LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

IMPLICĒTO VARBŪTĪBU BUTSTRAPS

KURSA DARBS

Autors: **Vineta Vītola**

Stud. apl. vv08073

Darba vadītājs: doc. Dr. math. Jānis Valeinis

RĪGA 2012

Saturs

Ievads

Kopš Hansena (1982) vispārinātā momentu metode (GMM) ir bijusi standartrīks epmīriskajai analīzei ekonometrikā. GMM nodrošina vienotu sistēmu statistiskiem slēdzieniem ekonometrikas modeļiem, kuri ir noteikti ar dažiem momentu nosacījumiem. Tomēr nesenie pētījumi rāda, ka ir ievērojamas problēmas ar GMM, īpaši tās ierobežotās izlases gadījumā un to, ka aproksimācija, kura balstās uz asimptotisko teoriju var dot sliktus rezultātus. Lai precizētu aproksimāciju GMM novērojumu sadalījumam un saistītajām testa statistikām, ir izstrādātas butstrapa metodes. Galvenais, lai izmantotu butstrapa metodi GMM kontekstā, ir nepieciešams uzlikt vairāk momentu nosacījumus butstrapa izlasēm nekā ir parametru vektora dimensija. Hall un Horowitz (1996) ieteica izmantot butstrapu ar atkārtotiem momentu nosacījumiem un izveidot augstākas kārtas uzlabojumu butstrapa asipmtotiskām aproksimācijām. Bet Brown un Newey (2002) ieteica nosvarot butstrapu ar implicētajām varbūtībām balstoties uz momentu nosacījumiem. Šīs implicētās varbūtības var iegūt balstoties uz GMM (Back un Brown, 1993), empīrisko ticamības metodi (Owen, 1988) vai vispārināto empīrisko ticamības metodi (Smith, 1997, Newey un Smith, 2004).

Lai pārbaudītu, vai implicēto varbūtību butstraps dod robustākus rezultātus, pārbaudīsim to uz simulēto datu piemēra - normālajam sadalījumam ar piejaukumu salīdzināsim ticamības intervāla pārklājuma precizitāti ar parastā butstrapa un procentīļu butstrapa ticamības intervāliem. Diviem datu piemēriem salīdzināsim implicēto varbūtību butstrapa ticamības intervāla garumus ar t-testa, parastā butstrapa, procentīļu butstrapa un pivotālā butstrapa ticamības intervāla garumiem.

1. daļā tiek aprakstīts, kas ir robusta statistika un kam tā ir nepieciešama. 2. daļā apskatu, kas ir butstraps un tā ticamības intervālus. 3. daļā apskatu specializēto butstrapa gadījumu ar implicētajām varbūtībām. Kā arī pārbaudu uz konkrētiem piemēriem, vai impilcēto varbūtību butstraps ir robustāks par parasto butstrapu pret izlecējdatiem.

1. Robusta statistika

Robusta statistika vienkaršā, netehniskā valodā attiecas uz faktu, ka daudzi pieņēmumi, kurus parasti veic statistikā (kā, piemēram, normalitāte, linearitāte, atkarība) ir tikai tuvi reālajiem. Viens iemesls tam ir lielu kļūdu pieļaušana, piemēram, kopējot vai pārrakstoties. Tās parasti parādās kā izlecēji, kuri ir tālu no datu vairākuma, un ir bīstami daudzām klasiskajām statistikas procedūram. Izlecēju problēma ir labi pazīstama un iespējams tikpat veca kā statistika, un metodes, kuras risina šo problēmu, kā, piemēram, to subjektīva noraidīšana vai jebkurš formāls noraidīšanas noteikums pieder robusta statistikai tās plašākajā nozīmē. Citi iemesli novirzēm no pieņēmuma par idealizēto modeļi ietver empīrisko raksturu daudziem modeļiem un aptuveno raksturu daudziem teorētiskajiem modeļiem.

Galvenie robusta statistikas mērķi ir [1]

1. aprakstīt piemērotāko struktūru datu vairākumam,
2. identificēt datu punktu novirzes (izlecējus).

Gadījumā, kad ir nebalansēta datu struktūra, kā tipiski tas ir regresijā, tad ir arī nepieciešams

3. identificēt un brīdināt par datu punktu augsto ietekmi.

Turklāt, tā kā ne tikai pieņemtais marginālais sadalījums, bet arī pieņemtā korelācijas struktūra (principā neatkarības pieņēmums) ir tikai tuvinājums, tad cits robusta statistikas uzdevums ir

4. tikt galā ar negaidītu sērijas korelāciju, vai plašāk, ar novirzi no pieņemtās korelācijas struktūras.

