LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKĀS TICAMĪBAS FUNKCIJA ROBUSTAJĀ STATISTIKĀ

MAĢISTRA DARBS

Autors: **Māra Vēliņa**

Stud. apl. nr. mv10047

Darba vadītājs: docents Dr. mat. Jānis Valeinis

RĪGA 2012

Anotācija

Maģistra darbā aplūkotas robustās statistikas pamatnostādnes un gludie M-novērtējumi. Darbā ieviesta robusta empīriskās ticamības (EL) metode divu gludu Hūbera M-novērtējumu starpībai, balstoties uz Qin un Zhao rezultātiem EL metodei divu izlašu gadījumā. Metode prasa noteiktus gluduma nosacījumus EL nenovirzītajiem vienādojumiem, un to pielietot Hūbera novērtējumam bija iespējams pateicoties nesen Hampel definētajam gludajam Hūbera novērtējumam, kuram šie nosacījumi izpildās. Metode tika pielietota simulāciju analīzē dažādiem piesārņotiem un nepiesārņotiem sadalījumiem un reālu datu piemēriem un tika salīdzināta ar EL metodi vidējo vērtību starpībai un divu izlašu t-testu. Tika secināts, ka jaunā metode darbojas līdzīgi EL metodei vidējo vērtību starpībai, bet tai ir priekšrocības situācijās, kad datos ir piesārņojums.

Atslēgas vārdi: robustā statistika, M-novērtējumi, gludie M-novērtējumi, gludais Hūbera novērtējums, empīriskās ticamības metode, divu izlašu problemātika

Abstract

Thesis outlines the basic ideas of robust statistics and smooth M-estimators. Empirical likelihood (EL) method for the difference of two smoothed Huber M-estimators has been implemented based on the results of Qin and Zhao. This method requires some smoothness conditions on the EL estimating equations which are satisfied by the smooth Huber M-estimator recently defined by Hampel. Simulation analysis was performed for various contaminated and uncontaminated distributions and several real data sets were analysed. The new method was compared to EL for the difference of two means and to two sample t-test. It was concluded that the new method works similarly to the EL for the difference of two means but outperforms it where contaminated data is involved.

Keywords: robust statistics, M-estimators, smooth M-estimators, smooth Huber estimator, empirical likelihood, two sample case

Saturs

Ievads

Pieņēmumi par normalitāti, linearitāti vai neatkarību parādās praktiski visās statistikas metodēs. Tomēr jāņem vērā, ka jebkura matemātiska modeļa pieņēmumi ir realitātes aptuvens apraksts. Daudzas parametriskās statistikas metodes ir konstruētas tā, lai tās būtu optimālās pie idealizētā modeļa, un to sniegums būtiski pasliktinās pat pie šķietami mazām novirzēm no šī modeļa. Šo problēmu pēta robustās statistikas teorija, kas aplūko novirzes no dažādiem parametrisko modeļu pieņēmumiem, pēta to efektu un meklē metodes, kas būtu pēc iespējas stabilas pret šādām novirzēm. Daži robustās statistikas jautājumi ir pazīstami jau kopš statistikas metožu pirmssākumiem, piemēram, jautājums par izlēcēju ietekmi un tās mazināšanas iespējām, bet formālas un visaptverošas robustības teorijas attīstība sākās XX gadsimta 60.-tajos un 70.-gados ar John Tukey (1960 [1], 1962 [2]), Peter J. Huber (1964 [3], 1967 [4]) un Frank Hampel (1971 [5], 1974 [6]) darbiem. Viens no pamatakmeņiem robustās statistikas teorijā bija Hūbera "Robust estimation of location"[3], kurā Hūbers definēja M-novērtējumu klasi, kas vispārināja maksimālās ticamības metodes ideju un vēlāk kalpoja par pamatu sarežģītāku robustu metožu, piemēram, robustas regresiju analīzes, korelāciju un variāciju analīzes izstrādāšanā. M-novērtējumu ietvarā Hūbers definēja kādu īpaši svarīgu novērtējumu – Hūbera novērtējumu, kuram piemīt optimālā dispersija minimax nozīmē normālā modeļa apkārtnē.

Robustā statistika reizēm tiek salīdzināta ar neparametrisko statistiku. Neparametriskās statistikas metodes balstās uz ļoti vispārīgiem pieņēmumiem par datiem, piemēram, to neatkarību un nepārtrauktību, bet neprasa veikt pieņēmumus par sadalījuma veidu. Starp abām statistikas nozarēm ir zināmas līdzības – daudzas neparametriskās procedūras ir izrādījušās robustas, turklāt neparametriskās statistikas metodes bieži radušās kā atbilde uz klasisko procedūru robustuma trūkumu. Tomēr robustā statistika ir parametriska statistika, jo apskata statistiku uzvedību konkrētu parametrisku modeļu apkārtnē.

Ir atrodami piemēri, kur abu statistikas nozaru metodes tikušas savienotas un savstarpēji papildinātas. Mūsdienu parametriskajā statistikā stabilu vietu ir ieņēmusi Owen radītā empīriskās ticamības metode [7], [8], [9], kuru, tāpat kā tās parametrisko līdzinieci– maksimālas ticamības funkcijas metodi, var izmantot parametru novērtēšanai, hipotēžu pārbaudei un novērtējumu ticamības intervālu konstruēšanai. Owen jau pirmajā empīriskās ticamības metodei veltītajā publikācijā [7] pierādīja, ka empīriskās ticamības intervālus iespējams konstruēt Hūbera definētajiem M-novērtējumiem. Turklāt veids, kādā definēti

M-novērtējumi, ļoti dabiski iekļaujas vispārīgajā Qin un Lawless [10] empīriskās ticamības metodes definīcijā ar nenovirzītiem vienādojumiem. Vēlāk Tsao un Zhou [11] pierādīja, ka uz robustiem M-novērtējumiem balstīti empīriskās ticamības intervāli saglabā punktveida novērtējumiem piemītošo robustumu un tādēļ ir īpaši interesanti situācijās, kad datos var parādīties izlēcēji.

Empīriskās ticamības metodi divu izlašu vidējo vērtību un sadalījuma funkciju starpībām vispirms aprakstīja Qin un Zhao [12]. Šajā metodē ir būtiski, lai nenovirzītajiem vienādojumiem izpildītos noteikti gluduma nosacījumi, taču parastajam Hūbera novērtējumam šie nosacījumi nav spēkā. Nesenā publikācijā [13] Hampel definēja gludināšanas principu M-novērtējumiem, un radās ideja šādam gludinātam Hūbera novērtējumam pielietot empīriskās ticamības metodi. Līdz ar to radās arī maģistra darba galvenais mērķis: īstenot empīriskās ticamības metodi divu gludu Hūbera novērtējumu starpībai un novērtēt tā sniegumu, salīdzinot ar citām klasiskām statistikas metodēm. Darba mērķis prasa sekojošo uzdevumu izpildi:

1. Iepazīties ar robustās statistikas teorijas pamatjautājumiem un metodēm, īpaši M-novērtējumiem.

2. Apskatīt gludos M-novērtējumus un salīdzināt to sniegumu ar parastajiem M- novērtējumiem.

3. Izveidot programmu divu gludu Hūbera novērtējumu starpības empīriskās ticamības metodes aprēķināšanai. Analizēt metodi, salīdzinot tās pārklājuma precizitāti un jaudu ar citām klasiskām statistikas metodēm reālu datu un simulāciju piemēros.

Empīriskās ticamības metode divu Hūbera novērtējumu starpībai tika prezentēta konferencē International Conference on Robust Statistics (autors Jānis Valeinis, līdzautori Māra Vēliņa un ar George Luta, University of Georgetown, Washington DC, USA), sagatavošanā ir arī publikācija. Divu izlašu problemātikai tika lietota Valeiņa un Cera izstrādātā R programma, kas publiski pieejama kā R paplašinājumprogramma *EL*.

Darbs sastāv no ievada, piecām nodaļām, secinājumiem un literatūras avotu saraksta. 1. nodaļā apskatīta robustās statistikas teorija, 2. nodaļā aprakstīti M-novērtējumi, 3. nodaļā tuvāk apskatīts gludais Hūbera novērtējums, 4. nodaļā izklāstīta empīriskās ticamības metode un 5. nodaļā atrodama simulāciju un datu analīze. Visi darbā veiktie aprēķini īstenoti programmā R.

1. Robustās statistikas teorija

Statistiskās metodes balstās ne tikai uz novērotajiem datiem, bet arī uz virkni tiešu vai netiešu pieņēmumu. Modeļa pieņēmumi atspoguļo statistiķim zināmo informāciju par risināmo problēmu, kā arī tie palīdz formulēt tādu matemātisku modeli, kuru būtu iespējams atrisināt no teorētiskā un izskaitļojamības viedokļa. Modeļi neizbēgami ir realitātes vienkāršojums, un tie labākajā gadījumā aptuveni raksturo reālo situāciju. Statistikā visplašāk lietotais pieņēmums ir par to, ka novērotie dati ir normāli sadalīti. Šis pieņēmums ir bijis pamatā visu klasisko regresiju analīzes, dispersiju analīzes un daudzdimensiju statistikas metožu konstruēšanā. Normālā sadalījuma lietošanu attaisno fakts, ka tas labi reprezentē lielu daļu reāli novērotu datu kopu, turklāt tas ir ērti lietojams no teorētiskā viedokļa, ļaujot atklātā veidā iegūt optimālu statistisko metožu formulas, kā piemēram, maksimālās ticamības metode vai maksimālās ticamības attiecības testi.

Bieži vien pieņēmums par normalitāti (piemēram, regresija ar normāli sadalītiem atlikumiem) ir spēkā tikai aptuveni, tādā nozīmē, ka normālais sadalījums labi raksturo lielāko daļu novērojumu, tomēr daži novērojumi atbilst kādai citai struktūrai, vai varbūt tiem nav saskatāma nekāda struktūra. Šādas netipiskas datu vienības sauc par *izlēcējiem*, un pat viens izlēcējs var kropļojoši ietekmēt klasiskas uz normalitāti balstītas statistikas metodes rezultātu.

Varētu cerēt, ka, ja šāda "aptuvena" normalitāte ir spēkā, tad uz normalitātes teoriju balstītajiem rezultātiem arī vajadzētu būt aptuveni pareiziem. Diemžēl tā tas nav. Ja tiek pieņemts, ka dati ir normāli sadalīti, bet patiesajam sadalījumam ir "smagas astes", tad uz maksimālās ticamības metodi balstītie novērtējumi ne tikai pārstāj būt optimālie, bet tiem var būt zema statistiskā efektivitāte (t.i., nepieņemami liela dispersija), ja astes ir simetriskas, vai arī pārlieku liels biass (parametra novērtējuma matemātiskās cerības atšķirība no tā patiesās vērtības), ja astes ir asimetriskas. Turklāt, klasiskie testi šādā gadījuma zaudē jaudu, savukārt klasiskie ticamības intervāli var būt neprecīzi vai pārāk plaši.

Kopš XX gadsimta vidus izpratne par to, ka daudzas no biežāk lietotajām statistikas procedūrām ir pārlieku jutīgas pret šķietami nenozīmīgām novirzēm no modeļa pieņēmumiem, arvien pieaugusi un tāpēc radīta virkne robustu procedūru. Terminu "robusts" Peter J. Huber [14] formulē sekojoši: *robustība nozīmē metodes nejutīgumu pret nelielām novirzēm no modeļa pieņēmumiem.* Robustās statistikas mērķis ir izstrādāt metodes, kas

dod uzticamus parametru novērtējumus un tiem atbilstošos testus un ticamības intervālus, ne tikai tad, kad dati precīzi atbilst noteiktajam sadalījumam, bet arī situācijā, kad tam iepriekš aprakstītajā nozīmē atbilst tikai aptuveni. Visbiežāk analizētais gadījums ir aptuveni normāli sadalījumi, tomēr šī pieeja labi darbojas arī citiem aptuveniem sadalījumiem, piemēram, aptuvenam gamma sadalījumam nesimetriski sadalītu datu gadījumā.

Robustās statistikas teorija sāka attīstīties XX gadsimta 60.-tajos un 70.-gados ar fundamentālajiem John Tukey (1960 [1], 1962 [2]), Peter J. Huber (1964 [3], 1967 [4]) un Frank Hampel (1971 [5], 1974 [6]) darbiem. Robusto metožu attīstība kļuva iespējama, pateicoties augošajai datoru pieejamībai un ātrumam, jo, atšķirībā no klasiskajām metodēm, robustu metožu novērtēšana bieži iekļauj nelineāru problēmu un optimizācijas uzdevumu risināšanu. Nozīmīgas grāmatas robustās statistikas teorijā sarakstījuši Huber (1981 [15]), Hampel, Ronchetti, Rousseeuw and Stahel (1986 [16]), Huber un Ronchetti (2009 [14]), uz praktiskām metodēm orientētu grāmatu ar S-PLUS pielietojumiem – Maronna, Martin un Yohai (2006 [17]).

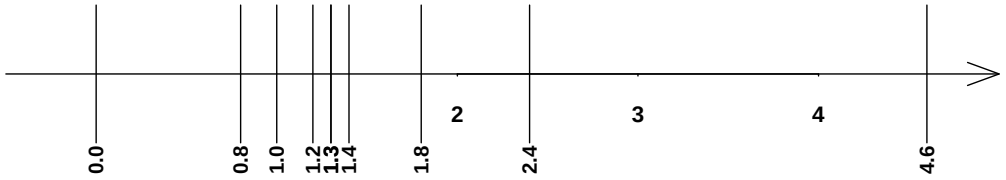
1.1. Datu diagnostika

Lai ilustrētu izlēcēju ietekmi uz klasiskās statistikas metodēm, aplūkosim Hampel [16] analizēto piemēru, kas satur datus par divu medikamentu miega paildzināšanas efektu.

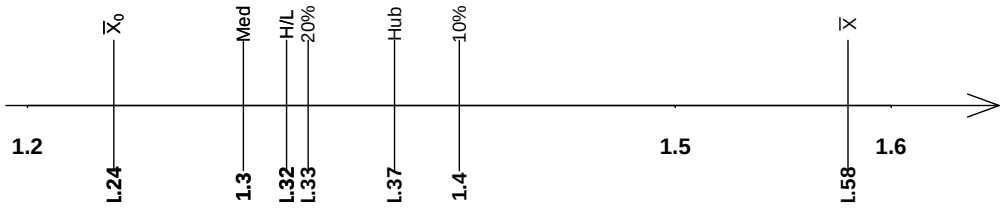
Piemērs 1. (Cushny and Peebles dati, [16, 78. lpp.]) Desmit personām reģistrētas atšķirības starp divu medikamentu miega paildzināšanas radītiem efektiem:

$$0.0,\ 0.8,\ 1.0,\ 1.2,\ 1.3,\ 1.3,\ 1.4,\ 1.8,\ 2.4,\ 4.6. \tag{1.1}$$

Kā norāda Hampel, šie dati ilgstoši dažādās grāmatās citēti kā normāli sadalītu datu paraugs un raksturots ar tā vidējo vērtību. Aplūkojot novērojumu vērtības 1.(a) attēlā, rodas aizdomas, ka novērojums 4.6 varētu būt izlēcējs. Savukārt 1.(b) attēlā redzami dažādi šīs izlases lokācijas parametra novērtējumi: parastā vidējā vērtība 1.58 un seši dažādi robusti novērtējumi: 10% nošķeltā vidējā vērtība tiek aprēķināta, izslēdzot 10% lielākos un 10% mazākos novērojumus no izlases un aprēķinot vidējo vērtību no atlikušajiem novērojumiem, 20% nošķeltā vidējā gadījumā attiecīgi tiek izslēgti 20% lielākie un mazākie novērojumi. Hodges-Lehmann novērtējumu definē kā mediānu no visu novērojumu pāru vidējām vērtībām: $(X_i + X_j)/2$, $i, j = 1, \ldots, 10$. Hūbera novērtējums ir robusts novērtējums, kas tiks definēts 2. nodaļā. Redzams, ka robustie novērtējumi pieņem vērtības no



(a)



(b)

1. att.: Cushnee an Peebles dati: (a) desmit novērojumu vērtības; (b) septiņi lokācijas parametra novērtējumi: $\bar{x}$ – vidējā vērtība, 10% – 10% nošķeltais vidējais, 20% – 20% nošķeltais vidējais, H/L – Hodges Lehmann novērtējums, Med – mediāna, Hub- Hūbera novērtējums, $\bar{X}_0$ – vidējā vērtība bez novērojuma 4.6.

1.24 līdz 1.4, un, pat ja tiem ir atšķirīgas robustuma pakāpes, tie visi atrodas samērā tālu no vidējās vērtības 1.58.

Jāatzīmē, ka vidējās vērtības $\bar{X}_0$ novērtējums izlasei bez ekstrēmās vērtības 4.6 varētu tikt iegūts, lietojot kādu labu *datu diagnostikas metodi.* Rodas jautājums, kādēļ nepieciešamas robustas metodes, ja iespējams lietot labi pazīstamās divu soļu procedūras: (1) attīrīt datu izlasi, lietojot kādu izlēcēju noteikšanas metodi; (2) pielietot klasiskās statistikas novērtēšanas un hipotēžu pārbaudes metodes atlikušajiem datiem. Tomēr divu soļu procedūrai ir dažādi trūkumi. Pirmkārt, ir situācijas, kad abus soļu nošķirt nav iespējams, piemēram, daudzparametru regresijā noteikt izlēcējus ir praktiski neiespējami,

ja nav iegūti robusti parametru novērtējumi. Otrkārt, attīrītie dati vienalga nebūs normāli sadalīti, jo iespējama gan kļūdaina datu izslēgšana, gan iekļaušana. Treškārt, to var uzskatīt par empīrisku faktu, ka labāko izlēcēju izslēgšanas procedūru sniegums nepanāk labāko robusto metožu sniegumu (skat. [16, 56.–71. lpp.]). Un visbeidzot, dažādi praktiski piemēri rāda, ka klasiskās izlecēju izslēgšanas metodes bieži nestrādā, ja izlasē ir vairāki izlēcēji: lielākie izlēcēji, būtiski palielinot izlases dispersiju, "noslēpj" mazāk nozīmīgos izlēcējus. Turklāt jāņem vērā arī fakts, ka tad, kad izlēcēji ir noteikti, statistiķim vēl ir jāpieņem subjektīvs lēmums, ko ar tiem darīt.

1.2. Robustu procedūru vēlamās īpašības

Pieņemsim, ka dots parametrisks modelis, un pieņemsim, ka ir pamats cerēt, ka tas pietiekoši labi apraksta reālo situāciju, tomēr netiek uzskatīts, ka pieņēmumi ir pilnīgi pareizi. Sekojot Huber [14] pieejai, šādā situācijā ikvienai robustai statistikas metodei būtu jāsasniedz sekojoši mērķi:

1. Efektivitāte. Metodei ir pietiekoši laba efektivitāte (dispersijas lieluma nozīmē) pie pieņēmumos aprakstītā modeļa – tā ir optimālā vai "gandrīz" optimālā procedūra.

2. Stabilitāte. Mazas novirzes no modeļa pieņēmumiem tikai nenozīmīgi pasliktina metodes sniegumu. Sniegums ir izmērāms ar, piemēram, asimptotisko dispersiju vai biasu punktveida novērtējumu gadījumā, vai izmēru un jaudu testu gadījumā.

3. Lūzums. Nozīmīgām novirzēm no modeļa nevajadzētu izrādīties katastrofālām. Šo īpašību raksturo statistikas *lūzuma punkts* – mazākais sliktu novērojumu īpatsvars izlasē, pie kura statistika var pieņemt patvaļīgi lielas vērtības. Matemātiski precīzāka definīcija tiks sniegta nākamajā nodaļā.

Visi trīs aspekti ir būtiski, tomēr vērts atzīmēt, ka rupjās kļūdas jeb neliels skaits novērojumu datos, kas mēra to pašu daudzumu, bet kuru kļūdas ir lielākas (t.i., tiem piemīt lielāka dispersija) nekā pārējiem novērojumiem, ir uzskatāmas par nelielu novirzi no modeļa. Tādēļ robustu procedūru pirmais mērķis ir nodrošināties pret rupjajām kļūdām. Jāņem arī vērā, ka robustība ir kompromiss, jeb, citējot Anscombe [18] metaforu, par robustību ir dabiski jāmaksā "apdrošināšanas prēmija" – zināma daļa procedūras efektivitātes pie modeļa.

