LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKĀS TICAMĪBAS FUNKCIJAS METODES VISPĀRINĀJUMS IZDZĪVOŠANAS DATIEM

MAĢISTRA DARBS

Autors: **Leonora Pahirko**

Stud. apl. lp06061

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2013

Anotācija

Zināms, ka empīriskās ticamības funkcijas metode pirmo reizi tika pielietota tieši ticamības intervālu konstruēšanai izdzīvošanas varbūtībām [1], tomēr ilgu laiku nebija vienotas pieejas metodes [2] pielietošanai citiem izdzīvošanas analīzes parametriem. Darbā aplūkota pielāgota empīriskās ticamības funkcijas metode izdzīvošanas analīzes datiem, kas ļauj konstruēt ticamības intervālus parametru klasei, kurus var izteikt kā funkcionāļus no izdzīvošanas funkcijas. Darbā aprakstīts arī empīriskās ticamības funkcijas metodes vispārinājums divu izlašu problēmām, un atrasts robežsadalījums pielāgotās metodes pielietojumam divu izlašu gadījumā. Apskatīti simulāciju rezultāti gan vienas, gan divu izlašu gadījumā, kā arī metodes pielietojums reālu datu piemēram.

Atslēgas vārdi: izdzīvošanas analīze, cenzēti dati, empīriskās ticamības funkcijas metode, divu izlašu problēmas

Abstract

It is known that the empirical likelihood method was first used to construct confidence intervals for survival probabilities [1], however for a long time there was no common approach how to use the method for other survival parameters. In this thesis the adjusted empirical likelihood method [2] has been considered which allows constructing confidence intervals for a class of functionals of a survival function. The extension of the empirical likelihood method for two-sample problems has been established, as well as the limiting distribution for the adjusted method in two-sample case has been derived. Some simulation studies have been done both for one and two-sample cases and implementation of the adjusted method was made to analyze a real data example.

Keywords: survival analysis, censoring, empirical likelihood method, two-sample problems

Saturs

Ievads

Izdzīvošanas datu analīzes pirmsākumi meklējami senā pagātnē, kad tika radītas pirmās mirstību tabulas, tāpēc arī šīs statistikas nozares terminoloģija balstās uz medicīnu. Tomēr Otrā Pasaules kara laikā izdzīvošanas datu analīze strauji attīstījās, pateicoties dažādiem pētījumiem inženierzinātnēs un ieroču rūpniecībā. Pašlaik dati, kas apraksta laiku līdz kādam pētniekus interesējošam notikumam, sastopami ne tikai medicīnā un inženierzinātnēs, bet arī tādās jomās kā bioloģija, ekonomika, sabiedrības veselība u.c.

Izdzīvošanas datu galvenā atšķirība no neatkarīgi un vienādi sadalītiem datiem ir novērojumu cenzēšana, kas notiek gadījumos, kad ir ierobežots pētījuma laiks vai pastāv citi apstākļi, kuru dēļ nav iespējams novērot atsevišķus subjektus visu pētījuma laiku. Rezultātā par šiem subjektiem ir zināms tikai fakts, ka līdz cenzēšanas brīdim pētāmais notikums vēl nebija iestājies, un arī tā ir noderīga informācija. Šī iemesla dēļ parasto statistikas metožu pielietošana var būt nepiemērota.

Empīriskās ticamības funkcijas metode ir kļuvusi par konkurētspējīgu un plaši pētītu neparametriskās statistikas metodi, kopš 1988. gadā to ieviesa Ovens [4], kas pašlaik tiek uzskatīts par empīriskās ticamības funkcijas metodes pamatlicēju. Lai gan pirmo reizi tā tika minēta tieši izdzīvošanas analīzes kontekstā izdzīvošanas varbūtību novērtēšanai Thomas un Grunkemeir darbā [1] jau 1975. gadā, tikai 1995. gadā Li savā publikācijā [5] sniedza rūpīgu teorētisko pamatojumu metodes pielietošanai izdzīvošanas datiem.

