LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKĀ TICAMĪBAS FUNKCIJA AR NOVĒRTĒTIEM PARAMETRIEM

DIPLOMDARBS

Autors: **Leonora Pahirko**

Stud. apl. lp06061

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2011

Saturs

Ievads **1**

1. Empīriskās ticamības funkcijas metode vienas izlases gadījumā 3
   1.1. Neparametriskā vislielākās ticamības funkcija 3
   1.2. EL metode ar novērtētiem parametriem vienas izlases gadījumā . . 6
2. Empīriskās ticamības funkcijas metode divu izlašu gadījumā 11
   2.1. EL ar novērtētiem parametriem divām izlasēm 14
3. EL metodes ar novērtētiem parametriem pielietojumi vienas izlases gadījumā 16
   3.1. Hodžes-Lēmaņa lokācijas novērtējums[1] 16
   3.2. Neparametriskās regresijas atlikumu sadalījumi 20
4. EL metodes ar novērtētiem parametriem pielietojumi divu izlašu gadījumā 25
   4.1. Strukturālo attiecību modeļi 25
   4.2. Citi piemēri 30

Secinājumi **33**

Izmantotā literatūra un avoti **34**

1. Pielikums 37

Anotācija

Darbs veltīts empīriskās ticamības funkcijas metodei, kas ir viena no populārākajām neparametriskās statistikas metodēm. Galvenais uzsvars likts uz empīriskās ticamības funkcijas metodes vispārinājumu, kas iegūts 2009.gadā, lai atļautu traucējošo parametru klātbūtni novērtējošajās funkcijās un noskaidrotu šīs metodes robežsadalījumu, kad parametri tiek novērtēti. Darbā apskatīta metode gan vienas izlases, gan divu izlašu gadījumā, ko definējis J.Valeinis [2] savā doktora disertācijā, kā arī tiek piedāvāti dažādi metodes pielietojumi. Empīriskās ticamības funkcijas metode ar novērtētiem parametriem vienas izlases gadījumā implementēta datorprogrammā R un teorētiskie rezultāti pielietoti reālu datu piemēriem. Apskatīti arī piemēri divu izlašu gadījumā, izmantojot J.Valeiņa un E.Cera [3] izstrādāto programmas R kodu.

Atslēgas vārdi: empīriskās ticamības funkcijas metode, novērtēti parametri, traucējošie parametri

Abstract

This thesis is dedicated to empirical likelihood method, which is one of the most widely used methods in nonparametric statistics. The main emphasis is on generalization of the empirical likelihood method to allow for plug-in estimates for nuisance parameters in estimating equations. In one sample case the plug-in empirical likelihood has been derived in 2009, but J. Valeinis defined method for two sample case in his PhD thesis [2]. There has been done an implementation of plug-in empirical likelihood for the program R to study various applications in one sample case. Some applications for two sample case also have been considered and several data examples were analyzed for both cases. The implementation of plug-in empirical likelihood for two sample case was already made by J.Valeinis and E.Cers [3].

Keywords: empirical likelihood function, plug-in estimates, nuisance parameters

Ievads

Pēdējā laikā statistikā aizvien lielāku popularitāti gūst neparametriskās metodes, dodot iespēju izvairīties no kļūdām, kas rodas lietojot neprecīzus pieņēmumus par datu sadalījumu. Arī empīriskās ticamības funkcijas (turpmāk EL) metode ir konkurētspējīga neparametriskā metode, kas modelē nezināmo datu sadalījumu, balstoties uz dotajiem novērojumiem.

Par empīriskās ticamības funkcijas metodes pamatlicēju tiek uzskatīts Owen ([4],[5]), kas piedāvāja EL metodes pielietojumu ticamības intervālu un reģionu konstruēšanai, taču šīs metodes pirmsākumi meklējami Thomas un Grunkemeier [6] darbā par ticamības intervālu novērtēšanu izdzīvošanas datu analīzē. Tomēr Owen vispārināja rezultātus, kas bija iegūti parametriskai vislielākās ticamības metodei, un parādīja, ka EL metodes statistika tiecas uz χ^2_p sadalījumu, kur p ir interesējošā parametra dimensija. Savukārt Qin un Lawless [7] 1994.gadā aprakstīja EL metodi vienas izlases gadījumā vispārējā formā, izmantojot novērtējošās funkcijas.

Pēdējos gados ir bijuši arī vairāki mēģinājumi vispārināt iegūtos rezultātus, lai atļautu traucējošo parametru klātbūtni novērtējošajās funkcijās. Vienas izlases gadījumā empīrisko ticamības metodi ar novērtētiem traucējošiem parametriem vispārējā formā definēja 2009. gadā Hjort, McKeague un Van Keilegom (skatīt [1]), piedāvājot arī vairākus pielietojumus tādās statistikas nozarēs kā piemēram neparametriskā regresija, izdzīvošanas datu analīze u.c.

2000. gadā Qin un Zhao [8] pirmo reizi aprakstīja empīriskās ticamības funkcijas metodi divu izlašu vidējo vērtību un sadalījuma funkciju starpībai. 2007. gadā Valeinis savā doktora disertācijā [2] parādīja, ka Qin un Zhao [8] rezultātus iespējams vispārināt, pielietojot tos tādām divu izlašu problēmām kā kvantiļu funkciju starpībām, varbūtību-varbūtību (P-P) un kvantiļu-kvantiļu (Q-Q) grafikiem, ROC līknēm un strukturālo attiecību modeļiem. Šī iemesla dēļ Valeinis [2] arī vispārināja Hjort, McKeague un Van Keilegom formulēto empīriskās ticamības funkcijas metodi ar novērtētiem traucējošiem parametriem divu izlašu gadījumā, lai konstruētu ticamības intervālus strukturālo attiecību modeļiem ar novērtētu traucējošo parametru h.

Šī darba galvenais mērķis ir iepazīties ar pieejamiem teorētiskajiem rezultātiem par empīriskās ticamības funkcijas metodi ar novērtētiem parametriem un apskatīt tās pielietojumus reāliem datu piemēriem, implementējot to programmā R.

Darbs sastāv no 4 nodaļām un pielikuma. Pirmā nodaļa veltīta empīriskās ticamības funkcijas definēšanai vienas izlases gadījumā, kā arī apskatīta metode ar novērtētiem traucējošiem parametriem. Otrajā nodaļā sniegts empīriskās funkcijas metodes un EL metodes ar novērtētiem parametriem apraksts divu izlašu gadījumā, savukārt trešajā un ceturtajā nodaļā apskatīti empīriskās ticamības funkcijas pielietojumi gan vienas, gan divu izlašu problēmām, un doti vairāki reālu datu piemēri. Pielikumā atrodams programmas R kods apskatītajiem piemēriem.

1. Empīriskās ticamības funkcijas metode vienas izlases gadījumā

Šajā nodaļā apskatīsim empīriskās ticamības funkcijas metodi vienas izlases gadījumā, kā arī definēsim empīriskās ticamības funkcijas metodi ar novērtētiem traucējošiem parametriem.

1.1. Neparametriskā vislielākās ticamības funkcija

Definīcija 1. $X_1,\ldots,X_n$ neatkarīgi un vienādi sadalīti (turpmāk *iid*), empīriskā sadalījuma funkcija tiek definēta kā

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}},$$

kur $-\infty < x < \infty$ un indikatorfunkcija

$$I_{\{X_i \le x\}} = \begin{cases} 1, & X_i \le x \\ 0, & X_i > x \end{cases}.$$

Definīcija 2. $X_1,\ldots,X_n$ *iid* ar $X_1 \sim F$. Tad funkcijas F neparametriskā (empīriskā) ticamības funkcija ir

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

kur $p_i = P(X = X_i)$.

Acīmredzami, ka $L(F)$ ir varbūtība iegūt tieši dotās izlases $X_1,\ldots,X_n$ novērojumu vērtības no sadalījuma funkcijas F. Turklāt, $L(F) = 0$, ja F ir nepārtraukts sadalījums.

Sekojošā teorēma pierāda, ka neparametrisko ticamības funkciju maksimizē empīriskā sadalījuma funkcija F_n, tas nozīmē, ka F_n ir sadalījuma funkcijas F neparametriskās vislielākās ticamības funkcijas novērtējums.

Teorēma 1. *(Owen,[5]) Pieņemsim, ka $X_1,\ldots,X_n$ ir iid gadījuma lielumi ar sadalījuma funkciju F, un F_n ir to empīriskā sadalījuma funkcija. Ja $F \neq F_n$, tad $L(F) < L(F_n)$.*

Tālāk cieši sekosim Qin un Lawless (1994, [7]) publikācijai, kurā tika vispārināti Owena iegūtie rezultāti empīriskās ticamības funkcijas metodei. Pieņemsim, ka doti d-dimensionāli *iid* gadījuma lielumi $X_1,\ldots,X_n$ ar nezināmu sadalījuma funkciju F, un mēs

interesējamies par kādu p-dimensionālu parametru θ, kas ir saistīts ar F. Visa informācija par θ un F ir pieejama $r \geq p$ funkcionāli neatkarīgu nenovirzītu novērtējošo funkciju formā, t.i., $m_j(x,\theta)$, $j = 1, 2, \ldots, r$, tā ka $E_F m_j(x,\theta) = 0$.

Definēsim empīriskās ticamības funkcijas attiecību

$$R(F) = \frac{L(F)}{L(\widehat{F}_n)} = \prod_{i=1}^{n} np_i$$

un profila empīrisko ticamības funkciju

$$\mathrm{EL}(\theta) = \max \left\{ \prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m(X_i,\theta) = 0 \right\}, \tag{1.1}$$

kur $m(x,\theta) = (m_1(x,\theta), \ldots, m_r(x,\theta))^T$ un $E_F m(x,\theta) = 0$. Ja parametrs θ ir zināms, tad eksistē viens vienīgs problēmas (1.1) atrisinājums, ja 0 pieder $m(X_1,\theta)$,...,$m(X_n,\theta)$ lineārajai čaulai, un to var atrast ar Lagranža reizinātāju metodi. Tad logaritmisko profila empīriskās ticamības attiecības funkciju var uzrakstīt formā

$$l_E(\theta) = \ln \mathrm{EL}(\theta) = \sum_{i=1}^{n} \ln\{1 + \lambda^T m(x_i,\theta)\}, \tag{1.2}$$

kur λ ir Lagranža reizinātāju vektors.

Qin un Lawless ([7]) savā darbā apskata nosacījumus, kuriem izpildoties, var pierādīt, ka empīriskās ticamības attiecības statistika hipotēzei $H_0 : \theta = \theta_0$ ir

$$W_E(\theta_0) = 2l_E(\theta_0) - 2l_E(\widehat{\theta}) \to_d \chi^2_p,$$

kur $\widehat{\theta}$ ir parametra θ empīriskās vislielākās ticamības novērtējums. Izmantojot šo rezultātu, iespējams konstruēt ticamības intervālus un veikt hipotēžu pārbaudi parametram θ.

Piemērs 1. Pieņemsim, ka mums ir doti X_1,...,X_n *iid* gadījuma lielumi ar nezināmu sadalījuma funkciju F. Lai konstruētu ticamības intervālus ar empīriskās ticamības metodi sadalījuma funkcijas F vidējai vērtībai $\mu = EX = \int_{-\infty}^{+\infty} x dF(x)$, izmantosim profila empīriskās ticamības attiecības funkciju

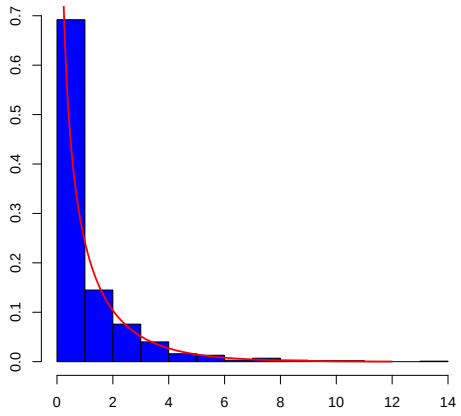
$$\mathrm{EL}(\mu) = \max \left\{ \prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu \right\}. \tag{1.3}$$

Pielietojot Lagranža reizinātāju metodi, iegūstam

$$\mathrm{EL}(\mu) = \prod_{i=1}^{n} \{1 + \lambda(X_i - \mu)\}^{-1}.$$

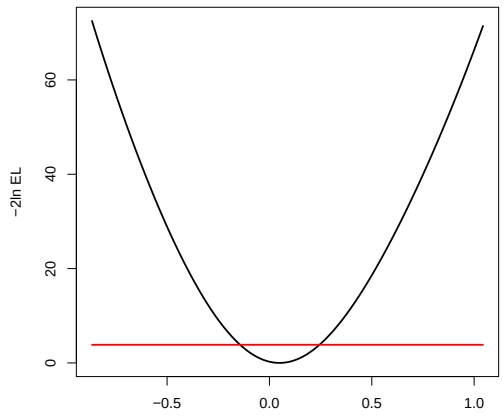
Pie zināmiem nosacījumiem Owen [4] pierādīja, ka $-2 \ln \mathrm{EL}(\mu_0) \to_d \chi^2_1$, kad $n \to \infty$.

Ticamības intervāli vidējai vērtībai



1. att.: Histogramma statistikai $-2\ln \mathrm{EL}(0)$ 1000 reižu ģenerētiem $N(0,1)$ datiem, n=100; χ^2_1 blīvuma funkcija

Ar simulāciju palīdzību tika pārbaudīta piemērā 1 aprakstītā problemātika, un 1. attēlā redzama histogramma 1000 reižu ģenerētai logaritmiskai empīriskās ticamības funkcijas statistikai $-2\ln \mathrm{EL}(\mu_0)$, kur $\mu_0 = 0$, $N(0,1)$ datiem apjomā $n = 100$. Redzams, ka χ^2_1 blīvuma funkcija labi aproksimē histogrammu.



2. att.: 95% ticamības intervāls $\mu_0 = 0$ ar EL metodi ģenerētiem N(0,1) datiem apjomā $n = 100$

2. attēlā redzams ticamības intervāls vidējai vērtībai $\mu_0 = 0$ standarta normāli sadalītiem simulētiem datiem, $\alpha = 0.05$. Tā kā īstā vērtība atrodas ticamības intervāla

1. tabula: Pārklājumu precizitāte 95% ticamības intervāliem vidējai vērtībai 10 000 reižu ģenerētiem $N(0,1)$ datiem apjomā n

	$n=20$	$n=50$	$n=100$
EL	0.9318	0.9524	0.949
t-tests	0.9501	0.9513	0.9469

2. tabula: Pārklājumu precizitāte 95% ticamības intervāliem 10 000 reižu ģenerētiem χ^2_1 datiem apjomā n, kuru pārbauda kā $N(0,1)$

	$n=20$	$n=50$	$n=100$
EL	0.8953	0.9297	0.9404
t-test	0.8445	0.8395	0.8355

robežās, kas ir [-0.1432,0.2445], tad varam domāt, ka EL metode strādā labi, bet, lai par to pārliecinātos, aplūkosim pārklājumu precizitāti ticamības intervāliem. 1. tabulā attēlota gan EL metodes, gan $t-$testa ticamības intervālu pārklājumu precizitāte izlases apjomiem $n = 20$, 50 un 100. Redzams, ka pie nelieliem izlašu apjomiem EL metode strādā nedaudz sliktāk, par t-testu, taču 10 000 reižu simulētiem datiem ar $n = 50$ un 100 pārklājumu precizitāte abiem testiem ir līdzīga, tātad šajā gadījumā varam secināt, ka EL metode strādā vienlīdz labi ar t-testu.

Savukārt 2. tabulā attēlota pārklājumu precizitāte 95% ticamības intervāliem konstruētiem ar empīriskās ticamības metodi un t-testu 10 000 reižu ģenerētiem χ^2_1 datiem ar dažādiem apjomiem n, izdarot pieņēmumu, ka dati sadalīti ar $N(0,1)$. Redzams, ka pie nelieliem izlašu apjomiem $n = 20$ un 50 EL metode strādā tikai nedaudz labāk par t-testu, bet pie $n = 100$ EL metode dod daudz labāku rezultātu nekā t-tests, kas skaidrojams ar neprecīzu pieņēmumu par datu sadalījumu.