Tā kā ticamības intervāli un testi robustiem novērtējumiem ir asimptotiski, tad aproksimācija vēlamajam var būt slikta, ja n ir mazs. Tas notiek īpaši tad, kad kļūdas sadalījumam ir ļoti smaga aste vai asimetrija. Labākus rezultātus var iegūt, izmantojot butstrapa metodi, kura novertē sadalījumu, ģenerējot liela skaitu izlašu ar aizvietošanu ("butstrapa izlases") no izlases un novērtējot katru no tām. [2]

2. Vispārinātā momentu metode (GMM)

Vispārināto momentu metodi izdomāja Lars Peter Hansen 1982. gadā kā vispārinājumu momentu metodei. Vispārinātās momentu metodes novertējumi bija svarīgs jaunināJums ekonometrikā. Tie nodrošināja noderīgu pieeju, lai atrisinātu novērtēšanas problēmas daudzos gadījumos, ieskaitot racionālas cerības modeļos, modeļos paneļu datiem, nepārtraukta laika modeļos un semiparametriskos modeļos.

Pieņemsim, ka z apzīmē datu novērojumus, β $p\times 1$ parametru vektoru, un $g(z,\beta)$ $m\times 1$ vektoru no funkcijām no datu novērojumiem un parametriem ar $m \geq p$. Vispārinātās momentu metodes novērtējums ir balstīts uz momentu ierobežojumiem formā

$$E[g(z,\beta_0)] = 0. \tag{2.1}$$

Bieži šis momentu ierobežojums ir daļa implikācijas dažiem modeļiem, kā arī tas var ietvert sevī visu pieejamo informāciju.[3]

Daudzos dokumentos ir izskatītas vispārinātās momentu metodes (GMM) informācijas teorētiskās alternatīvas. Idejas būtība ir izmantot momentu nosacījumus vairāk nekā parametru vektora dimensija, lai kopā novērtētu interesējošos parametrus un svaru kopu, katru pievienojot novērojumiem, kuri atspoguļo datu ģenerējoša sadalījuma novērtejumu. Novērtētie svari minimizē attālumu starp novērtēto sadalījumu un datu empīrisko sadalījumu, atbilstoši momentu nosacījumiem.

Imbens (1993, 1997) [4] ieteica novērtēt sadalījumu datiem kopā ar interesējošajiem orģinālajiem parametriem. Katram novērojumam ir piekārtoti svari, vai varbūtības, kur visi svari normalizēti, lai summā būtu viens. Tātad attiecīgie novērtējumi sadalījuma funkcijai izriet no

$$\hat{F}(z) = \sum_{i=1}^{N} 1\{z_i \leq z\} w_i, \tag{2.2}$$

kur $1\{.\}$ ir identifikatora funkcija. Dot katram novērojumam vienādus svarus, atbilst empīriskajam sadalījumam. Svari ir aprēķināti tā, ka svērtā izlase analogi visiem momentiem ir nulle

$$\sum_{i=1}^{N} w_i \psi_m(z_i, \hat{\theta}) = 0 \quad \forall m = 1, ..., M. \tag{2.3}$$

Bez turpmākiem ierobežojumiem, svaru vektors nav viennozīmīgi noteikts. Tāpēc svari ir izvēlēti no tiem, kuri implicē sadalījumu, kurš ir tuvākais empīriskajam sadalījumam, atbilstoši momentu nosacījumam, kas dots ar formulu 2.3. Ja mēs ierobežojumam mūsu

uzmanību uz attāluma mēru starp novērtēto un empīrisko sadalījumu formā $D(w) = \sum_{i=1}^{N} d(w_i)$, tad novērtēšanas problēmu var formulēt kā

$$\min_{w,\theta} \sum_{i=1}^{N} d(w_i) \quad \text{tā, ka} \sum_{i=1}^{N} w_i = 1 \quad \sum_{i=1}^{N} w_i \psi(z_i, \theta) = 0. \tag{2.4}$$

3. Butstraps

Pieņemsim, ka $X_1, X_2, ..., X_n \sim F$ ir neatkarīgi un vienādi sadalīti (iid) un $T = T(X_1, ..., X_n, F)$ ir kāds funkcionālis, t.i.,

$$T = T(X_1, ..., X_n, F) = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma},$$

kur $\mu = E_F(X_1)$ un $\sigma^2 = Var_F(X_1)$. Viena statistikas problēma ir, ka bieži jāzina kaut kas par izlases sadalījumu T sadalījumu, tas ir jāatrod $P_F(T(X_1, ...X_n, F) \leq t)$. Ja mēs atkārtoti veidojam izlases no populācijas, tad mēs varam novērtēt $P_F(T \leq t)$ izskaitot cik reižu $T_i' \leq t$. Bet statistikām izlasēm tā nedara. Mēs parasti neiegūstam atkārtoti veidotās izlases; mēs iegūstam vienu datu kopu no kāda izmēra n. Lielai izlasei no ierobežotas populācijas vajadzētu reprezentēt visu populāciju, tāpēc atkārtotās izlases no īstās izlases, kura būtu tikai iid izlase no empīriskās CDF F_n, varētu tikt uzskatīta par tuvinājumu aizvietotajām izlasēm no populācijas, pie nosacījuma, ka n ir liels. [5]