1.3. Robustības mēri

Kā tika norādīts iepriekš, parametriskās statistikas metožu mērķis ir definēt novērtējumus, kas ir optimāli pie kāda precīzi uzdota modeļa, bet robustā statistika meklē novērtējumus, kas pietiekoši labi darbojas kādā modeļa apkārtnē. Šajā nodaļā tiks sniegts matemātiski precīzāks skaidrojums šai idejai, aprakstot dažas metodes novērtējumu robustības pakāpes mērīšanai.

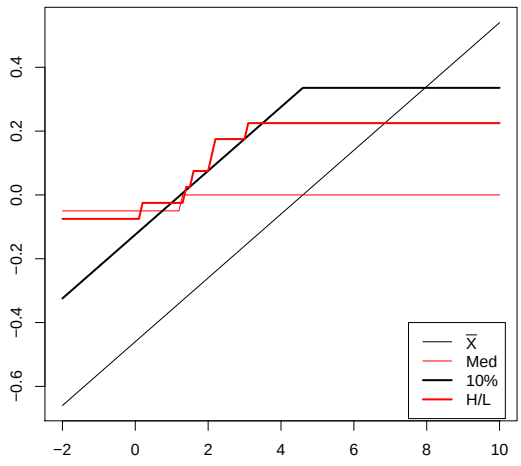
Ietekmes funkcija

Definīcija 1. Pieņemsim, ka dota novērojumu izlase $X_1, \ldots, X_n$. Par novērtējuma $\hat{\theta}$ *jutīguma līkni* izlasei $X_1, \ldots, X_n$ sauc funkciju

$$SC(x_0) = \hat{\theta}(X_1, \ldots, X_n, x_0) - \hat{\theta}(X_1, \ldots, X_n), \tag{1.2}$$

kas ir funkcija no izlēcēja x_0.

2. attēlā ir redzama jutīguma līkne Cushnee and Peebles datiem un dažiem no 1. piemērā aplūkotajiem novērtējumiem: vidējai vērtībai, 10% nošķeltajai vidējai vērtībai, mediānai un Hodges–Lehmann novērtējumam. Redzams, ka visām statistikām, izņemot vidējo vērtību, jutīguma līkne ir ierobežota funkcija. Tātad var ievērot, ka robustām statistikām ir raksturīga ierobežota jutīguma līkne.



2. att.: Jutīguma līkne Cushnee un Peebles datiem: $\bar{x}$ – vidējai vērtībai, 10% – 10% nošķeltajai vidējai vērtībai, H/L – Hodges–Lehmann novērtējumam un Med – mediānai.

Lai pārbaudītu šo novērojumu, nepieciešams analizēt statistiku asimptotisko uzvedību. Jutīguma līknes asimptotiskā versija ir *ietekmes funkcija* (angl. - *influence function*), bet lai to definētu, nepieciešams novertējumus izteikt kā funkcionāļus. Pieņemsim, ka doti viendimensionāli neatkarīgi un vienādi sadalīti (i.i.d.) novērojumi $X_1, X_2, \ldots, X_n$, kas pieder kādai novērojumu telpai $\mathcal{X}$, kas ir reālās taisnes $\mathbb{R}$ apakškopa, turklāt $\mathcal{X}$ var arī sakrist ar $\mathbb{R}$.

Parametriska modeļa ietvaros aplūko sadalījuma funkciju saimi F_θ, kur nezināmais parametrs θ pieder kādai parametru telpai Θ. Klasiskajā statistikā pieņem, ka novērojumi X_i sadalīti pilnīgi atbilstoši kādai no sadalījuma funkcijām F_θ. Piemēram, $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$ un F_θ ir normālais sadalījums ar vidējo vērtību θ un standartnovirzi 1; $\mathcal{X} = [0, \infty]$, $\Theta = (0, \infty)$ un F_θ ir eksponenciālais sadalījums ar matemātisko cerību θ.

Aplūkojam parametra θ novērtējumus, kas ir reālvērtīgas statistikas formā

$$T_n = T_n(X_1, \ldots, X_n) = T_n(G_n).$$

Mūs interesē novērtējumi, kas ir funkcionāļi, t.i.,

$$T_n(G_n) = T(G_n),$$

vai arī, kas ir asimptotiski aizstājami ar funkcionāļiem.

Definīcija 2. Pieņemsim, ka eksistē funkcionālis $T : \mathit{domain}(T) \to \mathbb{R}$ (kur $\mathit{domain}(T)$ ir visu sadalījumu $\mathcal{F}(\mathcal{X})$ kopa, kurā T definēts), kam spēkā

$$T_n(X_1, \ldots, X_n) \xrightarrow{n\to\infty} T(G)$$

pēc varbūtības, ja novērojumi ir i.i.d. sadalīti pēc sadalījuma funkcijas G, $G \in \mathit{domain}(T)$. Tādā gadījumā saka, ka $T(G)$ ir statistiku virknes $\{T_n; n \geq 1\}$ *asimptotiskā vērtība pie sadalījuma funkcijas* G.

Definīcija 3. Saka, ka statistikai $T(G)$ ir spēkā *asimptotiskā normalitāte*, ja izpildās

$$\mathcal{L}_G(\sqrt{n}\,[T_n - T(G)]) \xrightarrow[n\to\infty]{d} N(0, V(T, G)), \tag{1.3}$$

kur $\mathcal{L}_G$ nozīmē dotās statistikas sadalījumu pie novērojumu sadalījuma G. $V(T, G)$ sauc par virknes $\{T_n; n \geq 1\}$ *asimptotisko dispersiju* pie sadalījuma G. Ja funkcionālim T izpildās

$$T(F_\theta) = \theta \quad \forall \theta \in \Theta, \tag{1.4}$$

tad saka, ka funkcionālis T ir Fišera konsistents.

Definīcija 4. [16, 84. lpp.] Par funkcionāļa T *ietekmes funkciju* pie sadalījuma F sauc funkciju

$$IF(x;T,F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon}, \tag{1.5}$$

tajos $x \in \mathcal{X}$, kur šī robeža eksistē.

Ietekmes funkcija raksturo punktā x atrodošos nelielu daudzumu izlēcēju ietekmi uz novērtējuma vērtību. Lielums $T((1-\epsilon)F + \epsilon\Delta_x)$ ir funkcionāļa asimptotiskā vērtība situācijā, kad sadalījums ir F un izlēcēju īpatsvars ϵ ir vienāds ar x_0. Var teikt, ka ietekmes funkcija rāda, kāda ir statistikas asimptotiskās vērtības uzvedība mazā apkārtnē.

No funkcionālanalīzes viedokļa ietekmes funkcija ir funkcionāļa T atvasinājums pēc Gato.

Definīcija 5. Funkcionāli T sauc par *atvasināmu pēc Gato (Gateaux)* punktā $F \in domain(T)$ (kur F ir sadalījuma funkcija), ja eksistē reāla funkcija a_1 tāda, ka visiem $G \in domain(T)$ spēkā

$$\lim_{t \to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int a_1(x) dG(x), \tag{1.6}$$

jeb ekvivalenti,

$$\frac{\partial}{\partial t} \left[T((1-t)F + tG)\right]_{t=0} = \int a_1(x) dG(x). \tag{1.7}$$

Ievietojot (1.7) $G = \Delta_x$ (pie nosacījuma, ka $\Delta_x \in domain(T)$), iegūst ietekmes funkciju (1.5).

Ietekmes funkciju var uztvert kā jutīguma līknes robežu. Pievienojot izlasei $X_1, \ldots, X_n$ vienu izlēcēju, piesārņojuma īpatsvars ir $1/(n+1)$. Definē *standartizēto jutīguma funkciju*

$$\begin{aligned} SC_n(x_0) &= \frac{\hat{\theta}_{n+1}(X_1, \ldots, X_n, x_0) - \hat{\theta}_n(X_1, \ldots, X_n)}{1/(n+1)} = \\ &= (n+1)\hat{\theta}_{n+1}(X_1, \ldots, X_n, x_0) - \hat{\theta}_n(X_1, \ldots, X_n) \end{aligned} \tag{1.8}$$

un iegūst ietekmes funkcijas (1.5) galīgu izlašu versiju ar $\epsilon = 1/(n+1)$.

Eksistē sakarība starp funkcionāļa ietekmes funkciju un tā asimptotisko dispersiju no (1.3)

$$V(T,F) = \int IF(x;T,F)^2 dF(x). \tag{1.9}$$

Šī sakarība seko no T pie sadalījuma F izvirzījuma Teilora rindā punktā F_n, kur F_n ir empīriskā sadalījuma funkcija (heiristisku paskaidrojumu var skatīt, piemēram, [16, 85. lpp.]). Šī sakarība arī parāda, ka, lai statistika būtu robusta, nepieciešams, lai tās ietekmes

funkcija būtu ierobežota. Ja novērtējumu virknes $\{T_n; n \geq 1\}$ asimptotiskajai vērtībai T izpildās Fišera konsistence, tad ir spēkā Krāmera – Rao nevienādība.

Definīcija 6. Pieņemsim, ka f_θ ir F_θ blīvuma funkcija un $F_* := F_{\theta_*}$, kur θ_* ir kāds fiksēts Θ elements. Tad *Fišera informāciju* $J(F_*)$ pie sadalījuma F_* definē

$$J(F_*) = \int \left(\frac{\partial}{\partial \theta} \left[\ln f_\theta(x) \right]_{\theta_*} \right)^2 dF_*. \tag{1.10}$$

Atvasinot ietekmes funkciju pēc θ punktā θ_*, iegūst Kramera – Rao nevienādību

$$\int IF(x; T, F_*)^2 dF_*(X) = V(T, F_*) \geq \frac{1}{J(F_*)}, \tag{1.11}$$

kur vienādība ir spēkā tad un tikai tad, ja $IF(x; T, F_*)$ ir proporcionāls $\frac{\partial}{\partial \theta} \left[\ln f_\theta(x) \right]_{\theta_*}$.

Aplūkosim dažus piemērus. Pieņemsim, ka dota novērojumu telpa $\mathcal{X} = \mathbb{R}$, parametru telpa $\Theta = \mathbb{R}$ un dots *normālais lokācijas modelis* $F_\theta(x) = \Phi(x - \theta)$, kur Φ ir standartnormālā sadalījuma funkcija. Pieņemsim, ka $\theta_0 = 0$, tātad $F_{\theta_0} = \Phi$.

Piemērs 2. Vidējā vērtība. $T_n = (1/n) \sum_{i=1}^n X_i$ un atbilstošais funkcionālis $T(G) = \int u dG(u)$ ir definēts visiem varbūtību mēriem ar galīgu pirmo momentu. T ir Fišera konsistents, un no (1.5) seko

$$\begin{aligned} IF(x; T, \Phi) &= \lim_{\epsilon \downarrow 0} \frac{\int u d \left[(1-\epsilon)\Phi + \epsilon \Delta_x \right](u) - \int u d\Phi(u)}{\epsilon} = \\ &= \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon) \int u d\Phi(u) + \epsilon \int u d\Delta_x(u) - \int u d\Phi(u)}{\epsilon} = \\ &= \lim_{\epsilon \downarrow 0} \frac{\epsilon x}{\epsilon} \end{aligned} \tag{1.12}$$

(jo $\int u d\Phi(u) = 0$), līdz ar to

$$IF(x; T, \Phi) = x. \tag{1.13}$$

Acīmredzami $\int IF(x; T, \Phi) d\Phi(x) = 0$ un asimptotiskā dispersija $V(T, \Phi) =$
$= \int IF(x; T, \Phi)^2 d\Phi(x) = 1$. Tātad vidējā vērtība nav robusts novērtējums, jo tā ietekmes funkcija nav ierobežota.

Piemērs 3. Izlases mediāna. Ja n ir nepāra skaitlis, tad $T_n = X_{((n+1)/2)}$, pretējā gadījumā $T_n = X_{(n/2)} + X_{(n/2+1)}$. Atbilstošais funkcionālis ir $T(G) = G^{-1}(1/2)$. gadījumā, ja šī vērtība nav viena vienīga, izvēlas intervāla $\{t : G(t) = 1/2\}$ viduspunktu, un gadījumā, ja G ir lēciens pie $1/2$, izvēlas vērtību, kur atrodas šis lēciens.

Apgalvojums 1. *[16, 89. lpp.] T ir Fišera konsistents un*

$$IF(x;T,\Phi) = \frac{sgn(x)}{2\phi(0)}, \tag{1.14}$$

kur ϕ ir standartnormālā sadalījuma blīvuma funkcija. Redzams, ka ietekmes funkcija ir ierobežota, tātad mediāna ir robusts novērtējums. Arī mediānas gadījumā $\int IF(x;T,\Phi)d\Phi(x) = 0$. Asimptotiskā dispersija ir

$$V(T,\Phi) = \int IF(x;T,\Phi)^2 d\Phi(x) = (2\phi(0))^{-2} = \pi/2 \approx 1.571. \tag{1.15}$$

Lūzuma punkts

Parametra θ novērtējuma $\hat{\theta}$ lūzuma punkts ir maksimālais piesārņojuma (netipisku datu) daudzums, pie kura $\hat{\theta}$ joprojām spēj sniegt informāciju par θ, tas ir, par tipisko datu punktu sadalījumu. Arī lūzuma punktu var definēt gan galīgām izlasēm, gan asimptotiskajam gadījumam. Lūzuma punkta teorija tiks izklāstīta, balstoties uz [17, 3.2. nodaļa] izklāstu.

Pieņemsim, ka $\theta \in \Theta$. Lai $\hat{\theta}$ būtu informatīvs par θ, piesārņojuma ietekme nedrīkst likt novērtējumam $\hat{\theta}$ tiekties uz bezgalību vai arī uz Θ robežu $\partial\Theta$. Piemēram, dispersijas novērtējuma gadījumā $\Theta = [0,\infty]$, novērtējumam jāpaliek ierobežotam un tas nedrīkst arī sasniegt robežu 0.

Definīcija 7. Par novērtējuma $\hat{\theta}$ *asimptotisko lūzuma punktu* pie sadalījuma funkcijas F sauc lielāko vērtību $\epsilon^* \in (0,1)$ tādu, ka visiem $\epsilon < \epsilon^*$ funkcija $T((1-\epsilon)F+\epsilon G)$ kā funkcija no G ir ierobežota un tai nav kopīgu punktu ar Θ robežu. Tas ir, eksistē ierobežota un slēgta kopa $K \in \Theta$ tāda, ka $K \cap \partial\Theta = \emptyset$ un

$$T((1-\epsilon)F + \epsilon G) \in K \quad \forall \epsilon < \epsilon^* \text{ un } \forall G. \tag{1.16}$$

Dažreiz var būt noderīgāk apskatīt lūzuma punktu galīgai izlasei. Pieņemsim, ka $\hat{\theta}_n = \hat{\theta}_n(x)$ ir novērtējums, kas definēts izlasēm $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Definīcija 8. Par novērtējuma $\hat{\theta}_n$ *galīgu izlašu lūzuma punktu ar aizstāšanu* sauc maksimālo proporciju $\epsilon_n^*(\hat{\theta}_n, x)$ no novērojumiem, kurus aizstājot ar brīvi izvēlētiem izlēcējiem novērtējums $\hat{\theta}_n(y)$ paliek ierobežots un nešķeļas ar θ definīcijas kopas robežu $\partial\Theta$. Citiem vārdiem sakot, apzīmē ar $\mathcal{X}_m$ visas izlases $\mathbf{y}$, kam ar $\mathbf{x}$ sakrīt skaitā $n-m$ elementi:

$$\mathcal{X}_m = \{y : \#(\mathbf{y}) = n, \#(\mathbf{x} \cap \mathbf{y}) = n-m\}.$$

Tad, par galīgas izlases lūzuma punktu sauc maksimālo proporciju

$$\epsilon_n^*(\hat{\theta}_n, x) = \frac{m^*}{n},$$

kur $m^* = max\{m \geq 0 : \hat{\theta}_n(\mathbf{y})$ - ierobežots un nešķeļas ar $\partial\Theta \ \forall \mathbf{y} \in \mathcal{X}_m\}$.

Definīcija 9. Apzīmē ar $\mathcal{X}_m$ visas izlases $\mathbf{y}$ ar elementu skaitu $n+m$, kuras satur x:

$$\mathcal{X}_m = \{y : \#(\mathbf{y}) = n+m, \mathbf{x} \subset \mathbf{y}\}$$

Tad, par *galīgu izlašu lūzuma punktu ar papildināšanu* sauc maksimālo proporciju

$$\epsilon_n^{**}(\hat{\theta}_n, x) = \frac{m^{**}}{n+m},$$

kur $m^{**} = max\{m \geq 0 : \hat{\theta}_{n+m}(\mathbf{y})$ - ierobežots un nesakrīt ar $\partial\Theta \ \forall \mathbf{y} \in \mathcal{X}_m\}$.

Abas definīcijas ϵ^* un ϵ^{**} dod līdzīgas vērtības pie lieliem n, bet ϵ^* var izrādīties ērtāk lietojams, jo tas vienmēr aprēķināms pie viena un tā paša novērtējumu skaita n.

Piemērs 4. Vidējā vērtība. Skaidrs, ka, ja kaut vai vienu izlases elementu aizstāj ar izlēcēju $x^* \to \infty$, tad $\bar{x} \to \infty$. Līdz ar to vidējai vērtībai $\epsilon_n^* = 0$ un arī $\epsilon_n^{**} = 0$.

Piemērs 5. $\alpha_{\%}$ nošķeltā vidējā vērtība. Viegli redzēt, ka $m^* = [n\alpha]$ un līdz ar to $\epsilon_n^* = [\alpha n]/n$. Asimptotiskais lūzuma punkts $\epsilon^* = \alpha$.

Piemērs 6. Mediāna. Ievēro, ka mediāna ir vienāda ar 50% nošķelto vidējo vērtību. Līdz ar to, galīgu izlašu lūzuma punkts $\epsilon_n^* = [n/2]$, bet asimptotiskais lūzuma punkts $\epsilon^* = 0.5$.

Piemērs 7. Aplūko novērtējumu klasi $\hat{\mu}$, kam spēkā

$$\hat{\mu}(X_1 + c, \ldots, X_n + c) = \hat{\mu}(X_1, \ldots, X_n) + c. \tag{1.17}$$

Šādus novērtējumus sauc par lokācijas ekvivariantiem novērtējumiem.

Apgalvojums 2. *Lokācijas ekvivariantiem novērtējumiem ir spēkā*

$$\epsilon_n^* \leq \frac{1}{n}\left[\frac{n-1}{2}\right]. \tag{1.18}$$

Pierādījums [17, 76. lpp]. Ievēro, ka (1.18) ekvivalents tam, ka no $\epsilon < \epsilon^*$ seko $1-\epsilon > \epsilon^*$. Pieņemsim, ka $\epsilon < \epsilon^*$. Definē $F_t(x) = F(x-t)$ un

$$H_t = (1-\epsilon)F + \epsilon F_t \in \mathcal{F}_\epsilon, \ H_t^* = \epsilon F + (1-\epsilon)F_{-t} \in \mathcal{F}_{1-\epsilon}, \tag{1.19}$$

kur

$$\mathcal{F}_\epsilon = \{(1-\epsilon)F + \epsilon G : G \in \mathcal{G}\},$$

kur $\mathcal{G}$ ir visu sadalījuma funkciju klase. Apzīmē ar $\hat{\mu}_\infty(F)$ novērtējuma $\hat{\mu}$ asimptotisko vērtību pie sadalījuma funkcijas F, t.i.,
$\hat{\mu}_n \to_p \hat{\mu}_\infty(F)$. Ievērojot, ka $H_t(x) = H_t^*(x-t)$ un ņemot vērā ekvivariancei, seko

$$\hat{\mu}_\infty(H_t) = \hat{\mu}_\infty(H_t^*) + t \ \forall t.$$

Tā kā $\epsilon < \epsilon^*$, $\hat{\mu}_\infty(H_t)$ paliek ierobežots, kad $t \to \infty$, līdz ar to $\hat{\mu}_\infty(H_t^*)$ nav ierobežots un $H_t^* \in \mathcal{F}_{1-\epsilon}$, kas nozīme, ka $1-\epsilon > \epsilon^*$.

2. M-novērtējumi

2.1. M-novērtējumi kā maksimālās ticamības novērtējumu vispārinājums

1964. gadā P. J. Huber publicēja rakstu "Robust estimation of location", kurā tika vispārināts maksimālās ticamības novērtējumu jēdziens, aplūkojot to plašākas novērtējumu klases ietvaros, kura tika nosaukta par *M-novērtējumiem* (no angļu valodas – *generalized maximal likelihood*).