1.2. EL metode ar novērtētiem parametriem vienas izlases gadījumā

Šajā nodaļā vispārināsim empīriskās ticamības metodi vienas izlases gadījumā, lai pieļautu situāciju, ka novērtējošās funkcijas var saturēt "traucējošos" (*nuisance*) parametrus,

kurus nepieciešams novērtēt. Šāda traucējošo parametru aizvietošana ar novērtējumiem pirmo reizi tika izmantota vairākos izdzīvošanas datu analīzes kontekstos 2001. gadā (Qin un Jing [9], Wang un Jing [10] u.c.), taču autori Hjort, McKeague un Van Keilegom [1] vispārināja empīriskās ticamības funkcijas metodi ar novērtētiem galīgi vai bezgalīgi dimensionāliem traucējošiem parametriem, kas atļauj to pielietot ne tikai izdzīvošanas analīzē, bet arī citās statistikas apakšnozarēs.

Turpmāk vadīsimies pēc Hjort, McKeague un Van Keilegom publikācijas [1], lai definētu plašāk pielietojamo empīriskās ticamības funkcijas metodi ar novērtētiem traucējošiem parametriem.

Pieņemsim, ka mums ir līdzīga situācija kā 1.1. nodaļā - $X_1,\ldots,X_n$ ir d-dimensionāli iid gadījuma lielumi, interesējošais parametrs ir θ_0, par kuru informācija ir dota p-dimensionālas novērtējošās funkcijas $m_n(X,\theta,h)$ formā, kur h ir traucējošais parametrs ar nezināmu īsto vērtību h_0.

Kad h_0 ir zināms, mēs varam aizstāt h ar tā īsto vērtību profila EL attiecības funkcijā

$$\mathrm{EL}_n(\theta,h)=\max\left\{\prod_{i=1}^n np_i \mid p_i>0, \sum_{i=1}^n p_i=1, \sum_{i=1}^n p_i m_n(X_i,\theta,h)=0\right\},$$

un atrast ticamības intervālus parametram θ formā $\{\theta : \mathrm{EL}_n(\theta,h_0)>c\}$, kur statistikas robezsadalījums apskatīts [7].

Apskatīsim nosacījumus, kas nodrošina empīriskās ticamības funkcijas metodes statistikas ar novērtētiem traucējošiem parametriem nedeģenerētu robežsadalījumu. Ieviesīsim sekojošus apzīmējumus vektoriem v, $||v||$ - Eiklīda norma un $v^{\otimes 2}=vv^T$, un matricām $V=(v_{ij})$, $|V|=\max_{ij}|v_{ij}|$. $\{a_n\}$ - pozitīvu konstanšu virkne un U - nedeģenerēts p-dimensionāls gadījuma vektors. V_2 - $p\times p$ pozitīvi definīta kovariāciju matrica.

(A0) $P(\mathrm{EL}_n(\theta_0,\hat{h})=0)\to 0$.

(A1) $\sum_{i=1}^n m_n(X_i,\theta_0,\hat{h})\to_d U$.

(A2) $a_n\sum_{i=1}^n m_n^{\otimes 2}(X_i,\theta_0,\hat{h})\to_{pr} V_2$.

(A3) $a_n \max_{1\le i\le n}||m_n(X_i,\theta_0,\hat{h})||\to_{pr} 0$.

Kad nezināmais h_0 tiek aizstāts ar novērtējumu $\hat{h}$ un novērtējošai funkcijai ir atļauts būt atkarīgai no n, tad, izpildoties nosacījumiem $(A0)$ - $(A3)$, iespējams vispārināt Qin un Lawless [7] iegūtos rezultātus.

Teorēma 2. *[1] Ja* $(A0)$ - $(A3)$ *ir spēkā, tad* $-2a_n^{-1}\log EL_n(\theta_0,\widehat{h}) \to_d U^tV_2^{-1}U$.

Piezīme 3. *Ovena EL teorēma seko no Teorēmas 2, izvēloties* $a_n = 1$ *un* $m_n(X_i,\theta_0,h) = m(X_i,\theta_0,h)/\sqrt{n}$.

Piemērs 2. Apskatīsim nosacījumus $(A0)$ - $(A3)$ vienkāršākajā gadījumā – vidējai vērtībai, t.i., kad $m(X_i,\theta,h) = X_i - \mu$. Pieņemsim, ka $m_n(X_i,\theta_0,h) = m(X_i,\theta_0,h)/\sqrt{n}$ un $a_n = 1$, U ir normāli sadalīts gadījuma lielums ar vidējo vērtību 0 un dispersiju σ^2, un $V_2 = \sigma^2$.

Nosacījums $(A0)$ ir ekvivalents $P(0 \in C_n) \to 1$, kur C_n apzīmē $\{m_n(X_i,\theta_0,\hat{h}), i = 1,\dots,n\}$ lineāro čaulu. Tātad šis nosacījums izpildīsies vienmēr, arī vidējai vērtībai, jo tiek pieņemts, ka maksimizācijas problēmai eksistē atrisinājums.

$(A1)$ seko no Centrālās robežteorēmas [11]:

$$\sum_{i=1}^n m(X_i,\theta_0,\hat{h})/\sqrt{n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i-\mu) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n X_i - n\mu\right) = \sqrt{n}(\bar{X}-\mu) \to_d N(0,\sigma^2).$$

Kā redzams, arī $(A2)$ izpildās, izmantojot Lielo skaitļu likumu [11]:

$$\sum_{i=1}^n m^2(X_i,\theta_0,\hat{h})/n = \sum_{i=1}^n \frac{(X_i-\mu)^2}{n} = \widehat{\sigma}^2 \to_p \sigma^2.$$

Savukārt nosacījums $(A3)$ ir spēkā, jo

$$\max_{1\le i\le n} |m(X_i,\theta_0,\hat{h})/\sqrt{n}| = \max_{1\le i\le n} |(X_i-\mu)/\sqrt{n}|, \tag{1.4}$$

un tā kā $\max_{1\le i\le n}|X_i| = o_{pr}(\sqrt{n})$ [4], tad (1.4) pēc varbūtības tiecas uz 0.

Redzams, ka Teorēma 2 dod rezultātu formā $-2\log EL_n(\theta_0,\widehat{h}) \to_d U^2/\sigma^2$, kas sakrīt ar Owen [4] rezultātu, tā kā $N(0,1)^2 = \chi_1^2$.

Teorēmas 2 pierādījums [1]

Šis pierādījums ir atšķirīgs no literatūrā līdz šim sastopamajiem EL rezultātu pierādījumiem, kuri balstās uz Teilora izvirzījumiem (piemēram, Qin un Lawless [7], Valeinis [2]). Autori tā vietā izmanto duālo optimizācijas problēmas risinājumu (Christianini un Shawe-Taylor, [12]).

Apzīmēsim $X_{n,i} = m_n(X_i,\theta_0,\widehat{h})$.

No $(A0)$

$$\mathrm{EL}_n = \mathrm{EL}_n(\theta_0,\widehat{h}) = \prod_{i=1}^n (1+\widehat{\lambda}^T X_{n,i})^{-1},$$

kur p-dimensionāls Lagranža reizinātāju vektors $\widehat{\lambda}$ apmierina vienādojumu

$$\sum_{i=1}^{n} X_{n,i}/(1+\widehat{\lambda}^T X_{n,i}) = 0.$$

Tad EL statistiku varam izteikt duālā formā

$$-2\log \mathrm{EL}_n = G_n(\widehat{\lambda}) = \sup_{\lambda} G_n(\lambda), \tag{1.5}$$

kur $G_n(\lambda) = 2\sum_{i=1}^{n}\log(1+\lambda^T X_{n,i})$, un G_n definīcijas apgabals ir kopa, kurā tā ir definēta (attiecībā uz $\log x$, kas nav definēts $x \leq 0$). Jāatzīmē, ka G_n ir ieliekta un sasniedz maksimumu pie $\widehat{\lambda}$, tā kā $\nabla G_n(\widehat{\lambda}) = 0$.

Tālāk apskatīsim G_n kvadrātisko aproksimāciju

$$G_n^*(\lambda) = 2\lambda^T U_n - \lambda^T V_n \lambda, \quad \text{kur} \quad U_n = \sum_{i=1}^{n} X_{n,i}, \quad V_n = \sum_{i=1}^{n} X_{n,i}^{\otimes 2},$$

un G_n^* definīcijas apgabals ir $\mathbb{R}^p$.

Nedaudz tālāk pierādīsim, ka starpība starp maksimālajām G_n un G_n^* vērtībām ir ar kārtu $o_{pr}(a_n)$.

Tad no (1.5) un no fakta, ka G_n^* tiek maksimizēta pie $\lambda^* = V_n^{-1}U_n$, kad V_n apgriežama, seko, ka

$$-2a_n^{-1}\log \mathrm{EL}_n = a_n^{-1}\sup_{\lambda} G_n^*(\lambda) + o_{pr}(1) = U_n^T(a_n V_n)^{-1}U_n + o_{pr}(1), \tag{1.6}$$

kas pēc sadalījuma tiecas uz $U^T V_2^{-1} U$, ja pieņemam, ka $(A1)$ un $(A2)$ ir spēkā. No šī pierādījuma seko, ka teorēma 2 ir spēkā arī gadījumos, kad $(U_n, V_n) \to_d (U, V_2)$ (drīzāk ar gadījuma nekā ar fiksētu V_2).

Tātad, lai pabeigtu pierādījumu, vēl atlicis pierādīt, ka

$$\sup G_n - \sup G_n^* = o_{pr}(a_n).$$

Pirmkārt, noteiksim $\widehat{\lambda}$ stohastisko kārtu. Rakstīsim $\widehat{\lambda} = ||\widehat{\lambda}||u$, kur u vienības vektors. Tad ir spēkā

$$||\widehat{\lambda}||(u^T V_n u - D_n u^T U_n) \leq u^T U_n,$$

kur $D_n = \max_{i\leq n}||X_{n,i}||$. Bet $u^T V_n u \geq \min \mathrm{eig}(V_n) = o_{pr}(a_n^{-1})$, kur $\min \mathrm{eig}(V_n)$ ir matricas V_n minimālā īpašvērtība, $u^T U_n = o_{pr}(1)$ un $D_n u^T U_n = o_{pr}(a_n^{-1})$, tāpēc $||\widehat{\lambda}|| = o_{pr}(a_n)$. Turklāt $\lambda^* = V_n^{-1}U_n$, kad V_n apgriežama, tāpēc λ^* ir ar tādu pašu stohastisko kārtu $o_{pr}(a_n)$ kā $\widehat{\lambda}$.

Ir zināms, ka $\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 h(x)$, kur $|h(x)| \le 2$ un $|x| \le \frac{1}{2}$. No tā seko, ka katram $c > 0$ un $||\lambda|| \le c$,

$$G_n(\lambda) = 2\lambda^T U_n - \lambda^T V_n \lambda + r_n(\lambda),$$

kur

$$|r_n(\lambda)| \le \frac{2}{3}\sum_{i=1}^{n} |(\lambda^T X_{n,i})^3||h(\lambda^T X_{n,i})|$$

$$\le \frac{4}{3}||\lambda|| D_n \lambda^T V_n \lambda \le \frac{4}{3} c^3 D_n \max \mathrm{eig}(V_n),$$

kur $\max \mathrm{eig}(V_n)$ ir matricas V_n maksimālā īpašvērtība, kas nodrošina $cD_n \le \frac{1}{2}$.

Ar $T_{n,c}$ un $T^*_{n,c}$ apzīmēsim maksimālās G_n un G^*_n vērtības apgabalā $\Omega_n(c)$, kas ir vienāds ar $\{\lambda : ||\lambda|| \le ca_n\}$, un iegūstam

$$\left|\frac{T_{n,c}}{a_n} - \frac{T^*_{n,c}}{a_n}\right| \le \frac{1}{a_n} \max\{|r_n(\lambda)| : ||\lambda|| \le ca_n\}$$

$$\le \frac{4}{3} c^3 a_n D_n \max \mathrm{eig}(a_n V_n),$$

kamēr $ca_n D_n \le \frac{1}{2}$. Izvēlēsimies c pietiekami lielu, lai gan $\widehat{\lambda}$, gan λ^* pieder $\Omega_n(c)$ ar varbūtību lielāku par $1-\eta$, kādam $\eta > 0$. Tad

$$P\left\{\left|\frac{\sup G_n}{a_n} - \frac{\sup G^*_n}{a_n}\right| \ge \varepsilon\right\} \le P\left\{\frac{4}{3} c^2 a_n D_n \max \mathrm{eig}(a_n V_n) \ge \varepsilon\right\}$$

$$+P\left\{||\widehat{\lambda}|| > ca_n\right\} + P\left\{||\lambda^*|| > ca_n\right\} + P\left\{ca_n D_n > \frac{1}{2}\right\}.$$

Tātad varbūtību virkne kreisajā pusē ir ierobežota ar 2η. Tā kā η tika izvēlēts patvaļīgi, tad $\sup G_n/a_n$ un $\sup G^*_n/a_n$ ir jābūt vienādiem robežsadalījumiem, kas ir $U^T V_2^{-1} U$.■

Piezīme 4. *Speciālā gadījumā, kad traucējošais parametrs ir galīgdimensionāls, profila empīriskās ticamības statistika*

$$-2\ln\{\max_h EL_n(\theta_0, h) / \max_{\theta,h} EL_n(\theta, h)\} \to_d \chi^2_p,$$

izpildoties dažādiem regularitātes nosacījumiem no [7].

Šis rezultāts ļauj ērti konstruēt ticamības intervālus parametram θ un ir vieglāk pielietojams nekā EL metode ar novērtētiem parametriem, taču tas nav pielietojams bezgalīgi dimensionāiem traucējošajiem parametriem, turklāt novērtējošai funkcijai jābūt diferencējamai pret (θ, h).

2. Empīriskās ticamības funkcijas metode divu izlašu gadījumā

Kopš Owen [4] ieviesa empīriskās ticamības funkcijas metodi, ir bijuši vairāki mēģinājumi to vispārināt arī divu izlašu gadījumam. Vieni no pirmajiem bija Qin un Zhao [8], kas 2000. gadā iepazīstināja ar empīriskās ticamības metodi divu izlašu vidējo vērtību starpībai un sadalījuma funkciju starpībai kādā fiksētā punktā. Nedaudz vēlāk Valeinis savā doktora disertācijā [2] parādīja, ka Qin un Zhao rezultātus iespējams paplašināt, lai EL metodi pielietotu kvantiļu funkciju starpībām, P-P un Q-Q grafikiem, ROC līknēm un strukturālo attiecību modeļiem [2], bet vispārīgā formā empīrisko ticamības metodi divu izlašu gadījumā aprakstīja 2011. gadā [3].

Pieņemsim, ka ir dotas divas *iid* izlases $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ ar nezināmām sadalījuma funkcijām attiecīgi F_1 un F_2. Mūsu mērķis ir konstruēt ticamības intervālus kādam interesējošam parametram Δ, turklāt θ_0 ir parametrs, kas saistīts ar vienu no sadalījuma funkcijām F_1 vai F_2. Tāpat kā vienas izlases gadījumā, pieņemsim, ka visa informācija par θ_0, Δ, F_1 un F_2 ir pieejama nenovirzītu novērtējošo funkciju veidā, t.i.,

$$E_{F_1} m_1(X, \theta_0, \Delta) = 0, \tag{2.1}$$

$$E_{F_2} m_2(Y, \theta_0, \Delta) = 0. \tag{2.2}$$

Piemērs 3. Apskatīsim divu izlašu vidējo vērtību starpību. Apzīmēsim $\theta_0 = \int x dF_1(x)$ un $\Delta = \int y dF_2(y) - \int x dF_1(x)$. Šajā gadījumā kā novērtējošās funkcijas izvēlamies

$$m_1(X, \theta_0, \Delta) = X - \theta_0, \quad m_2(Y, \theta_0, \Delta) = Y - \theta_0 - \Delta,$$

lai (2.1) un (2.2) būtu spēkā.