Pieņemsim, ka $X_1, ..., X_n$ labi reprezentē īsto populācijas sadalījumu. Butstrapa ideja: simulēt daudz izlašu no dotās un aproksimēt statistikas T sadalījumu. Tas ir, izvēlēsimies no dotās izlases $X_1, ..., X_n$ jaunas iid izlases no empīriskās sadalījuma funkcijas $\hat{F}_n$. Visiem novērojumiem ir vienāda $1/n$ varbūtība tikt izvēlētiem. To arī sauc par neparametrisko, empīrisko vai dažkārt par parasto butstrapu.

Pieņemsim, ka mēs uzskatām, ka iegūtai izlasei ir kāds parametrisks sadalījums. Šādā gadījumā sadalījuma funkcija ir pilnībā noteikta ar parametriem $\theta : F = F_\theta$. Tad sadalījuma funkcija F nav jānovērtē kā pilnībā nezināma funkcija, bet ta vietā pietiek novērtēt parametru (vai vektoru) θ ar $\hat{\theta}$ un tad novērtēt F ar $\hat{F} = F_{\hat{\theta}}$. Attiecīgo butstrapa principu sauc par parametrisko butstrapu.

Apzīmēsim iegūtās butstrapa izlases $X_{11}^*, ..., X_{1n}^*$, $X_{21}^*, ..., X_{2n}^*$, ..., $X_{B1}^*, ..., X_{Bn}^*$, kur B apzīmē butstrapoto izlašu skaitu. Statistikā funkcionāļa T sadalījumu punktā t, tas ir, $P_F(T \leq t)$ aproksimēsim ar j skaits: $T_j^* t/B$, kur $T_1^*, T_2^*, ..., T_B^*$ apzīmē statistikas T vērtības B dažādajām butstrapa izlasēm. Piemēram, ja $T(X_1, ..., X - n, F) = \sqrt{n}(\hat{X}_n - \mu))/\sigma$, tad attiecīgā butstrapotā statistika ir $T(X_1, ..., X - n, F) = \sqrt{n}(\bar{X}_n^* - \bar{X}_n))/S$, kur S apzīmē izlases (empīrisko) dispersiju. Šis piemērs ir tipisks butstrapa metodei, kur īstās vertības μ un σ tiek aizstātas ar novērtētajām $\hat{X}_n$ un S no sākotnējiem datiem, jo mēs uzskatām, ka novērojumi labi reprezentē īstās popoulācijas sadalījumu. [5]

Butsrapu var izmantot arī, lai konstruētu ticamības intervālus.

4. Butstrapa ticamības intervāli

Atkarībā no vēlamās precizitātes un aprēķinu sarežģītības pakāpes ir dažādi veidi, kā konstruēt butstrapa ticamības intervālus. Visvienkāršākais ticamības intervālu konstruēšanas veids ir tā sauktie Normālie intervāli, kuri ir formā

$$T_n \pm z_{\alpha/2}\hat{se}_{boot}, \tag{4.1}$$

kur $\hat{se}_{boot}$ ir standartkļūdas butstrapa novērtējums statistiskajam funkcionālim T_n. Šie intervāli ir precīzi gadījumā, ja T_n sadalījums ir tuvs Normālajam.

Pivotālie intervāli. Pieņemsim, ka $\theta = T(F)$ un $\hat{\theta}_n = T(\hat{F}_n)$ un definēsim pivotu $R_n = \hat{\theta}_n - \theta$. Pieņemsim, ka $H(r)$ apzīmē pivotas kumulatīvo sadalījuma funkciju, tas ir,

$$H(r) = P_F(R_n \leq r)$$

. Pieņemsim, ka $C_n^* = (a, b)$, kur

$$a = \hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2}) \text{ un } b = \hat{\theta}_n - H^{-1}(\frac{\alpha}{2})$$

Seko, ka

$$\begin{aligned} P(a \leq \theta \leq b) &= P(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a) \\ &= H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\ &= H(H^{-1}(1 - \frac{\alpha}{2})) - H(H^{-1}(\frac{\alpha}{2})) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{aligned}$$

Tādējādi C_n^* ir precīzs $1 - \alpha$ ticamības intervāls funkcijai θ. Diemžēl a un b ir atkarīgi no H nezināmā sadalījuma, bet mēs varam veikt H butstrapa novērtējumu