Definīcija 10. [16, 100. lpp.] Parametra θ maksimālās ticamības (ML) novērtējums ir vērtība $T_n = T_n(X_1, X_2, \ldots, X_n)$, kas maksimizē $\prod_{i=1}^{n} f_{T_n}(X_i)$, jeb ekvivalenti,

$$\sum_{i=1}^{n} [-\ln f_{T_n}(X_i)] \to \min_{T_n}.$$

Aplūkosim sakarību

$$\sum_{i=1}^{n} \rho(X_i, T_n) \to \min_{T_n}, \tag{2.1}$$

kur ρ ir kāda funkcija telpā $\mathcal{X} \times \Theta$. Turklāt, ja ρ eksistē atvasinājums $\psi(x, \theta) = (\partial/\partial\theta)\rho(x, \theta)$, tad (2.1) ir ekvivalenta sakarība

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0. \tag{2.2}$$

Definīcija 11. [16, 101. lpp.] Jebkuru novērtējumu, kas definēts formā (2.1) vai (2.2) sauc par M-novērtējumu. Turklāt piezīmēsim, ka, ja G_n ir izlases ģenerētā empīriskā sadalījuma funkcija, tad (2.2) atrisinājumu T_n var izteikt arī formā $T(G_n)$, kur T ir funkcionālis, kuru uzdod sakarība

$$\int \psi(x, T(G)) dG(x) = 0 \tag{2.3}$$

visiem sadalījumiem G, kam šis integrālis ir definēts.

2.2. M-novērtējumu īpašības

Ievietojot sakarībā (2.3) G vietā $F_{t,x} = (1-t)F + t\Delta_x$ un atvasinot pēc t, iegūst M-novērtējuma ietekmes funkciju

$$IF(x; \psi, F) = \frac{\psi(x, T(F))}{-\int \frac{\partial}{\partial\theta} [\psi(y, \theta)]_{T(F)} \, dF(y)}. \tag{2.4}$$

Gadījumā, kad tiek apskatīts lokācijas modelis $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$, $F_\theta(x) = F(x-\theta)$, tad dabīgi pieņemt, ka ψ-funkcija uzdota veidā

$$\psi(x,\theta) = \psi(x-\theta), \tag{2.5}$$

kur, lai izpildītos Fišera konsistence (1.4), ir spēkā

$$\int \psi dF = 0 \quad \text{jeb} \quad E_F\psi = 0. \tag{2.6}$$

Tad ietekmes funkcija ir formā

$$IF(x;\psi,G) = \frac{\psi(x-T(G))}{\int \psi'(y-T(G))dG(y)}, \tag{2.7}$$

un pie modelī pieņemtā sadalījuma F iegūst

$$IF(x;\psi,F) = \frac{\psi(x)}{\int \psi' dF} \tag{2.8}$$

pie nosacījuma, ka $\int \psi' dF \neq 0$. No sakarības (2.8) iegūst M-novērtējuma dispersiju pie modeļa:

$$V(\psi,F) = \frac{\int \psi^2 dF}{(\psi' dF)^2}. \tag{2.9}$$

Var pierādīt (skatīt, piemēram [15, 49–52. lpp.]), ka nedilstošai ψ-funkcijai un pie noteiktiem regularitātes nosacījumiem par sadalījumu F, M-novērtējumiem ir spēkā asimptotiskā normalitāte ar dispersiju V no (2.9).

M-novērtējumiem spēkā Krāmera – Rao nevienādība (1.11) ar

$$J(F) = \int \left(\frac{f'}{f}\right)^2 dF. \tag{2.10}$$

2.3. Piemēri

Piemērs 8. Maksimālās ticamības novērtējums. No 10. definīcijas seko, ka izvēloties $\rho(x,\theta) = -\ln f(x,\theta)$ un $\psi(x,\theta) = -\frac{\partial}{\partial\theta}\ln f(x,\theta)$, (2.1) un (2.2) definē maksimālās ticamības novērtējumu, kas sasniedz mazāko iespējamo asimptotisko dispersiju $V(\psi,F)$, t.i., Fišera informācijas inverso funkciju $J(F)^{-1}$.

Pieņemsim, ka $F = \Phi$, tad iegūst aritmētisko vidējo: $\psi(x,\theta) = x-\theta$, tātad no (2.6) seko, ka $\theta = EX$. $\rho(x,\theta) = (x-\theta)^2$, $V(\psi,\Phi) = J(\Phi)^{-1} = 1$.

Piemērs 9. Ja F ir dubultais eksponenciālais sadalījums ar blīvuma funkciju $f(x) = 1/2\exp(-|x|)$, tad kā maksimālās ticamības novērtējumu iegūst mediānu. $\rho(x,\theta) = |x-\theta|$, ρ eksistē atvasinājums ψ pie $x \neq 0$ un

$$\psi(x,\theta) = sgn(x-\theta).$$

Ievēro, ka $sgn(x) = I_{\{x>0\}} - I_{\{x<0\}}$, tad (2.2) ir formā

$$\sum_{i=1}^{n} sgn(X_i - \theta) = \sum_{i=1}^{n} (I_{\{X_i>\theta\}} - I_{\{X_i<\theta\}}) = \tag{2.11}$$

$$= \#(X_i > \theta) - \#(X_i < \theta) = 0, \tag{2.12}$$

no kurienes seko, ka $\#(X_i > \theta) = \#(X_i < \theta)$, tas ir, θ ir mediāna. Asimptotiskā dispersija ir $V(\psi, F) = J(F)^{-1} = 1$.

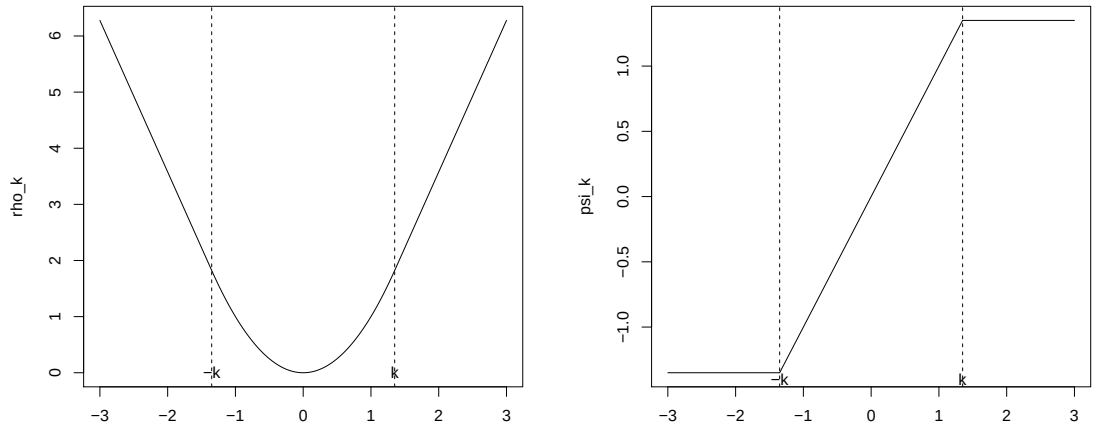
Piemērs 10. *Hūbera novērtējumu* definē ρ-funkcija

$$\rho(x) = \begin{cases} x^2, & |x| \le k \\ 2k|x| - k^2, & |x| > k, \end{cases} \tag{2.13}$$

kuras atvasinājums ir $2\psi(x)$, kur

$$\psi(x) = \begin{cases} k, & x \ge k \\ x, & -k \le x \le k \\ -k, & x \le -k, \end{cases} \tag{2.14}$$

kur konstante $k \in (0, \infty)$. Hūbera novērtējuma ρ un ψ funkcijas skatāmas 3. attēlā.



(a) (b)

3. att.: Hūbera novērtējuma ρ un ψ funkcijas, $k = 1.35$.

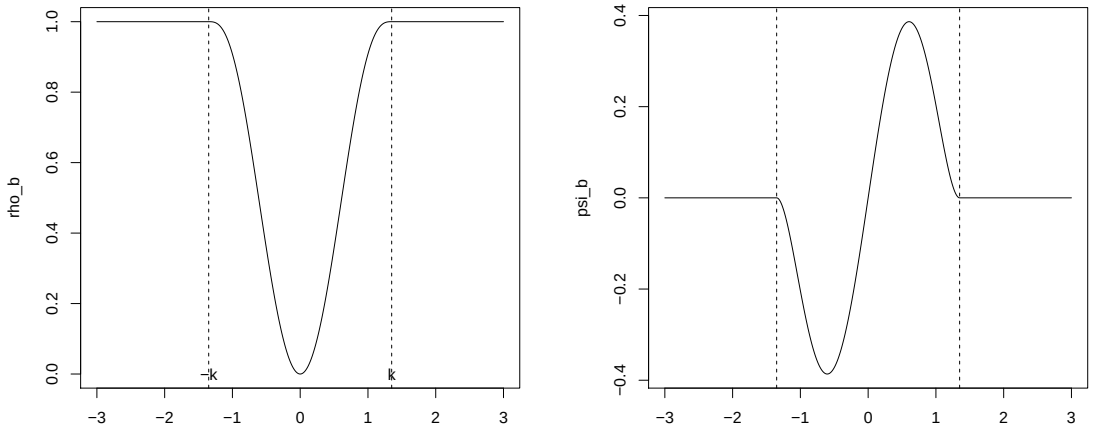
Piemērs 11. *Bi-square novērtējumu* definē, izvēloties

$$\rho(x) = \begin{cases} 1 - [1 - (x/k)^2]^3, & |x| < k \\ 1, & |x| \ge k, \end{cases} \tag{2.15}$$

atvasinājums $\rho'(x) = 6\psi(x)/k^2$, kur

$$\psi(x) = x\left[1 - \left(\frac{x}{k}\right)^2\right]^2 I_{\{|x|\leq k\}}. \tag{2.16}$$

Bi-square novērtējuma ψ-funkcija nav monotona (skat. 4. attēlu).



(a) (b)

4. att.: Bi-square novērtējuma ρ un ψ funkcijas, $k = 1.35$.

2.4. Hūbera minimax problēma

Tipiski, klasiskās statistikas procedūra ir optimālā procedūra pie kāda ideālā (visbiežāk, normālā) modeļa. Ja šī procedūra ir robusta un mēs vēlamies nodrošināties pret novirzēm no modeļa, tad par šo robustību ir jāmaksā ar novērtējuma efektivitāti. Līdz ar to jautājums ir, pret cik lielām novirzēm nepieciešams nodrošināties un cik daudz efektivitātes ir pieļaujams zaudēt. Viena no pieejām ir fiksēt noteiktu modeļa apkārtni un nodrošināties, lai novērtējums būtu robusts šajā apkārtnē. Tāda bija Hūbera [3] pieeja, kas lokācijas modeļa gadījumā noved pie minimax problēmas asimptotiskajai dispersijai vai biasam. Par kritēriju izvēloties asimptotisko dispersiju, iegūst vienkāršu, nerandomizētu minimax atrisinājumu. Vismazāk labvēlīgā situācija - sadalījums F_0 jeb dabas minimax stratēģija - minimizē Fišera informāciju dotajā apkārtnē.

Pieņemsim, ka īstā viendimensionālu kļūdu sadalījuma funkcija F atrodas kādā modeļa sadalījuma F_0 apkārtnē $\mathcal{P}_\epsilon$, novērojumi ir i.i.d. sadalīti ar sadalījuma funkciju $F(x-\theta)$ un ka nepieciešams novērtēt lokācijas parametru θ. Mērķis ir optimizēt šāda novērtējuma

robustības īpašības, minimizējot tā maksimālo dispersiju starp sadalījumiem $F \in \mathcal{P}_\epsilon$. Pieņemsim, ka mērķis ir nevis minimizēt novērtējuma atlikumu summu, bet gan sekojošu izteiksmi

$$\sum_{1=1}^{n} \rho(X_i - \theta), \tag{2.17}$$

kur ρ ir simetriska, augoša funkcija, kas aug lēnāk nekā kvadrātiski. Tad θ vērtība T_n, pie kuras izteiksme (2.17) sasniedz minimumu, ir vienādojuma

$$\sum_{i=1}^{n} \psi(X_i - T_n) = 0 \tag{2.18}$$

atrisinājums, kur $\psi = \rho'$. Pieņemsim, ka X_i ir i.i.d. sadalīti ar sadalījuma funkciju $F \in P_\epsilon$, kur

$$\mathcal{P}_\epsilon = \{F | F = (1-\epsilon)\Phi + \epsilon H, H \in \mathcal{M}, H - \text{simetrisks sadalījums.}\} \tag{2.19}$$

Definīcija 12. Sadalījumu klasi (2.19) sauc par normālo sadalījumu ar piesārņojumu, un ϵ sauc par piesārņojuma līmeni.

Šāds modelis rodas, piemēram, situācijā, kad tiek pieņemts, ka novērojumi sadalīti normāli ar vidējo vērtību 0 un dispersiju 1, bet ϵ-daļa no visiem datiem ir mērījumu kļūdas. Jāpiezīmē, ka ar piesārņojuma idejas palīdzību pirmoreiz statistikas teorijā tika aprakstīta kāda parametriska modeļa pilna apkārtne.

Izvirzot izteiksmi (2.18) Teilora rindā, iegūst, ka asimptotiski T_n ir spēkā

$$\sum \psi(X_i) - T_n \sum \psi'(X_i) = 0, \tag{2.20}$$

un, izmantojot centrālo robežteorēmu, var secināt, ka $\sqrt{n}T_n$ ir asimptotiski normālo sadalīts ar dispersiju

$$A(F,T) = \frac{E_F(\psi^2)}{(E_F\psi')^2}. \tag{2.21}$$

Šī ir pierādījuma skice, kas heiristiski pamato, kādēļ nepieciešams modeļa apkārtni definēt formā (2.19). Formālu pierādījumu var skatīt [14] 3.2.2. nodaļā.

Ievēro, ka lai $A(F,T)$ paliktu ierobežots kopā $F \in \mathcal{P}_\epsilon$, nepieciešams, lai ψ būtu ierobežots. Vienkāršākā ψ funkcija ir iegūstama, izvēloties par ρ izliektu funkciju

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le k \\ k|x| - \frac{1}{2}k^2, & |x| > k. \end{cases} \tag{2.22}$$

Tad ψ

$$\psi(x) = \begin{cases} k, & x \geq k \\ x, & -k \leq x \leq k \\ -k, & x \leq -k. \end{cases} \tag{2.23}$$

Ievēro, ka tikko izvēlētās ρ un ψ-funkcijas atbilst Hubera novērtējumu definējošajām funkcijām (2.13) un (2.14).

Tālāk ievēro, ka

$$A(F,T) \leq \frac{(1-\epsilon)E_\Phi\psi^2 + \epsilon k^2}{((1-\epsilon)E_\Phi\psi')^2}, \tag{2.24}$$

kur augšējā robeža tiek sasniegta pie tādiem sadalījumiem H, kuriem visa masa atrodas ārpus intervāla $[-k;k]$. Novērtējums, kuru definē sakarība (2.17), ir maksimālās ticamības novērtējums sadalījumam formā $f_0(x) = Ce^{-\rho(x)}$. Tātad, ja sakarībā (2.22) izvēlas k tādu, lai $C = (1-\epsilon)/\sqrt{2\pi}$, iegūst, ka k un ϵ saista sakarība

$$\frac{2\phi(k)}{k} - 2\Phi(-k) = \frac{\epsilon}{(1-\epsilon)}, \tag{2.25}$$

un iegūst, ka

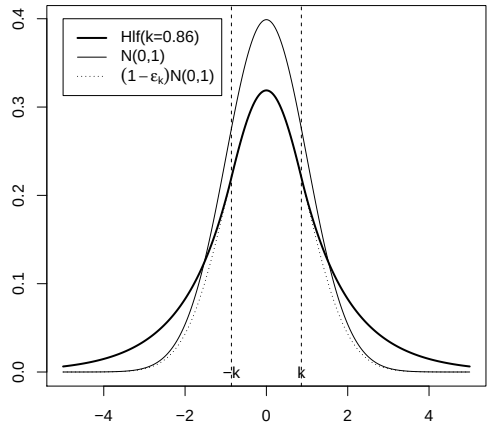
$$f_0 = \frac{1-\epsilon}{\sqrt{2\pi}} e^{-\rho(x)}. \tag{2.26}$$

f_0 atbilstošais sadalījums F_0 pieder apkārtnei $\mathcal{P}_\epsilon$, un seko, ka

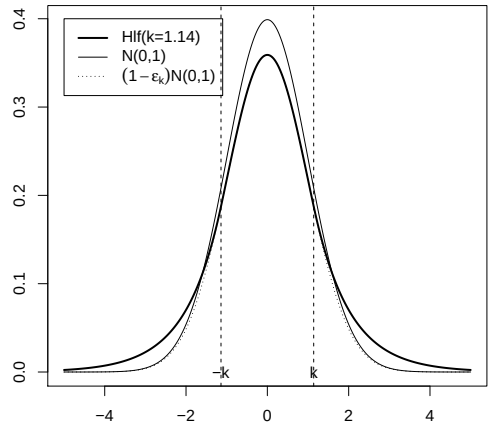
$$\sup_{F \in \mathcal{P}_\epsilon} A(F,T) = A(F_0,T). \tag{2.27}$$

Tātad ir iegūts, ka Hūbera novērtējums ir maksimālās ticamības novērtējums blīvuma funkcijai f_0 (2.26) atbilstošajam sadalījumam F_0, kuru sauc par *Hūbera vismazāk labvēlīgo sadalījumu* un tas minimizē F_0 asimptotisko dispersiju. Turklāt, Hūbera novērtējums minimizē asimptotisko dispersiju piesārņotu normālo sadalījumu klasē $\mathcal{P}_\epsilon$ (2.19). Hūbera vismazāk labvēlīgais sadalījums F_0 redzams 2.4. attēlā. Šim sadalījumam centrālajā daļā ir saspiests normālais sadalījums, un tam ir smagas astes, kas uzskatāmi parāda to, ka sadalījums konstruēts, lai raksturotu piesārņojuma ietekmi.

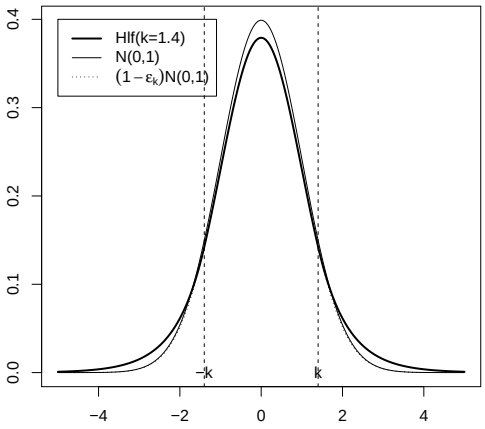
Hūbera novērtējums zināmā mērā ir vidusceļš starp diviem visbiežāk pielietotajiem lokācijas parametru novērtējumiem - vidējo vērtību un mediānu, un tas rod kompromisu starp pirmā efektivitāti un otrā robustumu. Aplūkojot konstantes k robežgadījumus, $k \to \infty$ un $k \to 0$, iegūst attiecīgi vidējo vērtību un mediānu, kas arī ir M-novērtējumi.



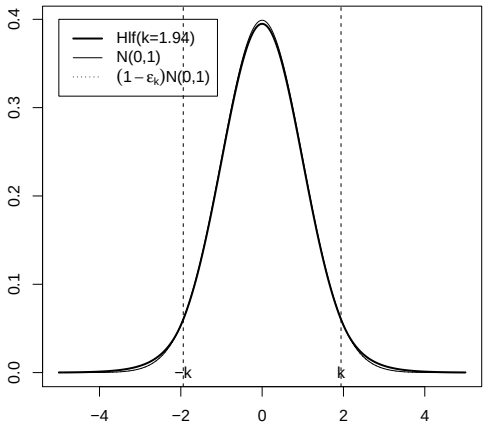
(a)



(b)



(c)



(d)

5. att.: Hūbera vismazāklabvelīgais sadalījums dažādiem ϵ piesārņojuma līmeņiem. Salīdzinājumam attēlota standartnormālā sadalījuma blīvuma funkcija un ar koeficientu $(1-\epsilon)$ saspiests standartnormālā sadalījuma blīvuma funkcijas grafiks, kas intervālā $|x| < k$ sakrīt ar Hūbera vismazāklabvēlīgā sadalījuma blīvuma funkcijas grafiku. (a)$\epsilon = 20\%$, $k = 0.86$; (b) $\epsilon = 10\%$, $k = 1.14$; (c)$\epsilon = 5\%$, $k = 1.4$; (d) $\epsilon = 1\%$, $k = 1.94$.

