmo centilni interval za naš primer.

```
# Uredimo tocke porazdelitve bootstrap in vzamemo b*alpha. ter b*(1-alpha). tocko.
> porazdelitev_b <- sort(porazdelitev_b)
> theta_sp <- porazdelitev_b[b * alpha]
[1] 2.664667
> theta_zg <- porazdelitev_b[b * (1 - alpha)]
[1] 3.462667
```

3.3 Interval zaupanja BCa

Izboljšana različica centilnega intervala je interval zaupanja BCa. Meje intervala še vedno sestavljajo centili, vendar so ti odvisni od oblike porazdelitve bootstrap. Kratica BCa namreč pomeni *bias-corrected and accelerated*, interval je odvisen od parametra pospeška $\hat{a}$ in faktorja $\hat{z}_0$, ki izničuje vpliv pristranskosti naše ocene $\hat{\theta}$.

Naj bo $z^{(\alpha)}$ 100α. centil standardne normalne porazdelitve in $\Phi(\cdot)$ standardna normalna kumulativna porazdelitvena funkcija. Interval zaupanja BCa s stopnjo zaupanja $100(1-2\alpha)$ % je enak

$$[\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}],$$

pri čemer

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

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

Parameter $\hat{z}_0$ meri neskladje med mediano $\hat{\theta}^*$ in vrednostjo $\hat{\theta}$ ter je enak 0 takrat, ko je natančno polovica vrednosti $\hat{\theta}^*(b)$ manjša ali enaka $\hat{\theta}$:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{1}{b}\sum_{i=0}^{b} I(\hat{\theta}^*(i) < \hat{\theta})\right).$$

Parameter pospeška $\hat{a}$ ocenjuje hitrost spremembe standardne napake ocene $\hat{\theta}$ glede na pravo vrednost parametra θ:

$$\hat{a} = \frac{\sum_{i=0}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left(\sum_{i=0}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\right)^{3/2}},$$

pri čemer je $\hat{\theta}_{(i)}$ ocena *jackknife*[1] naše statistike in $\hat{\theta}_{(\cdot)} = \sum_{i=0}^{n}\hat{\theta}_{(i)}/n$. Opazimo, da je v primeru, ko sta $\hat{z}_0$ in $\hat{a}$ enaka 0, interval BCa enak centilnemu.

Interval zaupanja BCa je točen in natančen drugega reda, invarianten ter ohranja razpon [7, 10, 12]. Efron in Tibshirani [10] ga navajata kot najbolj zanesljiv interval zaupanja za neparametrične probleme.

[1]Oceno jackknife $\hat{\theta}_{(i)}$ dobimo iz originalnega vzorca $\mathbf{x}$, če pri izračunu izpustimo i-to točko.

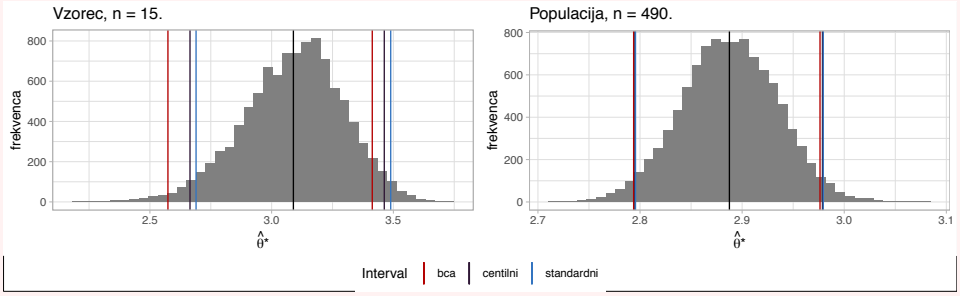
Primer 1.1 (nadaljevanje)**.** Izračunajmo interval BCa za naš primer.

```
# Faktor z0.
> z0 <- qnorm(sum(porazdelitev_b < mean(vzorec))/b)
[1] -0.06521854

Jackknife ocene nasih podatkov.
> h <- rep(0, length(vzorec))
> for (i in 1:length(vzorec) {
+         h[i] <- mean(vzorec[-i])
+ }

# Faktor pospeska a.
> a <- (sum((mean(h) - h)^3)) / (6 * (sum((mean(h) - h)^2))^(3/2))
[1] -0.04830405

# Katere centile bomo uporabili za interval.
> alpha1 <- pnorm(z0 + (z0 + qnorm(alpha)) / (1 - a * (z0 + qnorm(alpha))))
[1] 0.01044421
> alpha2 <- pnorm(z0 + (z0 + qnorm(1 - alpha)) / (1 - a * (z0 + qnorm(1 - alpha))))
[1] 0.9526049

# Uredimo tocke porazdelitve bootstrap.
> porazdelitev_b <- sort(porazdelitev_b)

# Z izracunanimi alphami konstruiramo interval.
> theta_sp <- porazdelitev_b[alpha1 * b]
[1] 2.574
> theta_zg <- porazdelitev_b[alpha2 * b]
[1] 3.413333
```

	Simetričnost	Invariantnost	Ohranjanje razpona	Natančnost/ točnost
Standardni	da	ne	ne	1. reda
Centilni	ne	da	da	1. reda
BCa	ne	da	da	2. reda

Tabela 3.1: Lastnosti intervalov zaupanja bootstrap.

Primer 1.1 (nadaljevanje)**.** Izračunali smo vse meje intervalov, zdaj pa si jih poglejmo še na grafu.



Slika 3.1: Histograma statistike bootstrap $\hat{\theta}^*$ vzorca in populacije. S črno navpično črto na sredini je označena pričakovana vrednost $\hat{\theta}$.

Vzorec. Na levem histogramu slike 3.1 opazimo rahlo nagnjenost v desno, kar lahko razberemo tudi iz negativne vrednosti faktorja z_0 pri intervalu BCa. Standardni interval tega ne upošteva, centilni interval naredi manjši popravek. Standardni in centilni interval sta tudi bolj optimistična od intervala BCa: njuna dolžina je 0,80, dolžina intervala BCa pa je 0,84.