Piemērs 4. Ja interesējamies par kvantiļu – kvantiļu jeb Q-Q grafiku, tad $\theta_0 = F_2(t)$ un $\Delta = F_1^{-1}(F_2(t))$. Nenovirzītas novērtējošās funkcijas ir formā

$$m_1(X, \theta_0, \Delta) = I_{\{X \le \Delta\}} - \theta_0, \quad m_2(Y, \theta_0, \Delta) = I_{\{Y \le t\}} - \theta_0.$$

Definīcija 3. $X_1, \ldots, X_n \sim F_1$ *iid* un $Y_1, \ldots, Y_m \sim F_2$ *iid*, neparametriskā (empīriskā) ticamības funkcija ir

$$L(F_1, F_2) = \prod_{i=1}^{n} p_i \prod_{j=1}^{m} q_i = \prod_{i=1}^{n} P(X = X_i) \prod_{j=1}^{m} P(Y = Y_j).$$

Arī divu izlašu gadījumā empīriskās ticamības funkciju $L(F_1, F_2)$ maksimizē izlašu empīriskās sadalījuma funkcijas F_{1n} un F_{1m} [8]. Tālāk varam definēt empīriskās ticamības attiecības funkciju

$$R(F_1, F_2) = L(F_1, F_2)/L(F_{1n}, F_{2m}) = \prod_{i=1}^{n}(np_i)\prod_{j=1}^{m}(mq_i).$$

Lai konstruētu ticamības intervālus parametram Δ, definēsim arī profila empīrisko ticamības attiecības funkciju

$$\mathrm{EL}(\Delta, \theta) = \sup_{\theta,p,q}\prod_{i=1}^{n}(np_i)\prod_{j=1}^{m}(mq_j), \tag{2.3}$$

kur $p = (p_1, \ldots, p_n)$ un $q = (q_1, \ldots, q_m)$ uzlikti ierobežojumi

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i m_1(X_i, \theta, \Delta) = 0,$$

$$q_i \geq 0, \quad \sum_{i=1}^{m} q_j = 1, \quad \sum_{i=1}^{m} q_j m_2(Y_j, \theta, \Delta) = 0.$$

Ja θ ir dots, tad eksistē viens vienīgs (2.4) atrisinājums, ja 0 pieder $m_1(X_i, \theta, \Delta)$ izliektajai čaulai un $m_2(Y_j, \theta, \Delta)$ izliektajai čaulai.

Maksimizācijas problēmu var atrisināt ar Lagranža reizinātāju metodi [8], no kurienes varam iegūt profila empīriskās ticamības attiecības funkciju parametram Δ izskatā

$$\mathrm{EL}(\Delta, \theta) = \prod_{i=1}^{n}\left\{\frac{1}{1 + \lambda_1(\theta)m_1(X_i, \theta, \Delta)}\right\}\prod_{j=1}^{m}\left\{\frac{1}{1 + \lambda_2(\theta)m_2(Y_j, \theta, \Delta)}\right\},$$

kur λ_1 un λ_2 ir Lagranža reizinātāji un atrodami no vienādojumiem

$$\frac{1}{n}\sum_{i=1}^{n}\frac{m_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta)m_1(X_i, \theta, \Delta)} = 0,$$

$$\frac{1}{m}\sum_{j=1}^{m}\frac{m_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta)m_2(Y_j, \theta, \Delta)} = 0.$$

Savukārt logaritmiskā empīriskā ticamības attiecības statistika ir

$$-2\ln\mathrm{EL}(\Delta, \theta) = 2\sum_{i=1}^{n}\ln(1 + \lambda_1(\theta)m_1(X_i, \theta, \Delta)) + 2\sum_{j=1}^{m}\ln(1 + \lambda_2(\theta)m_2(Y_j, \theta, \Delta)).$$

Piemērs 5. Ja interesējamies par kvantiļu – kvantiļu jeb Q-Q grafiku, tad $\theta_0 = F_2(t)$ un $\Delta = F_1^{-1}(F_2(t))$. Nenovirzītas novērtējošās funkcijas ir formā

$$m_1(X, \theta_0, \Delta) = I_{\{X \leq \Delta\}} - \theta_0, \quad m_2(Y, \theta_0, \Delta) = I_{\{Y \leq t\}} - \theta_0.$$

Definīcija 4. $X_1, \ldots, X_n \sim F_1$ *iid* un $Y_1, \ldots, Y_m \sim F_2$ *iid*, neparametriskā (empīriskā) ticamības funkcija ir

$$L(F_1, F_2) = \prod_{i=1}^{n} p_i \prod_{j=1}^{m} q_i = \prod_{i=1}^{n} P(X = X_i) \prod_{j=1}^{m} P(Y = Y_j).$$

Arī divu izlašu gadījumā empīriskās ticamības funkciju $L(F_1, F_2)$ maksimizē izlašu empīriskās sadalījuma funkcijas F_{1n} un F_{1m} [8]. Tālāk varam definēt empīriskās ticamības attiecības funkciju

$$R(F_1, F_2) = L(F_1, F_2)/L(F_{1n}, F_{2m}) = \prod_{i=1}^{n}(np_i) \prod_{j=1}^{m}(mq_i).$$

Lai konstruētu ticamības intervālus parametram Δ, definēsim arī profila empīrisko ticamības attiecības funkciju

$$\mathrm{EL}(\Delta, \theta) = \sup_{\theta,p,q} \prod_{i=1}^{n}(np_i) \prod_{j=1}^{m}(mq_j), \tag{2.4}$$

kur $p = (p_1, \ldots, p_n)$ un $q = (q_1, \ldots, q_m)$ uzlikti ierobežojumi

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i m_1(X_i, \theta, \Delta) = 0,$$

$$q_i \geq 0, \quad \sum_{i=1}^{m} q_j = 1, \quad \sum_{i=1}^{m} q_j m_2(Y_j, \theta, \Delta) = 0.$$

Ja θ ir dots, tad eksistē viens vienīgs (2.4) atrisinājums, ja 0 pieder $m_1(X_i, \theta, \Delta)$ izliektajai čaulai un $m_2(Y_j, \theta, \Delta)$ izliektajai čaulai.

Maksimizācijas problēmu var atrisināt ar Lagranža reizinātāju metodi [8], no kurienes varam iegūt profila empīriskās ticamības attiecības funkciju parametram Δ izskatā

$$\mathrm{EL}(\Delta, \theta) = \prod_{i=1}^{n} \left\{ \frac{1}{1 + \lambda_1(\theta) m_1(X_i, \theta, \Delta)} \right\} \prod_{j=1}^{m} \left\{ \frac{1}{1 + \lambda_2(\theta) m_2(Y_j, \theta, \Delta)} \right\},$$

kur λ_1 un λ_2 ir Lagranža reizinātāji un atrodami no vienādojumiem

$$\frac{1}{n} \sum_{i=1}^{n} \frac{m_1(X_i, \theta, \Delta)}{1 + \lambda_1(\theta) m_1(X_i, \theta, \Delta)} = 0,$$

$$\frac{1}{m} \sum_{j=1}^{m} \frac{m_2(Y_j, \theta, \Delta)}{1 + \lambda_2(\theta) m_2(Y_j, \theta, \Delta)} = 0.$$

Savukārt logaritmiskā empīriskā ticamības attiecības statistika ir

$$-2 \ln \mathrm{EL}(\Delta, \theta) = 2 \sum_{i=1}^{n} \ln(1 + \lambda_1(\theta) m_1(X_i, \theta, \Delta)) + 2 \sum_{j=1}^{m} \ln(1 + \lambda_2(\theta) m_2(Y_j, \theta, \Delta)).$$

Teorēma 5. *(Qin un Zhao, [8]) Izpildoties standarta nosacījumiem,*

$$-2\ln EL(\Delta,\hat{\theta}) \to_d \chi_1^2,$$

kur $\hat{\theta}$ ir parametra θ_0 novērtējums, kas maksimizē (2.4).

Izmantojot šo rezultātu, varam konstruēt uz EL metodes balstītus ticamības intervālus interesējošajam parametram Δ formā $\Delta : \{\mathrm{EL}(\Delta,\hat{\theta}) > c\}$.

2.1. EL ar novērtētiem parametriem divām izlasēm

Šajā nodaļā apskatīsim vispārinātu EL metodi ar novērtētiem traucējošiem parametriem divu izlašu gadījumā, kuru ieviesa Valeinis (2007) savā doktora disertācijā, lai konstruētu ticamības intervālus strukturālo attiecību modeļiem. Ir zināmi ļoti daudzi pielietojumi EL metodei ar novērtētiem parametriem vienas izlases gadījumā, taču, lai arī metode ir vispārināta divu izlašu gadījumam, pagaidām ir zināmi tikai daži praktiski pielietojami piemēri, kurus apskatīsim 3. nodaļā.

Pieņemsim, ka mums ir līdzīga situācija, kā aprakstīts EL metodei divu izlašu gadījumā bez novērtētiem traucējošiem parametriem. Taču tāpat kā vienas izlases gadījumā novērtējošās funkcijas var saturēt traucējošo parametru h ar nezināmu īsto vērtību h_0, tāpēc novērtējošās funkcijas ir izskatā

$$Em_1(X,\theta_0,\Delta_0,h_0) = 0 \text{ un } Em_2(Y,\theta_0,\Delta_0,h_0) = 0,$$

bet, lai saīsinātu pierakstu, lietosim $m_1(X,\theta_0,h) := m_1(X,\theta_0,\Delta,h)$ un $m_2(Y,\theta_0,h) := m_2(Y,\theta_0,\Delta,h)$. Tālāk definēsim EL metodi divu izlašu gadījumā, kur nezināmais h_0 tiek aizstāts ar tā novērtējumu $\hat{h}$, lai noteiktu konstanti c sekojoši konstruētos ticamības intervālos parametram Δ_0: $\{\Delta : \mathrm{EL}(\Delta,\hat{\theta},\hat{h} > c)\}$.

Apzīmēsim

$$M_{1n}(\theta,h) := \frac{1}{n}\sum_{i=1}^{n} m_1(X_i,\theta,h), \quad S_{1n}(\theta,h) := \frac{1}{n}\sum_{i=1}^{n} m_1^2(X_i,\theta,h),$$

$$M_{2m}(\theta,h) := \frac{1}{m}\sum_{j=1}^{m} m_2(Y_j,\theta,h), \quad S_{2m}(\theta,h) := \frac{1}{m}\sum_{j=1}^{m} m_1^2(Y_j,\theta,h).$$

Līdzīgi kā vienas izlases gadījumā, formulēsim pieņēmumus pirmajai izlasei $X_1,\ldots,X_n$.

(B0) $P(\mathrm{EL}(\theta_0,\hat{h}) = 0) \to 0$,kad $n \to \infty$.

(B1) $n^{1/2} M_{1n}(\theta_0, \hat{h}) \to_d U_1$, kur $U_1 \sim N(0, V_1)$.

(B2) $S_{1n}(\theta_0, \hat{h}) \to_p V_2$.

(B3) $n^{-1} \sum_{i=1}^{n} \alpha_1(X_i, \theta_0, \Delta, \hat{h}) \to_p V_3$, kur $\alpha_1(X_i, \theta_0, \Delta, \hat{h}) = \partial m_1(X_i, \theta_0, \hat{h})/\partial\theta_0$.

(B4) $\max_{1 \le i \le n} |m(X_i, \theta_0, \hat{h})| \to_p 0$

Pieņemsim, ka arī otrajai izlasei $Y_1, \ldots, Y_m$ ir spēkā $(B0)$ - $(B4)$ funkcijām M_{2m} un S_{2m}, aizstājot V_1, V_2 un V_3 ar T_1, T_2 un T_3, turklāt $\alpha_2(Y_i, \theta_0, \Delta, \hat{h}) = \partial m_2(Y_i, \theta_0, \hat{h})/\partial\theta_0$.

Teorēma 6. *(Valeinis,[2]) Ja* $(B0)$ - $(B4)$ *spēkā abām izlasēm, tad*

$$-2 \ln EL(\Delta, \hat{\theta}, \hat{h}) \to_d \frac{V_3^2 T_2 + k T_3^2 V_2}{T_1 V_3^2 + k V_1 T_3^2} \chi_1^2. \tag{2.5}$$

Piezīme 7. *Kā jau tika minēts, pierādījuma tehnika divu izlašu gadījumā empīriskās ticamības funkcijas metodei ar novērtētiem parametriem ir atšķirīga no [1] izmantotās, taču šeit pierādījumu neapskatīsim. Ar to var iepazīties [2].*

3. EL metodes ar novērtētiem parametriem pielietojumi vienas izlases gadījumā

Šajā nodaļā apskatīsim dažus piemērus ar $d = 1$, kuros izmantota EL metode ar novērtētiem parametriem. Kā jau iepriekš norādījām, pieņemsim, ka $a_n = 1$ un $U \sim N_p(0, V_1)$, kā arī $m_n(X_i, \theta, h) = m(X_i, \theta, h)/\sqrt{n}$, taču vispārinājums var būt noderīgs citos pielietojumos.

3.1. Hodžes-Lēmaņa lokācijas novērtējums[1]

Šajā nodaļā pieskarsimies rangu testiem, kas ir robusti pret izlēcēju klātbūtni izlasē, pretēji t-testam vai F-testam. Tomēr šo testu asimptotiskais sadalījums, protams, ir atkarīgs no dotās izlases sadalījuma, kādēļ rodas nepieciešamība to novērtēt.

Pieņemsim, ka mums ir dotas divas izlases $X_1, \ldots, X_n$ *iid* un $Y_1, \ldots, Y_m$ *iid*, un ir zināms, ka $P(X_i \leq x) = F(x)$ un $P(Y_i \leq x) = F(x - \Delta)$. Hodžes-Lēmaņa (turpmāk tekstā HL) lokācijas novērtējums ir

$$\hat{\Delta} = \operatorname{med}(X - Y),$$

kur $X - Y$ apzīmē nm starpības $X_i - Y_j$, $i = 1, \ldots, n$ un $j = 1, \ldots, m$.

Lai konstruētu ticamības intervālus lokācijas parametram Δ, nepieciešams zināt tā robežsadalījumu. Hodges un Lehmann publikācijā [13] pierādīja, ka $\hat{\Delta}$ ir asimptotiski normāli sadalīts ar vidējo vērtību 0 un dispersiju

$$\sigma^2 = \frac{1}{12\gamma(1-\gamma)\left(\int f^2(x)\mathrm{d}x\right)^2},$$

kur $n/(n+m) \to \gamma$, kad $n + m \to \infty$, un $f(x)$ ir funkcijas $F(x)$ blīvuma funkcija.

σ^2 novērtējums ar EL metodi

Lai novērtētu σ^2, interesēsimies par parametru $\theta = \int f^2 \mathrm{d}x$ ar īsto vērtību $\theta_0 = \int f_0^2 \mathrm{d}x$, jo HL novērtējuma asimptotiskā dispersija ir proporcionāla $1/\theta^2$ (skatīt arī [14] un [15]). Pieņemsim, ka funkcijas $F(x)$ teorētiskā blīvuma funkcija f_0 ir vienmērīgi nepārtraukta, bet nav vienmērīgā sadalījuma.

Kā novērtējošo funkciju izvēlēsimies $m(X, \theta, f) = f(X) - \theta$ ar traucējošo parametru f, kura novērtēšanai izmantosim $\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} k_b(X_i - x)$, kur $k_b(\cdot) = k(\cdot/b)/b$, kas ir pazīstamais kodolu blīvuma funkcijas novērtējums ar joslas platumu b.