Līdz ar to k var interpretēt kā saskaņošanas konstanti, kas nosaka novērtējuma robustuma pakāpi.

Aprēķinot asimptotiskās dispersijas augšējo robežu dažādiem sadalījumiem no saimes $\mathcal{P}_\epsilon$ [3, 1. tabula], mainot ψ-funkcijas k vērtību un piesārņojuma līmeni ϵ, var secināt, ka k izvēlei nav kritiskas nozīmes. Hūbera secinājums bija, ka jebkura vērtība $1 \leq k \leq 2$ dod apmierinošus rezultātus piesārņojuma līmeņiem $\epsilon \leq 0.2$. Motivāciju šim secinājumam var rast 1. tabulā, kas ir fragments no Hūbera oriģinālās tabulas: redzams, ka jebkuram $0 \leq \epsilon \leq 0.2$ (kas, turklāt, atbilst vairumam saprātīgu praktisku situāciju) asimptotiskās dispersijas robeža, izvēloties $1 \leq k \leq 2$, nemainās tik strauji, lai novērtējums kļūtu neefektīvs.

No šīs tabulas iespējams arī nolasīt katram piesārņojuma līmenim optimālo k. Piemēram, $\epsilon = 0.05$ vismazāko dispersiju $V = 1.047$ sasniedz pie $k = 1.4$, pie $\epsilon = 0.2$ mazāko dispersiju $V = 2.046$ sasniedz pie aptuveni $k = 0.9$ (precīzās vērtības var aprēķināt no sakarības (2.25)). Šīs sakarības var izmantot situācijā, kad piesārņojuma līmenis ir zināms vai nojaušams.

Klasiska pieeja k izvēlei situācijā, kad piesārņojuma līmenis nav zināms, ir izvēlēties fiksētu efektivitātes zudumu (jeb "apdrošināšanas prēmijas lielumu") pie normālā modeļa. 2.4. tabulā redzama attiecība starp robstību un efektivitāti dažādiem k gadījumā, kad sadalījums ir jaukts normālais sadalījums $F = (1 - \epsilon)N(0, 1) + \epsilon N(0, 10)$. Ja izvēlēsimies nodrošināties pret novirzēm no modeļa, lietojot Hūbera novērtējumu ar $k = 1.4$, bet dati izrādīsies precīzi standartnormāli sadalīti, tad iegūtais novērtējums būs par 4.7% mazāk efektīvs, nekā vidējā vērtība. Šo efektivitātes zuduma apmēru katram k var nolasīt tabulas kolonnā, kur $\epsilon = 0$. Literatūrā minēts klasisks ieteikums (piemēram, [17]) ir izvēlēties $k = 1.35$, kas atbilst 5% efektivitātes zudumam pie normālā modeļa.

1. tabula: $\sqrt{n}T_n$ asimptotiskās dispersijas augšējās robežas (2.24) pie modeļa $F = (1 - \epsilon)\Phi + \epsilon H$, kur H - simetrisks sadalījums, T_n - Hūbera novērtējums no (2.18) ar ψ-funkciju no (2.14).

k	$E_\Phi\psi'$	$E_\Phi\psi^2$	$\epsilon = 0$	0.01	0.02	0.05	0.1	0.2	0.5
0.0	0.0000	0.0000	1.571	1.603	1.636	1.741	1.939	2.454	6.283
0.1	0.0797	0.0095	1.492	1.523	1.556	1.658	1.853	2.358	6.137
0.2	0.1585	0.0358	1.423	1.454	1.485	1.586	1.778	2.276	6.030
0.3	0.2358	0.0758	1.362	1.393	1.424	1.524	1.714	2.209	5.961
0.4	0.3108	0.1265	1.309	1.339	1.370	1.470	1.659	2.154	5.930
0.5	0.3829	0.1851	1.263	1.293	1.324	1.423	1.613	2.111	5.935
0.6	0.4515	0.2491	1.222	1.252	1.284	1.384	1.576	2.079	5.976
0.7	0.5161	0.3160	1.187	1.217	1.249	1.351	1.546	2.058	6.053
0.8	0.5763	0.3840	1.156	1.187	1.220	1.324	1.522	2.047	6.166
0.9	0.6319	0.4511	1.130	1.162	1.195	1.302	1.506	2.046	6.317
1.0	0.6827	0.5161	1.107	1.140	1.175	1.284	1.495	2.055	6.506
1.1	0.7287	0.5777	1.088	1.122	1.158	1.272	1.490	2.072	6.734
1.2	0.7699	0.6352	1.072	1.107	1.144	1.263	1.491	2.099	7.003
1.3	0.8064	0.6880	1.058	1.095	1.134	1.258	1.496	2.135	7.314
1.4	0.8385	0.7358	1.047	1.086	1.126	1.256	1.507	2.179	7.669
1.5	0.8664	0.7785	1.037	1.078	1.121	1.258	1.522	2.233	8.069
1.6	0.8904	0.8160	1.029	1.073	1.118	1.262	1.542	2.296	8.517
1.7	0.9109	0.8487	1.023	1.069	1.116	1.270	1.567	2.367	9.012
1.8	0.9281	0.8767	1.018	1.066	1.117	1.280	1.595	2.448	9.558
1.9	0.9426	0.9006	1.014	1.065	1.119	1.292	1.628	2.537	10.154
2.0	0.9545	0.9205	1.010	1.065	1.122	1.307	1.665	2.635	10.802
2.5	0.9876	0.9776	1.002	1.078	1.156	1.410	1.905	3.255	14.821
3.0	0.9973	0.9950	1.000	1.103	1.209	1.554	2.229	4.078	20.098

2. tabula: Hūbera M-novērtējuma asimptotiskās dispersijas pie modeļa $F = (1-\epsilon)N(0,1) + \epsilon N(0,10)$. [17, 27. lpp.]

k	ϵ=0	ϵ=0.05	ϵ=0.10
0	1.571	1.722	1.897
0.7	1.187	1.332	1.501
1.0	1.107	1.263	1.443
1.4	1.047	1.227	1.439
1.7	1.023	1.233	1.479
2.0	1.010	1.259	1.550
∞	1	5.95	10.9

2.5. Robusti mēroga novērtējumi

Definīcija 13. Par *mēroga parametra novērtējumu* sauc jebkuru pozitīvu statistiku S_n, kas ir ekvivarianta pret mērogošanu:

$$S_n(aX_1, aX_2, \ldots, aX_n) = aS_n(X_1, X_2, \ldots, X_n), \text{ kur } a > 0. \tag{2.28}$$

Robustajā statistikā mēroga parametri parādās kā traucējošie parametri. Piemēram, lokācijas parametra M-novērtējumi nav invarianti attiecībā pret mērogu, taču šo problēmu var pārvarēt, uzdodot M-novērtējumu T_n formā

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{S_n}\right) = 0, \tag{2.29}$$

kur S_n ir kāds mēroga parametra novērtējums. Dabīgi būtu novērtēt S_n ar kādu robustu novērtējumu. Zināms, ka visparastākais mēroga novērtējums - izlases standartnovirze (SD) s, kuru definē

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2, \tag{2.30}$$

nav robusta. Tukey (1960) piedāvātā alternatīva bija vidējā absolūtā novirze (angl. – *mean absolute deviation*), kas arī ir jutīga pret izlēcējiem, bet mazāk neka standartnovirze:

$$MD = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{x}|. \tag{2.31}$$

Visbiežāk lietotais robustais mēroga novērtējums ir MAD (angl. – *median absolute deviation*)

$$MAD = median(|X_i - median(X_i)|). \tag{2.32}$$

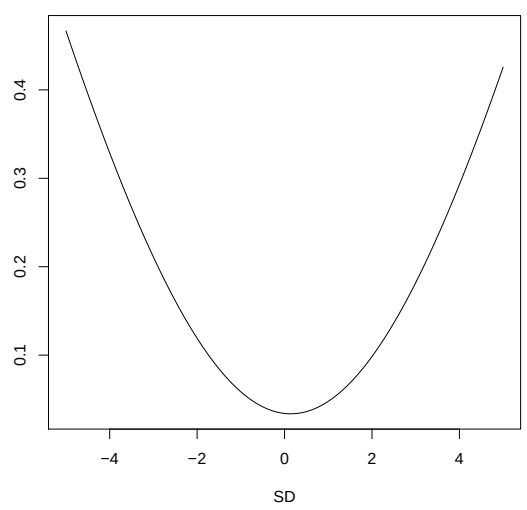
Atšķirībā no cita labi zināma mēroga novērtējuma - starpkvartiļu novērtējuma

$$IQR = X_{(n-m+1)} - X_{(m)},\ m = [n/4]\,, \tag{2.33}$$

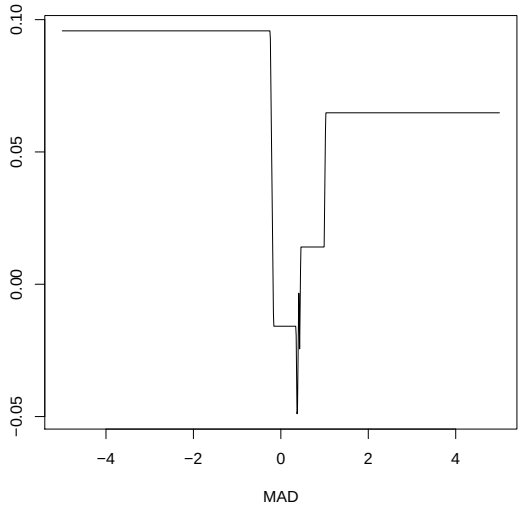
MAD piemīt augstāks lūzuma punkts $\epsilon^* = 0.5$ (IQR tas ir $\epsilon^* = 0.25$). Parasti mēroga novērtējumus nepieciešams standartizēt, lai pie modeļa tie būtu Fišera konsistenti. Piemēram, lai MAD būtu konsistents pie normālā modeļa, to nepieciešams izdalīt ar $\Phi^{-1}(3/4) = 0.6745$.

6. attēlā konstruētas jutīguma līknes izkliedes novērtējumiem. MAD un IQR jutīguma līkne ir ierobežota, tātad šie novērtējumi patiešām ir robusti. Savukārt standartnovirzei SD un vidējai absolūtajai novirzei MD jutīguma līkne nav ierobežota, kas parāda to robustuma trūkumu.

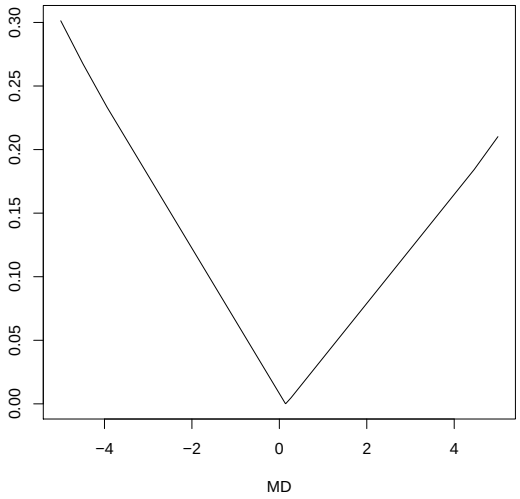
Lai atrisinātu (2.29), parasti ir pietiekoši novērtēt S_n vispirms un tad ievietot (2.29). Alternatīvi, var arī novērtēt T_n un S_n vienlaicīgi, taču tas ir skaitliski sarežģītāks uzdevums; par abu pieeju atbilstību konkrētām situācijām sīkāk aprakstīts, piemēram, Huber [3] "Proposal 2".



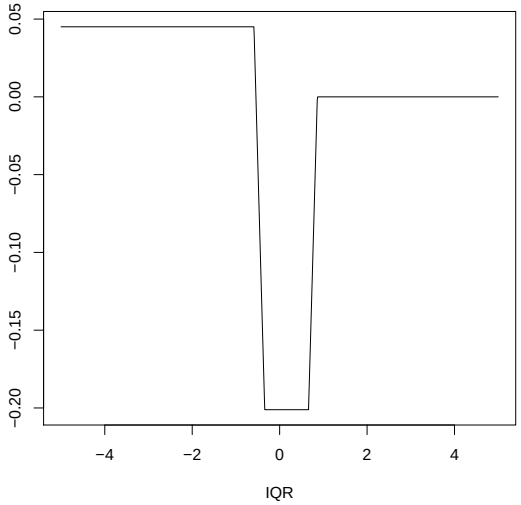
(a)



(b)



(c)



(d)

6. att.: Jutīguma līkne dažādiem izkliedes novērtējumiem, $N(0,1)$ izlasei: n=20. (a) standartnovirzei (SD), (b) mediānas absolūtajai novirzei (MAD), (c) vidējai absolūtajai novirzei (MD), (d) starpkvartiļu novērtējumam (IQR).

3. Gludais Hūbera novērtējums

Nesenā publikācijā [13] Hampel definēja gludināšanas principu M-novērtējumiem, kur gludināšanas pakāpe ir atkarīga no izlases lieluma. Jaunais novērtējums saglabā M-novērtējuma asimptotiskās īpašības, un, kā liecina Hampel simulāciju eksperimenta re-

zultāti, mazu izlašu gadījumā sniedz precīzākus rezultātus nekā parastais M-novērtējums, īpaši sadalījumu astēs.

Gludināšanas principu definē M-novērtējuma ψ funkcijai. Vispārīgai M-novērtējuma ψ funkcijai definē skores funkciju

$$\tilde{\psi}(x) = \int \psi(x+u) dQ_n(u), \tag{3.1}$$

kur Q_n var izvēlēties kā sākotnējā M-novērtējuma sadalījuma funkciju pie n i.i.d. novērojumiem no modelī pieņemtā sadalījuma. Izņemot vidējo vērtību un mediānu, M-novērtējumu sadalījuma funkcija galīgiem izlašu apjomiem nav izsakāma atklātā veidā. Tomēr, kā minēts (2.2.) nodaļā, M-novērtējumiem ir spēkā asimptotiskā normalitāte, līdz ar to Q_n var aizstāt ar $N(0, V/n)$, kur V ir M-novērtējuma asimptotiskā dispersija. Jāpiezīmē, ka gludināšanas principu var pielietot arī M-novērtējumiem, kuru ψ funkcijas jau ir gludas.

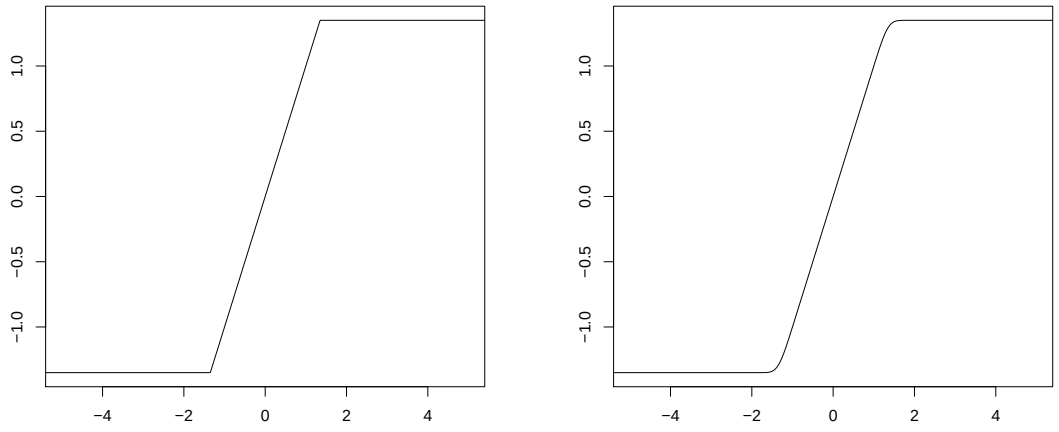
Ja M-novērtējums ir arī maksimālās ticamības novērtējums, tad par Q_n var izvēlēties atbilstošo sadalījumu, pie kura novērtējums ir asimptotiski optimāls. Hūbera novērtējuma gadījumā tas ir Hūbera vismazāk labvēlīgais sadalījums (2.26). Izvēloties par Q_n blīvuma funkciju f_0 no izteiksmes (2.26), ψ-funkcija, kas definē gludo M-novērtējumu, izsakāma atklātā formā

$$\begin{aligned}\tilde{\psi}_k(x) = {} & k\Phi\left(\frac{x-k}{\sigma_n}\right) - k\left(1 - \Phi\left(\frac{x+k}{\sigma_n}\right)\right) + x\left(\Phi\left(\frac{x+k}{\sigma_n}\right) - \Phi\left(\frac{x-k}{\sigma_n}\right)\right) \\ & + \sigma_n\left(\phi\left(\frac{x+k}{\sigma_n}\right) - \phi\left(\frac{x-k}{\sigma_n}\right)\right),\end{aligned} \tag{3.2}$$

kur $\sigma_n = \sqrt{V/n}$. Gludinātā funkcija ψ_k, salīdzinājumā ar Hūbera novērtējuma ψ funkciju, redzama 3. attēlā.

Hampel [13] simulāciju eksperimentā pētīja datus no dažādiem sadalījumiem: normālā sadalījuma, Hūbera vismazāk labvēlīgā sadalījuma ar parametru $k = 0.862$ (kas, saskaņā ar (2.25), atbilst $\epsilon = 0.2$), dubultā eksponenciālā sadalījuma un Košī sadalījuma. Visos gadījumos Hampel izvēlējās $V = 2.046$ no (3.2), kas ir atbilstošā minimax dispersija piesārņojuma līmenim $\epsilon = 0.2$.

Ja nepieciešams aprēķināt gludo Hūbera novērtējumu reāliem datiem, piesārņojuma līmenis ϵ nav zināms, līdz ar to nav iespējams izvēlēties k un aprēķināt asimptotisko dispersiju. Tādēļ reāliem datiem šī darba ietvaros tika nolemts V novērtēt ar izlases dispersiju, izmantojot kvantiļu butstrapa metodi.



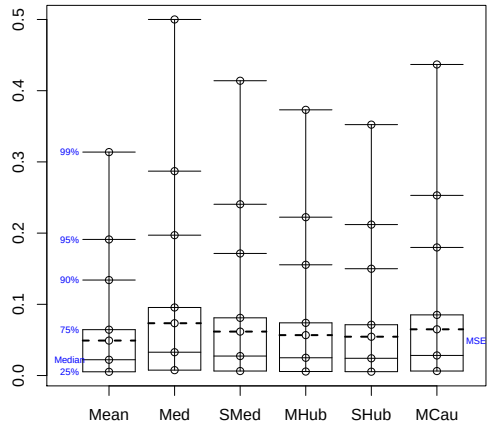
(a) (b)

7. att.: (a) Hūbera novērtējuma ψ funkcija, (b) gludinātā Hūbera novērtējuma $\tilde{\psi}_k$ funkcija. $k = 1.35$.

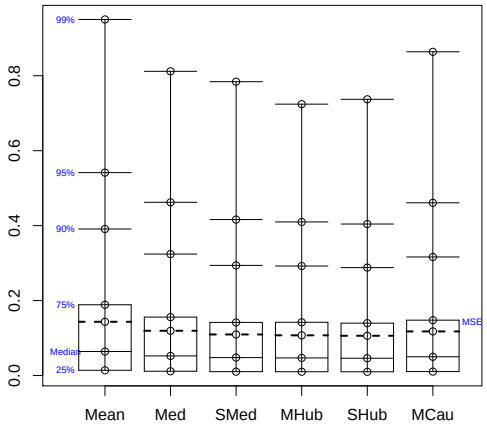
3.1. Negludo un gludo M-novērtējumu salīdzinājums - simulāciju piemēri

Lai salīdzinātu gludos un parastos M-novērtējumus, līdzīgi kā [13] tika veiktas simulācijas normālajam, Hūbera vismazāk labvēlīgajam (ar $k = .862$), dubultajam eksponenciālajam un Košī sadalījumiem. Tika novērtētas sekojošas statistikas: vidējā vērtība (attēlos apzīmēta kā Mean), mediāna (Med), gludinātā mediāna (SMed), Hūbera novērtējums (MHub), gludais Hūbera novērtējums (SHub), Košī novērtējums (Cau) un gludais Košī novērtējums (SCau). Visas statistikas tika aprēķinātas, pieņemot, ka skalas parametrs ir zināms ($s = 1$). Katrai statistikai tika analizēts kvadrātisko kļūdu sadalījums pie $N = 1000$ simulācijām. Tika aprēķināta kvadrātisko kļūdu sadalījuma vidējā vērtība (MSE), 99% kvantile un mediāna. Lai salīdzinātu novērtējumus, tika konstruēti vidējo kvadrātisko kļūdu (MSE) grafiks (9. attēls), 99% kvantiles grafiks (10. attēls) un mediānas efektivitātes grafiks (11. attēls) maziem izlašu apjomiem, $n = 3, 4, 5, 8, 20$. Kā arī 25%, 50%, 75%, 90%, 95% un 99% kvantiļu grafiki katra sadalījuma veida vienam izlases apjomam – $n = 20$ vai $n = 8$ (8. attēls). Gludinātie M-novērtējumi tika aprēķināti, izmantojot R paplašinājumprogrammu *smoothmest*, kuru saistībā ar publikāciju [13] izstrādājis Hampel un līdzautori.