Populacija. Na desnem histogramu slike 3.1, kjer smo uporabili vse podatke, lahko prepoznamo znano obliko normalne porazdelitve – n je dovolj velik, da velja centralni limitni izrek in meje intervalov sovpadajo. Bolj smo prepričani o oceni $\hat{\theta}$, dolžina vseh treh intervalov zaupanja je 0,18.

4

Zakaj metoda bootstrap deluje

Splošno znano je, da večji ko je vzorec, bolj bo podoben populaciji. Enako velja za porazdelitev bootstrap. Metoda bootstrap deluje, če je porazdelitev bootstrap vedno bolj podobna pravi vzorčni porazdelitvi, ko $n \to \infty$. Čeprav se zdi intuitivno, so formalni dokazi zahtevni in različni za različne statistike. Obstaja veliko literature, v kateri se lahko poglobimo v dokaze (npr. [6, 12]), tukaj pa opišemo dokaz delovanja metode bootstrap, kadar nas zanima pričakovana vrednost. Zgledujemo se po [6].

Za začetek definirajmo nekaj pojmov. Zanima nas, kako sploh ocenimo podobnost dveh funkcij. Najprej izberemo ustrezno metriko. V našem primeru je to metrika Kolmogorov-Smirnov, ki je najbolj standardna, lahko pa bi uporabili tudi katero drugo, recimo ℓ^2. Zatem pa moramo vedeti, kaj je želena podobnost oz. k čemu stremimo. Definiramo skoraj gotovo konvergenco.

Definicija 4.1 (metrika Kolmogorov-Smirnov)**.** *Metrika Kolmogorov-Smirnov za dani kumulativni porazdelitveni funkciji F in G je enaka*

$$K(F,G) = \sup_x |F(x) - G(x)|.$$

Definicija 4.2 (skoraj gotova konvergenca)**.** *Zaporedje slučajnih spremenljivk X_n skoraj gotovo konvergira k slučajni spremenljivki X, če velja*

$$P\left(\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right) = 1.$$

To zapišemo $X_n \xrightarrow{a.s.} X$.

Želimo pokazati, da sta si porazdelitev bootstrap in prava vzorčna porazdelitev statistike podobni. Za potrebe razumevanja dokaza pa najprej navajamo nekaj definicij in izrekov, povzamemo jih iz [6].

Definicija 4.3 (konvergenca v porazdelitvi)**.** *Naj bosta X_n in X slučajni spremenljivki, definirani na istem verjetnostnem prostoru. Pravimo, da X_n konvergira v porazdelitvi proti X, če velja $\lim_{n\to\infty} P(X_n \leq X) = P(X \leq x)$, pri čemer sta obe porazdelitvi zvezni. To zapišemo $X_n \xrightarrow{\mathcal{D}} X$.*

Izrek 4.1 (Polya)**.** *Če $F_n \xrightarrow{\mathcal{D}} F$ in je F zvezna kumulativna porazdelitvena funkcija, potem velja $\sup_x |F_n(x) - F(x)| \to 0$, ko $n \to \infty$.* Dokaz najdemo v [15].

Izrek 4.2 (zvezne preslikave)**.** *Naj X_n konvergira proti X in naj bo g zvezna funkcija. Potem tudi $g(X_n)$ konvergira proti $g(X)$.* Dokaz najdemo v [8, str. 87].

Izrek 4.3 (Berry-Esseen)**.** *Za vzorec $X_1, ..., X_n \overset{iid}{\sim} F$ z upanjem $E[X_1] = \mu$, varianco $Var[X_1] = \sigma^2$ in tretjim momentom $E|X_1 - \mu|^3 < \infty$ obstaja univerzalna konstanta c, neodvisna od n ali porazdelitve X_i, da velja:*

$$\sup_x \left| P\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq x\right) - \Phi(x) \right| \leq \frac{c \cdot E|X - \mu|^3}{\sigma^3\sqrt{n}},$$

kjer je Φ kumulativna porazdelitvena funkcija normalne porazdelitve. Dokaz najdemo v [8, str. 118].

Lema 4.1 (zakon o velikih številih Zygmund-Marcinkiewicz)**.** *Naj bodo* $X_1, X_2, ..., X_n \overset{iid}{\sim} F$ *slučajne spremenljivke, za katere velja* $E|X_1|^\delta < \infty$ *za neko spremenljivko* $0 < \delta < 1$*. Potem velja*

$$n^{-1/\delta} \sum_{i=0}^{n} X_i \xrightarrow{a.s.} 0.$$

Dokaz najdemo v [5, str. 356].

Trditev 4.1. *Če ima vzorec* $X_1, ..., X_n \overset{iid}{\sim} F$ *končno varianco, potem porazdelitev spremenljivke* $\sqrt{n}(\overline{X} - \mu)$ *konvergira k normalni, ko* $n \to \infty$*, kar pomeni, da verjetnost* $P(\sqrt{n}(\overline{X} - \mu) \leq x)$ *lahko zapišemo tudi* $\Phi(\frac{x}{s})$*.*

Izrek 4.4. *Imamo vzorec* $X_1, X_2, ..., X_n \overset{iid}{\sim} F$ *s končnim upanjem* $E[X_1^2] < \infty$*. Naj bo* $\hat{\theta} = \sqrt{n}(\overline{X} - \mu)$*. Potem velja* $K(H, H^*) \xrightarrow{a.s.} 0$*, ko* $n \to \infty$*, pri čemer sta*

$$H^*(x) = P(\hat{\theta}^* \leq x) \quad \text{in} \quad H(x) = P(\hat{\theta} \leq x)$$

kumulativni porazdelitveni funkciji statistike bootstrap $\hat{\theta}^*$ *in prave statistike* $\hat{\theta}$*.*

Dokaz.