Nosacījumu $(A0)$ **-** $(A3)$ **pārbaude**

Vispirms pārliecināsimies, ka novērtējošā funkcija $m(X,\theta,f)=f(X)-\theta$ ir nenovirzīta, t.i.,

$$E(f(X)-\theta)=E(f(X))-E\theta=\int_{-\infty}^{+\infty}f(x)f(x)\mathrm{d}x-\theta=0.$$

Lai pārliecinātos, ka ir spēkā nosacījumi $(A0)$ - $(A3)$, definēsim

$$V=\int(f_0-\theta_0)^2f_0\mathrm{dx}=\int f_0^3\mathrm{d}x-\left(\int f_0^2\mathrm{d}x\right)^2,$$

kas ir $\sum_{i=1}^n m(X_i,\theta_0,f_0)/\sqrt{n}$ asimptotiskā dispersija, un ir pozitīva, tā kā f_0 nav vienmērīgā sadalījuma. Pārliecināsimies, ka $(A2)$ ir spēkā, kad $V_2=V$. Tātad

$$n^{-1}\sum_{i=1}^n m^2(X_i,\theta_0,\hat{f})=n^{-1}\sum_{i=1}^n\left(\hat{f}(X_i)-\theta_0\right)^2=\int\hat{f}^2\mathrm{d}F_n-2\theta_0\widehat{\theta}+\theta_0^2,$$

kur F_n empīriskā sadalījuma funkcija un $\widehat{\theta}=n^{-1}\sum_{i=1}^n\hat{f}(X_i)=\int\hat{f}\mathrm{d}F_n$. Tad $\int\hat{f}\mathrm{d}F_n$ un $\int\hat{f}^2\mathrm{d}F_n$ pēc varbūtības konverģē attiecīgi uz $\int f_0^2\mathrm{d}x$ un $\int f_0^3\mathrm{d}x$, kad $b\to0$ un $nb\to\infty$.

Lai pārbaudītu $(A1)$, nepieciešama rūpīgāka izpēte izteiksmei

$$\widehat{\theta}=n^{-1}\sum_{i=1}^n\widehat{f}(X_i)=n^{-2}\sum_{i,j}k_b(X_i-X_j)=\frac{k(0)}{nb}+\frac{n-1}{n}\widehat{g}.$$

Šeit $\widehat{g}=\widehat{g}(0)$, kur $\widehat{g}(y)=(\frac{n}{2})^{-1}\sum_{i<j}\bar{k}_b(Y_{i,j},y)$ ir dabiskais kodolu novērtējums blīvuma funkcijai $g(y)=\int f(y+x)f(x)\mathrm{d}x$ no starpības $Y_{i,j}=X_i-X_j$, un $\bar{k}_b(Y_{i,j},y)=\frac{1}{2}\{k_b(Y_{i,j}-y)+k_b(Y_{i,j}+y)\}$. Hjort [16] parādīja, ka $\widehat{g}(y)$ vidējā vērtība ir $g(y)+\frac{1}{2}b^2g''(y)\int u^2k(u)\mathrm{d}u+o(b^2)$ un dispersija $\frac{4}{n}\{g^*(y)-g^2(y)\}$ plus bezgalīgi mazas funkcijas ar zemāku kārtu, kur $g^*(y)=\frac{1}{4}\{\bar{g}(y,y)+\bar{g}(y,-y)+\bar{g}(-y,y)+\bar{g}(-y,-y)\}$ un $\bar{g}(y_1,y_2)$ ir kopējā blīvuma funkcija divām saistītām starpībām (X_2-X_1,X_3-X_1). No tā seko, ka izteiksmei

$$n^{-1/2}\sum_{i=1}^n m(X_i,\theta_0,\widehat{f})=\sqrt{n}(\widehat{\theta}-\theta_0)$$

ir vidējā vērtība ar kārtu $O(1/(\sqrt{n}b)+\sqrt{n}b^2)$ un dispersija, kas tiecas uz $4V$. Tas apstiprina $(A1)$ ar $U\sim N(0,4V)$, pie nosacījumiem $\sqrt{n}b\to\infty$ un $\sqrt{n}b^2\to0$.

Lai pārbaudītu $(A3)$, ievērosim, ka $\widehat{f}(x)\le b^{-1}k_{\max}$ katram x, kur $k_{\max}$ ir $k(u)$ maksimālā vērtība. Tātad $\max_{i\le n}|\hat{f}(X_i)-\theta_0|$ ir ierobežots ar $b^{-1}k_{\max}+\theta_0$, kas nozīmē, ka $(A3)$ ir spēkā, ja $\sqrt{n}b\to\infty$.

Visbeidzot, pierādot $(A0)$, mums jāparāda, ka

$$P\left\{\min_{1\le i\le n}m(X_i,\theta_0,\hat{f})<0<\max_{1\le i\le n}m(X_i,\theta_0,\hat{f})\right\}\to1.$$

Vispirms apskatīsim

$$\max_{1\le i\le n} m(X_i,\theta_0,\widehat{f}) \ge \max_{1\le i\le n} f_0(X_i) - \max_{1\le i\le n} |\widehat{f}(X_i) - f_0(X_i)| - \theta_0.$$

Ievērosim, ka $\max_{1\le i\le n} |\hat{f}(X_i) - f_0(X_i)| \to 0$ g.d. tā kā $\hat{f}$ vienmērīgi nepārtraukta ar piemērotu kodolu, piemēram, standarta normālo blīvuma funkciju, jo pieņēmām, ka f_0 ir vienmērīgi nepārtraukta (Teorēma A no [17]). Turklāt, $\max_{1\le i\le n} f_0(X_i) \to \sup_t f_0(t) > \theta_0$ g.d., tā kā f_0 nepārtraukta un nav vienmērīgā sadalījuma, tāpēc $P\{\max_{1\le i\le n} m(X_i,\theta_0,\widehat{f}) > 0\} \to 1$. Līdzīgā veidā varam apskatīt $\min_{1\le i\le n} m(X_i,\theta_0,\hat{f})$ un, tā kā pierādījām, ka $(A0)$ - $(A3)$ ir spēkā, tad varam secināt, ka

$$-2\ln \mathrm{EL}_n(\theta_0,\hat{f}) \to_d 4\chi_1^2.$$

Piemērs 6. Normāla ķermeņa temperatūra

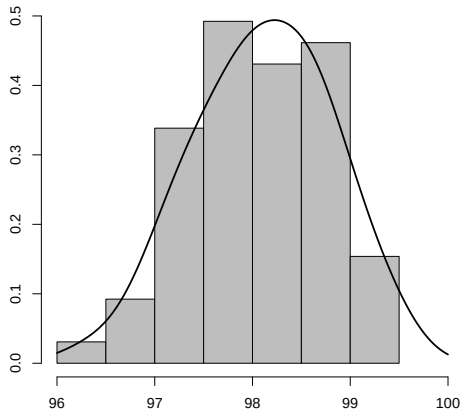
Pielietosim reāliem datiem iepriekš apskatīto EL metodi HL lokācijas novērtējuma dispersijai, ar mērķi konstruēt ticamības intervālu lokācijas parametram. Visus aprēķinus veiksim datorprogrammā R, kuras kods atrodams pielikumā.

1868. gadā Karls Vunderlihs veica pētījumu, kura rezultātā ieviesa normālas ķermeņa temperatūras amplitūdu (temperatūras mērījumi veikti mutē zem mēles). 1992. gadā tika veikts atkārtots pētījums, kura mērķis bija novērtēt līdz tam iegūtos rezultātus un noteikt medicīnā tik svarīgo vesela cilvēka ķermeņa temperatūras augšējo kritisko robežu. Dati atrodami Mackowiak, Wasserman, un Levine darbā [18], un tie satur mērījumus (Fārenheita grādos) par 130 pacientiem, no kuriem 65 bija sievietes un tieši tikpat daudz vīriešu. Šie autori ne tikai pierādīja savas aizdomas, ka augšējā kritiskā robeža un vidējā ķermeņa temperatūra ir zemāka, bet arī novēroja, ka sievietēm vidēji ir nedaudz augstāka normāla ķermeņa temperatūra nekā vīriešiem.

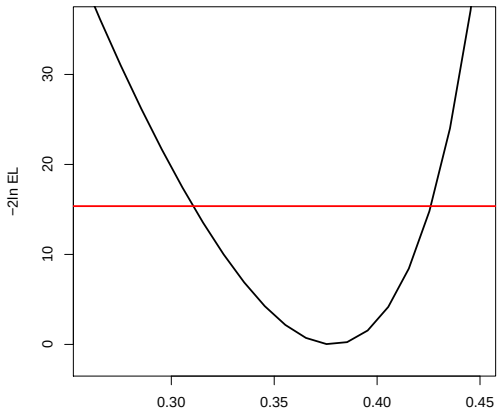
No dotā novērojuma mums ir zināms, ka $P(X_i \le x) = F(x)$ un $P(Y_i \le x) = F(x - \Delta)$, kur $X_i,\ldots,X_n$ satur datus par vīriešiem, bet $Y_1,\ldots,Y_m$ par sievietēm. Atrodam HL lokācijas novērtējumu un tas ir $\hat{\Delta} = -0.3$, kas liecina, ka sievietēm ķermeņa temperatūra ir vidēji par 0.3 Fāreinheita grādiem augstāka nekā vīriešiem.

Tālāk veiksim nepieciešamos aprēķinus, lai novērtētu σ^2. 3. attēlā redzama histogramma vīriešu ķermeņa temperatūras mērījumiem un kodolu blīvuma funkcijas novērtējums $\hat{f}$ traucējošajam parametram f. Ar EL metodi iegūstam $\hat{\theta} = 0.378$ un 95% ticamības intervālu $[0.311, 0.426]$ parametram θ_0. 4. attēlā redzama EL metodes statistika $-2\ln \mathrm{EL}(\theta,\hat{f})$, kas nošķelta ar $4\chi_1^2$ 0.95-to kvantili.

Ievērosim, ka mūsu piemērā $n = m$, tāpēc γ novērtēsim ar 0.5, pieņemot, ka apskatām tikai datus ar vienādiem apjomiem. Iegūstam $\hat{\sigma}^2 = 2.33$, un 95% ticamības intervālu $[-0.611, 0.011]$. Salīdzināšanai izmantosim t-testu divu izlašu vidējo vērtību starpībai. Rezultātā iegūstam $\bar{X} - \bar{Y} = -0.289$ un ticamības intervālu $[-0.54, -0.04]$, kas ir šaurāks nekā ar EL metodi iegūtais intervāls. t-tests strādā ļoti labi, tā kā dati nesatur izlēcējus, jo cilvēka ķermeņa temperatūra var svārstīties tikai noteiktā amplitūdā, tāpēc būtu interesanti salīdzināt rezultātus datiem ar lielu izlēcēju klātbūtni.



3. att. Histogramma un kodolu blīvuma funkcijas novērtējums $\hat{f}$ izlasei $X_1, \ldots, X_n$



4. att. EL statistika θ un 95% ticamības intervāls parametram θ_0

3.2. Neparametriskās regresijas atlikumu sadalījumi

Regresija ir viena no statistikas centrālajām problēmām. Joprojām ļoti populārs un bieži sastopams ir lineārais regresijas modelis, kurā tiek pieņemts, ka regresijas atlikumi ir normāli sadalīti. Tomēr ir gadījumi, kad statistiķiem nākas saskarties ar eksperimentu rezultātiem, kuros normalitātes pieņēmums atlikumiem var nebūt spēkā. Daudzi autori uzsver, ka pirms lineārā modeļa pielāgošanas, lai cik pievilcīgs tas arī nešķistu, būtu svarīgi veikt hipotēžu pārbaudi par atlikumu sadalījumu. Turklāt, lietojot neparametriskās regresijas modeli, papildus zināšanas par atlikumu sadalījumu var tikai uzlabot statistiskās analīzes efektivitāti, piemēram, zinot, ka neparametriskās regresijas modelis ar normāli sadalītiem atlikumiem ir asimptotiski ekvivalents Gausa baltā trokšņa modelim (Brown un Low, [19]), var veikt dažādus papildus testus.

Atlikumu sadalījuma novērtēšanai tiek piedāvātas tādas metodes kā empīriskā sadalījuma funkcija (Akritas un Van Keilegom, [20]), butstraps (Neumeyer, Dette un Nagel, [21]), empīriskie procesi (Khmaladze un Koul, [22]) u.c., savukārt Hjort, McKeague un Van Keilegom [1] piedāvā novērtēt atlikumu sadalījuma funkciju ar empīriskās ticamības funkcijas metodi.

Pieņemsim, ka doti datu pāri $(X_1, Y_1),\ldots,(X_n, Y_n)$ *iid*. Aplūkosim neparametriskās regresijas modeli

$$Y = \mu(X) + \varepsilon,$$

kur X un ε ir neatkarīgi, ε sadalījuma funkcija F_ε nav zināma, un $\mu(\cdot)$ ir nezināma regresijas funkcija. Kā jau iepriekš tika minēts, interesējošais parametrs ir $\theta_0 = F_\varepsilon(z) \in (0,1)$ kādā fiksētā punktā z. Ieviesīsim tādus pašus pieņēmumumus kā Akritas un Van Keilegom publikācijā [20] - F_ε ir nepārtraukta, $\mu(\cdot)$ ir gluda funkcija, un X ir ierobežota. Vienkāršībai ierobežosim X intervālā $(0,1)$.

Par novērtējošo funkciju izvēlamies

$$m(X, Y, \theta, \mu) = I_{\{Y-\mu(X)\le z\}} - \theta,$$

kur regresijas funkcija μ ir traucējošais parametrs un tā novērtēšanai izmantosim populāro Nadaraya-Watson novērtējumu $\widehat{\mu}(x) = \sum_{i=1}^n W_{n,i}(x; b_n)Y_i$, kur svari $W_{n,i}(x; b_n) = k_{b,x}(X_i)/\sum_{j=1}^n k_{b,x}(X_j)$, un $k_{b,x}(u) = b^{-1}k((u-x)/b)$.

Nosacījumu pārbaude

Tāpat kā iepriekš pārbaudīsim, vai funkcija $m(X, Y, \theta, \mu) = I_{\{Y-\mu(X)\le z\}} - \theta$ ir nenovirzīta:

$$E(I_{\{Y-\mu(X)\le z\}} - \theta_0) = E(I_{\{Y-\mu(X)\le z\}}) - E\theta_0 = E(I_{\{\varepsilon\le z\}}) - \theta_0 = F_\varepsilon(z) - \theta_0 = 0.$$

Tālāk pārbaudīsim nosacījumus $(A0)$ - $(A3)$, kas nepieciešami, lai varētu pielietot Teorēmu 2 ticamības intervālu konstruēšanai. Pirmkārt $(A1)$ seko no $\widehat{\theta} = n^{-1}\sum_{i=1}^n I_{\{\hat{\varepsilon}_i\le z\}}$, kur $\hat{\varepsilon}_i = Y_i - \widehat{\mu}(X_i)$, asimptotiskās normalitātes, kas aprakstīta [20]: $\sqrt{n}\{\widehat{F}_\varepsilon(z) - F_\varepsilon(z)\} = n^{-1/2}\sum_{i=1}^n m(X_i, Y_i, \theta_0, \widehat{\mu}) \to_d N(0, V_1)$, kur

$$V_1 = E(I_{\{\varepsilon\le z\}} - F_\varepsilon(z) + \varphi(X, Y, z))^2,$$

kur

$$\varphi(x, y, z) = -f_\varepsilon(z)\int (I_{\{y\le v\}} - F(v|x))J(F(v|x))(1 + zv - z\mu(x))dv,$$

kur $J(s)$ zināma skores *score* funkcija, kas apmierina $\int_0^1 J(s)ds = 1$, un $F(\cdot|x)$ ir nosacītā sadalījuma funkcija.

Nosacījums $(A2)$ ir spēkā ar $V_2 = \theta_0(1-\theta_0)$, ja $0 < \theta_0 < 1$. Arī $(A3)$ izpildās, tā kā funkcija $\sqrt{n}m_n$ ir vienmērīgi ierobežota ar 1. Visbeidzot, $(A0)$ ir tūlītējas sekas no fakta, ka $P\{Y - \widehat{\mu}(X) \le z\}$ konverģē uz $F_\varepsilon(z)$, kas seko no Teilora izvirzījuma un $\widehat{\mu}$ vienmērīgās nepārtrauktības. Tā kā $F_\varepsilon(z)$ ir strikti ierobežots $(0, 1)$, tad

$$P\{\exists \quad 1 \le i, j \le n : Y_i - \widehat{\mu}(X_i) \le z \text{ un } Y_j - \widehat{\mu}(X_j) > z\} \to 1,$$

kas apstiprina $(A0)$.