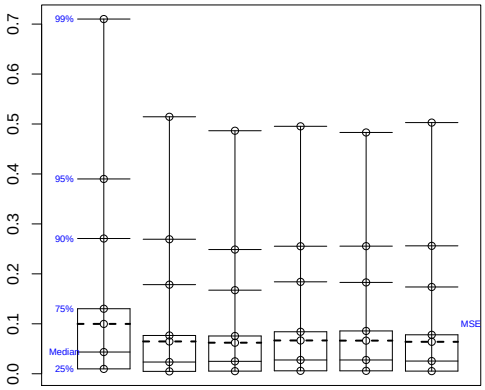
Kā jau sagaidāms, katram sadalījumam labāko rezultātu (mazāko dispersiju) sasniedz pie novērtējuma, kas ir tam atbilstošais maksimālās ticamības novērtējums, ka to var redzēt 8. attēlā. Normālajam sadalījumam vislabāko rezultātu sasniedz vidējā vērtība gan MSE, gan 99% kvantilei, gan mediānai. Hūbera vismazāk labvēlīgajam sadalījumam mediānas un 99% kvantiles gadījumā efektīvākais novērtējums ir gludais Hūbera novērtējums, kuram seko parastais Hūbera novērtējums (kas liecina, ka gludais novērtējums darbojas labāk), savukārt kvadrātisko kļūdu mediānas gadījumā līdzīgi labu rezultātu dod arī gludinātā mediāna. Dubultajam eksponenciālajam sadalījumam efektīvākā MSE un 99% kvantiles gadījumā ir gludā un parastā mediāna. Košī sadalījumam matemātiskā cerība neeksistē, un $\bar{x} \to \infty$, kad $n \to \infty$. Līdz ar to, vidējās vērtības novērtējums salīdzinājumā ar pārējiem bija tik slikts, ka grafikos nav attēlots. Visefektīvākais ir Košī novērtējums, turklāt gludais darbojas labāk par parasto.



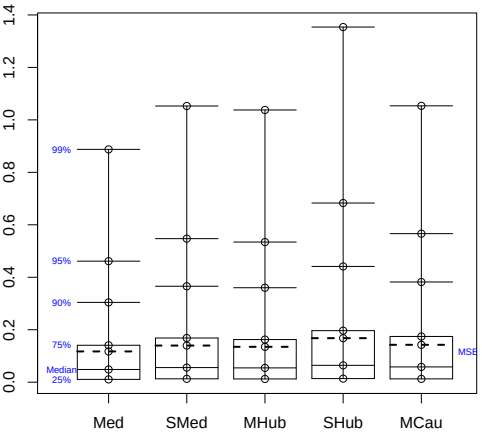
(a)



(b)

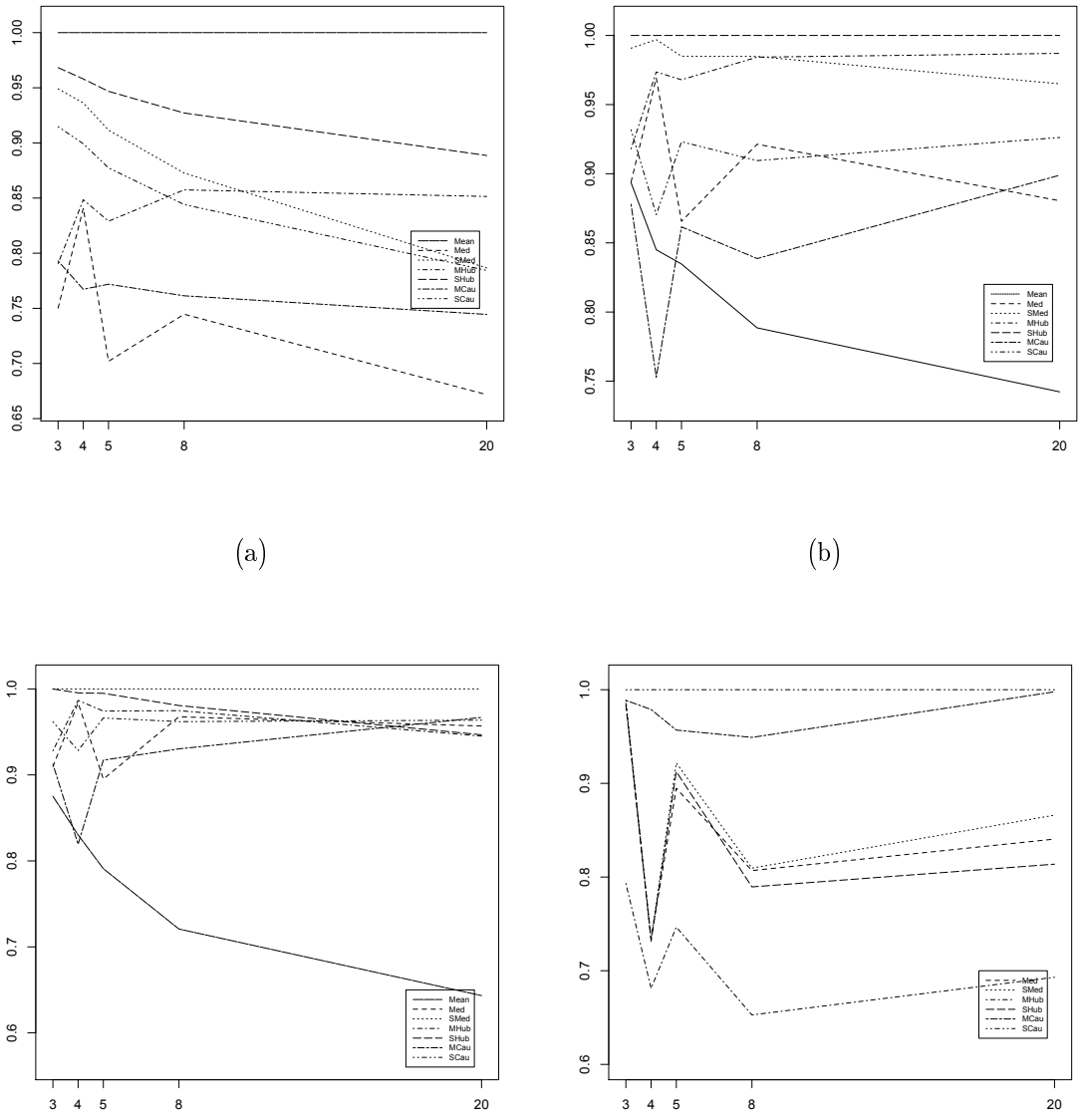


(c)

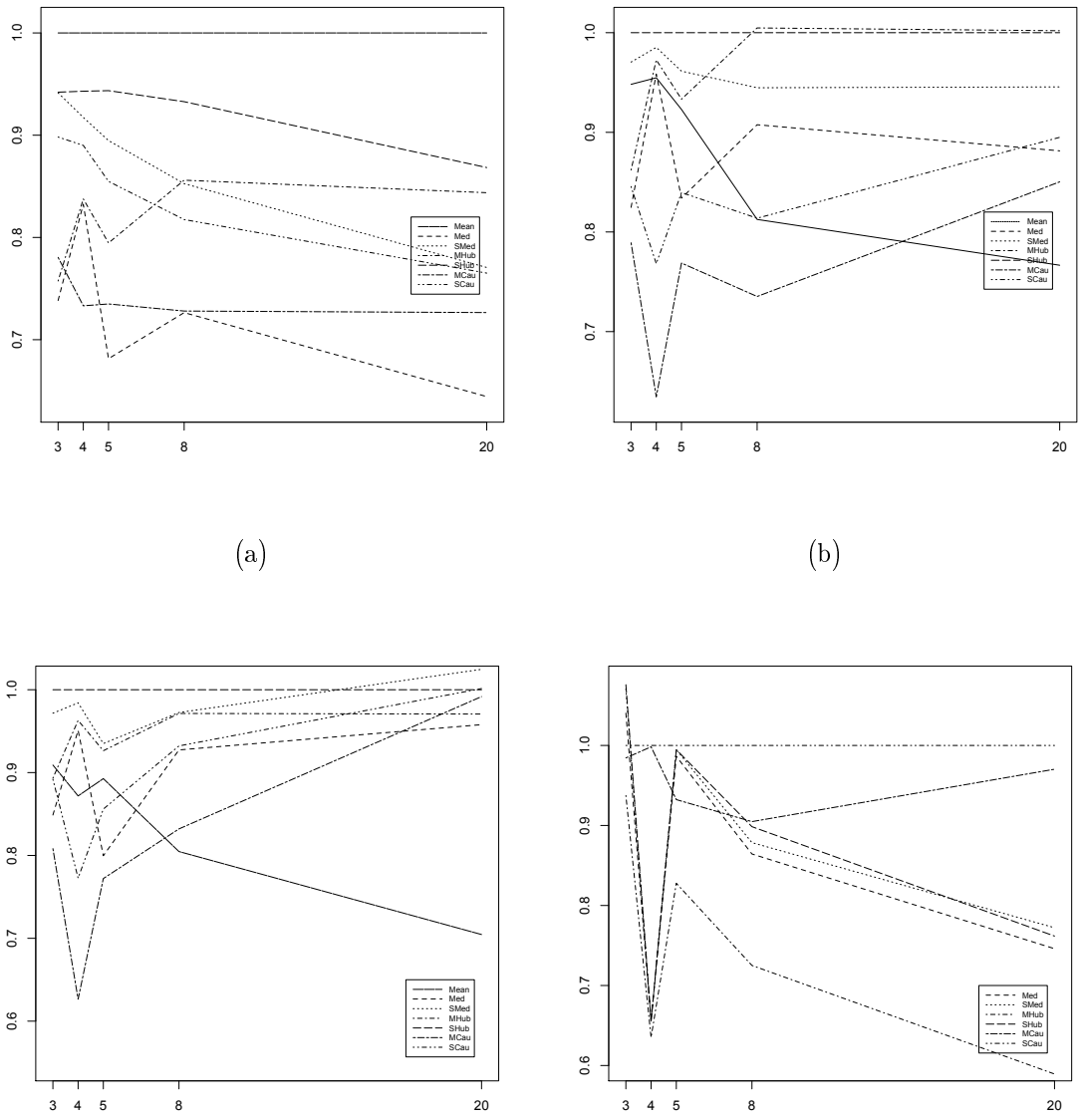


(d)

8. att.: Lokācijas parametra M-novertējumu kvadrātisko kļūdu sadalījuma kvantiles. (a) Standartnormālais sadalījums, $n = 20$, (b) Hūbera vismazāk labvēlīgais sadalījums ar $k = 0.862$, $n = 8$, (c) dubultais eksponenciālais sadalījums ar $\mu = 0$, $\lambda = 1$, $n = 20$, (d) Košī sadalījums ar $\mu = 0$, $\sigma = 1$, $n = 20$.



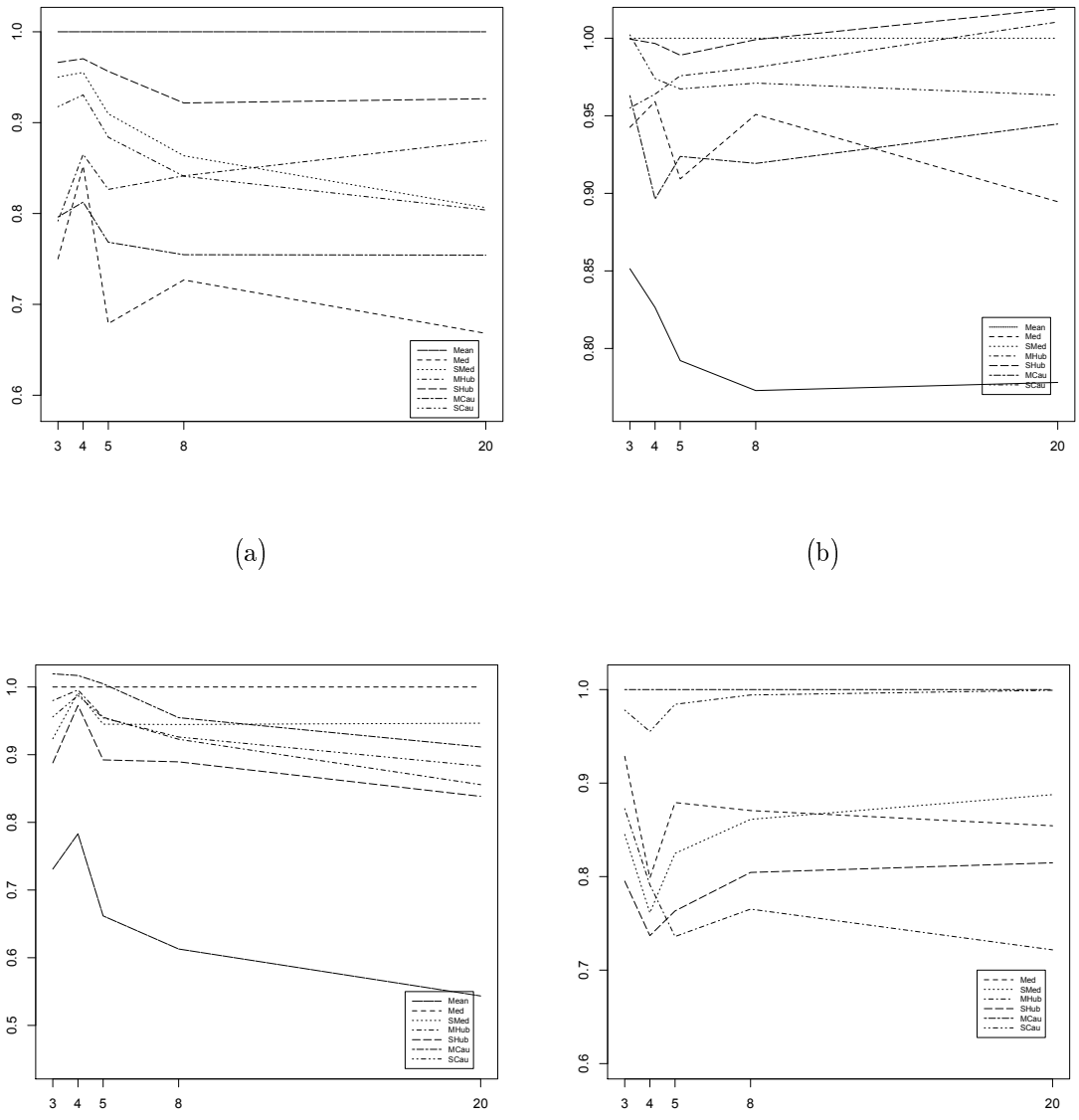
9. att.: Lokācijas parametra M-novertējumu vidējās kvadrātiskās kļūdas efektivitāte dažādiem izlašu apjomiem. (a) Standartnormālais sadalījums (b) Hūbera vismazāk labvēlīgais sadalījums ar $k = 0.862$, (c) dubultais eksponenciālais sadalījums ar $\mu = 0$, $\lambda = 1$, (d) Košī sadalījums ar $\mu = 0$, $\sigma = 1$.



(a) (b)

(c) (d)

10. att.: Lokācijas parametra M-novertējumu kvadrātisko kļūdu 99% kvantiles efektivitāte dažādiem izlašu apjomiem. (a) Standartnormālais sadalījums (b) Hūbera vismazāk labvēlīgais sadalījums ar $k = 0.862$, (c) dubultais eksponenciālais sadalījums ar $\mu = 0$, $\lambda = 1$, (d) Košī sadalījums ar $\mu = 0$, $\sigma = 1$.



(a) (b)

(c) (d)

11. att.: Lokācijas parametra M-novertējumu kvadrātisko kļūdu mediānas efektivitāte dažādiem izlašu apjomiem. (a) Standartnormālais sadalījums (b) Hūbera vismazāk labvēlīgais sadalījums ar $k = 0.862$, (c) dubultais eksponenciālais sadalījums ar $\mu = 0$, $\lambda = 1$, (d) Košī sadalījums ar $\mu = 0$, $\sigma = 1$.

4. Empīriskās ticamības metode

Empīriskās ticamības (angl. - *empirical likelihood, EL*) metode ir mūsdienu neparametriskās statistikas metode, kuru izstrādāja Owen [7], [8], [9]. Metodes būtība ir aproksimēt datus ar diskrētiem sadalījumiem, kam piemīt punktveida varbūtības datu punktos. Empīriskās ticamības metodi, tāpat kā tās parametrisko līdzinieci maksimālas ticamības funkcijas metodi, var izmantot parametru novērtēšanai, hipotēžu pārbaudei un novērtējumu ticamības intervālu konstruēšanai, tomēr tai ir vairākas priekšrocības. Piemēram, empīriskie ticamības intervāli automātiski atspoguļo novērotās datu kopas īpašības un tiem nav jābūt simetriskiem.

Oriģinālajā publikācijā Owen [7] parādīja, ka empīriskās ticamības intervālus iespējams konstruēt vidējai vērtībai, diferencējamiem statistiskajiem funkcionāļiem, tajā skaitā kvantilēm, un noteiktiem M-novērtējumiem. Qin un Lawless [10] formulēja empīrisko ticamības metodi vispārīgā formā ar nenovirzītiem vienādojumiem, kas ietver visus iepriekš minētos piemērus. Qin un Lawless aplūko $d-$dimensionālus i.i.d. gadījuma lielumus $X_1, \ldots, X_n$ ar nezināmu sadalījuma funkciju F un p-dimensionālu parametru θ. Šīs nodaļas ietvaros vienkāršības labad tiks pieņemts, ka $p = 1$ un $d = 1$.

Pieņemsim, ka informācija par F un θ dota nenovirzītas funkcijas $g(X, \theta)$ formā, kur $E\{g(X, \theta)\} = 0$. Piemēram, vidējās vērtības gadījumā $g(X_i, \theta) = X_i - \theta$. Aplūkojot kvantiles $\theta_q = F^{-1}(q)$, iegūst $g(X_i, \theta_q) = I_{\{x \leq \theta_q\}} - q$.

Definīcija 14. Par sadalījuma F *neparametrisko ticamības funkciju* $L(F)$ sauc funkciju

$$L(F) = \prod_{i=1}^{n} p_i, \tag{4.1}$$

kur $p_i = P(X = X_i)$ un $\sum_{i=1}^{n} p_i = 1$.