$$\begin{aligned}
K(H, H^*) &= \sup_x \left| P(\hat{\theta} \leq x) - P(\hat{\theta}^* \leq x) \right| = \\
&= \sup_x \left| P\left(\frac{\hat{\theta}}{\sigma} \leq \frac{x}{\sigma}\right) - P\left(\frac{\hat{\theta}^*}{s} \leq \frac{x}{s}\right) \right| = \\
&= \sup_x \left| P\left(\frac{\hat{\theta}}{\sigma} \leq \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) + \Phi\left(\frac{x}{s}\right) - \right. \\
&\quad \left. - P\left(\frac{\hat{\theta}^*}{s} \leq \frac{x}{s}\right) \right| \leq \\
&\leq \underbrace{\sup_x \left| P\left(\frac{\hat{\theta}}{\sigma} \leq \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) \right|}_{A} + \underbrace{\sup_x \left| \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) \right|}_{B} + \\
&\quad + \underbrace{\sup_x \left| \Phi\left(\frac{x}{s}\right) - P\left(\frac{\hat{\theta}^*}{s} \leq \frac{x}{s}\right) \right|}_{C}.
\end{aligned}$$

Iz izreka 4.1 sledi $A \to 0$. Ker vzorčna varianca s^2 skoraj gotovo konvergira k pravi varianci σ^2 (glej prilogo B), enako velja tudi za vzorčni standardni odklon s in pravi standardni odklon σ (izrek 4.2). Ker je Φ zvezna funkcija, skoraj gotovo lahko trdimo $B \to 0$. Da pokažemo $C \to 0$, uporabimo izrek 4.3:

$$\begin{aligned}
\sup_x \left| \Phi\left(\frac{x}{s}\right) - P\left(\frac{\hat{\theta}^*}{s} \leq \frac{x}{s}\right) \right| &\leq \frac{4}{5\sqrt{n}} \cdot \frac{E|{X_1}^* - \overline{X}_n|^3}{Var[{X_1}^*]^{3/2}} = \\
&= \frac{4}{5\sqrt{n}} \cdot \frac{\sum_{i=0}^{n} |X_i - \overline{X}_n|^3}{ns^3} \leq \\
&\leq \frac{4}{5n^{3/2}s^3} \cdot 2^3 \left(\sum_{i=0}^{n} |X_i - \mu|^3 + n|\mu - \overline{X}_n|^3 \right) = \\
&= \frac{32}{5s^3} \left(\frac{\sum_{i=0}^{n} |X_i - \mu|^3}{n^{3/2}} + \frac{|\overline{X}_n - \mu|^3}{\sqrt{n}} \right).
\end{aligned}$$

Ker vzorčni standardni odklon s skoraj gotovo konvergira proti pravemu standardnemu odklonu σ in povprečje $\overline{X}_n$ skoraj gotovo konvergira proti pričakovani vrednosti μ (zakon o velikih številih), velja $|\overline{X}_n - \mu|^3/(\sqrt{n}s^3) \xrightarrow{a.s.} 0$. Tudi prvi člen gre proti 0. Naj bo $\delta = 2/3$, potem velja

$$E|X_i - \mu|^{3 \cdot 2/3} = Var[X_1] < \infty.$$

Iz leme 4.1 sledi

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} |X_i - \mu|^3 \xrightarrow{a.s.} 0,$$

ko $n \to \infty$.

Pokazali smo, da velja $A + B + C \xrightarrow{a.s.} 0$, iz česar sledi $K(H, H^*) \xrightarrow{a.s.} 0$. □

5

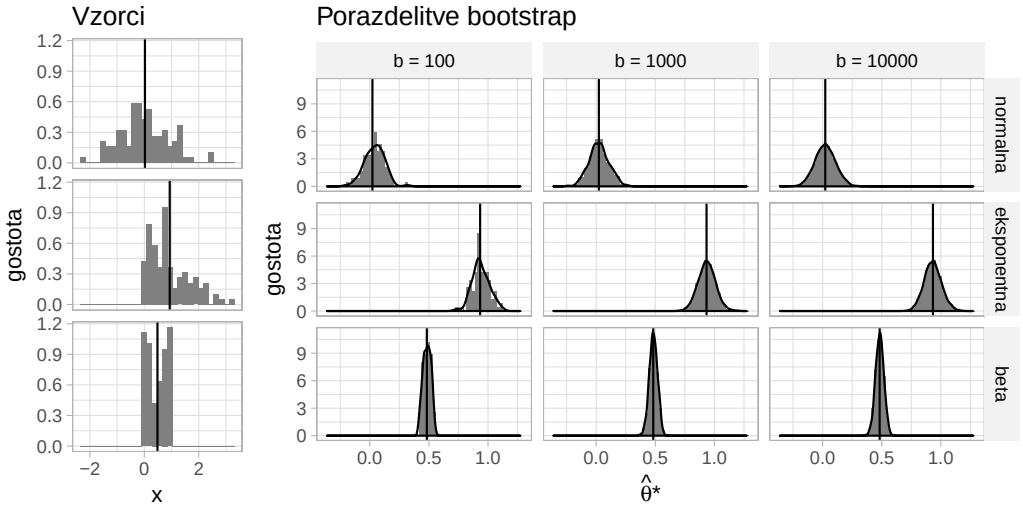
Empirični del

V tem poglavju odgovorimo na vprašanja, ki se pojavijo pri uporabi metode bootstrap. S simulacijami utemeljimo teoretične argumente o zadostnem število vzorcev bootstrap in v različnih situacijah med seboj primerjamo intervale zaupanja. Predstavimo tudi nekaj primerov, v katerih metoda odpove.

5.1 Število vzorcev bootstrap

Ob uporabi metode bootstrap se pojavi pomembno vprašanje, koliko vzorcev bootstrap zadostuje, da bo napaka Monte Carlo zanemarljiva. Efron in Tibshirani [10] trdita, da z $b = 25$ dobimo že dovolj dobro oceno za standardno napako, za intervale zaupanja pa naj bi zadostoval $b = 1000$. Hesterberg [13] zagovarja večje število ponovitev, predlaga vsaj $b = 10^4$. Kot razlog navaja hitrost in zmogljivost današnjih računalnikov.