Tagad atliek novērtēt V_1 un V_2. Ievērosim, ka $\hat{V}_2 = \hat{\theta}(1-\hat{\theta})$ ir V_2 būtisks novērtējums, taču acīmredzami, ka novērtēt V_1 ir daudz sarežģītāk. Šī iemesla dēļ tiks izmantots butstrapots sadalījums, lai aproksimētu U.

Visbeidzot $100(1-\alpha)\%$ ticamības intervāls parametram $\theta_0 = F_\varepsilon(z)$ ir formā $\{\theta : -2\ln \mathrm{EL}_n(\theta, \mu) \ge e_{1-\alpha}\}$, kur $e_{1-\alpha}$ ir $100(1-\alpha)$ kvantile sadalījumam

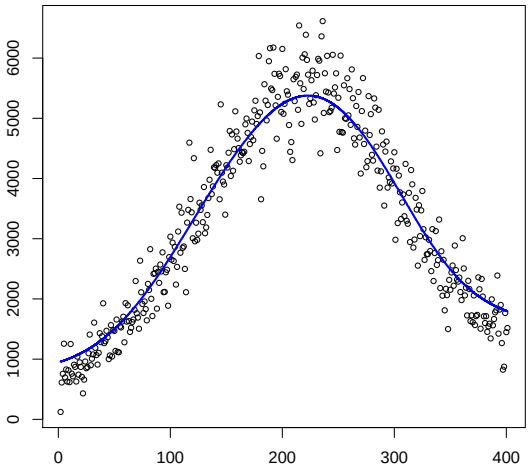
$$\frac{n\left(n^{-1}\sum_{i=1}^n I_{\{Y_i^* - \widehat{\mu}(X_i^*)\le z\}} - \widehat{\theta}\right)^2}{\widehat{\theta}(1-\widehat{\theta})}, \tag{3.1}$$

kur Y_i^* un X_i^* butstrapotās izlases elementi (skatīt [1]).

Piemērs 7. Neparametriskā regresija kosmiskās radiācijas datiem

Par visuma sākumu parasti tiek uzskatīts "Lielais sprādziens", kaut gan ir maldinoši domāt, ka tas notika tukšā vietā. Pirmatnējais visums patiesībā bija karstā, blīvā stāvoklī, un tikai kopš lielā sprādziena tas sāka atdzist un izplesties. Taču lielā sprādziena paliekas joprojām ir pamanāmas un tiek sauktas par kosmisko mikroviļņu izcelsmes radiāciju. Dati satur informāciju par visumu 13 bilionus gadu senā pagātnē (379 000 gadus pēc Lielā sprādziena) un attēlo temperatūras svārstību intensitāti dažādās frekvencēs, kas nodrošina kosmosa pētniekus ar svarīgu informāciju par pirmatnējo visumu.

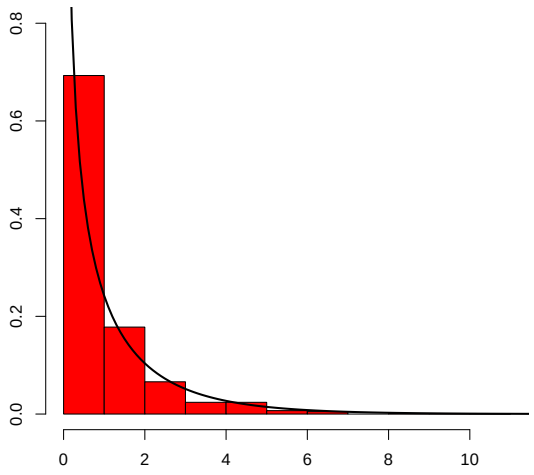
Dati *cmb* atrodami [23], izmantosim tikai pirmos 400 novērojumus, lai novērtētu regresijas funkciju un atlikumu sadalījumu, kā arī konstruētu ticamības intervālu atlikumu sadalījuma funkcijas vērtībai punktā $z = 0$. 5. attēlā uz x ass attēlotas frekvences, bet uz y ass attēlotas temperatūras svārstību intensitātes, kā arī Nadaraya-Watson neparametriskās regresijas funkcijas novērtējums.



5. att. Nadaraya-Watson novērtējums regresijas funkcijai μ datiem *cmb*

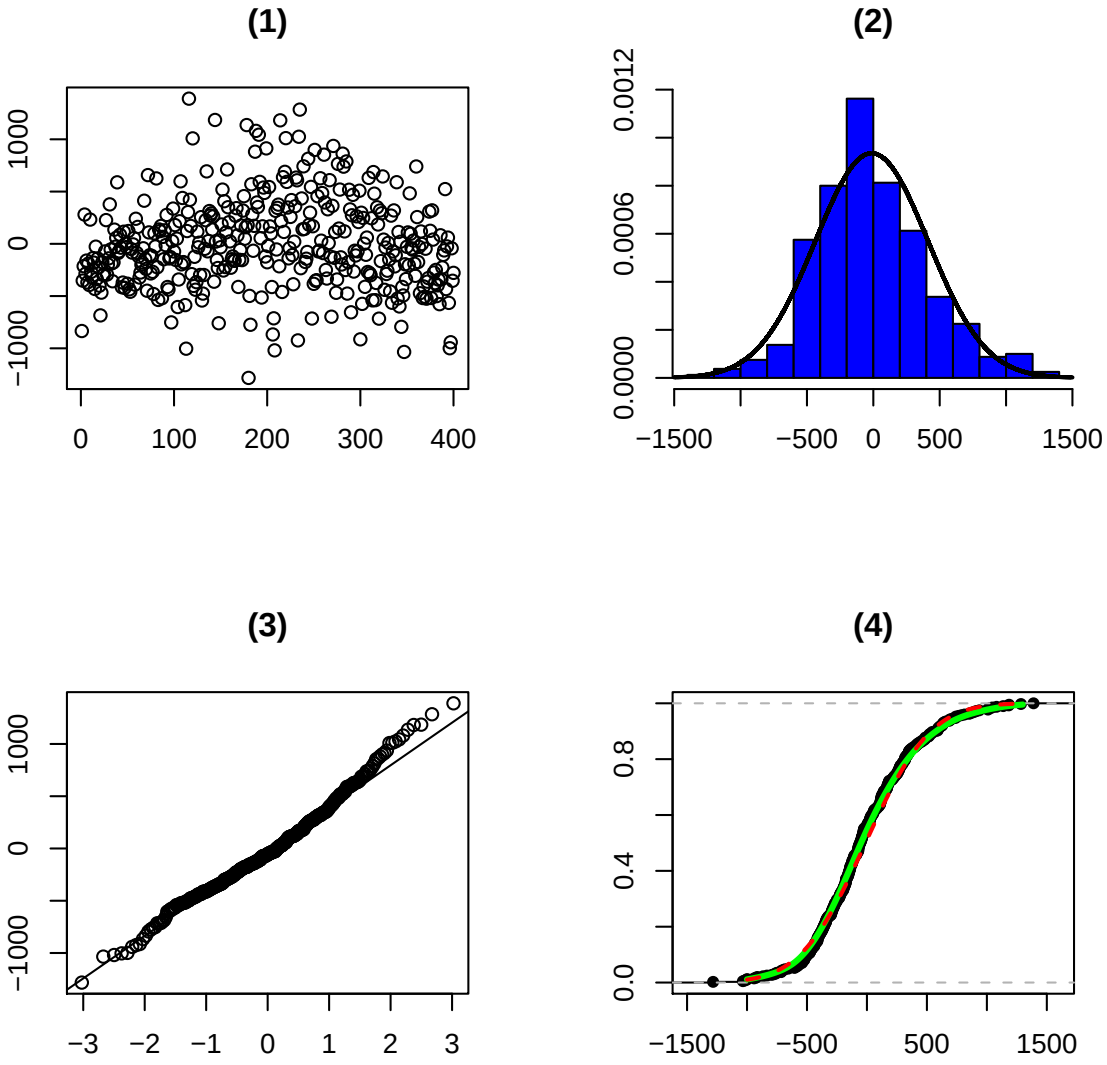
6. attēlā redzama histogramma 1000 reižu butstrapotai statistikai, kas definēta (3.1), $z = 0$, un, kā redzams, χ^2_1 blīvuma funkcija labi aproksimē histogrammu butstrapotajai izlasei. Punktā $z = 0$ interesējošā parametra $\theta_0 = F_\varepsilon(0)$ novērtējums ir $\hat{\theta} = 0.546$ un 95% ticamības intervāls $[0.504, 0.588]$, kvantile butstrapotajam sadalījumam $e_{95} = 3.287$. Iegūtos rezultātus varam salīdzināt ar butstrapa 95% ticamības intervālu $[0.534, 0.586]$, kas ir nedaudz novirzīts pa labi, taču mazliet šaurāks nekā ar EL metodi iegūtais. Tālāk

konstruēsim sadalījuma funkcijas novērtējumu ar EL metodi visos punktos, lai varētu izdarīt secinājumus par sadalījuma likumu.



6. att. Bootstrapotais sadalījums statistikai (3.1) datiem *cmb*, χ^2_1 blīvuma funkcija

7. attēlā grafikā (1) redzami neparametriskās regresijas atlikumi $\hat{\varepsilon}_i$, savukārt grafikā (2) attēlota atlikumu histogramma ar pievienotu normālā sadalījuma ar novērtētu vidējo vērtību $\hat{\mu} = -12.072$ un dispersiju $\hat{\sigma}^2 = 181992.3$ blīvuma funkciju, savukārt grafikā (3) redzams kvantiļu-kvantiļu grafiks atlikumiem pret teorētiskajām normālā sadalījuma ar novērtētiem parametriem $\hat{\mu}$ un $\hat{\sigma}^2$ kvantilēm. Grafikā (4) attēlots atlikumu sadalījuma funkcijas novērtējums ar EL metodi zaļā krāsā un pievienota normālā sadalījuma funkcija ar vidējo vērtību $\hat{\mu}$ un dispersiju $\hat{\sigma}^2$ sarkanā krāsā, kā arī atlikumu empīriskā sadalījuma funkcija melnā krāsā. No 7. attēla grafika (4) atlikumiem varētu būt $N(\hat{\mu}, \hat{\sigma}^2)$ sadalījums, lai arī histogrammā redzamas novirzes no normālā sadalījuma blīvuma funkcijas, tāpēc pielietosim programmā R iebūvētās komandas `lillie.test` un `shapiro.test`, lai pārbaudītu saliktu hipotēzi par neparametriskās regesijas atlikumu normalitāti. Lilliefora teta statistikas vērtība ir $D = 0.062$ un p-vērtība ir 0.000856, kā arī Šapiro-Vilksa testa statistikas vērtība $W = 0.9869$ un p-vērtība ir 0.001146. Tātad skaidrs, ka mums ir jānoraida nulles hipotēze, ka regresijas atlikumiem varētu būt normālais sadalījums $N(\hat{\mu}, \hat{\sigma}^2)$.



7. att.: (1) Neparametriskās regresijas atlikumi datiem *cmb*, (2) Histogramma atlikumiem un $N(\hat{\mu}, \hat{\sigma}^2)$ blīvuma funkcija, (3) Q-Q grafiks atlikumiem pret $N(\hat{\mu}, \hat{\sigma}^2)$ teorētiskajām kvantilēm, (4) Atlikumu sadalījuma funkcijas novērtējums ar EL metodi (zaļā krāsa), atlikumu empīriskā sadalījuma funkcija (melnā krāsā) un $N(\hat{\mu}, \hat{\sigma}^2)$ sadalījuma funkcija (sarkanā krāsā); $\hat{\mu}$ ir novērtēta atlikumu vidējā vērtība, $\hat{\sigma}^2$ ir novērtēta atlikumu dispersija

4. EL metodes ar novērtētiem parametriem pielietojumi divu izlašu gadījumā

Kā jau tika minēts empīriskās ticamības funkcijas metodei ir vairāki zināmi pielietojumi, kā piemēram kvantiļu funkciju starpībām, P-P un Q-Q grafikiem, ROC līknēm, savukārt pagaidām ir ļoti maz pielietojumu EL metodei ar novērtētiem parametriem, turklāt, ja traucējošais parametrs h ir ar galīgu dimensiju un novērtējošās funkcijas ir diferencējamas, tad ticamības intervālus var konstruēt, balstoties uz pamata empīriskās ticamības funkcijas teoriju.

Viens no zināmajiem EL metodes ar novērtētiem parametriem pielietojumiem ir strukturālo attiecību modeļiem, kad jānovērtē galīgdimensionāls traucējošais parametrs h, šo problemātiku apskatīsim 4.1. nodaļā un pielietosim arī trīs dažādiem datu piemēriem. 4.2. nodaļā apskatīsim divus citus piemērus. Pirmajam no tiem praktisks pielietojums nav zināms, taču, vadoties pēc līdzīgas shēmas, iespējams konstruēt divu izlašu problēmu no jebkuras zināmas vienas izlases problēmas. savukārt otrs piemērs ļoti jauns empīriskās ticamības funkcijas ar novērtētiem parametriem pielietojums gludo Hūbera novērtējumu starpībai, kura implementācija ir ļoti sarežģīta, tāpēc joprojām nav veikta.

4.1. Strukturālo attiecību modeļi

Strukturālo attiecību modeļi sevī iekļauj labāk zināmos lokācijas-skalēšanas modeļus (Hettmansperger, [24]) un Lēmaņa alternatīvu modeļus (Lehmann, [25]). Taču vispārējā formā strukturālo attiecību modeļus pirmo reizi 2005.gadā savā publikācijā aprakstīja vācu matemātiķi Gudrun Freitag un Axel Munk [26].

Definīcija 5. Klasisks lokācijas skalēšanas modelis divām neatkarīgām izlasēm $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ ar sadalījuma funkcijām attiecīgi F_1 un F_2 tiek definēts

$$F_1(t) = F_2\left(\frac{t-\mu}{\sigma}\right) =: F_2(t,h), \quad t \in \mathbb{R} \tag{4.1}$$

kādam parametram $h = (\mu, \sigma)$ un $\sigma > 0$.

Šo pašu attiecību var izteikt arī ar kvantiļu funkcijām

$$F_1^{-1}(u) = F_2^{-1}(u)\sigma + \mu, \quad u \in [0,1]. \tag{4.2}$$

Ja $\sigma \equiv 1$, tad starp izlasēm pastāv *šifta* jeb lokācijas modelis, savukārt pastāv tikai skalēšanas modelis, ja $\mu \equiv 0$.

Definīcija 6. Divu neatkarīgu izlašu $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ sadalījuma funkcijas F_1 un F_2 pieder Lēmaņa alternatīvu modelim, ja

$$F_1(t) = 1 - (1 - F_2(t))^{(1/h)} =: F_2(t, h), \quad t \in \mathbb{R},$$

kur $h > 0$.

Atkal, izmantojot kvantiļu funkcijas, varam šo attiecību definēt sekojoši

$$F_1^{-1}(u) = F_2^{-1}(1 - (1 - u)^h), \quad u \in [0, 1]. \tag{4.3}$$

Ja starp dotām izlasēm pastāv Lēmaņa alternatīvu modelis, tad šīs izlases atbilst proporcionālā riska nosacījumam, kas ir nozīmīgs un bieži pielietots medicīniskajā statistikā un izdzīvošanas datu analīzē ([27]).

Divu izlašu modeļi, tādi kā (4.2) un (4.3), kas ir izsakāmi ar kvantiļu funkcijām, var tikt pierakstīti kā strukturālo attiecību modeļi vispārējā formā. Pieņemsim, ka divu *iid* gadījuma lielumu X un Y realizāciju sadalījuma funkcijas F_1 un F_2 pieder kopai

$$\mathcal{F}^2 := \{F : F \text{ ir sadalījuma funkcija un } \int t^2 dF(t) < \infty\}.$$

Definīcija 7. [26] Pieņemsim, ka $\mathcal{H} \subseteq \mathbb{R}^l$ un $\phi_1 : \mathbb{R} \times \mathcal{H} \to \mathbb{R}$, $\phi_2 : [0, 1] \times \mathcal{H} \to [0, 1]$. Funkcijas F_1 un F_2 ir saistītas ar strukturālu attiecību, kuru veido ϕ_1 un ϕ_2, ja $(F_1, F_2) \in \mathcal{U}_{\phi_1,\phi_2} =: \mathcal{U}$, kur modeļu klase $\mathcal{U}$ tiek definēta kā

$$\mathcal{U} := \{(F_1, F_2) \in \mathcal{F}^2 \times \mathcal{F}^2 \mid \exists h \in \mathcal{H} : F_1(\phi_1(F_2^{-1}(\phi_2(u, h))), h), \quad u \in [0, 1]\}.$$

Strukturālo attiecību var izteikt arī, kvantiļu funkciju vietā lietojot sadalījuma funkcijas,

$$F_1(t) = \phi_2^-(F_2(\phi_1^-(t, h)), h) =: F_2(t, h),$$

kur ϕ_1^- un ϕ_2^- attiecīgi ϕ_1 un ϕ_2 inversās funkcijas.