Funkcija $L(F)$ sasniedz maksimumu, kad $p_i = 1/n$, t.i., kad F ir empīriskā sadalījuma funkcija $F_n(x) = n^{-1} \sum_{i=1}^{n} I_{\{x_i < x\}}$. Līdz ar to sadalījuma funkcijai F var definēt neparametrisko jeb empīrisko ticamības attiecību

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i. \tag{4.2}$$

Funkciju $L(F)$ maksimizē pie nosacījumiem

$$p_i \geq 0, \ \sum_i p_i = 1, \ \sum_i p_i g(X_i, \theta) = 0. \tag{4.3}$$

$L(F)$ eksistē viens vienīgs maksimums, pie nosacījuma, ka 0 atrodas izliektas čaulas iekšpusē, kuru veido punkti $g(X_1,\theta)$, $g(X_2,\theta)$, ..., $g(X_n,\theta)$. Atrisinājumu atrod ar Lagranža reizinatāju palīdzību. Definē

$$H = \sum_i \log p_i + \lambda_0 \left(1 - \sum_i p_i\right) - n\lambda \sum_i p_i g(X_i,\theta), \tag{4.4}$$

kur λ un λ_0 ir Lagranža reizinātāji. Atvasinot H attiecībā pret p_i, iegūst

$$\begin{aligned} \frac{\partial H}{\partial p_i} &= \frac{1}{p_i} - \lambda_0 - n\lambda g(X_i,\theta) = 0, \\ \sum_i p_i \frac{\partial H}{\partial p_i} &= n - \lambda_0 = 0 \Rightarrow \lambda_0 = n \end{aligned} \tag{4.5}$$

un

$$p_i = \left(\frac{1}{n}\right) \frac{1}{1+\lambda g(X_i,\theta)}. \tag{4.6}$$

Izmantojot trešo ierobežojumu no (4.3), iegūst

$$0 = \sum_i p_i g(X_i,\theta) = \frac{1}{n} \sum_i \frac{g(X_i,\theta)}{1+\lambda g(X_i,\theta)}, \tag{4.7}$$

līdz ar to λ ir nosakāma kā funkcija no θ. No $0 \le p_i \le 1$ seko, ka $1+\lambda g(X_i,\theta) \ge 1/n$ visiem i. Fiksētam θ aplūko kopu $D_\theta = \{\lambda : 1+\lambda g(X_i,\theta) \ge 1/n\}$. D_θ ir izliekta un slēgta, un ierobežota, ja 0 pieder punktu $g(X_i,\theta)$ izliektās čaulas iekšienei. Turklāt, atvasinot (4.7) pēc λ, iegūst

$$\frac{\partial}{\partial\lambda}\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{g(X_i,\theta)}{1+\lambda g(X_i,\theta)}\right\} = -\frac{1}{n}\sum_{i=1}^{n} \frac{(g(X_i,\theta))^2}{(1+\lambda g(X_i,\theta))^2}, \tag{4.8}$$

kas vienmēr ir negatīvs. Saskaņā ar inversās funkcijas teorēmu, $\lambda = \lambda(\theta)$ nepārtraukti diferencējama funkcija pēc θ.

Empīriskās ticamības attiecības funkcija parametram θ ir formā

$$L(\theta) = \prod_{i=1}^{n} \left\{ \left(\frac{1}{n}\right) \frac{1}{1+\lambda(\theta) g(X_i,\theta)} \right\}, \tag{4.9}$$

un empīriskās ticamības attiecība

$$R(\theta) = \prod_{i=1}^{n} np_i = \prod_{i=1}^{n} \frac{1}{1+\lambda g(X_i,\theta)}. \tag{4.10}$$

Ērtāk lietojama ir logaritmiskā empīriskā ticamības attiecība,

$$l(\theta) = -\log R(\theta) = \sum_{i=1}^{n} \log\left[1+\lambda(\theta) g(X_i,\theta)\right]. \tag{4.11}$$

Minimizējot $l(\theta)$, iegūst parametra θ empīriskās ticamības novērtējumu $\arg\min_\theta l(\theta) = \tilde{\theta}$. Qin un Lawless parādīja, ka pie zināmiem nosacījumiem, logaritmiskajai empīriskās ticamības attiecībai ir spēkā neparametriskā Vilksa teorēma [10, 307. lpp.]

$$W(\theta_0) = -2\log R(\theta_0) \xrightarrow{d} \chi^2_1, \tag{4.12}$$

kad $n \to \infty$ un $H_0 : \theta = \theta_0$ ir spēkā. Šī teorēma ļauj sekojoši konstruēt empīriskās ticamības novērtējumu intervālus: nosaka c tādu, ka

$$P(\chi^2_1 \leq c) = 1 - \alpha, \tag{4.13}$$

kur $(1-\alpha)$ ir izvēlētā pārklājuma precizitāte. Tad empīriskais ticamības intervāls ir

$$\mathcal{R}_c = \{\theta : W(\theta) \leq c\}, \tag{4.14}$$

un, saskaņā ar Vilksa teorēmu (4.12), intervāla asimptotiskā pārklājuma precizitāte ir

$$P(\theta_0 \in \mathcal{R}_c) = P\{W(\theta_0) \leq c\} = 1 - \alpha. \tag{4.15}$$

Piemērs 12. Vidējā vērtība μ. Novērtējošā funkcija ir $g(X_i;\theta) = X_i - \mu$ un (4.7) ir formā

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0, \tag{4.16}$$

no kurienes nosaka λ vērtību. Logaritmiskā empīriskās ticamības attiecības statistika ir

$$W(\mu) = 2\sum_{i=1}^{n} \log\left\{1 + \lambda(X_i - \mu)\right\}. \tag{4.17}$$

4.1. Empīriskās ticamības metode M-novērtējumiem

Owen [7] parādīja, ka empīriskās ticamības metode ir pielietojama arī noteiktiem M-novērtējumiem. Aplūkosim M-novērtējumus, kas ir uzdoti funkcionālajā formā (2.3)

Teorēma 3. *[7, 243. lpp.] Pieņemsim, ka $\tau = T(F)$ ir M-novērtējumu definējošā vienādojuma*

$$\int \psi(X, \tau) dF(x) = 0 \tag{4.18}$$

atrisinājums, un pieņemsim, ka $X_1, X_2, \ldots, X_n$ ir i.i.d. sadalīti gadījuma lielumi, $X_i \sim F_0$. Aplūko viena argumenta funkcijas $\psi_{\cdot t}(x)$ un $\psi_{x\cdot}(t)$, kur

$$\psi_{\cdot t}(x) = \psi(x, t) = \psi_{x\cdot}(t). \tag{4.19}$$

Pieņemsim, ka funkcijai $\psi(x,t)$ spēkā: (i) $T(F_0)=\tau$ eksistē un ir viens vienīgs, (ii) $\psi_{\cdot\tau}(x)$ ir mērojama, (iii) $D\{\psi(X_1,\tau)\}>0$, (iv) $E\{|\psi_{\cdot\tau}(X)|^3\}<\infty$.

Pozitīvai konstantei c un empīriskās ticamības attiecībai R, kas definēta (4.2), aplūko $\mathcal{F}_{c,n}=\{F|R(F)\geq c, F\ll F_n\}$, un

$$S_{c,n}=\bigcup_{F\in\mathcal{F}_{c,n}}\left\{t\Big|\int\psi(x,t)dF(x)=0\right\}. \tag{4.20}$$

Tad

$$P(T(F_0)\in S_{c,n})\to P\left\{\chi_1^2\leq -2\log c\right\},$$

kad $n\to\infty$. Turklāt, ja (v) $\psi_{x,t}$ ir neaugoša pēc t, tad $S_{c,n}$ ir intervāls.

Piemērs 13. Hūbera novērtējums [7]. Hūbera novērtējuma ψ-funkcijai (2.14) izpildās 3. teorēmas nosacījumi, un $S_{c,n}$ ir intervāls. $g(X_i,\theta)=\psi\{(X_i-\mu)/\hat{\sigma}\}$, kur $\hat{\sigma}$ ir mēroga novērtējums, un empīriskās ticamības attiecība ir formā

$$R(\mu)=\sup_p\Big\{\prod_{i=1}^n np_i|p_i\geq 0,\sum_{i=1}^n p_i=1,\sum_{i=1}^n p_i\psi\left(\frac{X_i-\mu}{\hat{\sigma}}\right)=0\Big\}. \tag{4.21}$$

Apzīmē $Z_i=\psi(X_i,\mu)$. Tad (4.21) sasniedz maksimumu pie

$$p_i=\{n(1+\lambda Z_i)\}^{-1}, \tag{4.22}$$

un λ ir vienādojuma

$$\frac{1}{n}\sum_{i=1}^n\frac{Z_i}{1+\lambda Z_i}=0 \tag{4.23}$$

sakne, un tā atrodas intervālā $(-Z_{(n)}^{-1},-Z_{(1)}^{-1})$. Owen norāda, ka λ ierobežojošās vērtības var atrast, atrisinot vienādojumu $\{n(1+\lambda z)\}^{-1}=2$, kur z pieņem vērtības $Z_{(1)}$ un $Z_{(n)}$.

Viens no ticamības intervāla kvalitātes mēriem ir tā garums, tādēļ ir svarīgi pārbaudīt, vai intervāla garums ir robusts. Skaidrs, ka ticamības intervāls vidējai vērtībai, kas balstīts uz centrālo robežteorēmu, nav robusts, jo pat viens izlēcējs datos var likt dispersijai eksplodēt un var veidoties ļoti plats intervāls. Empīriskās ticamības intervāls vidējai vērtībai ir balstīts tikai uz novērotajiem datiem, un tā galapunkti ir datu punktu svērtais vidējais. Tā kā pēc metodes nosacījumiem visi svari ir pozitīvi, tad intuitīvi var spriest, ka intervāla garumu būtiski var ietekmēt izlēcēju klātbūtne datos.

Rodas jautājums, vai robusta M-novērtējuma empīriskās ticamības intervāls saglabā punktveida novērtējuma robustuma īpašību. Izrādās, ka asimptotiski tā patiešām ir, kā to pierādīja Tsao un Zhou [11]. Lai to pētītu, tāpat kā punktveida novērtējumam, arī ticamības intervāla garumam nepieciešams definēt lūzuma punktu.

Definīcija 15. [11] Pieņemsim, ka doti novērojumi $X_1, \ldots, X_n$, $|X_i| < \infty$, kuriem iegūts ticamības intervāls ar garumu L_n. Tad L_n galīgas izlases lūzuma punktu ϵ_n definē

$$\epsilon_n(L_n; X_1, \ldots, X_n) = \frac{1}{n} \min \left\{ m \Big| \max_{i_1,\ldots,i_m} \sup_{Y_1,\ldots,Y_m} \{L_n(Z_1, \ldots, Z_n)\} = \infty \right\}, \tag{4.24}$$

kur $Z_1, \ldots, Z_n$ ir izlase, kas iegūta, aizstājot skaitā m datu punktus $X_{i_1}, \ldots, X_{i_m}$ ar brīvi izvēlētām vērtībām $Y_1, \ldots, Y_m$. Šeit X_i un Y_i nav gadījuma lielumi.

Galīgu izlašu lūzuma punkts parāda, kāds izlasē ar apjomu n ir minimālais izlēcēju skaits m, pie kura ticamības intervāla garums tieksies uz bezgalību. Tsao un Zhou [11] pierādīja, ka empīriskās ticamības intervālam vidējai vērtībai $\epsilon_n = 1/n$, tātad pietiek ar vienu izlēcēju, lai intervāla garums brīvi palielinātos līdz bezgalībai.

Definēsim galīgu izlašu apakšējo lūzuma punktu ϵ_n^U intervāla garumam L_n kā

$$\epsilon_n^U = \frac{1}{n} \min \left\{ m \Big| \sup_{Y_{(1)},\ldots,Y_{(m)}} \left\{ L_n(X_{(1)}, \ldots, X_{(n-m)}, Y_{(1)}, \ldots, Y_{(n)}) \right\} = \infty \right\}. \tag{4.25}$$

Šajā gadījumā skaitā m lielākās izlases vērtības X_i tiek aizstātas ar lielām vērtībām Y_i, kam spēkā $X_{(n-m)} < Y_{(1)}, \ldots, < Y_{(m)}$.

Teorēma 4. *[11, 133. lpp.] Hūbera novērtējuma* $(1-\alpha)\%$ *empīriskās ticamības intervāla galīgu izlašu augšējais lūzuma punkts ir*

$$\epsilon_n^U = \min \{m |\ c(m) \geq c\} / n, \tag{4.26}$$

kur n *ir izlases apjoms,*

$$c(m) = \left(\frac{n}{2m}\right)^m \left(\frac{n}{2(n-m)}\right)^{(n-m)}, \tag{4.27}$$

kur $c = \exp(-\chi^2_{1,\alpha}/2)$.

Atzīmēsim, ka $c(m)$ ir Hūbera novērtējuma empīriskās ticamības attiecības $R(\mu)$ no (4.21) maksimālā vērtība situācijā, kad izlase satur skaitā m augšējos izlēcējus, turklāt tā tiek sasniegta pie svariem $p_i = 1/(2m)$, kad i atbilst izlēcēju datiem, un $p_i = 1/\{2(n-m)\}$ pārējos gadījumos.

Lai noteiktu ϵ_n^U lielumu fiksētiem n un c, uz brīdi pieņemsim, ka ϵ_n^U var pieņemt jebkuru vērtību no intervāla $(0, 1]$. Pieņemsim, ka $\epsilon = m/n$ ir augšējo izlēcēju īpatsvars izlasē ar lielumu n un definēsim

$$l(\epsilon) = \{c(m)\}^{1/n} = \frac{1}{2} \left(\frac{1}{\epsilon}\right)^{\epsilon} \left(\frac{1}{1-\epsilon}\right)^{1-\epsilon}. \tag{4.28}$$

Tad

$$\epsilon_n^U = \inf\{\epsilon | l(\epsilon) \geq c^{1/n}\}. \tag{4.29}$$

No (4.28) un (4.29) seko, ka asimptotiskais lūzuma punkts Hūbera novērtējuma empīriskās ticamības intervālam ir 0.5, tātad tāds pats kā Hūbera punktveida novērtējumam. 3. tabulā aprēķinātas ϵ_n^U vērtības 90% ticamības intervālam. Praktiskos uzdevumos jāņem vērā, ka $\epsilon_n^U \in \{1/n, \ldots, n/n\}$, piemēram, ja $n = 20$, tad, 90% ticamības intervālam īstā ϵ_n^U vērtība ir $[20 \times 0.318]/n$, kur $[nx]$ apzīmē skaitļa veselo daļu, tas ir, $\epsilon_n^U = 7/20$.

12. attēlā konstruēta logaritmiskā empīriskā ticamības attiecība $-2 \log R(\mu)$ vidējai vērtībai un Hūbera novērtējumam pie normālā modeļa $N(0, 3)$ un piesārņotā normālā modeļa $0.95N(0, 3) + 0.05N(20, 3)$. Redzams, ka nepiesārņotā modeļa gadījumā abas metodes dod ļoti līdzīgus ticamības intervālus, bet piesārņojuma gadījumā Hūbera novērtējuma intervāls ir daudz šaurāks. Šis piemērs ir ilustrācija tam, ka Hūbera novērtējum EL ticamības intervāli saglabā robustumu, bet vidējās vērtības EL ticamības intervāli – nē.

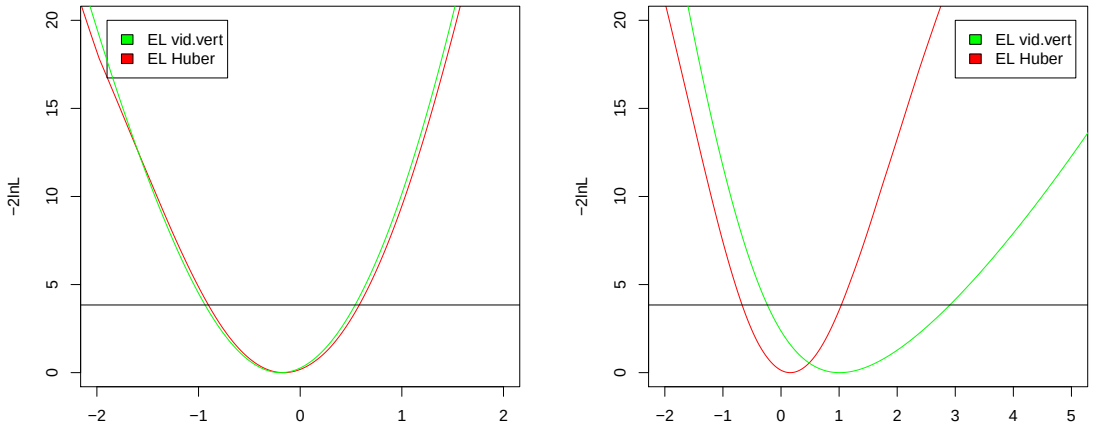
Zhang [19] pierādīja, ka Teorēmā 3 definētā M-novērtējuma ticamības intervāla pārklājuma kļūda, izmantojot χ^2 aproksimāciju, ir ar kārtu $O(n^{-1})$. Zhang arī pierādīja, ka, tāpat kā vidējās vērtības gadījumā, Hūbera novērtējuma gadījumā ir iespējams pielietot Bartleta korekciju un samazināt pārklājuma kļūdu līdz $O(n^{-2})$. Šī darba ietvaros Bartleta korekcijas aspekts netiks apskatīts.

3. tabula: Galīgu izlašu augšējais lūzuma punkts Hūbera novērtējuma 90% empīriskās ticamības intervālam. Tsao un Zhou, [11]

n	10	20	30	100	500	∞
ϵ_n^U	0.246	0.318	0.351	0.419	0.463	0.5

4.2. Empīriskās ticamības metode divu izlašu gadījumā

Empīriskās ticamības metodi divu izlašu vidējo vērtību un sadalījuma funkciju starpībām aprakstīja Qin un Zhao [12]. Šajā rakstā izvirzītie rezultāti tika pierādīti, balstoties uz Qin un Lawless [10] rezultātiem par empīrisko ticamības metodi vispārīgā formā. Tomēr, formulējumos iesaistītajām nenovirzītām funkcijām jāizpildās noteiktiem gluduma nosacījumiem, lai pierādījumos varētu lietot Teilora izvirzījumus.



(a) (b)

12. att.: Logaritmiskā ticamības attiecība $-2 * \ln R(\mu)$, kur μ ir vidējā vērtība (EL) un Hūbera novērtējums (EL Huber), pie modeļiem (a) $N(0,3)$ (b) $0.95N(0,3)+0.05N(20,3)$. Taisne attēlo χ^2_1 sadalījuma 95% kvantili.

Aplūkosim divu izlašu problēmu, kur $X_1, \ldots, X_{n_1}$ ir i.i.d sadalīti gadījuma lielumi ar nezināmu sadalījuma funkciju F_1, un $Y_1, \ldots, Y_{n_2}$ ir i.i.d. sadalīti ar nezināmu sadalījuma funkciju F_2, un ka funkcijas F_1 un F_2 ir atkarīgas attiecīgi no nezināmiem viendimensionāliem parametriem θ_0 un θ_1. Mēs interesējamies par parametru starpību $\Delta = \theta_1 - \theta_0$. Pieņemsim, ka informācija par patiesajiem parametriem Δ un θ_0 ir dota nenovirzītu funkciju formā

$$\begin{aligned} E_{F_1} w_1(X, \theta_0, \Delta) &= 0, \\ E_{F_2} w_2(Y, \theta_0, \Delta) &= 0. \end{aligned} \tag{4.30}$$

Lai iegūtu ticamības intervālus parametram Δ, definē empīriskās ticamības attiecību

$$R(\Delta, \theta) = \sup_{p,q} \prod_{i=1}^{n_1} (n_1 p_i) \prod_{j=1}^{n_2} (n_2 q_j), \tag{4.31}$$

kur $p = (p_1, \ldots, p_{n_1})$ un $q = (q_1, \ldots, q_{n_2})$ nosaka ierobežojumi

$$\begin{aligned} p_i \geq 0, i = 1, \ldots, n_1, \sum_{i=1}^{n_1} p_i = 1, \sum_{i=1}^{n_1} p_i w_1(X_i, \theta, \Delta) = 0, \\ q_j \geq 0, j = 1, \ldots, n_2, \sum_{j=1}^{n_2} q_j = 1, \sum_{j=1}^{n_2} q_j w_2(Y_j, \theta, \Delta) = 0. \end{aligned} \tag{4.32}$$

(4.31) eksistē viens vienīgs atrisinājums pie nosacījuma, ka 0 pieder izliektajai čaulai, kuru veido punkti $w_1(X_i, \theta, \Delta)$ un izliektajai čaulai, kuru veido $w_2(Y_j, \theta, \Delta)$. Maksimumu atrod, izmantojot Lagranža reizinātāju metodi, un iegūst, ka

$$p_i = \frac{1}{n_1(1 + \lambda_1(\theta)\omega_1(X_i, \theta, \Delta))}, \; i = 1, \ldots, n_1, \tag{4.33}$$

$$q_j = \frac{1}{n_2(1 + \lambda_2(\theta)\omega_2(Y_j, \theta, \Delta))}, \; j = 1, \ldots, n_2, \tag{4.34}$$

kur Lagranža reizinātājus $\lambda_1(\theta)$ un $\lambda_2(\theta)$ nosaka no vienādojumiem

$$\sum_{i=1}^{n_1} \frac{w_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta) w_1(X_i, \theta, \Delta)} = 0, \tag{4.35}$$

$$\sum_{j=1}^{n_2} \frac{w_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta) w_2(Y_j, \theta, \Delta)} = 0. \tag{4.36}$$

Definē empīrisko logaritmisko ticamības attiecību

$$W(\Delta, \theta) = -2 \log R(\Delta, \theta) = 2 \sum_{i=1}^{n_1} \log(1 + \lambda_1(\theta) w_1(X_i, \theta, \Delta)) + 2 \sum_{j=1}^{n_2} \log(1 + \lambda_2(\theta) w_2(Y_j, \theta, \Delta)). \tag{4.37}$$

Novērtējumu $\hat{\theta} = \hat{\theta}(\Delta)$, kas minimizē $R(\Delta, \theta)$ fiksētam parametram Δ, nosaka no vienādojuma

$$\frac{\partial W(\Delta, \theta)}{\partial \Delta} = \sum_{i=1}^{n_1} \frac{\lambda_1(\theta)\alpha_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta) w_1(X_i, \theta, \Delta)} + \sum_{j=1}^{n_2} \frac{\lambda_2(\theta)\alpha_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta) w_2(Y_j, \theta, \Delta)} = 0, \tag{4.38}$$

kur

$$\alpha_1(X_i, \theta, \Delta) = \frac{\partial w_1(X_i, \theta, \Delta)}{\partial \theta} \text{ un } \alpha_2(Y_j, \theta, \Delta) = \frac{\partial w_2(Y_j, \theta, \Delta)}{\partial \theta}. \tag{4.39}$$

Teorēma 5. *[12] Pieņemsim, ka*

(i) $\theta_0 \in \Omega$, *un* Ω *ir vaļējs intervāls.*

(ii) $E_{F_1} w_1^2(X, \theta, \Delta) > 0$ *un* $E_{F_2} w_2^2(Y, \theta, \Delta) > 0$, $\alpha_1(X, \theta, \Delta)$ *un* $\alpha_2(Y, \theta, \Delta)$ *ir nepārtrauktas* θ_0 *apkārtnē,* $\alpha_1(X, \theta, \Delta)$ *un* $w_1^3(X, \theta, \Delta)$ *šajā apkārtnē ir ierobežotas ar kādu integrējamu funkciju* $G_1(X)$, *un* $\alpha_2(Y, \theta, \Delta)$ *un* $w_2^3(Y, \theta, \Delta)$ *šajā apkārtnē ir ierobežotas ar kādu integrējamu funkciju* $G_2(Y)$. *Pieņemsim arī, ka* $E_{F_1}\alpha_1(X, \theta, \Delta)$ *un* $E_{F_2}\alpha_2(Y, \theta, \Delta)$ *nav vienādas ar nulli.*

(iii) $\frac{n_1}{n_2} \to k$ *(kad* $n_1, n_2 \to \infty$*) un* $0 < k < \infty$.