Iz slike 5.1 je razvidno, kako se napaka Monte Carlo manjša z večanjem števila vzorcev b. Uporabimo metodo bootstrap na treh vzorcih velikosti $n = 100$ iz standardne normalne porazdelitve, eksponentne porazdelitve s parametrom $\lambda = 1$ in porazdelitve beta s parametroma $\alpha = \beta = 0,5$. Zanima nas pričakovana vrednost. Opazimo, da se oblika porazdelitve bootstrap ne spreminja – usrediščena je okoli vrednosti $\hat{\theta}$. Pomembno je, da se zavedamo, da je natančnost metode odvisna od reprezentativnosti vzorca, z uporabo metode bootstrap namreč nikoli ne bomo izboljšali prvotne ocene $\hat{\theta}$.



Slika 5.1: Porazdelitve bootstrap velikosti 100, 1000 in 10000, dobljene iz treh vzorcev različnih porazdelitev (standardna normalna, eksponentna s parametrom $\lambda = 1$ in beta s parametroma $\alpha = \beta = 0,5$). Navpične črte označujejo povprečno vrednost posameznega vzorca.

5.2 Primerjava intervalov zaupanja pri različnih velikostih vzorca

Opisane intervale zaupanja bomo primerjali v različnih situacijah, obenem pa spremljali, kako na rezultate vpliva velikost začetnega vzorca. Pripravili smo dva primera: z umetno ustvarjenimi podatki in s podatki iz prakse. Najprej oba predstavimo in na koncu skupaj interpretiramo rezultate.

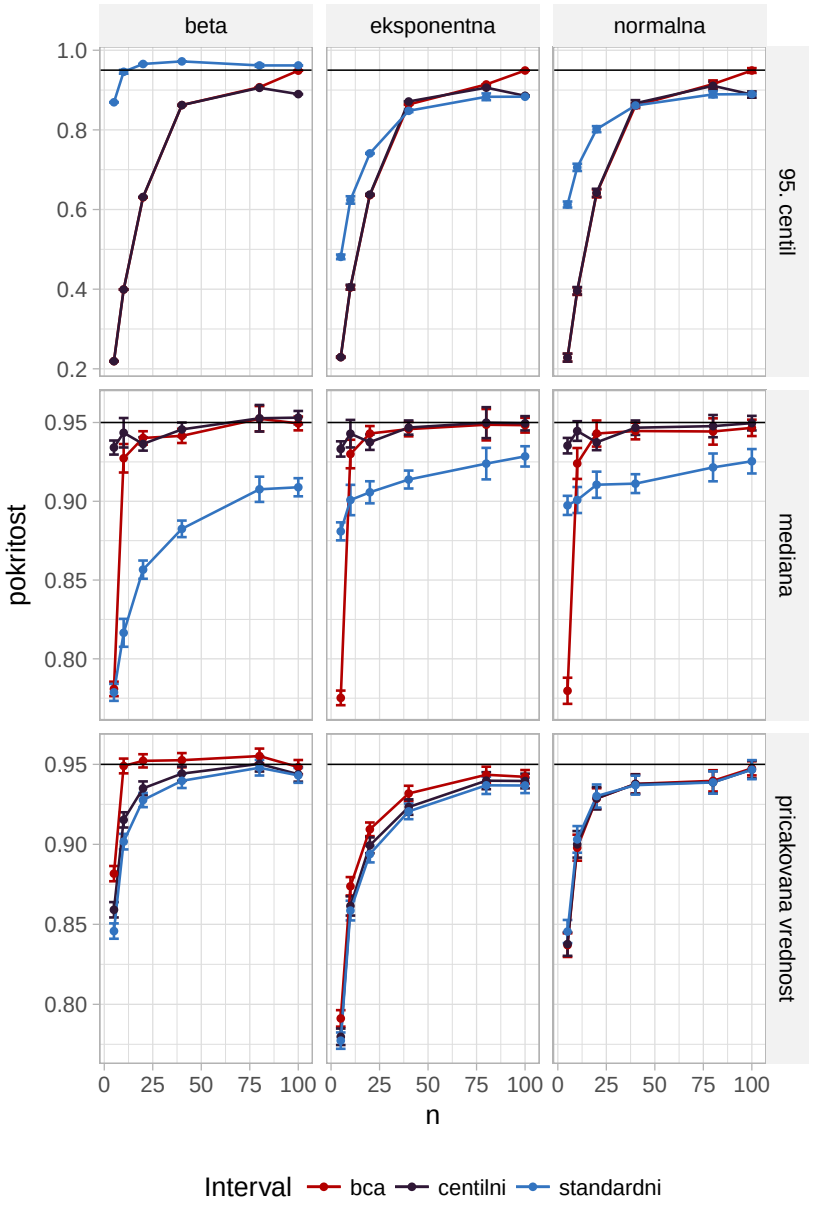
5.2.1 Umetno ustvarjeni podatki

Izbrali smo tri različne porazdelitve, iz katerih dobimo začetni vzorec: standardna normalna porazdelitev, eksponentna porazdelitev s parametrom $\lambda = 1$ in porazdelitev beta s parametroma $\alpha = \beta = 0,5$. Ustvarimo 10^4 vzorcev bootstrap in opazujemo pokritost 95 % intervalov zaupanja za pričakovano vrednost, mediano in 95. centil za različno velikost vzorca n. Da zmanjšamo variabilnost Monte Carlo, so rezultati plod 10^4 ponovitev. Slika 5.2 prikazuje rezultate.

Teoretičen primer služi kot ponazoritev lastnosti intervalov za različen n. V praksi je velikost vzorca bootstrap običajno kar enaka velikosti začetnega vzorca, le tako bodo namreč standardna napaka in intervali zaupanja odražali dejanske podatke. Velikost vzorca bootstrap spreminjamo le, kadar nas zanima, kako bi se te vrednosti spreminjale na hipotetični večji ali manjši množici podatkov.

5.2.2 Podatki iz prakse

Prednosti metode bootstrap bi v praksi lahko izkoristili na vseh področjih empirične znanosti. Za primer vzemimo strojno učenje, pri katerem pogosto na eni ali več množicah podatkov ocenjujemo napovedne modele in jih med seboj primerjamo glede na izbrano napako. Za izračun na-



Slika 5.2: Pokritost intervalov zaupanja bootstrap za različne velikosti vzorcev n, kadar nas zanimajo pričakovana vrednost, mediana in 95. centil. Populacijo predstavljajo normalna, eksponentna in porazdelitev beta. Točkam na grafih smo dodali še 95 % interval zaupanja.

pake navadno uporabimo prečno preverjanje (ki prav tako spada med metode samovzorčenja), za ocenjevanje negotovosti pa intervale zaupanja, največkrat kar standardnega. Pripravili smo primer, na katerem bomo videli, da je ta izbira velikokrat napačna.