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r),$$

kur $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$. Pieņemsim, ka r_β^* apzīmē β izlases kvantili no $(R_{n,i}^*, ..., R_{n,B}^*)$ un θ_β^* apzīmē β izlases kvantili no $(\theta_{n,1}^*, ..., \theta_{n,B}^*)$. Atzīmēsim, ka $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$. Seko, ka $1 - \alpha$ ticamības intervāls ir $C_n = (\hat{a}, \hat{b})$, kur

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}(1 - \frac{\alpha}{2}) = \hat{\theta}_n - r_{1-\frac{\alpha}{2}}^* = 2\hat{\theta}_n - \theta_{1-\frac{\alpha}{2}}$$

,

$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}(\frac{\alpha}{2}) = \hat{\theta}_n - r_{\frac{\alpha}{2}}^* = 2\hat{\theta}_n - \theta_{\frac{\alpha}{2}}$$

.

Apkopojot: $1-\alpha$ butstrapa pivotālais ticamības intervāls ir

$$C_n = (2\hat{\theta}_n - \hat{\theta}_{((1-\frac{\alpha}{2})B)}, 2\hat{\theta}_n - \hat{\theta}_{((\frac{\alpha}{2})B)}). \tag{4.2}$$

Procentīļu intervāli. Butstrapa procentīļu intervāli ir definēti šādi

$$C_n = (T^*_{(B\alpha/2)}, T^*_{(B(1-\alpha/2))}), \tag{4.3}$$

tas ir, izmanto $\alpha/2$ un $1-\alpha/2$ kvantiles butstrapa izlasei.

Pamatojums. Pieņemam, ka eksistē monotona transformācija $U = m(T)$ tāda, ka $U \sim N(\phi, c^2)$, kur $\phi = m(\theta)$. Pieņem, ka $U^*_b = m(T^*_b)$. Ievērojam, ka $U^*_{(B\alpha/2)} = m(T^*_{(B\alpha/2)})$, jo monotona trasformācija saglabā kvantiles. Tā kā $U \sim N(\phi, c^2)$, tad U $\alpha/2$ kvantile ir $\phi - z_{\alpha/2}c$ Tātad $U^*_{(B\alpha 2)} = \phi - z_{\alpha/2}c \approx U - z_{\alpha/2}c$ un $U^*_{(B(1-\alpha/2))} \approx U + z_{\alpha/2}c$. Tādēļ

$$\begin{aligned} P(T^*_{B\alpha/2} \le \theta \le T^*_{B(1-\alpha/2)}) &= P(m(T^*_{B\alpha/2}) \le m(\theta) \le m(T^*_{B(1-\alpha/2)})) \\ &= P(U^*_{B\alpha/2} \le \phi \le U^*_{B\alpha/2)}) \\ &\approx P(U - cz_{\alpha/2} \le \phi \le U + cz_{\alpha/2}) \\ &= P(-z_{\alpha/2} \le \frac{U-\phi}{c} \le z_{\alpha/2}) = 1-\alpha \end{aligned}$$

Pārsteidzoši, ka mums nekad nevajadzēs zināt m. Diemžēl, precīza normalizēšanas trasnformēšana pastāvēs reti, bet var eksistēt aptuvena normalizācijas transformēšana. [6]

5. Implicēto varbūtību butstraps

Implicēto varbūtību butstraps, ko ierosināja Brown un Newey (2002), pārkārto datus ar dažādām varbūtībām, kuras ir definētas kā implicētas varbūtības no momentu nosacījumiem, un izmanto kvantiles no pārkārtotās statistikas balstoties uz momentu nosacījumiem bez atkārtotas centrēšanas.

Meklējot pieņemamus butstrapa svarus, kuri apmierina interesējošos momentu nosacījumus, informācijai par teorētiskajiem argumentiem ir liela nozīme. Īpaši, mēs koncentrējamies uz informācijas projekciju par empīrisko sadalījumu, lai uzstādītu sadalījumam atbilstošos momentu nosacījumus un izmantotu projekciju, kuras sauksim par implicētajām varbūtībām, kā butstrapa svarus. Šīs implicētās varbūtības var tikt aprēķinātas balstoties uz Boltzmann-Shannon entropiju iegūstot eksponenciālās nolieces svarus(Kitamura and Stutzer, 1997, un Imbens, Spady un Johnson, 1998), Burga entropiju iegūstot empīriskās ticamības svarus (Owen, 1988), Fišera informāciju iegūstot GMM-tipa svarus (Back un Brown,1993), vai to variantus (Smith, 1997, un Newey un Smith, 2004).