Piezīme 8. *No strukturālo attiecību modeļu definīcijas vispārējā formā viegli iegūt iepriekš definēto lokācijas-skalēšans modeli, izvēloties* $\phi_1(t, (\mu, \sigma)^T) = \mu + \sigma t$ *un* $\phi_2 \equiv u$. *Savukārt, Lēmaņa alternatīvu modeli, var iegūt, izvēloties* $\phi_1(t, h) \equiv t$ *un* $\phi_2(u, h) = 1 - (1 - u)^h$.

Lai novērtētu parametru h, var tikt izmantots Mallova attālums, kas sīkāk apskatīts [28]. Vispārīgā gadījumā strukturālo attiecību modelim teorētiskais parametrs h_0 tiek aprēķināts kā

$$h_0 = \text{argmin}_{h \in \mathcal{H}} \left\{ \frac{1}{b - a} \int_a^b F_1^{-1}(u) - \phi_1(F_2^{-1}(\phi_2(u, h)))^2 du \right\},$$

parametra h novērtējumu $\hat{h}$ iegūst, aizvietojot sadalījuma funkcijas ar empīriskajām sadalījuma funkcijām.

Ticamības intervāli ar EL metodi

Konstruēsim ticamības intervālus interesējošajam parametram

$$\Delta := \Delta(t) = F_1(\phi_1(F_2^{-1}(\phi_2(t,h)),h)). \tag{4.4}$$

(4.4) var tikt uzskatīts arī par P-P grafika vispārinājumu. Šajā gadījumā novērtējošās funkcijas ir formā

$$m_1(X,\theta_0,\Delta,t,\hat{h}) = I_{\{X\le\theta_0\}} - \Delta,$$

$$m_2(Y,\theta_0,\Delta,t,\hat{h}) = I_{\{Y\le\phi_1^-(\theta_0,\hat{h})\}} - \phi_2(t,\hat{h}),$$

kur $\theta_0 = \phi_1(F_2^{-1}(\phi_2(t,\hat{h})),\hat{h})$ un galīgdimensionālā traucējošā parametra h novērtējums $\hat{h}$ tiek atrasts ar Mallova attāluma palīdzību.

Valeinis (2007,[2]) pierādīja, ka pieņēmumi $(B0)$ - $(B4)$ ir spēkā strukturālo attiecību modeļiem ar

$$V_1 = V_2 = \Delta(1-\Delta)$$

un

$$T_1 = T_2 = \phi_2(t,h_0)(1-\phi_2(t,h_0)),$$

tādējādi $-2\ln \mathrm{EL}(\Delta,\hat{\theta},t,\hat{h}) \to_d \chi_1^2$.

Piezīme 9. *Lokācijas skalēšanas modeļiem novērtējošās funkcijas ir*

$$m_1(X,\theta,\Delta,t,\hat{h}) = I_{\{X\le\theta\}} - \Delta,$$

$$m_2(Y,\theta,\Delta,t,\hat{h}) = I_{\{Y\le(t-\hat{\mu})/\hat{\sigma}\}} - t,$$

kur $h = (\mu,\sigma)^T$, $\hat{\mu}$ lokācijas parametra novērtējums, savukārt $\hat{\sigma}$ skalēšanas parametra novērtējums.

Turklāt, lai pielietotu EL metodi praktiski, novērtējošās funkcijas nepieciešams nogludināt

$$\tilde{m}_1(X,\theta,\Delta,t,\hat{h}) = H_{b_1}(\theta - X) - \Delta,$$

$$\tilde{m}_2(Y,\theta,\Delta,t,\hat{h}) = H_{b_2}(\theta - \hat{h} - Y) - t,$$

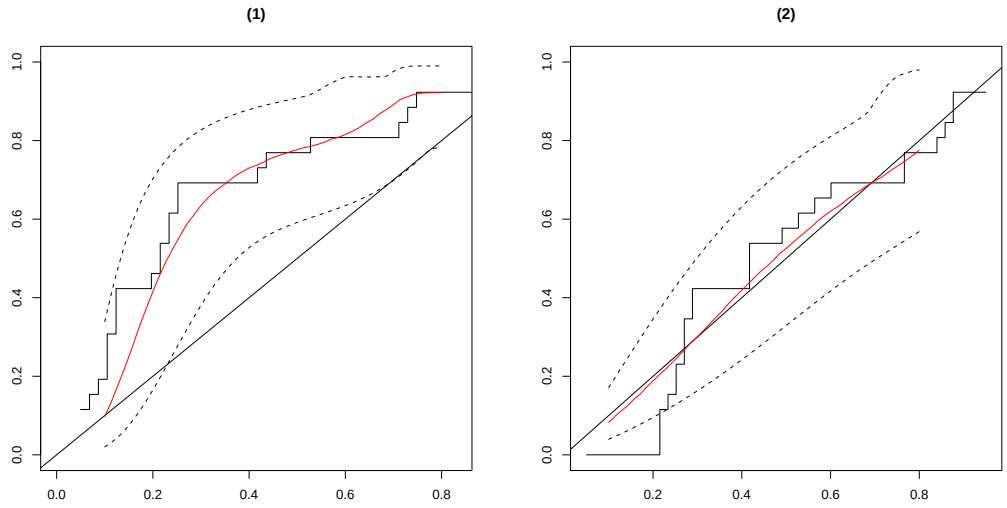
kur $H_b(t) = \int_{u\le t} K(u)du$ un K ir kāda kodola funkcija ar joslas platumu b [29].

Datu piemēri

Šeit apskatīsim trīs reālu datu piemērus, kas atbilst lokācijas-skalēšanas modeļiem. Pārbaudīsim gan vienkāršas hipotēzes par datu sadalījuma likuma vienādību, t.i., kad $\mu = 0$ un $\sigma = 1$, gan saliktas hipotēzes ar novērtētiem lokācijas un skalēšanas parametriem. Konstruēsim 95% vienlaicīgās ticamības joslas ar empīriskās ticamības funkcijas metodi divu izlašu problēmām, izmantojot programmas R kodu, kuru izstrādāja un implementēja J. Valeinis un E. Cers [3]. H_0 tiks noraidīta, ja ticamības joslas kaut vienā punktā neiekļaus taisni $y = x$.

Piemērs 8. Sudraba nitrāta ietekme uz nokrišņu daudzumu

1975. gadā tika veikts eksperiments, kura laikā 26 nejauši izvēlētus mākoņus apsēja ar sudraba nitrātu. Dati *clouds* satur nokrišņu daudzuma mērījumus no apsētajiem un 26 neapsētiem mākoņiem, un ir atrodami [23]. Pārbaudīsim, vai starp izlasēm pastāv lokācijas-skalēšanas modelis, konstruējot P-P grafiku un ticamības joslas ar EL metodi, lai secinātu, vai mākoņu apsēšana ar sudraba nitrātu var ietekmēt nokrišņu daudzuma palielināšanos.



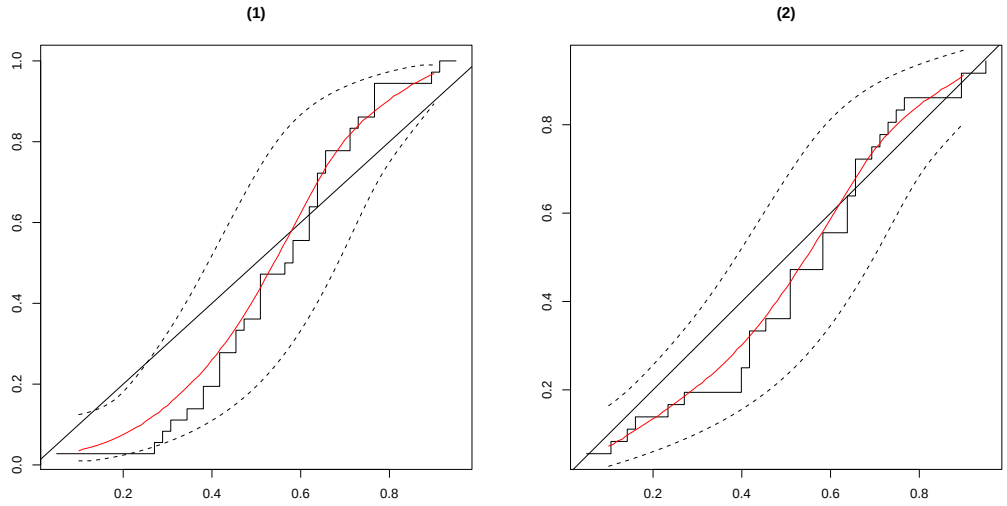
8. att. (1) $H_0 : \mu = 0, \sigma = 1$, (2) $H_0 : \mu = -18.09, \sigma = 0.413$

8. attēlā redzams P-P grafiks abām izlasēm ar 95% EL ticamības joslām un P-P grafiks neapsētajiem mākoņiem pret modificētu apsēto mākoņu izlasi formā $\hat{\sigma}Y_i + \hat{\mu}$, kur Y_i ir apsēto mākoņu izlases elementi. Kā redzams 8. attēlā pirmajā grafikā, noraidām hipotēzi, ka abām izlasēm sadalījuma likumi ir vienādi. Novērtēsim lokācijas parametru $\hat{\mu} = -18.09$

un skalēšanas parametru $\hat{\sigma} = 0.413$. 8. attēlā otrajā grafikā, redzams, ka nevar noraidīt hipotēzi, ka izlases pieder lokācijas-skalēšanas modelim, respektīvi, $F_1(x) = F_2(\frac{x+18.09}{0.413})$, kur F_1 ir neapsēto mākoņu sadalījuma funkcija, bet F_2 apsēto mākoņu sadalījuma funkcija. Varam secināt, ka mākoņu apsēšana ar sudraba nitrātu palielina nokrišņu daudzumu no katra apsētā mākoņa.

Piemērs 9. Peļu leikēmijas dati

1979. gadā veikta pētījuma laikā tika iegūti dati, kas satur dzīves ilgumus pelēm, sirgstošām ar aizkrūts dziedzera leikēmiju. Dati atrodami [30]. Zhang un Li (1996,[31]) savā publikācijā raksta, ka, pielietojot Vilkoksona testu šiem pašiem datiem, neizdodas uzrādīt atšķirības starp izdzīvošanas sadalījuma funkcijām 40 sieviešu un 36 vīriešu kārtas pelēm, kas sirgst ar aizkrūts dziedzera leikēmiju. Autori to skaidro ar testa zemo jūtīgumu atklāt atšķirības sadalījuma funkcijās, kas vienlaicīgi atbilst gan lokācijas, gan skalēšanas modelim, tāpēc, līdzīgi kā iepriekšējā piemērā, konstruēsim P-P grafiku izlasēm un ticamības joslas ar EL metodi.



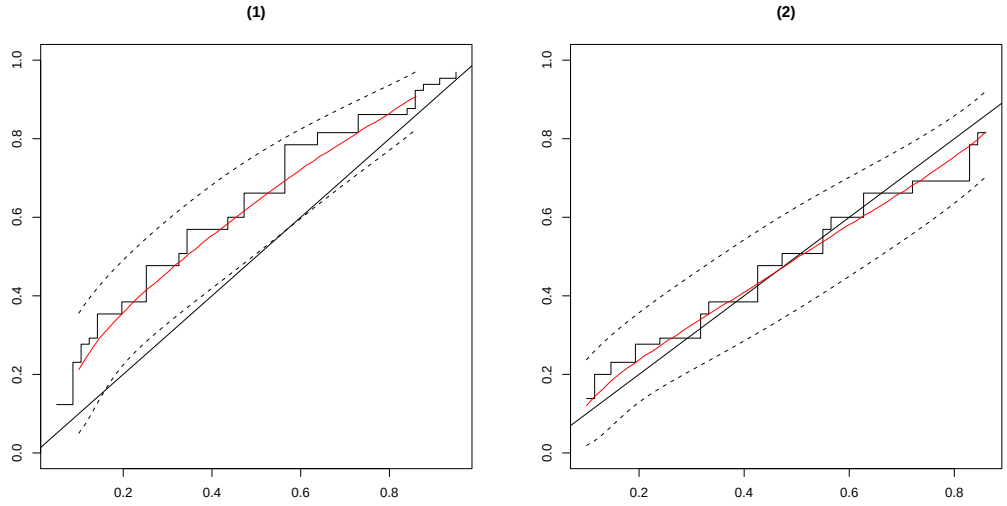
9. att. (1) $H_0 : \mu = 0, \sigma = 1$, (2) $H_0 : \mu = 101.192, \sigma = 0.773$

Līdzīgi kā iepriekšējā piemērā pārbaudīsim, vai sieviešu kārtas un vīriešu kārtas pelēm ar leikēmiju ir atšķirīgas izdzīvošanas sadalījuma funkcijas. 9. attēlā pirmajā grafikā redzams,ka tomēr jānoraida hipotēzi par abu dzimumu peļu sadalījuma likumu vienādību. Lokācijas parametra novērtējums ir $\hat{\mu} = 101.192$ un skalēšanas parametra novērtējums ir $\hat{\sigma} = 0.773$. 9. attēlā otrajā grafikā, redzams, ka nenoraidām hipotēzi, ka izlases pieder lokācijas-skalēšanas modelim - $F_1(x) = F_2(\frac{x+101.192}{0.773})$, kur F_1 ir vīriešu kārtas peļu izdzīvo-

šanas sadalījuma funkcija, bet F_2 sieviešu kārtas peļu izdzīvošanas sadalījuma funkcija. Varam secināt, ka izdzīvošanas sadalījuma funkcijas abu dzimumu pelēm atšķiras.

Piemērs 10. Normāla ķermeņa temperatūra

Atgriezīsimies pie 3.1. nodaļā apskatītā piemēra par veselu cilvēku normālas ķermeņa temperatūras mērījumiem. Šajā piemērā pārbaudīsim, vai izlases pieder lokācijas modelim, pieņemot, ka skalēšanas parametrs $\sigma = 1$.



10. att. (1) $H_0 : \mu = 0, \sigma = 1$, (2) $H_0 : \mu = -0.289, \sigma = 1$

10. attēlā redzams, ka atkal noraidām hipotēzi par sadalījumu likumu vienādību, bet hipotēzi, ka pastāv lokācijas modelis ar $\hat{\mu} = -0.289$ noraidīt nevaram. Tas apstiprina jau iepriekš iegūtos rezultātus ar Hodžes-Lēmana novērtējumu, ka sievietēm ķermeņa temperatūra ir vidēji par aptuveni 0.3 Fārenheita grādiem augstāka nekā vīriešiem.

4.2. Citi piemēri

Divu blīvuma funkciju integrāļu starpība

Šajā nodaļā konstruēsim piemēru divu izlašu problēmai, kuram varētu pielietot empīriskās ticamības funkcijas metodi ar novērtētiem traucējošiem parametriem. Balstīsimies uz vienas izlases problēmu, kuru apskatījām 3.1. nodaļā, t.i., par Hodžes-Lēmana asimptotiskās dispersijas novērtējumu. Tā kā šim piemēram nav praktiska pielietojuma, tas ir samērā nenoderīgs, bet, sekojot šeit apskatītajai konstrukcijai, līdzīgā veidā ir iespējams

konstruēt jebkuru citu divu izlašu problēmu, kurai pamatā ir kāda vienas izlases problēma EL metodei ar novērtētiem parametriem.