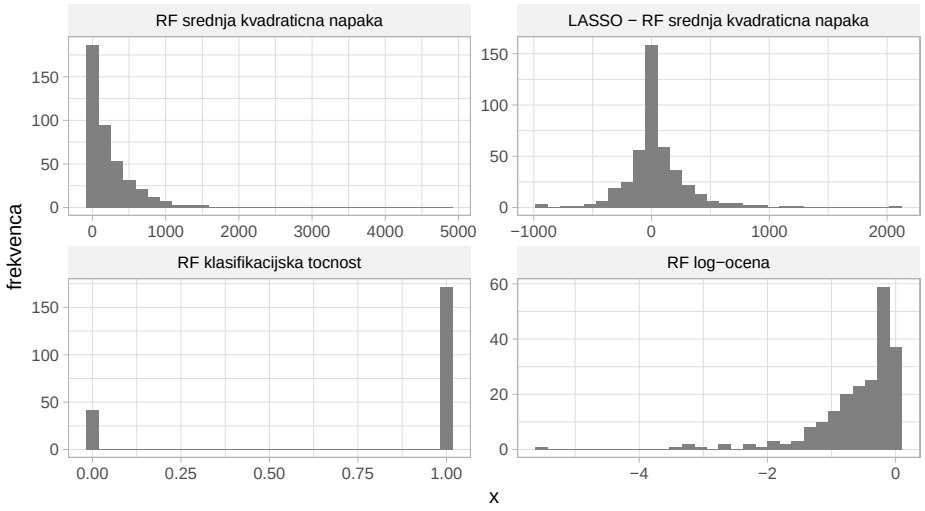
Na sliki 5.3 so histogrami naših podatkov. Zgornja grafa prikazujeta srednjo kvadratno napako (MSE) modelov, zgrajenih iz podatkov o ozonu [4] (425 primerov). MSE smo izračunali z metodo prečnega preverjanja *izloči enega* (LOOCV). Levi graf prikazuje MSE naključnih gozdov, desni pa MSE razlike napovedi regresijske metode Lasso in naključnih gozdov. Na spodnjih grafih pa sta histograma klasifikacijske točnosti in log-ocene naključnih gozdov, zgrajenih na podatkih o identifikaciji stekla [11] (214 primerov). Vidimo, da so v praksi podatki redko porazdeljeni normalno. MSE in log-ocena sta na eni strani omejena in imata eksponentno (oz. gama) porazdelitev. Klasifikacijska točnost je porazdeljena po Bernoulliju, MSE razlike dveh modelov pa spominja na Laplaceovo porazdelitev.

Ustvarimo 10^4 vzorcev bootstrap za različne velikosti vzorcev n in izračunamo intervale zaupanja. Tokrat se bomo osredotočili le na pričakovano vrednost. Rezultate predstavimo na sliki 5.4, spet so pridobljeni na podlagi 10^4 ponovitev.

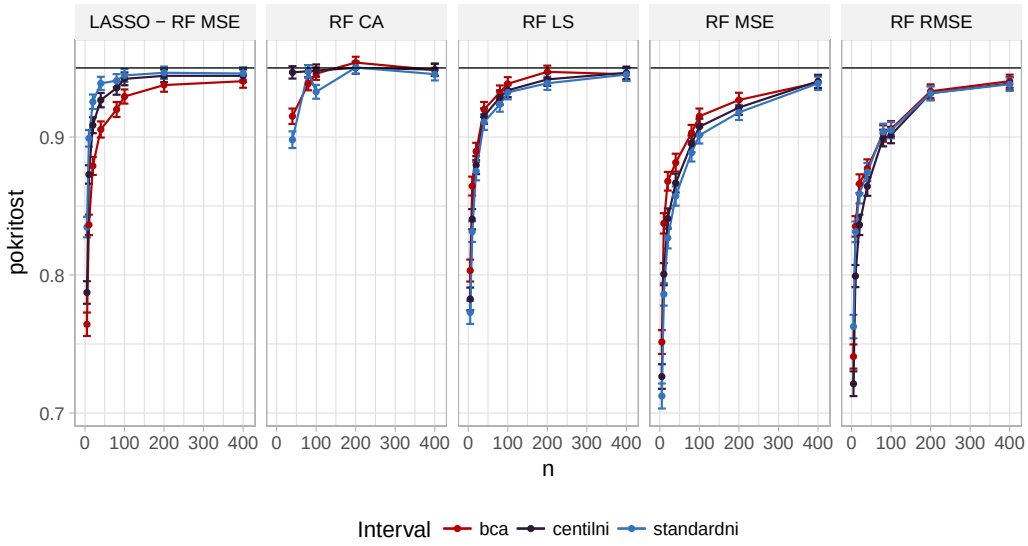
5.2.3 Interpretacija rezultatov

Majhni vzorci

Iz rezultatov je očitno, da metoda bootstrap ne more nadomestiti pomanjkljivosti premajhnega vzorca, ki slabo predstavlja populacijo. Izraz *premajhen vzorec* ima različen pomen za različne statistike. Kadar nas zanima neka bolj ekstremna lastnost (npr. 95. centil), za dobre rezultate



Slika 5.3: Histogrami podatkov, ki so v navedenem praktičnem primeru populacija za metodo bootstrap, predstavljajo napake modelov iz dveh različnih množic podatkov.