Brown un Newey (2002) apgalvoja, ka implicēto varbūtību butstraps piedāvā augstākās kārtas uzlabojumu virs pirmās kārtas asimptotiskās aproksimācijas pie noteiktiem regularitātes nosacījumiem tikpat labu kā butstraps (ar atkārtotu centrēšanu). Svarīga iezīme implicētajām varbūtībām, ko uzsver literatūrā ir tā, ka tās nodrošina semiparametriski kvalitatīvu novērtējumu sadalījuma funkcijai un tā momentiem pie momentu nosacījumiem.

Back un Brown (1993) [7] parādīja, ka pie šādas palīginformācijas $E[g(X_i, \theta)] = 0$, kur $g : \mathrm{R} \to \mathrm{R}$ funkcija ar skalārām vērtībām, $\bar{g} = \frac{1}{n}\sum_{i-1}^{n} g(X_i, \hat{\theta})$, funkcijas X sadalījumu var efektīvi novērtet, izmantojot šādas implicētās varbūtības:

$$\pi_i = \frac{1}{n} - \frac{1}{n}\frac{(g(X_i, \hat{\theta}) - \bar{g})\bar{g}}{\frac{1}{n}\sum_{i=1}^{n} g(X_i, \hat{\theta})^2}, \tag{5.1}$$

kur $\bar{\theta}$ ir θ novērtējums, $i = 1, ..., n$. π_i otro daļu var interpretēt kā sodu par novirzēm no palīginformācijas: ja $|g(X_i, \hat{\theta})|$ kļūst lielāks, tad $(g(X)_i, \hat{\theta}) - \bar{g})\bar{g}$ ir tendence būt pozitīvai un svariem π_i tendence samazināties.

Lai nodrošinātu, ka visas implicētās varbūtības ir nenegatīvas, izmanto pieeju, kuru ieteica Antoine, Bonnal un Renault (2007) (t.i., $\hat{\pi_i} = \frac{1}{1+\epsilon_n}\pi_i + \frac{\epsilon_n}{1+\epsilon_n}\frac{1}{n}$ ar $\epsilon_n = -n\min\{\min_{1\leq i\leq n}\pi_i, 0\}$). Kā norādīja Antoine, Bonnal un Renault, šī pieeja saglabā implicēto varbūtību secību, neietekmē to, ka implicētās varbūtības jau ir nenegatīvas, un

pieškir nulles varbūtību tikai novērojumam, kurš saistīts ar mazāko varbūtību, kad tā ir negatīva.[7]

Normālais sadalījums ar piejaukumu. Pieņemam, ka proporcija $1-\epsilon$ no novērojumiem ir ģenerēti kā normālais modelis, kamēr proporcija ϵ ģenerēta pēc nezināma mehānisma. Piemēram, atkārtoti mērījumi ir veikti dažādā apjomā, kuri ir 95% ir pareizi, bet 5% aparāts kļūdās vai eksperimentētajs veic nepareizus pierakstus. Tas varētu būt pierakstīts sekojoši [2]

$$F = (1-\epsilon)G + \epsilon H, \tag{5.2}$$

kur $G = N(\mu, \sigma^2)$ un H var būt jebkurš sadalījums; piemēram normālais ar lielu dispersiju un iespējams ar atšķirīgu vidējo vertību. To sauc par normālo sadalījumu ar piejaukumu. Pieņemsim, ka A būs gadījums, kad āparāts kļūdas", kur $P(A) = \varepsilon$, un A' tā papildinājums. Mēs pieņemam, ka mūsu mainīgajam x ir sadalījums G pie nosacījuma A' un H pie nosacījuma A. Tad

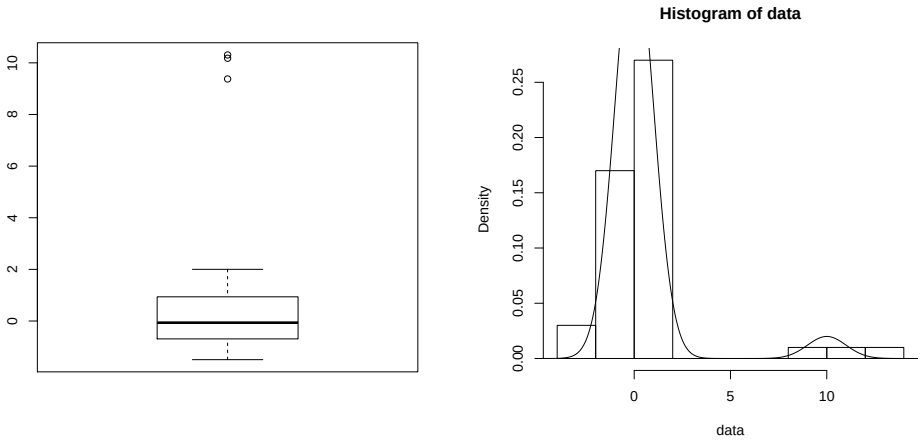
$$F(t) = P(x \le t) = P(x \le t|A')P(A') + P(x \le t|A)P(A) = G(t)(1-\varepsilon) + H(t)\varepsilon.$$