Pieņemsim, ka mums ir divas izlases $X_1,\ldots,X_n \sim F_1$ un $Y_1,\ldots,Y_m \sim F_2$. Interesēsimies par parametru Δ, kas ir divu izlašu blīvuma funkciju kvadrātu integrāļu starpība

$$\Delta = \int f_2^2(y)dy - \int f_1^2(x)dx,$$

un iegūstam novērtējošās funkcijas formā

$$m_1(X,\theta_0,\Delta) = f_1(X) - \theta_0 \text{ un } m_2(Y,\theta,\Delta) = f_2(Y) - \theta_0 - \Delta,$$

kur $\theta_0 = \int f_1^2(x)dx$ un traucējošie parametri ir f_1 un f_2, kuru novērtēšanai izmantosim neparametrisko kodolu blīvuma novērtējumu tāpat kā 3.1.

$$\widehat{f_1}(x) = n^{-1}\sum_{i=1}^{n} k_b(X_i - x) \text{ un } \widehat{f_2}(x) = m^{-1}\sum_{j=1}^{m} k_b(Y_j - x),$$

kur $k_b(\cdot) = k(\cdot/b)/b$ un b ir joslas platums.

Gludu Hūbera novērtējumu starpība

Huber [32] 1964. gadā ieviesa Hūbera lokācijas parametra novērtējumu, kas pieder robusto M-novērtējumu klasei. Kā zināms, robustās statistikas metodes ir labi piemērotas datiem, kas satur lielu skaitu izlēcēju, un tradicionālās metodes šādos gadījumos dod ļoti sliktus rezultātus. Tā piemēram, izlases vidējā vērtība var tikt nozīmīgi ietekmēta pat ar vienu izlēcēju, savukārt mediāna, kas ir robusts novērtējums, netiek būtiski ietekmēta pat ar vairāku izlēcēju klātbūtni.

Lai apskatītu Hūbera novērtējumu, vispirms definēsim M-novērtējumu. Pieņemsim, ka $X_1,\ldots,X_n$ *iid* ar sadalījuma funkciju F.

Definīcija 8. M-novērtējums ir statistiskais funkcionālis T_n, kas noteiktai funkcijai ρ minimizē

$$\sum_{i=1}^{n} \rho(X_i,t) = \int \rho(x,t)\mathrm{d}F_n,$$

kur F_n ir empīriskā sadalījuma funkcija.

Ja ρ ir diferencējama pēc t, tad $\sum_{i=1}^{n}\rho(X_i,t)$ minimizē vienādojuma $\sum_{i=1}^{n}\psi(X_i,t) = 0$ sakne, kur $\psi(x,t) = \partial\rho(x,t)/\partial t$. Savukārt Huber [32] piedāvāja ψ vietā lietot

$$\psi(x,t) = \psi_0(x-t),$$

kur

$$\psi_0(z) = \begin{cases} c, & z \geq c \\ z, & |z| < c \\ -c, & z \leq -c. \end{cases}$$

Ievērosim, ka, ja $c \to \infty$, tad ψ_0 iegūstam kā vidējo vērtību, savukārt, ja $c \to 0$, iegūstam mediānu. Hampel, Henning un Roncheti [33] 2011. gadā definēja gludināšanas principu M-novērtējumiem un parādīja, ka gludināšana uzlabo rezultātus.

Tālāk definēsim empīriskās ticamības funkcijas metodi gludu Hūbera novērtējumu $\tilde{\psi}$ (skatīt [33]) starpībai, lai konstruētu ticamības intervālus. Pieņemsim, ka ir dotas divas *iid* izlases $X_1,\ldots,X_n$ un $Y_1,\ldots,Y_m$ ar sadalījuma funkcijām attiecīgi F_1 un F_2. Interesējošais parametrs ir $\Delta = \theta_1 - \theta_0$. Tad novērtējošās funkcijas ir formā

$$m_1(X_i, \theta_0, \Delta) = \tilde{\psi}\left(\frac{X_i - \theta_0}{\hat{\sigma}_1}\right)$$

$$m_2(Y_i, \theta_0, \Delta) = \tilde{\psi}\left(\frac{Y_i - \Delta + \theta_0}{\hat{\sigma}_2}\right),$$

kur σ_1 un σ_2 ir traucējošie parametri, kurus nepieciešams novērtēt.

Kā jau tika minēts, šis pielietojums ir pavisam jauns, un pie tā joprojām strādā Valeinis, Vēliņa un Luta, bet līdz ar tā implementāciju tiks pavērtas plašākas empīriskās ticamības funkcijas metodes ar novērtētiem parametriem pielietošanas iespējas.

Secinājumi

Darbā apskatītā empīriskās ticamības funkcijas metode ir ļoti vienkārša pēc būtības, taču konkurētspējīga ar parametriskajām metodēm, it īpaši gadījumos, kad ir dotas samērā nelielas izlases vai arī to teorētiskais sadalījums ir grūti nosakāms. EL metode ar novērtētiem parametriem paplašina empīriskās ticamības funkcijas metodes pielietošanas iespējas gan vienas, gan divu izlašu problēmām.

Vienas izlases gadījumā apskatīti divi piemēri, no kuriem pirmais attiecās uz Hodžes-Lēmana lokācijas novērtējuma asimptotiskās dispersijas novērtēšanu, lai konstruētu ticamības intervālu. Ticamības intervāls tika salīdzināts ar t-testa rezultātiem, un ir iespējams secināt, ka EL metode ar novērtētiem parametriem strādā labi, lai gan būtu interesanti šo pašu pieeju pielietot datiem ar lielu skaitu izlēcēju, lai ņemtu vērā Hodžes-Lēmana novērtējuma robustību. Savukārt otrā piemērā tika apskatīta neparametriskās regresijas atlikumu sadalījuma funkcijas novērtēšana, un rezultāti salīdzināti ar butstrapa metodes rezultātiem fiksētā punktā. Arī šajā piemērā EL strādā labi. Turpmāk būtu nepieciešams līdzīgā veidā apskatīt pielietojumus izdzīvošanas datu analīzē, jo, kā zināms, tā ir viena no plašākajām empīriskās ticamības funkcijas ar un bez novērtētiem parametriem pielietojumu klasēm.

Diemžēl divu izlašu problēmai ir zināms ļoti maz EL metodes ar novērtētiem parametriem pielietojumu, iespējams, tāpēc ka šī metode ir diezgan jauna un vēl maz pielietota. Taču tika atrasti reālu datu piemēri strukturālo attiecību modeļiem, kuriem tika konstruētas 95% ticamības joslas ar EL metodi programmā R, pielietojot J.Valeiņa un E.Cera izstrādāto programmas kodu [3]. Tika apskatīti piemēri lokācijas-skalēšanas modeļu klasei un veiktas hipotēžu pārbaudes par datu sadalījumu likumiem.

Ir arī zināmi pavisam jauni rezultāti EL metodes ar novērtētiem parametriem pielietojumiem robustajā statistikā, taču šajā darbā tam pieskārāmies tikai nedaudz, jo vēl nav pieejamas nevienas publikācijas par šo tēmu. Tā kā EL metode plaši tiek pielitota nesen, tad gaidāmi ne tikai metodes uzlabojumi, kurus veicot, EL būtu vēl spēcīgāks konkurents parametriskajām metodēm, bet arī plašākas pielietošanas iespējas divu izlašu problēmām, kas arī varētu būt šī darba turpinājums.

Izmantotā literatūra un avoti

[1] N.L. Hjort, I.W. McKeague, and I. Van Keilegom. Extending the scope of empirical likelihood. *Ann. Statist*, 37(3):1079–1111, 2009.

[2] J. Valeinis. *Confidence bands for structural relationship models*. PhD thesis, Niedersachsische Staats-und Universitatsbibliothek Gottingen, 2007.

[3] J. Valeinis and E. Cers. Extending the two-sample empirical likelihood method. Preprint. 2011.

[4] A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[5] A.B. Owen. *Empirical likelihood*. CRC press, 2001.

[6] D.R. Thomas and G.L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, pages 865–871, 1975.

[7] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994.

[8] Y.S. Qin and L.C. Zhao. Empirical likelihood ratio confidence intervals for various differences of two populations. *Systems Sci Math Sci*, 13:23–30, 2000.

[9] G. Qin and B.Y. Jing. Empirical likelihood for censored linear regression. *Scandinavian journal of statistics*, 28(4):661–673, 2001.

[10] Q.H. Wang and B.Y. Jing. Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53(3):517–527, 2001.

[11] R.J. Serfling. Approximation theorems of mathematical statistics. *New York*, 1980.

[12] N. Christianini and S.J. Taylor. An introduction to support vector machines (and othre kernel-based learning methods). 2000.

[13] J.L. Hodges Jr and E.L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, 1963.

[14] E.L. Lehmann and G. Casella. *Theory of point estimation*. Springer Verlag, 1998.

[15] N. Inagaki. The asymptotic representation of the Hodges-Lehmann estimator based on Wilcoxon two-sample statistic. *Annals of the Institute of Statistical Mathematics*, 25(1):457–466, 1973.

[16] N.L. Hjort and U. I Oslo. *Towards semiparametric bandwidth selectors for kernel density estimators*. Department of Mathematics, University of Oslo, 1999.

[17] B.W. Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, 6(1):177–184, 1978.

[18] P.A. Mackowiak, S.S. Wasserman, and M.M. Levine. A critical appraisal of 98.6 f, the upper limit of the normal body temperature, and other legacies of carl reinhold august wunderlich. *JAMA: The Journal of the American Medical Association*, 268(12):1578, 1992.

[19] L.D. Brown and M.G. Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

[20] M.G. Akritas and I. Van Keilegom. Non-parametric Estimation of the Residual Distribution. *Scandinavian Journal of Statistics*, 28(3):549–567, 2001.

[21] N. Neumeyer, H. Dette, and E.R. Nagel. Bootstrap tests for the error distribution in linear and nonparametric regression models. *Australian & New Zealand Journal of Statistics*, 48(2):129–156, 2006.

[22] E.V. Khmaladze and H.L. Koul. Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics*, 32(3):995–1034, 2004.

[23] L. Wasserman. *All of nonparametric statistics*. Springer-Verlag New York Inc, 2006.

[24] T.P. Hettmansperger and J.W. McKean. Statistical inference based on ranks. *Psychometrika*, 43(1):69–79, 1978.

[25] E.L. Lehmann. The power of rank tests. *The Annals of Mathematical Statistics*, 24(1):23–43, 1953.

[26] G. Freitag and A. Munk. On hadamard differentiability in k-sample semiparametric models–with applications to the assessment of structural relationships. *Journal of multivariate analysis*, 94(1):123–158, 2005.

[27] J.P. Klein and M.L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Verlag, 2003.

[28] G. Freitag. *Validierung von Modellen in der Uberlebenszeitanalyse*. PhD thesis, Ruhr-Universitat Bochum, Universitatsbibliothek, 2000.

[29] S.X. Chen and P. Hall. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181, 1993.

[30] J.D. Kalbfleisch, R.L. Prentice, and JD Kalbfleisch. *The statistical analysis of failure time data*, volume 5. Wiley New York:, 1980.

[31] Z. Zhang and G. Li. A simple quantile approach to the two-sample location-scale problem with random censorship. *Journal of Nonparametric Statistics*, 6(4):323–335, 1996.

[32] P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[33] F. Hampel, C. Hennig, and E. Ronchetti. A smoothing principle for the huber and other location m-estimators. *Computational Statistics & Data Analysis*, 55(1):324–337, 2011.

1. Pielikums

Programmas R kods Owena teorēmas pārbaudei

```
R_FF<-c()
n<-100
mu<-0
for (k in 1:1000)
{
izl<-rnorm(n,0,1)
izl_sort<-c()
izl_sort<-sort(izl)
lambda_l<-(1-1/n)/(mu-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(mu-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
sum((izl-mu)/(1+lambda*(izl-mu)))
}
f.lam2<-Vectorize(f.lam)
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(izl-mu)))
R_FF[k]<--2*log(prod(n*p))
}
hist(R_FF,prob=TRUE,main="",xlab="-2ln EL",col="blue")
xx<-seq(0,12,by=0.01)
lines(xx,dchisq(xx,1),col="red",lwd=2)
```

Programmas R kods Hodžes-Lēmaņa lokācijas novērtējumam

```
library(sm)

n<-100
izl<-rnorm(n,0,1)
b<-hcv(izl) #gludinosais parametrs
```

```
###Atrast isto theta parametru
d<-function(x) dnorm(x)^2
theta0<-integrate(d,-5,5)
theta0
plot(X,Y)
help(plot)
###Butstrapa ticamibas intervali
theta1<-function(dati)
{
f00<-function(x) #kodola blivuma f-jas novertejums
{
1/n/b*sum(dnorm((dati-x)/b))
}
f111<-Vectorize(f00)
d1<-function(x) f111(x)^2
integrate(d1,-5,5)$value
}
B<-1000
theta.boot<-replicate(B,theta1(sample(izl,replace=TRUE)))
se.boot<-var(theta.boot) #S^2

apak.rob.boot<-theta1(izl)-qnorm(0.95)*sqrt(se.boot)
aug.rob.boot<-theta1(izl)+qnorm(0.95)*sqrt(se.boot)

apak.rob.boot
aug.rob.boot
####Intervaali
R_FF<-function(theta)
{
f0<-function(x) #kodola blivuma f-jas novertejums
{
```

```
1/n/b*sum(dnorm((izl-x)/b))
}
f1<-Vectorize(f0)
izl_sort<-c()
izl_sort<-sort(f1(izl))
lambda_l<-(1-1/n)/(theta-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(theta-izl_sort[1])  #Lambda augseja robeza

f.lam<-function(lambda)
{
sum((f1(izl)-theta)/(1+lambda*(f1(izl)-theta)))
}
f.lam2<-Vectorize(f.lam)

#lam<-seq(lambda_l,lambda_u,by=0.1)
#plot(lam,f.lam2(lam),type="l")

lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root

p<-1/(n*(1+lambda*(f1(izl)-theta)))

-2*log(prod(n*p))
}

R_FF2<-Vectorize(R_FF)



r.ff<-seq(min(f1(izl))+0.099,max(f1(izl))-0.037,by=0.01)
plot(r.ff,R_FF2(r.ff),type="l",ylab="-2ln EL",
xlab="",lwd=2,ylim=c(-2,130),xlim=c(0.15,0.41))
abline(h=4*qchisq(0.95,1),lwd=2,col="red")
R_FF3<-function(theta)
```

```
{
R_FF3<-R_FF(theta)-4*qchisq(0.95,1)
}
R_FF4<-Vectorize(R_FF3)
lines(r.ff,R_FF4(r.ff),col="blue")
abline(h=0)
apak.rob<-uniroot(R_FF3,c(min(f1(izl))+0.1,
optimize(R_FF,c(min(f1(izl)),max(f1(izl))))$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(f1(izl)),
max(f1(izl))))$minimum,max(f1(izl))-0.01))$root
apak.rob
aug.rob
abline(v=0.28209,col="gray")
abline(v=apak.rob.boot,col="green",lwd=2)
abline(v=aug.rob.boot,col="green",lwd=2)
EL<-optimize(R_FF,c(min(f1(izl)),max(f1(izl))))$minimum
EL #Neparametriskas ticamibas metodes novertejums
```

HL novērtējums datiem body temperature

```
####Ticamibas intervali Hodges Lehmann novertejumam
gamma<-0.5 ####length(dati1)/(length(dati1)+length(dati2))

hl<-matrix(0,n,n)

for (i in 1:n)
{
for (j in 1:n)
{
hl[i,j]<-dati1[i]-dati2[j]
}
}
```

```
hl.nov<-median(hl)
hl.nov
dati2.nov<-dati2-hl.nov
mean(dati1)-mean(dati2)

sigma<-1/(12*gamma*(1-gamma)*EL^2)
sigma

apak.rob.hl<-hl.nov-qnorm(0.95)*sqrt(sigma)/sqrt(n)
aug.rob.hl<-hl.nov+qnorm(0.95)*sqrt(sigma)/sqrt(n)
apak.rob.hl
aug.rob.hl

mean(dati1)-mean(dati2)
t.test(dati1,dati2)$conf.int

> apak.rob.hl
[1] -0.6114009
> aug.rob.hl
[1] 0.01140088
>
> mean(dati1)-mean(dati2)
```