Slika 5.4: Pokritost intervalov zaupanja bootstrap za različne velikosti vzorcev n, kadar nas zanima pričakovana vrednost. Populacijo predstavljajo podatki o napakah modelov. Točkam na grafih smo dodali še 95 % interval zaupanja.

potrebujemo večji vzorec. Pri klasifikacijski točnosti premajhen vzorec (v našem primeru $n < 40$) lahko metodi celo povzroča težave, saj dobimo vzorce z ničelno varianco in ne moremo izračunati intervalov zaupanja. Na splošno so intervali zaupanja, izračunani z metodo bootstrap, preozki, zato moramo biti na to pozorni, kadar imamo opraviti z majhnim n. Kadar delamo s tako majhnimi vzorci, je uporaba neparametrične metode bootstrap lahko tvegana in je bolje, da naredimo nekaj predpostavk in uporabimo parametrično različico metode bootstrap.

Primerjava intervalov

Rezultati kažejo, da v večini primerov z naraščanjem n vsak interval doseže želeno pokritost $0,95$, vprašanje je le, kolikšen n je zadosten za to. Pričakovano se interval BCa v večini primerov izkaže najbolje, slabše rezultate vrača le pri podatkih, porazdeljenih po Laplaceu. Podobne rezultate vidimo tudi v [17, 19]. Metoda za izračun mej centilnega intervala je veliko preprostejša od tiste za izračun intervala BCa, ampak, zanimivo, centilni interval ne zaostaja veliko za intervalom BCa. Standardni interval zaupanja predvidljivo vrača najboljše rezultate, ko je porazdelitev bootstrap vsaj približno normalna (ko je naša statistika pričakovana vrednost), nezanesljiv je za mediano ali 95. centil. Izjemo opazimo pri 95. centilu porazdelitve beta, ko interval precenjuje pravo vrednost. Porazdelitev bootstrap je v tem primeru omejena in vrednosti so zgoščene ob zgornji meji, standardni interval pa ne ohranja razpona in vrača nesmiselne zgornje meje intervala, ki nam dajejo zavajajoč občutek dobre pokritosti.

5.3 Kdaj metoda bootstrap odpove

V prejšnjih podpoglavjih smo videli, da se na metodo bootstrap ne moremo zanesti, kadar je vzorec premajhen. To pravzaprav ni le njena omejitev, temveč večine metod, s katerimi se srečujemo na področju statistike

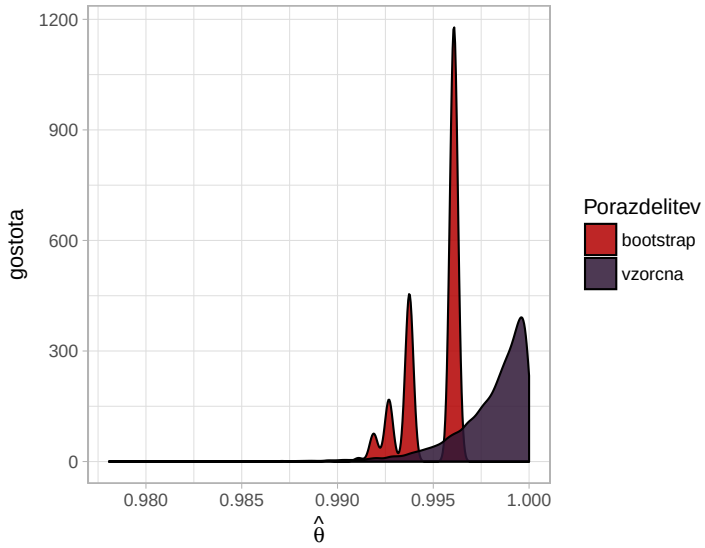
in strojnega učenja. Obstaja pa še nekaj drugih okoliščin, v katerih metoda bootstrap odpove. To se najpogosteje zgodi, kadar centralni limitni izrek ne velja. V tem podpoglavju predstavimo tri takšne primere, več o tem pa si lahko preberemo v [1, 3, 6]. Težave običajno rešujemo z uporabo katere druge različice metode bootstrap.

5.3.1 Ekstremi

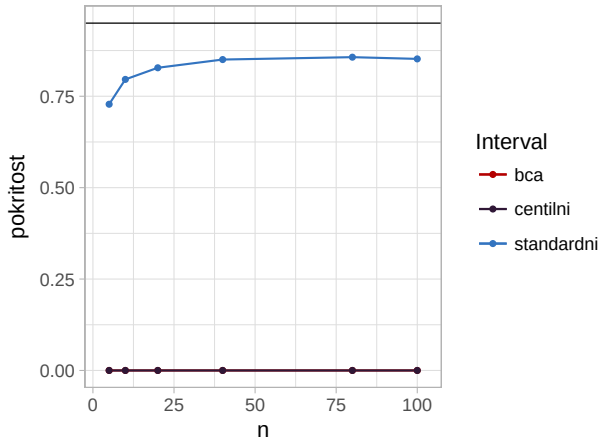
Metoda bootstrap odpove, kadar nas zanimajo ekstremi, recimo maksimum. Empirično predstavimo problem, matematično ozadje pa najdemo v [6, str. 476].

Za slučajni vzorec $x_1, x_2, ..., x_{500}$ iz zvezne enakomerne porazdelitve $U(0,1)$ nas zanima maksimalna vrednost $\hat{\theta} = max(x_1, x_2, ..., x_n)$. Generiramo $b = 10000$ vzorcev bootstrap. Če metoda deluje, naj bi bila porazdelitev bootstrap podobna pravi vzorčni porazdelitvi. To porazdelitev bomo simulirali tako, da iz $U(0,1)$ vzorčimo še 10000-krat in za vsak vzorec izračunamo vrednost $\hat{\theta}$.