Ja G un H blīvuma funkcijas ir g un h, tad attiecīgi f ir sekojoša blīvuma funkcija

$$f = (1-\varepsilon)g + \varepsilon h.$$

6. Piemēri

Salīdzināju parastā butstrapa, procentīļu intervāla butstrapa un impliecēto varbūtību butstrapa pārklājuma precizitāti simulētajam normālajam sadalījumam ar piejaukumu $N(0,1) * (0.95) + 0.05 * N(10,1)$. Dati apkopoti tabulā 1.. Pēc tabulas rezultātiem redzams, ka implicēto varbūtību butstraps dod labāku rezultātu nekā parastais butstraps pie $n = 30$, $n = 50$ un $n = 100$. Tika salīdzināti arī ticamības intervāli vidējai vērtībai



(a) Kastu grafiks (b) Histogramma

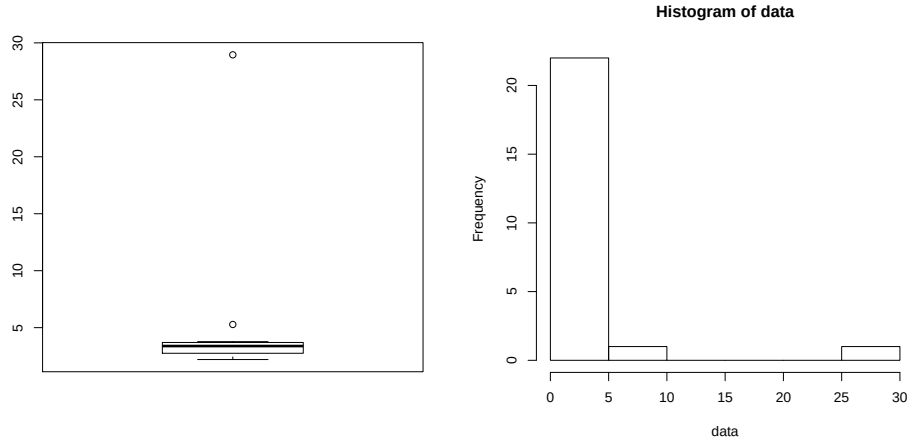
1. att.: 1.(a) Kastu grafiks un 1.(b) histogramma ar normālo līkni simulētajiem datiem ar $N(0,1) * (0.95) + 0.05 * N(10,1)$ pie $n = 50$

1. tabula: Parastā butstrapa (apz.B), procentīļu butstrapa (apz.PB) un implicēto varbutību butstrapa (apz.IVB) pārklājma precizitātes salīdzinājums simulētajiem datiem ar $N(0,1) * (0.95) + 0.05 * N(10,1)$ pie dažādiem n un N

	N=200			N=500			N=1000		
n	B	PB	IVB	B	PB	IVB	B	PB	IVB
20	0.93	0.86	0.915	0.94	0.854	0.916	0.934	0.863	0.913
30	0.885	0.73	0.925	0.868	0.81	0.894	0.9	0.79	0.895
50	0.735	0.725	0.77	0.758	0.668	0.78	0.786	0.649	0.81
100	0.475	0.37	0.525	0.392	0.378	0.442	0.451	0.364	0.465

diviem datu piemēriem ar dažādām metodēm: t-testu, parasto butstrapu, pivotālie, procentīļu un implicēto varbūtību butstrapu. Implicētajām varbūtībām kā vidējās vērtības novērtējumu izmantoju mediānu, kas ir robusts novērtējums. Pēc tabulas 2. redzams,

ka visšaurākais ticamības intervāls vidējai vērtībai ir implicēto varbūtību butstrapam, tas ir, aptuveni 1.99, kas diez gan atšķiras no procentīļu ticamības intervāla, kurš dod otro labāko rezultātu, tas ir, aptuveni 3.64.