Tad vienādojumam (4.38) eksistē sakne $\hat{\theta}$*, kas ir konsistents* θ_0 *novērtējums,* $R(\Delta, \theta)$ *sasniedz savu lokālo maksimuma vērtību punktā* $\hat{\theta}$*, un*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \frac{\beta_1 \beta_2}{\beta_2 \beta_{10}^2 + k\beta_1 \beta_{20}^2}\right), \tag{4.40}$$

kur $k < \infty$ *ir tāda pozitīva konstante, ka izpildās* $n_2/n_1 \to k$*, kad* $n_1, n_2 \to \infty$*, un*

$$-2 \log R(\Delta_0, \hat{\theta}) \xrightarrow{d} \chi_1^2, \tag{4.41}$$

kad $n_1, n_2 \to \infty$*, un*

$$\begin{aligned} &\beta_1 = E_{F_1} w_1^2(X, \theta_0, \Delta_0),\ \beta_2 = E_{F_2} w_2^2(Y, \theta_0, \Delta_0), \\ &\beta_{10} = E_{F_1} \alpha_1(X, \theta_0, \Delta_0),\ \beta_{20} = E_{F_2} \alpha_2(Y, \theta_0, \Delta_0). \end{aligned} \tag{4.42}$$

Piemērs 14. (Qin un Zhao, [12]) Divu vidējo vērtību starpība. Apzīmē $\theta_0 = \int x dF_1(x)$ un $\Delta_0 = \int y dF_2(y) - \int x dF_1(x)$. (4.30) iegūst, izvēloties

$$w_1(X, \theta_0, \Delta_0) = X - \theta_0, \quad w_2(Y, \theta_0, \Delta_0) = Y - \theta_0 - \Delta_0. \tag{4.43}$$

Piemērs 15. Divu gludo Hūbera novērtējumu starpība. Ievērosim, ka gludais Hūbera novērtējums (3.1) atbilst Teorēmas 5 nosacījumiem. Pieņemsim, ka θ_0 un θ_1 ir attiecīgi izlašu X un Y gludie Hūbera novērtējumi. Tad, $\Delta_0 = \theta_1 - \theta_0$ un

$$w_1(X, \theta_0, \Delta_0) = \tilde{\psi}\left(\frac{X - \theta_0}{\hat{\sigma}_1}\right), \quad w_2(Y, \theta_0, \Delta_0) = \tilde{\psi}\left(\frac{Y - \Delta_0 - \theta_0}{\hat{\sigma}_2}\right), \tag{4.44}$$

kur $\tilde{\psi}$ atbilst gludā Hūbera novērtējuma ψ-funkcijai, kas definēta (3.2), un $\hat{\sigma}_1$ un $\hat{\sigma}_2$ ir attiecīgi izlašu X un Y mēroga novērtējumi.

5. Rezultāti

Analīzes mērķis bija pārbaudīt, kā strādā jaunā uz gludo Hūbera novērtējumu balstītā empīriskās ticamības metode divām izlasēm. Kā tika aprakstīts 4.1. nodaļā, empīriskās ticamības metode Hūbera novērtējumam vienas izlases gadījumā ir robusta metode un gadījumos, kad datos sastopami izlēcēji, darbojas labāk par citām metodēm. Tāpēc tika izvirzīta hipotēze, ka situācijās, kad nepieciešama robusti novērtējumi, arī divu izlašu gadījumā uz gludo Hūbera novērtējumu balstītā empīriskās ticamības metode ir pārāka par alternatīvām metodēm. Pirmkārt, tika apskatītās reālu datu piemēri, kas satur izlēcējus. Tika pārbaudīta nulles hipotēze, ka abu populāciju lokācijas parametri ir vienādi. Otrkārt, tika veikta simulāciju analīze, ar mērķi pētīt ticamības intervālu pārklājuma precizitāti, kā arī veikt jaudas analīzi. Gan reālo, gan simulēto datu gadījumā jaunā metode tika salīdzināta ar klasisko t-testu divām izlasēm un empīriskās ticamības metodi divu vidējo vērtību starpībai.

5.1. Izmantoto datu piemēru apraksts

Šajā apakšnodaļā tiks analizētas sešas divu izlašu datu kopas no publikācijām [20], [21] un [22]. **IQ dati,** Heritier et al. [20]. Datu kopa satur 94 piecus gadus vecu bērnu IQ rādītājus. Piecpadsmit bērnu mātes cieš no pēcdzemdību depresijas sindroma (1. izlase), 79 bērnu mātes ir veselas (2. izlase). Pārbauda H_0, ka IQ rādītāju sadalījumu lokācijas parametri starp grupām neatšķiras. Lielākā daļa no IQ vērtībām abās grupās ir starp 80 un 144, izņemot divas mazas vērtības vienam bērnam katrā grupā, attiecīgi 22 un 48.

Gaismas pārvietošanās ilguma dati, Stigler [21]. Apskatām datu kopas nr. 9., nr. 10 un nr. 11 ($n_1 = 20$, $n_2 = 20$, $n_3 = 26$), kas satur Newcombe gaismas pārvietošanās ilguma trešās mērījumu sērijas datus, kas tika iegūti 1882. gadā. Šajā gadījumā ir zināma īstā vērtība 33.02. Lielākā daļa mērījumu ir starp 20 un 35, izņemot divus ievērojamus izlēcējus, attiecīgi -2 un -44 datu kopā nr. 9. Visiem salīdzinātajiem izlašu pāriem pārbauda H_0, ka lokācijas parametrs starp sērijām neatšķiras.

LOS jeb uzturēšanās ilguma dati, Marazzi [22]. Pirmā izlase satur datus par 315 pacientu uzturēšanās ilgumiem slimnīcās Beļģijā, 1988. gadā, saistībā ar noteiktiem nervu sistēmas traucējumiem. Otrā datu kopa satur datus par 32 uzturēšanās ilgumiem slimnīcās Šveicē tajā pašā gadā un saistībā ar to pašu slimību. Šveicei atbilstošā izlase satur divas ekstrēmas vērtības, attiecīgi 374 un 198 dienas. Tika apskatīta arī atvasināta datu

kopa (apzīmēta kā LOS*), kura iegūta, izslēdzot no Šveices datu kopas divas ekstrēmās vērtības. Pārbauda H_0, ka uzturēšanās ilgumu sadalījumu lokācijas parametri Beļģijā un Šveicē neatšķiras.

Šim datu piemēram ir būtiska praktiska nozīme. LOS ir nozīmīgs indikators, kuru izmanto slimnīcu izmaksu novērtēšanai, un LOS vidējās vērtības medicīniski homogēnām pacientu grupām bieži tiek izmantotas par pamatu resursu sadalei slimnīcām. Taču LOS statistiskais sadalījums nav simetrisks, turklāt dati satur izlēcējus, kuru vērtības starp gadiem būtiski mainās, līdz ar to vidējā vērtība ir nepiemērots novērtējums, un tā jāaizstāj ar kādu robustu procedūru. Tāpat praktiska nozīme ir arī divu izlašu problēmai, jo var būt nepieciešams salīdzināt robustus LOS novērtējumus starp dažādām slimnīcām vai periodiem.

5.2. Datu analīze

Hipotēze, ka divi lokācijas parametri ir vienādi, katrai datu kopai tiek pārbaudīta, lietojot trīs metodes: (1) divu izlašu t-testu vidējo vērtību starpībai, (2) empīriskās ticamības testu divu izlašu vidējo vērtību starpībai un iepriekšējā nodaļā definēto (3) empīriskās ticamības testu divu gludu Hūbera novērtējumu starpībai. Testiem atbilstošās p-vērtības apkopotas 5. tabulā, un atbilstošie ticamības intervāli apkopoti 6. tabulā.

Tiek aplūkotas divas t-testa versijas, pieņemot attiecīgi vienādas dispersijas (panelis *"v.eq=T"*) un atšķirīgas dispersijas (panelis *"v.eq = F"*) starp izlasēm. Lietojot gludos Hūbera novērtējumus, vispirms ar butstrapa metodi tika novērtēta parastā Hūbera novērtējuma dispersija V un rezultāti apkopti 4. tabulā. Butstrapa dispersijas no 4. tabulas paneļa $\hat{\sigma}=MAD$ tika izmantotas, lai novērtētu gludos Hūbera novērtējumus un tiem atbilstošās testu p-vērtības 5. tabulā un ticamības intervālus 6. tabulā (panelis *EL Hūbera, V=novērtēts*). Šie rezultāti tika salīdzināti ar p-vērtībām un ticamības intervāliem, kas iegūti, izmantojot Hampel [13] ieteikto dispersiju $V = 2.046$ (panelis *EL Hūbera, V=2.046* 5. un 6. tabulā). Papildus tam, gludie Hūbera novērtējumi abos paneļos tika aprēķināti, izmantojot trīs dažādus mēroga novērtējumus: $\hat{\sigma} = 1$, izlases standartnovirzi sd un normalizēto mediānas absolūto novirzi MAD $= (1/0.6745)\, Med\{|X - Med(X)|\}$.

Kā tika aprakstīts 2.5. nodaļā par mēroga novērtēšanu, aprēķinot Hūbera novērtējumu, vēlams lietot robustu parametra σ no (2.14) novērtējumu. Iegūtie rezultāti apstiprina, ka tam patiešām ir nozīme. Datu kopām LOS un LOS*, lietojot MAD, tiek iegūtas būtiski

4. tabula: Hūbera novērtējuma (negludā) dispersijas V, novērtētas balstoties uz $10,000$ neparametriskā butstrapa izlasēm. Hūbera novērtējums aprēķināts ar $k = 1.35$, izmantojot dažādus mēroga novērtējumus ($\hat{\sigma} = 1$, standartnovirze sd un mediānas absolūtā novirze MAD)

	$\hat{\sigma} = 1$		$\hat{\sigma}$ =MAD		$\hat{\sigma}$ =sd	
Dati	izlase 1	izlase 2	izlase 1	izlase 2	izlase 1	izlase 2
IQ	167.42	318.89	189.19	357.86	173.97	465.76
Stigler9-10	59.06	37.86	58.04	32.23	160.48	33.08
Stigler10-11	37.86	13.14	32.23	17.51	33.08	17.04
Stigler9-11	59.06	13.14	58.04	17.51	160.48	17.04
LOS	15.46	9.85	42.41	14.73	74.25	2690.92
LOS*	16.67	6.95	42.41	8.28	74.06	106.48

5. tabula: p-vērtības hipotēžu pārbaudei H_0: lokācijas parametri ir vienādi. Gludais Hūbera novērtējums aprēķināts ar $k = 1.35$ un dažādiem mēroga parametriem ($\hat{\sigma} = 1$, SD un MAD)

	t-tests		EL	EL Hūbera, V novērtēts			EL Hūbera, V=2.046		
Dati	V.eq=F	V.eq=T	tests	$\hat{\sigma} = 1$	$\hat{\sigma}$ = MAD	$\hat{\sigma}$ =sd	$\hat{\sigma} = 1$	$\hat{\sigma}$ = MAD	$\hat{\sigma}$ =sd
IQ	0.122	0.016	0.052	0.052	0.052	0.052	0.066	0.068	0.055
Stigler9-10	0.112	0.106	0.019	0.024	0.019	0.019	0.201	0.073	0.031
Stigler10-11	0.632	0.626	0.620	0.619	0.620	0.620	0.640	0.606	0.625
Stigler9-11	0.147	0.097	0.033	0.041	0.034	0.033	0.280	0.119	0.050
LOS	0.192	0.000	0.023	0.558	0.449	0.024	0.932	0.792	0.023
LOS*	0.953	0.933	0.951	0.016	0.210	0.998	0.663	0.252	0.785

atšķirīgas p-vērtības, kas noved pie atšķirīgiem lēmumiem par H_0 patiesumu. Rezultāti liecina, ka nozīme ir arī V izvēlei. Novērtējot V, tiek iegūtas atšķirīgas p-vērtības nekā gadījumā, kad $V = 2.046$, turklāt, datu kopas Stigler9-11 gadījumā V novērtēšana pat noved pie pretēja lēmuma par H_0 patiesumu. Var arī ievērot, ka rezultāti ar novērtēto V ir līdzīgi tiem, ko dod EL metode vidējo vērtību starpībai. Izņemot piemēru LOS, kur metode ar novērtēto V ir vienīgā metode, kas nenoraida H_0, ka lokācijas parametri ir vienādi. Tas nozīmē, ka šai metodei vienīgajai ir spēja "atpazīt" piesārņojumu. Līdzīga sakarība vērojama arī ticamības intervālu gadījumā tabulā 6.: pirmajiem četriem datu

6. tabula: Ticamības intervāli hipotēžu pārbaudei H_0: lokācijas parametri ir vienādi. Gludais Hūbera novērtējums ar $k = 1.35$, mēroga novērtējums MAD

Dati	t-test v.eq=F	t-tests v.eq=T	EL tests	EL Hūbera, V novērtēts	EL Hūbera, V =2.046
IQ	(-3.49, 26.927)	(2.272, 21.165)	(-0.105, 29.356)	(-0.105, 29.356)	(-0.769, 24.371)
Stigler9-10	(-15.309, 1.709)	(-15.11, 1.51)	(-17.871, -0.912)	(-17.831, -0.905)	(-9.37, 0.345)
Stigler10-11	(-2.243, 3.651)	(-2.188, 3.596)	(-2.101, 3.546)	(-2.099, 3.545)	(-2.062, 3.542)
Stigler9-11	(-14.504, 2.312)	(-13.342, 1.15)	(-17.104, -0.387)	(-17.062, -0.381)	(-8.477, 0.778)
LOS	(-44.517, 9.325)	(-26.985, -8.206)	(-55.189, -1.313)	(-12.947, 2.561)	(-1.851, 1.563)
LOS*	(-7.976, 7.522)	(-5.523, 5.069)	(-11.97, 4.213)	(-1.943, 4.657)	(-0.756, 1.964)

pāriem EL vidējo vērtību starpībai un EL Hūbera novērtējumu starpībai ar novērtētu V dod ļoti līdzīgus ticamības intervālus, bet LOS datu kopām EL Hūbera novērtējumu starpībai dod ievērojami mazākus ticamības intervālus. Īsāks intervālu garums pierāda, ka metode ir robustāka par pārējām.

5.3. Simulāciju analīze

Tika veikta simulāciju analīze divu gludo Hūbera novērtējumu starpībai datiem no gamma sadalījuma ar un bez piesārņojuma, Košī sadalījuma, dubultā eksponenciālā sadalījuma un Hūbera vismazāk labvēlīgā sadalījuma (2.26). Tiek pārbaudīta pārklājuma precizitāte novērtējumam $\Delta = \theta_2 - \theta_1$ no 15. piemēra. Gamma sadalījumu simulācijas iepriekš analizēja Marazzi [22] kā asimetrisku sadalījumu piemēru un šajā darba sekots [22] pieejai. Ar gamma sadalījumu iespējams modelēt datus, kas raksturo izmaksu mainību, un divu izlašu problēmās gamma sadalījumiem bieži vien ir atšķirīgi skalas parametri. Tāpēc simulāciju sadalījumi konstruēti tā, lai abām izlasēm būtu vienādas vidējās vērtības, bet skalas parametri būtu atšķirīgi: $F_1 = \text{Gamma}(\text{a} = \sigma; \text{s} = 1)$; $F_2 = \text{Gamma}(\text{a} = 1; \text{s} = 1/\sigma)$, kur gamma sadalījuma vidējā vērtība ir a/s un tiek apskatītas dažādas σ vērtības. Lielas izlēcēju vērtības datos modelētas, pievienojot sadalījumiem vienmērīgā sadalījuma Unif$[0, 50]$ piesārņojumu. Tika veiktas simulācijas ar $N = 10,000$ atkārtojumiem diviem piesārņojuma līmeņiem $\epsilon = 0.2$ un $\epsilon = 0.06$ un diviem izlašu apjomiem, $n_1 = n_2 = 50$ un $n_1 = n_2 = 100$.

Ticamības intervālu pārklājuma precizitāte apkopota 8. – 13. tabulās un salīdzinātas četras metodes: divu izlašu t-tests, empīriskās ticamības metode vidējo vērtību starpībai

(panelis "EL") un empīriskās ticamības metode divu gludo Hūbera novērtējumu starpībai, ar fiksētu un novērtētu asimptotisko dispersiju V. Visi Hūbera novērtējumi aprēķināti, izmantojot MAD mēroga novērtējumu. V izvēlēts atbilstoši Hampel [13] norādījumiem kā Hūbera novērtējuma asimptotiskās minimax dispersija atbilstošajam piesārņojuma līmenim : pie $\epsilon = 0.06$ $V = 1.302$, un pie $\epsilon = 0.2$, $V = 2.046$. Metodē ar novērtētu V, V iegūts, aprēķinot simulētajiem sadalījumiem negludo Hūbera novērtējumu, un iegūtajam sadalījumam robusti novērtējot dispersiju ar MAD.

8. un 9. tabulā modeļiem bez piesārņojuma redzams, ka metode Hūbera novērtējumu starpībai ar novērtētu V nedaudz atpaliek no t-testa, bet darbojas ļoti līdzīgi EL metodei. Savukārt Hūbera novērtējums ar fiksētu V dod sliktāku pārklājuma precizitāti. 10. un 11. ar vienmērīgā sadalījuma piesārņojuma līmeni $\epsilon = 0.2$ EL un Hūbera novērtējums ar novērtētu V atkal dod līdzīgus rezultātus, savukārt rezultāti pie fiksēta V ir nekonsekventi: dažos gadījumos metodes pārklājuma precizitāte tā ir būtiski labāka, bet dažos gadījumos sliktāka. Arī 12. un 13. tabulās piesārņojuma līmenim $\epsilon = 0.06$ V fiksēšana dod nekonsekventus rezultātus, savukārt metode ar novērtētu V uzrāda būtiski labākus rezultātus nekā EL metode. Turklāt, pie lielākām skalas atšķirībām starp modeļiem F_1 un F_2 ($\sigma = 7$, $\sigma = 10$, $\sigma = 20$) Hūbera novērtējums ar novērtētu V pārsniedz arī t-testa rezultātus.