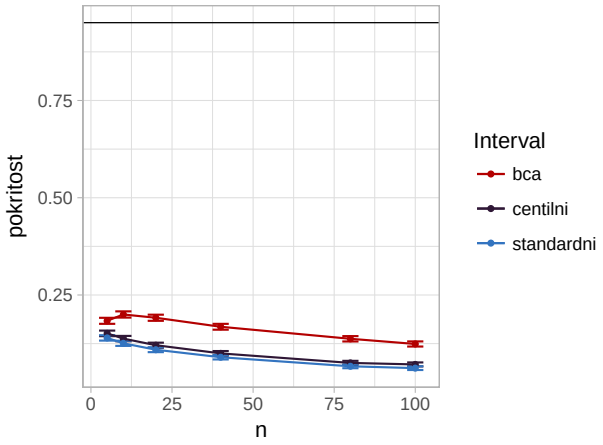
S slike 5.5 je očitno, da si porazdelitvi nista podobni in tokrat premajhen vzorec ni razlog. Izračunali smo še intervale zaupanja bootstrap, rezultate vidimo na sliki 5.6. Verjetnost, da bo začetni vzorec vseboval maksimum porazdelitve, je zelo majhna, in če vzorec ne vsebuje te vrednosti, je ne more vsebovati noben vzorec bootstrap. Pokritost intervala BCa in centilnega intervala je torej enaka 0. Pokritost standardnega intervala zaupanja pa razložimo z enakim argumentom kot v prejšnjem podpoglavju – standardni interval nima lastnosti ohranjanja razpona, zato vrača nerealne meje intervala, ki so zunaj zaloge vrednosti naših podatkov, zaradi česar sicer navidezno vsebuje pravo vrednost, vendar pokritost ni realna.



Slika 5.5: Porazdelitev bootstrap in prava vzorčna porazdelitev maksimalne vrednosti zvezne enakomerne porazdelitve na intervalu $[0, 1]$.



Slika 5.6: Pokritost intervalov zaupanja bootstrap za različne velikosti vzorcev n, kadar nas zanima maksimalna vrednost zvezne enakomerne porazdelitve $U(0, 1)$.



Slika 5.7: Pokritost intervalov zaupanja bootstrap za različne velikosti vzorcev n, kadar nas zanima pričakovana vrednost porazdelitve Pareto s parametrom merila $m = 1$ in parametrom oblike $\alpha = 1.5$.

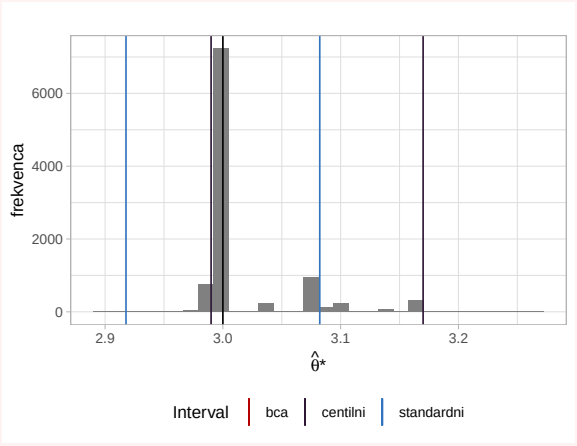
5.3.2 Porazdelitve z neskončno varianco

Za metodo bootstrap so težava tudi podatki, ki imajo porazdelitev z neskončno varianco oz. *z dolgim repom*. Podrobno je to težavo proučil Athreya v svojem članku [1], mi pa jo prikažemo z vzorcem iz porazdelitve Pareto[1].

Za slučajni vzorec $x_1, x_2 \ldots x_n$ iz porazdelitve Pareto s parametrom merila $m = 1$ in parametrom oblike $\alpha = 1,5$ želimo izračunati pričakovano vrednost in intervale zaupanja. Slika 5.7 prikazuje rezultate. Pokritost intervalov je daleč od želene in z naraščanjem velikosti vzorca n pada, kar se nam iz prejšnjih primerov morda ne zdi intuitivno. Vendar z večjim n se verjetnost, da vzorec vsebuje ekstremne vrednosti, poveča, kar potem metodo »uniči«. Centralni limitni izrek ne velja za porazdelitve z neskončno varianco in metoda bootstrap odpove.

[1]Porazdelitev Pareto s parametrom oblike $1 < \alpha \leq 2$ ima neskončno varianco.

Primer 1.1 (nadaljevanje)**.** Tokrat nas zanima mediana. Naš vzorec naj bo kar celotna populacija, da bo n zagotovo dovolj velik.



Slika 5.8: Histogram $\hat{\theta}^*$ celotne populacije ($n = 490$), kadar nas zanima mediana. Ta je označena s črno navpično črto ($\hat{\theta} = 3$).

Za mediano centralni limitni izrek sicer velja (prek metode delta [6, str. 40]) in metoda bootstrap ne bi smela imeti težav (kot smo videli v podpoglavju 5.2), vendar mora biti izpolnjen pogoj zveznosti. V Sloveniji imamo 490 restavracij, ki ponujajo obroke na bone, vendar vse skupaj le 72 unikatnih cen. Porazdelitev populacije je torej diskretna in centralni limitni izrek ne velja. Na sliki 5.8 lahko opazimo, da mej intervala BCa ni. Zaradi majhne variabilnosti cen namreč ne moremo izračunati vrednosti parametra pospeška $\hat{a}$, saj so vse ocene *jackknife* enake 3 in dobimo v imenovalcu 0.

6

Sklepne ugotovitve

V diplomskem delu smo predstavili metodo bootstrap, ki spada v družino metod samovzorčenja. Opisali smo njene lastnosti in glavno idejo z namenom, da jo bralec spozna kot alternativo tradicionalnim metodam za ocenjevanje negotovosti. Povzemimo nekaj prednosti metode:

- bolj preprosta in bolj intuitivna,
- nismo omejeni s številnimi predpostavkami,
- lahko jo uporabimo za skoraj vsako statistiko (ne da bi se postopek spremenil),
- preprosta implementacija.

Dokazali smo njeno pravilnost za pričakovano vrednost in predstavili tri intervale zaupanja: standardnega normalnega z uporabo standardne napake bootstrap in dva klasična intervala bootstrap: centilnega ter BCa.

V empiričnem delu smo s simulacijami odgovarjali na praktična vprašanja, ki se lahko pojavijo pri uporabi metode bootstrap. Zaradi zmogljivosti današnjih računalnikov zagovarjamo čim večje število vzorcev bootstrap, sicer pa simulacije pokažejo, da $b = 1000$ zadostuje. Interval zaupanja BCa se pričakovano najbolje odreže v večini situacij. Centilni ne zaostaja veliko, uporabo standardnega pa odsvetujemo, razen če smo prepričani o normalnosti porazdelitve bootstrap. Na metodo bootstrap se ne moremo zanesti, kadar je začetni vzorec premajhen in/ali nas zanima neka ekstremnejša lastnost (npr. 95. centil), težave pa povzročajo tudi situacije, v katerih centralni limitni izrek ne velja (kadar nas zanima maksimum, imamo opraviti s porazdelitvijo z neskončno varianco ...).