(a) Kastu grafiks (b) Histogramma

2. att. Datu piemēram ar 24 novērojumiem

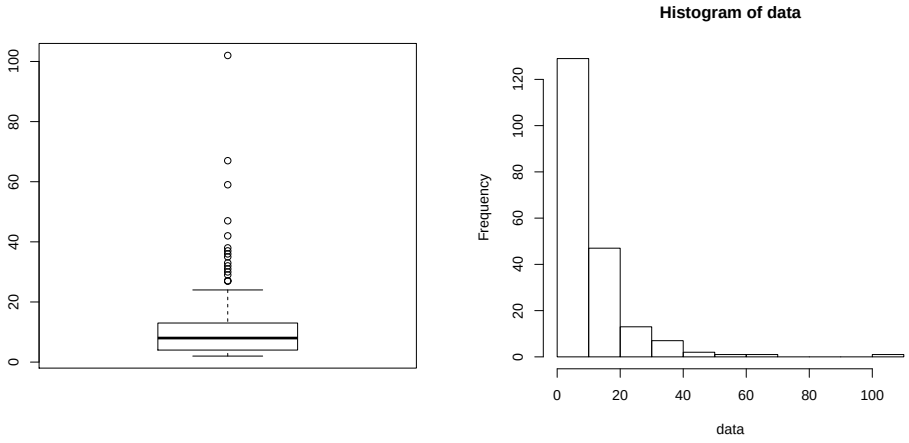
2. tabula: Ticamības intervālu salīdzināšana dažādām metodēm. Dati ir 24 novērojumi par vara sastāvu rupja maluma miltos (daļās uz miljona), sakārtoti augošā secībā (Analytical Methods Committe, 1989).

metode	a	b	garums
t tests	2.043523	6.517311	4.473788
Parastais butstraps	2.184072	6.376761	4.192689
pivotālie intervāli	2.059167	5.543333	3.484166
procentīļu intervāli	3.007917	6.649583	3.641666
implicēto varbūtību butstraps	3.285342	5.275491	1.990149

Otrajam datu piemēram, tas ir, datiem *los* vissliktāko rezultātu dod parastā butstrapa ticamības intervāls, tas ir, aptuveni 3.36, bet vislabāko implicēto varbūtību butstraps, tas ir, aptuveni 2.39. Otrais labākais rezultāts ir pivotālajiem ticamības intervāliem.

3. tabula: Ticamības intervālu salīdzināšana dažādām metodēm. Dati ir uzturēšanās ilgums 201 pacientam, kas palika Lozannas Universitātes slimnīcā 2000. gadā

metode	a	b	garums
t tests	9.603189	12.89432	3.291134
Parastais butstraps	9.567872	12.92964	3.361768
pivotālie intervāli	9.656716	12.75124	3.094524
procentīļu intervāli	9.701493	12.97512	3.273627
implicēto varbūtību butstraps	10.05564	12.44187	2.38623



(a) Kastu grafiks (b) Histogramma

3. att. Datiem los

Secinājumi

Kursa darbā tika aplūkots implicēto varbūtību butstraps. Implicētās varbūtības tika meklētas pēc Back un Brown (1993) formulas. Sniegts neliels ieskats par robustu statistiku, visparināto momentu metodi (GMM), butstrapu un tā dažādajiem ticamības intervāliem.

Praktiskajā daļā programmā R tika salīdzināts implicēto varbūtību butstrapa ticamības intervāli vidējai vērtībai diviem datu piemēriem un pārklājuma precizitāte simulētajiem datiem, kuriem ir normālais sadalījums ar piejaukumu. Dotajiem datu piemēriem implicēto varbūtību butstraps ievērojami samazina ticamības intrevāla garumu vidējai vērtībai.

Programmas R kodi

Datu piemēriem ticamības intervālu konstruēšana.

```
#data<-scan(file="piemers.txt")
library(robustbase)
data<-los
n<-length(data)
m<-mean(data)
m
alfa<-0.05
# t tests
t.test(data)$conf.int
##########################################
# 1. metode - "normalie" butstrapa intervali
##########################################
Disp.nov<-function(N,T.fun,dati)
{
resamples<-lapply(1:N,function(i)
sample(dati,replace=T))
r.mean<-sapply(resamples, T.fun)
var(r.mean)
}
alpha<-0.05
d.mean<-Disp.nov(1000,function(x) mean(x),data)
mean(data)-qnorm(1-alpha/2)*sqrt(d.mean)
mean(data)+qnorm(1-alpha/2)*sqrt(d.mean)

##########################################
#2. metode - pivotālie butstrapa intervali
##########################################
T.but<-function(N,T.fun,dati)
{
resamples<-lapply(1:N,function(i)
```

```
sample(dati,replace=T))
r.mean<-sapply(resamples, T.fun)
r.mean
}
N<-1000
but.mean<-T.but(N,function(x) mean(x), data)
2*mean(data)-sort(but.mean)[(1-alpha/2)*N]
2*mean(data)-sort(but.mean)[(alpha/2)*N]

##############################################
# 3. metode - procentīļu butstrapa intervāli
##############################################
N<-1000
but.mean<-T.but(N,function(x) mean(x), data)
sort(but.mean)[(alpha/2*N)]
sort(but.mean)[(1-alpha/2)*N]

#####################################################
#implied probability butstrap
#####################################################
varb<-function(data)
{

n<-length(data)
g.fun<-function(x,vid) x-vid
g.sv<-mean(g.fun(data,median(data)))
pii<-1/n-1/n*(g.fun(data,median(data))-g.sv)*g.sv/mean(g.fun(data,median(data))^2)
eps<--n*min(pii,0)
pii.vil<-(1/(1+eps))*pii+(eps/(1+eps))*1/n
pii.vil
}
```

```
Disp.nov<-function(N,T.fun,dati,pi)
{
resamples<-lapply(1:N,function(i)
sample(dati,replace=T,prob=pi))
r.mean<-sapply(resamples, T.fun)
var(r.mean)
}
alpha<-0.05
d.mean<-Disp.nov(1000,function(x) mean(x),data,varb(data))
mean(data)-qnorm(1-alpha/2)*sqrt(d.mean)
mean(data)+qnorm(1-alpha/2)*sqrt(d.mean)
```