Kopumā rezultāti liek domāt, ka V fiksēšana dod nepareizus rezultātus, un svarīgi ir V novērtēt. Metode ar novērtētu V ļoti labi konkurē ar EL metodi vidējai vērtībai un t-testu, turklāt situācijās, kur sadalījumam ir piesārņojums un skalas parametri ir stipri atšķirīgi starp modeļiem, parādās šīs metodes priekšrocības.

Papildus pārklājuma precizitātes analīzei tikai veikta jaudas analīze. Testē H_0, ka lokācijas parametri starp izlasēm neatšķiras, dati simulēti no sadalījumiem F_1 = (1-ϵ)Gamma(a = 5; s =1)+ $\epsilon\, Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$), $\sigma = 5, 6, 7, 8, 9$. Pēc sadalījuma konstrukcijas, H_0 nav spēkā pie $\sigma \neq 5$. Pirmajā solī tika ģenerēts H_0 sadalījums testa statistikai $-2 \log R(\Delta, \theta)$, izmantojot EL metodi divu vidējo vērtību starpībai un divu Hūbera novērtējumu starpībai ar fiksētu un novērtētu V. Sadalījums simulēts $N = 10,000$ izlasēm ar apjomu $n_1 = n_2 = 50$ no F_1 = Gamma(a = 5; s =1); F_2 = Gamma(a = 1; s = $1/\sigma$). Statistiku sadalījums redzams 13. attēlā un tās kritiskās vērtības redzamas 7. tabulā.

Saskaņā ar 5. Teorēmu, $-2 \log R(\Delta, \theta)$ statistikai jābūt sadalītai pēc $\rightarrow_d \chi^2_1$ sadalī-

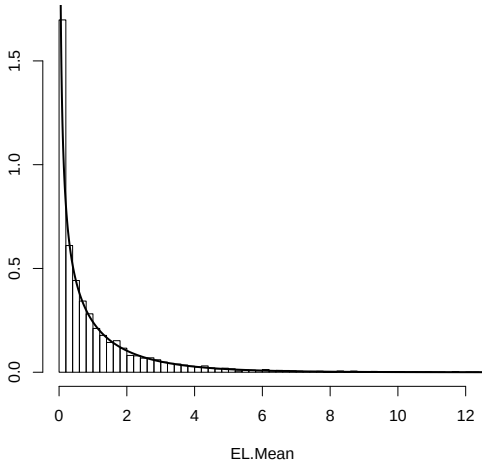
7. tabula: Kritiskās vērtības H_0 sadalījumam testa statistikai $-2\log R(\Delta,\theta)$, simulēts $N = 10,000$ izlasēm no sadalījumiem F_1 = Gamma(a = 5; s =1); F_2 = Gamma(a = 1; s = $1/\sigma$), un χ^2_1 sadalījuma kvantiles.

	$\alpha = 0.10$	$\alpha = 0.05$
EL	2.893	4.176
Huber, $V = 2.046$	4.058	5.621
Huber, V novērtēts	2.885	4.174
χ^2_1	2.706	3.842

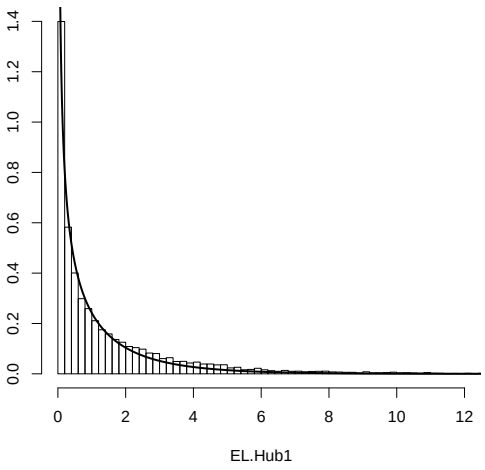
juma. 13. attēlā redzams, ka šis nosacījums izpildās EL vidējo vērtību starpībai un EL Hūbera novērtējumu starpībai ar novērtētu asimptotisko dispersiju V. Fiksēta V gadījumā tiek iegūts sadalījums, kam ir smagāka aste nekā χ^2_1, arī testa kritiskā vērtība ir lielāka nekā χ^2 atbilstošā kvantile. Iespējams, ka šī metode konverģē uz robežsadalijumu lēnāk nekā pārējās EL metodes, un šis novērojums sniedz vel vienu argumentu par labu V novērtēšanai.

Testa rezultāti apkopti 14. tabulā. Modeļos bez piesārņojuma EL metode vidējai vērtībai un Hūbera novērtējumam ar novērtētu V jauda ir ļoti līdzīga, t-testa jauda nedaudz atpaliek. Modeļos ar piesārņojumu Hūbera novērtējums ar novērtētu V ir visjaudīgākais un lielākiem σ būtiski pārsniedz EL vidējo vērtību starpībai sniegumu un ari t-testa sniegumu. Pie $\epsilon = .06$ un $\sigma = 5$ visas metodes noraida H_0 biežāk nekā pieņemtajā līmenī. Tas nozīmē, ka neviena no metodēm tomēr nespēj efektīvi atpazīt vienmērīgo sadalījumu kā piesārņojumu. Tomēr Hūbera metodei ir zemāks H_0 noraidīšanas līmenis nekā pārējām metodēm. Visbeidzot, arī jaudas analīzē V fiksēšana rada nekonsekvenci rezultātos.

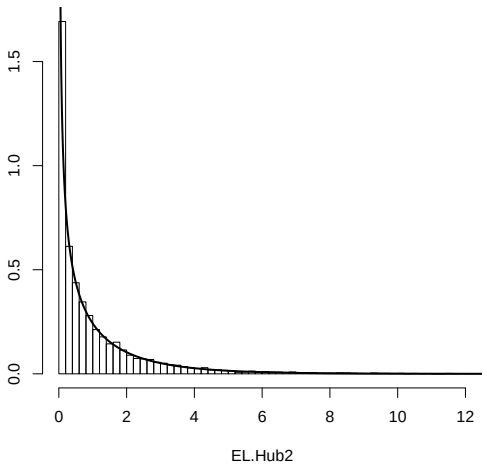
15.-17. tabulās apkopota pārklājuma precizitāte Košī, dubultajam eksponenciālajam un Hūbera vismazāk labvēlīgajam parametram. Šie sadalījumi tika izvēlēti, jo tie tika analizēti Hampel [13] publikācijā par gludiem M-novērtējumiem. Redzams, ka V novērtēšana iespaido rezultātus Košī sadalījuma gadījumā, bet Hūbera vismazāk labvēlīgajam sadalījumam V iespējams darbojas kā saskaņošanas parametrs un rezultāti V novērtējot vai V fiksējot ir ļoti līdzīgi.



(a)



(b)



(c)

13. att.: H_0 sadalījums testa statistikai $-2\log R(\Delta, \theta)$, simulēts $N = 10,000$ izlasēm no sadalījumiem $F_1 = \text{Gamma(a} = 5; \text{ s} = 1)$; $F_2 = \text{Gamma(a} = 1; \text{ s} = 1/\sigma)$. Salīdzinājumam attēlota χ^2_1 sadalījuma blīvuma funkcija. (a) EL metode vidējai vērtībai, (b) EL metode divu Hūbera novērtējumu starpībai, $V = 2.046$, (c) EL metode divu Hūbera novērtējumu starpībai, V novērtēts.

8. tabula: Pārklājuma precizitāte modelim F_1 = Gamma(a = σ; s = 1); F_2 = Gamma(a = 1; s = $1/\sigma$), $n_1 = n_2 = 50$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	3.3	7.3	0.942	0.938	0.894	0.935
4	4.1	12.7	0.945	0.938	0.894	0.938
5	5.8	20.2	0.944	0.941	0.887	0.941
6	6.8	28.9	0.941	0.935	0.883	0.936
7	8.2	39.2	0.941	0.938	0.881	0.938
10	11.8	81.1	0.939	0.937	0.877	0.937
20	24.6	318.3	0.936	0.935	0.866	0.935

9. tabula: Pārklājuma precizitāte modelim F_1 = Gamma(a = σ; s = 1); F_2 = Gamma(a = 1; s = $1/\sigma$), $n_1 = n_2 = 100$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	3.3	7.3	0.946	0.944	0.879	0.941
4	4.1	12.7	0.946	0.945	0.867	0.943
5	5.8	20.2	0.945	0.946	0.862	0.946
6	6.8	28.9	0.944	0.944	0.858	0.944
7	8.2	39.2	0.948	0.950	0.851	0.949
10	11.8	81.1	0.943	0.944	0.840	0.944
20	24.6	318.3	0.945	0.945	0.832	0.945

10. tabula: Pārklājuma precizitāte modelim F_1 = (1-ϵ)Gamma(a = σ; s = 1) + $\epsilon Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$) ar $\epsilon = 0.2$, $n_1 = n_2 = 50$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	8.7	8.0	0.146	0.056	0.353	0.108
4	10.2	13.5	0.211	0.119	0.405	0.162
5	13.4	21.3	0.285	0.199	0.459	0.232
6	16.0	31.2	0.373	0.307	0.511	0.328
7	16.8	42.0	0.448	0.404	0.556	0.420
10	22.0	82.6	0.653	0.655	0.659	0.660
20	36.8	344.3	0.911	0.924	0.834	0.924

11. tabula: Pārklājuma precizitāte modelim F_1 = (1-ϵ)Gamma(a = σ; s = 1) + $\epsilon Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$) ar $\epsilon = 0.2$, $n_1 = n_2 = 100$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	8.7	8.0	0.005	0.002	0.109	0.007
4	10.2	13.5	0.016	0.010	0.156	0.017
5	13.4	21.3	0.037	0.026	0.206	0.037
6	16.0	31.2	0.076	0.065	0.253	0.078
7	16.8	42.0	0.136	0.125	0.299	0.140
10	22.0	82.6	0.401	0.417	0.450	0.424
20	36.8	344.3	0.908	0.924	0.760	0.925

12. tabula: Pārklājuma precizitāte modelim F_1 = (1-ϵ)Gamma(a = σ; s = 1) + $\epsilon Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$) ar $\epsilon = 0.06$, $n_1 = n_2 = 50$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	4.2	8.6	0.839	0.646	0.781	0.842
4	5.7	14.9	0.845	0.723	0.779	0.841
5	7.0	23.4	0.852	0.780	0.776	0.845
6	8.3	34.5	0.859	0.818	0.773	0.856
7	9.7	46.6	0.871	0.853	0.782	0.870
10	13.0	93.5	0.893	0.901	0.785	0.904
20	24.9	375.1	0.927	0.930	0.803	0.930

13. tabula: Pārklājuma precizitāte modelim F_1 = (1-ϵ)Gamma(a = σ; s = 1) + $\epsilon Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$) ar $\epsilon = 0.06$, $n_1 = n_2 = 100$, $N = 10,000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	4.2	8.6	0.839	0.646	0.781	0.842
4	5.7	14.9	0.845	0.723	0.779	0.841
5	7.0	23.4	0.852	0.780	0.776	0.845
6	8.3	34.5	0.859	0.818	0.773	0.856
7	9.7	46.6	0.871	0.853	0.782	0.870
10	13.0	93.5	0.893	0.901	0.785	0.904
20	24.9	375.1	0.927	0.930	0.803	0.930

14. tabula: Simulētā jauda hipotēžu pārbaudei H_0: nav atšķirības starp lokācijas parametriem. Modelis F_1 = (1-ϵ)Gamma(a = 5; s =1)+ $\epsilon Unif[0, 50]$; F_2 = Gamma(a = 1; s = $1/\sigma$), $n_1 = n_2 = 50$, $N = 10,000$, $\epsilon = 0$ un $\epsilon = 0.06$. Gludā Hūbera novērtējuma mēroga novērtējums ar MAD.

		$\epsilon = 0$				$\epsilon = 0.06$			
σ	Līmenis	t-tests	EL	Huber, V=2.046	Huber, V novērtēts	t-tests	EL	Huber, V=1.302	Huber, V novērtēts
5	0.10	0.1039	0.1073	0.1023	0.1077	0.2605	0.2968	0.2081	0.2243
	0.05	0.0559	0.0531	0.0535	0.0541	0.1482	0.196	0.1286	0.1381
6	0.10	0.2623	0.3149	0.1400	0.3212	0.1041	0.1256	0.0884	0.1213
	0.05	0.1488	0.2042	0.0803	0.2094	0.0477	0.0681	0.0428	0.0617
7	0.10	0.6266	0.6886	0.4306	0.6944	0.1864	0.1869	0.2525	0.2906
	0.05	0.4685	0.5657	0.315	0.5725	0.1133	0.1225	0.1619	0.1864
8	0.10	0.8713	0.9077	0.7342	0.9115	0.3811	0.3582	0.5274	0.5455
	0.05	0.7666	0.8400	0.6264	0.8451	0.2694	0.2532	0.4008	0.4143
9	0.10	0.9674	0.9776	0.9042	0.9786	0.5815	0.5503	0.7611	0.7578
	0.05	0.9235	0.9601	0.8391	0.9618	0.4607	0.4242	0.6635	0.6451

15. tabula: Pārklājuma precizitāte modelim F_1 = Cauchy(0; s = σ); F_2 = Cauchy(0; s = $1/\sigma$), $N = 1000$, $n_1 = n_2 = 50$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
3	25	0.311	0.971	0.764	0.654	0.625
5	75	0.117	0.981	0.731	0.618	0.498
7	156	0.063	0.974	0.758	0.668	0.588
10	278	0.031	0.980	0.792	0.732	0.687
20	1147	0.008	0.975	0.931	0.918	0.911

16. tabula: Pārklājuma precizitāte modelim $F_1 = \text{Doublexp}(0;\, \sigma)$; $F_2 = \text{Doublexp}(0;\, 1)$, $n_1 = n_2 = 50$, $N = 1000$.

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
0.1	0.01	1.12	0.960	0.940	0.955	0.945
0.2	0.05	1.19	0.955	0.935	0.950	0.950
0.5	0.30	1.19	0.940	0.940	0.945	0.945
2	4.78	1.19	0.955	0.935	0.965	0.960
5	29.85	1.19	0.965	0.940	0.960	0.945
10	119.39	1.19	0.965	0.945	0.960	0.940

17. tabula: Pārklājuma precizitāte modelim $F_1 = \text{Hlf}(0;\, \sigma)$, $F_2 = \text{Hlf}(0;\, 1)$, $n_1 = n_2 = 50$, $N = 1000$

σ	Huber V1	Huber V2	t-tests	EL	Huber, V=2.046	Huber, V novērtēts
0.2	0.08	2.15	0.95	0.936	0.938	0.941
0.5	0.49	2.25	0.951	0.945	0.946	0.944
1	1.95	2.25	0.949	0.943	0.941	0.942
2	7.81	2.25	0.955	0.944	0.946	0.944
5	48.83	2.25	0.954	0.947	0.950	0.945
10	195.32	2.25	0.964	0.947	0.954	0.948

Secinājumi

Maģistra darbā tika aplūkotas robustās statistikas pamatnostādnes, kā arī pielietota empīriskās ticamības funkcijas metode nesen ieviestajiem gludajiem Hubera M-novērtējumiem.

Simulāciju piemēros vienas izlases gadījumā tika noskaidrots, ka gludie Hūbera novērtējumi darbojas labāk, salīdzinājumā ar to negludinātajām versijām. Tika ieviesta empīriskās ticamības metode divu gludo Hūbera novērtējumu starpībai un metodes sniegums tika analizēts salīdzinājumā ar divu izlašu t-testu un empīriskās ticamības metodi vidējai vērtībai. Metodes ieviešanas gaitā izrādījās, ka Hampel [13] norādījums par fiksētas asimptotiskās dispersijas V izvēli gludā Hūbera novērtējuma aprēķinā empīriskās ticamības metodes gadījumā dod pretrunīgus rezultātus. Tāpēc tika secināts, ka šāda pieeja nav pareiza, un pareizāk ir V novērtēt.

Kopumā var secināt ka empīriskās ticamības metode divu gludo Hūbera novērtējumu starpībai ar novērtētu V darbojas līdzīgi EL metodei divu vidējo vērtību starpībai. Simulāciju piemēros vērojama tendence, ka EL metodei vidējai vērtībai darbojas labāk tajos piemēros, kur dati nav piesārņoti. Savukārt EL metodei, kura balstīta uz gludo Hūbera novērtējumu, bija vērojamas nelielas priekšrocības divos gadījumos. Pirmkārt, situācijās, kur datos parādās piesārņojums. Šāds rezultāts atbilst robustības–efektivitātes kompromisa principam, kas nosaka, ka robustās metodes pie modeļa ir mazāk efektīvas nekā klasiskās statistikas metodes. Otrkārt, jaunā metode pārspēja pārējās metodes situācijās pie asimetriskiem sadalījumiem ar ļoti atšķirīgiem skalas parametriem.

Arī reālo datu piemēros bija vērojams, ka EL metode vidējai vērtībai un Hubera novērtējumam ar novērtētu V darbojās līdzīgi. Savukārt V fiksēšana vairākos piemēros lika pieņemt pretēju lēmumu par H_0 nekā pārējās EL metodes. Šāda tendence arī norāda uz to, ka metode nedarbojas pareizi. Rezultāti liecina, ka empīriskās ticamības metode divu Hūbera novērtējumu starpībai ir vēl viena metode, kuru iespējams lietot sadalījumu lokācijas parametru salīdzināšanai divu izlašu problēmās. Īpaši gadījumos, kad dotie dati ir ar piesārņojumu, to būtu noderīgi lietot paralēli klasiski lietotajām statistikas metodēm.

Saistībā ar šī darba rezultātiem sagatavošanā ir publikācija. Turklāt, robustā statistika ir izrādījies interesants darba lauks, un tālāk tiek plānots aplūkot empīrisko ticamības metodi divu nošķeltu vidējo vērtību starpībai, balstoties uz Qin un Tsao [23] rezultātu vienas izlases gadījumā.

Izmantotā literatūra un avoti

[1] J.W. Tukey. *A survey of sampling from contaminated distributions*, pages 221–223. Stanford University Press, Stanford, CA, 1960.

[2] J. W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, (33):1–67, 1962.

[3] P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, (35):73–101, 1964.

[4] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard condions. *Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics Probability*, (1):221–223, 1967.

[5] F. R. Hampel. Ageneral definition of qualitative robustness. *The Annals of Mathematical Statistics*, (42):1887 – 1896, 1971.

[6] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of American Statistical Association*, (69):383 – 393, 1974.

[7] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, (75):237–249, 1988.

[8] A. B. Owen. Empirical likelihood confidence regions. *The Annals of Statistics*, 18:90–120, 1990.

[9] A. B. Owen. Empirical likelihood for linear models. *The Annals of Statistics*, (19):1725–1747, 1991.

[10] J. Qin and J. F. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, (22):300–325, 1994.

[11] M. Tsao and J. Zhou. On the robustness of empirical likelihood ratio confidence intervals for location. *Canadian Journal of Statistics*, (29).

[12] Y. Qin and L. Zhao. Empirical likelihood ratio confidence intervals for various differences of two populations. *Syst. Sci. Math. Sci*, (13):23–30, 2000.

[13] F. Hampel, C. Hennig, and E.A. Ronchetti. A smoothing principle for the huber and other location m-estimators. *Computational Statistics & Data Analysis*, (55):324–337, 2011.

[14] P. J. Huber and E. Ronchetti. *Robust Statistics, 2nd edition.* Wiley, New York, 2009.

[15] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[16] F. R. Hampel, E. M. Roncheti, P. J. Rousseeuw, and A. W. Stahel. *Robust statistics. The approach based on influence functions.* Wiley, New York, 1986.

[17] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics. Theory and Methods.* Wiley, New York, 2006.

[18] E. J. Anscombe. Rejection of outliers. *Technometrics*, (2):123–147, 1960.

[19] B. Zhang. On the accuracy of empirical likelihood confidence intervals for m-functionals. *Journal of Nonparametric Statistics*, (29).

[20] S. Heritier, E. Cantoni, S. Copt, and M. P. Victoria Feser. *Robust Methods in Biostatistics.* 2009.

[21] S. M. Stigler. Do robust estimators work with real data? *The Annals of Statistics*, (5):1055–1098, 1977.

[22] A. Marazzi. Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational statistics & data analysis*, (49):503–528, 2002.

[23] G. Qin and M. Tsao. Empirical likelihood ratio confidence interval for the trimmed mean. *Communicatons in Statistics*, 31(12).