V tem delu smo se osredotočili na neparametrično različico metode, s katero se v praksi največkrat srečamo, poznamo pa še veliko drugih. Ideja ostaja enaka, razlikujejo se predvsem v načinu vzorčenja. Raziskave v zadnjem času kažejo opazen premik proti Bayesovskim pristopom pri primerjavi modelov in interpretaciji rezultatov [2], zato bi izpostavili predvsem Bayesovsko različico metode bootstrap [16], ki pa se je v tem delu nismo dotaknili, saj bi se srečali z istimi vprašanji kot se pri frekventističnem pristopu.

Literatura

[1] K. B. Athreya. Bootstrap of the Mean in the Infinite Variance Case. *The Annals of Statistics*, 15(2):724–731, 1987.

[2] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

[3] P. J. Bickel and D. A. Freedman. Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

[4] R. Češnovar and E. Štrumbelj. Bayesian Lasso and multinomial logistic regression on GPU. *PLoS ONE*, 12(6):e0180343, 2017.

[5] Y. S. Chow and H. Teicher. *Probability Theory*. Springer-Verlag, 1978.

[6] A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer-Verlag, 2008.

[7] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, 1996.

[8] R. Durret. *Probability: Theory and Examples*. Cambridge University Press, 2010.

[9] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 01 1979.

[10] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, Inc., 1993.

[11] B. German. Glass identification data set. `https://archive.ics.uci.edu/ml/datasets/glass+identification`. Dostopano: 10.07.2018.

[12] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, 1992.

[13] T. C. Hesterberg. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician*, 69(4):371–386, 2015.

[14] E. Parzen. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365):105–121, 1979.

[15] G. Pólya. Über den zentalen grenzwertsatz der wahrscheinlichkeitsrechnung und momenten problem. *Mathematische Zeitschrift*, 8:171–181, 1920.

[16] D. B. Rubin. The Bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

[17] K. Samart, N. Jansakul, and M. Chongcheawchamnan. Exact bootstrap confidence intervals for regression coefficients in small samples. *Communications in Statistics - Simulation and Computation*, 2017.

[18] L. Wasserman. *All of Statistics*. Springer-Verlag, 2010.

[19] Y. Zhu and J. Kolassa. Assessing and comparing the accuracy of various bootstrap methods. *Communications in Statistics - Simulation and Computation*, 2017.

Priloge

A Konvergenca Studentove t-porazdelitve proti standardni normalni

Izrek A.1. *Studentova t-porazdelitev konvergira proti standardni normalni, ko* $n \to \infty$.

Dokaz. Za lažje razumevanje dokaza se najprej spomnimo limitnega zapisa Eulerjeve konstante

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n \tag{1}$$

in Sterlingove aproksimacije za funkcijo Γ

$$\Gamma(x) = \sqrt{2\pi} x^{x-\frac{1}{2}} e^{-x} (1 + O(x^{-1})). \tag{2}$$

Dokaz razdelimo na dva dela:

$$\lim_{n\to\infty} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} = \underbrace{\lim_{n\to\infty} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}}_{A} \underbrace{\lim_{n\to\infty} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}}_{B}.$$

Začnimo z delom B. Za izračun limite uporabimo Eulerjevo konstanto (enačba 1):

$$\lim_{n\to\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} = \lim_{n\to\infty}\left(\left(1+\frac{1}{\frac{n}{x^2}}\right)^{\frac{n}{x^2}}\right)^{-\frac{(n+1)x^2}{2n}} = \\ = e^{-\frac{x^2}{2}\lim_{n\to\infty}\frac{n+1}{n}} = e^{-\frac{x^2}{2}}.$$

Izraz dela A poenostavimo z uporabo Sterlingove aproksimacije (enačba 2):

$$\lim_{n\to\infty}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} = \lim_{n\to\infty}\frac{\sqrt{2\pi}(\frac{n+1}{2})^{\frac{n}{2}}e^{-\frac{n+1}{2}}}{\sqrt{n\pi}\sqrt{2\pi}(\frac{n}{2})^{\frac{n}{2}-\frac{1}{2}}e^{-\frac{n}{2}}} = \\ = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}\lim_{n\to\infty}\left(\frac{n+1}{n}\right)^{\frac{n}{2}} = \frac{1}{\sqrt{2\pi}}.$$

Dokazali smo, da Studentova t porazdelitev konvergira k standardni normalni, ko $n \to \infty$:

$$\lim_{n\to\infty}\frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

□

B Konvergenca vzorčne variance s^2 proti pravi varianci σ^2

Izrek B.1. *Vzorčna varianca s^2 konvergira proti pravi varianci σ^2, ko $n \to \infty$.*

Dokaz. Naj bodo $X_1, ..., X_n$ slučajne spremenljivke z upanjem $E[X_i] = \mu$ in

končno varianco $Var[X_i] = \sigma^2 < \infty$. Z s^2 označimo vzorčno varianco:

$$\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=0}^{n}(X_i - \overline{X})^2 = \\
&= \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu + \mu - \overline{X})^2 = \\
&= \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})\frac{1}{n}\sum_{i=0}^{n}(X_i - \mu) + (\mu - \overline{X})^2 = \\
&= \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 + 2(\mu - \overline{X})(\overline{X} - \mu) + (\mu - \overline{X})^2 = \\
&= \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 + (\mu - \overline{X})(2(\overline{X} - \mu) + (\mu - \overline{X})) = \\
&= \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 - (\mu - \overline{X})^2.
\end{aligned}$$

Po krepkem zakonu o velikih številih velja $\overline{X} \xrightarrow{a.s.} \mu$, iz česar sledi

$$s^2 \xrightarrow{a.s.} \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 \xrightarrow{a.s.} E[(X_i - \mu)^2] = \sigma^2.$$

□