Programmas R kods regresijas atlikumu sadalījumiem

```
library(sm)
n<-100
X<-runif(n,0,1)
eps<-rnorm(n,0,0.1)
Y<-X^2+eps
scatterplot(X,Y)
pnorm(0,0,0.1)
###mu noveerteetais
```

```
b1<-hcv(X) #gludinosais parametrs
mu<-function(x)
{
sum((1/b1*dnorm((X-x)/b1))/sum(1/b1*dnorm((X-x)/b1))*Y)
}
mu1<-Vectorize(mu)
m<-seq(min(X),max(X),by=0.01)
plot(m,mu1(m),type="l",lwd=2,col="blue",ylab="mu(X)",xlab="")
m1<-function(x) x^2
points(m,m1(m),type="l",lwd=2)
eps.nov<-Y-mu1(X)
z<-0
b2<-hcv(eps.nov)
H<-pnorm((z-eps.nov)/b2) #Sadaliijuma f-jas noveerteejums

izl_sort<-c()
izl_sort<-sort(H)

R_FF<-function(theta)
{
lambda_l<-(1-1/n)/(theta-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(theta-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
sum((H-theta)/(1+lambda*(H-theta)))
}
f.lam2<-Vectorize(f.lam)
#lam<-seq(lambda_l,lambda_u,by=0.1)
#plot(lam,f.lam2(lam),type="l")
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(H-theta)))
-2*log(prod(n*p))
```

```
}
R_FF2<-Vectorize(R_FF)
r.ff<-seq(min(H)+0.1,max(H)-0.1,by=0.01)
plot(r.ff,R_FF2(r.ff),type="l",lwd=2,main="",
ylab="-2ln EL",xlab="",ylim=c(-5,120))

abline(h=qchisq(0.95,1),col="red",lwd=2)
R_FF3<-function(theta)
{
R_FF3<-R_FF(theta)-qchisq(0.95,1)
}
R_FF4<-Vectorize(R_FF3)
lines(r.ff,R_FF4(r.ff),col="blue")
abline(h=0)
apak.rob<-uniroot(R_FF3,c(min(H)+0.1,optimize(R_FF,c(min(H),max(H)))
$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(H),max(H)))
$minimum,max(H)-0.1))$root
apak.rob
aug.rob
theta_nov<-optimize(R_FF,c(min(H),max(H)))$minimum
theta_nov

###Butstrapa ticamibas intervali
z<-0
w<-0
mu<-function(t,v)
{
k<-function(x)sum((1/b1*dnorm((t-x)/b1))
/sum(1/b1*dnorm((t-x)/b1))*v)
k1<-Vectorize(k)
eps.nov.boot<-Y-k1(X)
```

```
for (i in 1:n)
if (eps.nov.boot[i]<=0) w<-w+1
w/n
}
B<-1000
theta.boot<-replicate(B,mu(sample(X,replace=TRUE),
sample(Y,replace=TRUE)))
se.boot<-var(theta.boot) #S^2
apak.rob.boot<-mu(X,Y)-qnorm(0.95)*sqrt(se.boot)
aug.rob.boot<-mu(X,Y)+qnorm(0.95)*sqrt(se.boot)
apak.rob.boot
aug.rob.boot
abline(v=pnorm(0),col="gray")
abline(v=apak.rob.boot,col="green",lwd=2)
abline(v=aug.rob.boot,col="green",lwd=2)
################Butstrapotais sadalijums

z<-0
theta.nov<-ecdf(eps.nov)(z)
for (k in 1:10000)
{
X.boot<-sample(X,replace=TRUE)
Y.boot<-c()

for (i in 1:n)
{
for(j in 1:n)
if (X.boot[i]==X[j]) Y.boot[i]<-Y[j]
}
eps.nov.boot<-Y.boot-mu1(X.boot)
theta.nov.boot<-ecdf(eps.nov.boot)(z)
e[k]<-(n*(theta.nov.boot-theta.nov)^2)/(theta.nov*(1-theta.nov))
```

```
}

hist(e,prob=TRUE,main="",ylab="",xlab="",col="2",ylim=c(0,0.8))
e.sort<-sort(e)
e.sort[9500]
l<-seq(0,12,by=0.1)
points(l,dchisq(l,1),type="l",lwd=2)

qchisq(0.95,1)
################Butstrapa ticamibas intervali
z<-0
w<-0

mu<-function(t,v)
{
k<-function(x)sum((1/b1*dnorm((t-x)/b1))/sum(1/b1*dnorm((t-x)/b1))*v)
k1<-Vectorize(k)
eps.nov.boot<-Y-k1(X)
for (i in 1:n)
if (eps.nov.boot[i]<=0) w<-w+1
w/n
}


B<-1000
theta.boot<-replicate(B,mu(sample(X,replace=TRUE),
sample(Y,replace=TRUE)))
se.boot<-var(theta.boot) #S^2


apak.rob.boot<-mu(X,Y)-qnorm(0.95)*sqrt(se.boot)
aug.rob.boot<-mu(X,Y)+qnorm(0.95)*sqrt(se.boot)
```

```
apak.rob.boot
aug.rob.boot

abline(v=pnorm(0),col="gray")
abline(v=apak.rob.boot,col="green",lwd=2)
abline(v=aug.rob.boot,col="green",lwd=2)
```

Datu piemers cmb

```
y.data<-read.table(file="CMB.txt",header=TRUE)[,2]
x.data<-read.table(file="CMB.txt",header=TRUE)[,1]
par(mfrow=c(1,2))
n<-length(x.data)
plot(x.data,y.data,cex=.7,xlab="",ylab="")
title(main=list('(1)',cex=1,font=1))
library(KernSmooth)
h<-dpill(x.data,y.data)
fit<-locpoly(x.data,y.data,bandwidth=h,degree=0)
lines(fit,lwd=1.5)
#ar 400 noverojumiem
xx.data<-c()
yy.data<-c()
for (i in 1:400){
xx.data[i]<-x.data[i]
yy.data[i]<-y.data[i] }
plot(xx.data,yy.data,cex=.7,xlab="",ylab="")


####EL atlikumu sadalijumiem neparametriskaja regresija

library(sm)
```

```
n<-length(xx.data)
X<-xx.data
Y<-yy.data

plot(X,Y,cex=.7)

###mu noveerteetais
b1<-hcv(X) #gludinosais parametrs

mu<-function(x)
{
sum((1/b1*dnorm((X-x)/b1))/sum(1/b1*dnorm((X-x)/b1))*Y)
}

mu1<-Vectorize(mu)
m<-seq(min(X),max(X),by=0.01)
points(m,mu1(m),type="l",lwd=2,col="blue",ylab="mu(X)",xlab="")

eps.nov<-Y-mu1(X)
z<-0

for (i in 0:2300)
{
z<-i-1000
b2<-bw.nrd(eps.nov)
H<-pnorm((z-eps.nov)/b2) #Sadaliijuma f-jas noveerteejums

izl_sort<-c()
izl_sort<-sort(H)

R_FF<-function(theta)
```

```
{
lambda_l<-(1-1/n)/(theta-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(theta-izl_sort[1])  #Lambda augseja robeza

f.lam<-function(lambda)
{
sum((H-theta)/(1+lambda*(H-theta)))
}
f.lam2<-Vectorize(f.lam)

#lam<-seq(lambda_l,lambda_u,by=0.1)
#plot(lam,f.lam2(lam),type="l")

lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root

p<-1/(n*(1+lambda*(H-theta)))

-2*log(prod(n*p))
}

R_FF2<-Vectorize(R_FF)

R_FF3<-function(theta)
{
R_FF3<-R_FF(theta)-4
}

apak.rob[i]<-uniroot(R_FF3,c(min(H)+0.001,optimize(R_FF,c(min(H),max(H)))$minimum))
aug.rob[i]<-uniroot(R_FF3,c(optimize(R_FF,c(min(H),max(H)))$minimum,max(H)-0.001))$

theta_nov[i]<-optimize(R_FF,c(min(H),max(H)))$minimum
}
```

```
plot.ecdf(eps.nov,xlab="",lwd=1,cex=0.8,main="(4)",ylab="")
xx<-seq(-1000,1299,by=1)
points(xx,theta_nov,type="l",lwd=3,lty=1,xlab="",ylab="",main="(4)",col="green")
points(xx,pnorm(xx,mean(eps.nov),sd(eps.nov)),lwd=2,lty=2,ylab="",xlab="",type="l",
points(xx,apak.rob,type="l",col="red",lty="dashed")
points(xx,aug.rob,type="l",col="red",lty="dashed")
apak.rob[1000]
aug.rob[1000]
plot.ecdf(eps.nov)

################Butstrapotais sadalijums

z<--700
theta.nov<-ecdf(eps.nov)(z)
for (k in 1:1000)
{
X.boot<-sample(X,replace=TRUE)
Y.boot<-c()

for (i in 1:n)
{
for(j in 1:n)
if (X.boot[i]==X[j]) Y.boot[i]<-Y[j]
}
eps.nov.boot<-Y.boot-mu1(X.boot)
theta.nov.boot<-ecdf(eps.nov.boot)(z)
e[k]<-(n*(theta.nov.boot-theta.nov)^2)/(theta.nov*(1-theta.nov))
}

hist(e,prob=TRUE,main="",ylab="",xlab="",col="2",ylim=c(0,0.8))
e.sort<-sort(e)
```

```
e.sort[950]
l<-seq(0,12,by=0.1)
points(l,dchisq(l,1),type="l",lwd=2)

qchisq(0.95,1)

##########Atlikumu analize
hist(eps.nov,prob=TRUE,main="(2)",xlab="",ylab="",col="blue")
xx<-seq(-1500,1500,by=0.01)
lines(xx,dnorm(xx,mean(eps.nov),sd(eps.nov)),lwd=2)

qqnorm(eps.nov,main="(3)",xlab="",ylab="")
qqline(eps.nov)

shapiro.test(eps.nov)
library(nortest)
lillie.test(eps.nov)

plot(eps.nov,main="(1)",xlab="",ylab="")
atl<-read.table(file="atlikumi.txt")
```

EL divu izlašu gadījumā

```
###Datu piemers clouds: location - scale modelis
par(mfrow=c(1,2))
dati<-read.table("clouds.txt",header=TRUE)

dati1<-dati$Unseeded_Clouds
dati2<-dati$Seeded_Clouds
x<-dati1
y<-dati2
```

```
tt <- seq(0.05, 0.95, length=50)
plot(tt,ecdf(dati1)(quantile(dati2,tt)),type="s",
xlab="",xlim=c(0,0.83),ylim=c(0,1),main="(1)",ylab="")
abline(0,1)

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
#s<-1 # location
m<-mean(x)-s*mean(y)
y.mod<-y*s+m

qqplot(x,y.mod,main="Clouds")
abline(0,1)

###Ar empirical likelihood: clouds
tt <- seq(0.1, 0.7999, length=50)
plot(tt,ecdf(dati1)(quantile(dati2,tt)),type="s",ylim=c(0,1))
abline(0,1)

plot(tt,ecdf(x)(quantile(y.mod,tt)),type="s",
ylim=c(0,1),xlab="",main="(2)",ylab="")
abline(0,1)

zz <- EL.curve(x, y.mod, tt, type="pp", conf.level=0.95,
 sim.conf.level=0.95)
lines(tt, zz$estim, xlab="", ylab="", main="P-P plot",
 type='l',col="red")
lines(tt, zz$conf.int[1,], lty="dashed")
lines(tt, zz$conf.int[2,], lty="dashed")
lines(tt, zz$simult.conf.int[1,], lty="dotted")
lines(tt, zz$simult.conf.int[2,], lty="dotted")
```

```
###Datu piemers body temperature: location - scale modelis
par(mfrow=c(1,2))
dati11<-read.table("body_temperature.txt",header=TRUE)
dati22<-read.table("body_temperature_female.txt",header=TRUE)

dati1<-dati11$temperature #Kermena t viriesiem
dati2<-dati22$temperature #Kermena t sievietem
x<-dati1
y<-dati2

tt <- seq(0.05, 0.95, length=50)
plot(tt,ecdf(dati1)(quantile(dati2,tt)),
type="s",ylab="",xlab="",ylim=c(0,1),main="(1)")
abline(0,1)

plot(tt,ecdf(dati2)(quantile(dati1,tt)),type="s",
ylim=c(0,1),main="Body temperature")
abline(0,1)

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
s<-1 # location
m<-mean(x)-s*mean(y)
y.mod<-y*s+m

qqplot(y.mod,x,main="Body temperature")
abline(0,1)

###Ar empirical likelihood: body temperature
tt <- seq(0.1, 0.86, length=50)
plot(tt,ecdf(dati1)(quantile(dati2,tt)),type="s",ylim=c(0,1))
abline(0,1)
```

```
plot(tt,ecdf(x)(quantile(y.mod,tt)),type="s",
ylim=c(0,1),ylab="",xlab="",main="(2)")
abline(0,1)

zz <- EL.curve(x, y.mod, tt, type="pp", conf.level=0.95,
 sim.conf.level=0.95)
lines(tt, zz$estim, xlab="", ylab="", main="P-P plot",
 type='l',col="red")
lines(tt, zz$conf.int[1,], lty="dashed")
lines(tt, zz$conf.int[2,], lty="dashed")
lines(tt, zz$simult.conf.int[1,], lty="dotted")
lines(tt, zz$simult.conf.int[2,], lty="dotted")

###Datu piemers mice leukemia: location - scale modelis
par(mfrow=c(1,2))
dati11<-read.table("mice_leukemia_male.txt",header=TRUE)
dati22<-read.table("mice_leukemia_female.txt",header=TRUE)

dati1<-dati11$surv_time #Kermena t viriesiem
dati2<-dati22$surv_time #Kermena t sievietem
x<-dati1
y<-dati2

tt <- seq(0.05, 0.95, length=50)
plot(tt,ecdf(dati1)(quantile(dati2,tt)),type="s",
ylab="",main="(1)",xlab="")
abline(0,1)

plot(tt,ecdf(dati2)(quantile(dati1,tt)),type="s",
ylim=c(0,1),main="Mice leukemia")
abline(0,1)
```

```
#sigma sastavdalas un mu
a<-function(tt) as.vector(quantile(x,tt)*quantile(y,tt))
a1<-integrate(a,0.05,0.95,subdivisions=1000,abs.tol=0.001)$value

b<-function(tt) quantile(x,tt)
c1<-function(tt) quantile(y,tt)
bc1<-integrate(b,0.05,0.95)$value*integrate(c1,0.05,0.95)$value

d<-function(tt) quantile(y,tt)*quantile(y,tt)
d1<-integrate(d,0.05,0.95)$value
e<-integrate(c1,0.05,0.95)$value^2

sigma<-(a1-bc1)/(d1-e)
mu<-integrate(b,0.05,0.95)$value
-sigma*integrate(c1,0.05,0.95)$value

s<-sigma
m<-mu  ### vai m<-mean(x)-s*mean(y)
y.mod<-y*s+m

plot(tt,ecdf(x)(quantile(y.mod,tt)),type="s",
main="Mice leukemia modified",ylab="")
abline(0,1)

###Ar empirical likelihood: mice leukemia
xx <- seq(0.1, 0.9, by=0.01)

plot(tt,ecdf(x)(quantile(y.mod,tt)),type="s",
main="(2)",ylab="",xlab="")
abline(0,1)

zz <- EL.curve(x, y.mod, xx, type="pp",
```

```
conf.level=0.95, sim.conf.level=0.95)
lines(xx, zz$estim, xlab="", ylab="",
main="P-P plot", type='l',col="red")
lines(xx, zz$conf.int[1,], lty="dashed")
lines(xx, zz$conf.int[2,], lty="dashed")
lines(xx, zz$simult.conf.int[1,], lty="dotted")
lines(xx, zz$simult.conf.int[2,], lty="dotted")
```

Diplomdarbs “Empīriskā ticamības funkcija ar novērtētiem parametriem” izstrādāts LU Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Leonora Pahirko

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: Sandra Vucāne

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Diplomdarbs aizstāvēts valsts gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretāre: asoc. prof. Dr.math. Inese Bula _____

(paraksts)