Simulētajiem datiem pārklājuma precizitātes noteikšana

```
alfa<-0.05
N<-200
n<-50
# simulesana
eps<-0.05
set.seed(1)
m<-0
F<-function(x) dnorm(x,0,1)
G<-function(x) dnorm(x,10,1)
fun<-function(y) integrate(function(x) F(x)*(1-eps)+G(x)*eps, -20,y)$value
data.gen<-function(n)
{
  X<-runif(n)
  data<-c()
for (i in 1:n)
{
f<-function(x) (fun(x)-X[i])
data[i]<-uniroot(f, c(-30,30))$root
}
 data
```

```
}

# parastais butstraps
but.int<-function(dati,N,m)
{
resamples<-lapply(1:N,function(i) sample(dati,replace=T))
r.mean<-sapply(resamples, mean)
d.mean<-var(r.mean)
a<-mean(dati)-qnorm(1-alfa/2)*sqrt(d.mean)
b<-mean(dati)+qnorm(1-alfa/2)*sqrt(d.mean)
(b-m)*(a-m)
}

rez<-replicate(N,but.int(data.gen(n),1000,0))
length(rez[rez<0])/N

#procentīļu intervāli
T.but<-function(dati,N,m)
{
resamples<-lapply(1:N,function(i)
sample(dati,replace=T))
r.mean<-sapply(resamples, mean)
a<-sort(r.mean)[(alfa/2*N)]
b<-sort(r.mean)[(1-alfa/2)*N]
(b-m)*(a-m)
}

rez<-replicate(N,T.but(data.gen(n),1000,0))
length(rez[rez<0])/N

# implied probability bootstrap
varb<-function(data)
```

```
{
n<-length(data)
g.sv<-mean(data-median(data))
pii<-1/n-1/n*((data-median(data))-g.sv)*g.sv/mean((data-median(data))^2)
eps<-(-1)*n*min(pii,0)
pii.viln<-(1/(1+eps))*pii+(eps/(1+eps))*(1/n)
for (i in 1:n) {
if( (pii.viln[i] <0) ) pii.viln[i]<-0 }
pii.viln
}

im.but.int<-function(dati,N,m,pi)
{
resamples<-lapply(1:N,function(i) sample(dati,replace=T, prob=(pi)))
r.mean<-sapply(resamples, mean)
d.mean<-var(r.mean)

a<-mean(dati)-qnorm(1-alfa/2)*sqrt(d.mean)
b<-mean(dati)+qnorm(1-alfa/2)*sqrt(d.mean)
(b-m)*(a-m)
}
rez<-replicate(N,im.but.int(data.gen(n),1000,0,varb(data.gen(n))))
length(rez[rez<0])/N
```

Izmantotā literatūra un avoti

[1] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust statistics:The approach based on influence.* John Wiley & Sons, 2011.

[2] R.A Maronna, R.Douglas Martin, and V.J. Yohai. *Robust Statistic, Theory and Methods.* John Wiley & Sons, 2006.

[3] Bryan W. Brown and Whitney K. Newey. Gmm, efficient bootstrapping, and improved inference. *Journal of Business & Economic Statistics*, 20(4), 2001.

[4] Aviv Nevo. Sample selection and information-theoretic alternatives to gmm. *Journal of Econometrics*, (107), 2002.

[5] Anirban DasGupta. *Asymptotic theory of statistics and probability.* Springer New York, 2008.

[6] L. Wasserman. *All of nonparametric statistics.* Springer New York, 2006.

[7] Lorenzo Camponovo and Taisuke Otsu. Breakdown point theory for implied probability bootstrap. *Cowles fooundation discussion paper*, (1793), 2011.