Taču joprojām nebija vienotas pieejas, kā EL metodi vispārināt cenzētiem novērojumiem. 2001. gadā Wang un Jing [2] tomēr pielāgoja empīriskās ticamības funkcijas metodi funkcionāļu klasei, kas ļauj konstruēt ticamības intervālus gan izdzīvošanas varbūtībām, gan arī vidējam izdzīvošanas laikam un citiem parametriem, kurus var izteikt ar izdzīvošanas funkcijas palīdzību.

2009. gadā Hjort, McKeague un Van Keilegom [6] vispārināja Ovena ieviesto empīriskās ticamības funkcijas metodi, lai pieļautu traucējošo parametru aizvietošanu ar to novērtējumiem,

turklāt kā viens no pielietojumiem tika minēts Wang un Jing [2] rezultāts, kurā izdzīvošanas funkcijas tika aizstātas ar to Kaplan-Meier novērtējumiem, kā rezultātā mainās statistikas robežsadalījums.

Šī darba galvenais mērķis ir atrast robežsadalījumu vispārējā formā divu izlašu problēmām, balstoties uz Wang un Jing [2] pieeju. Mērķa sasniegšanai tika izvirzīti sekojoši uzdevumi.

1. Iepazīties sīkāk ar izdzīvošanas datu analīzi, galveno parametru novērtējumiem un dažādām cenzēšanas shēmām.

2. Rūpīgi iepazīties ar Wang un Jing publikāciju [2], kā arī izpētīt rezultātu pierādījumu un tajā pielietotās metodes.

3. Izprast Qin un Zhao [7] empīriskās ticamības funkcijas metodes pierādījumu divu izlašu problēmām.

4. Veikt simulācijas datorprogrammā R gan vienas, gan divu izlašu gadījumā, lai pārbaudītu teorētiskos rezultātus empīriski.

5. Apskatīt metodes pielietojumu reāliem datiem.

Maģistra darbs sastāv no ievada, 5 nodaļām, nobeiguma, izmantotās literatūras saraksta un pielikuma. Pirmajā nodaļā aprakstīta izdzīvošanas datu analīze, galvenie izdzīvošanas datu sadalījumu raksturlielumi un biežāk pielietotās cenzēšanas shēmas, savukārt otrā nodaļa veltīta empīriskās ticamības funkcijas metodei vienas izlases gadījumā. Trešā nodaļa sastāv no divām apakšnodaļām, kurās apskatītas dažādas pieejas empīriskās ticamības funkcijas metodes pielietošanai izdzīvošanas datiem - tradicionālā pieeja, kuru ieviesa Thomas un Grunkemeier [1], un metode ar novērojumu modifikāciju [2]. Ceturtajā nodaļā apskatīta empīriskās ticamības funkcijas metode divu izlašu problēmām, arī tās vispārinājums ar traucējošiem parametriem, kur kā piemērs iekļauta lokācijas modeļu problemātika, un galvenie rezultāti metodes pielietojumam izdzīvošanas datiem. Piektajā nodaļā aprakstīti simulāciju rezultāti, kā arī empīriskās ticamības funkcijas metodes pielietojums primārās biliārās aknu cirozes (pbc) pacientu datiem. Pielikumā iekļauts izveidoto programmu kods datorpaketē R.

1 Izdzīvošanas datu analīze

Šajā nodaļā apskatīsim, kas īsti ir izdzīvošanas datu analīze, ar ko tā atšķiras no citām statistikas nozarēm un kādas metodes tajā galvenokārt tiek pielietotas. Par literatūras avotu pamatā tiks izmantota Klein un Moeschberger grāmata "Survival analysis. Techniques for Censored and Truncated Data" [3].

Izdzīvošanas datu analīze ir statistikas nozare, kas pēta pozitīvus novērojumus, kuri apraksta laiku līdz kāda notikuma iestāšanās dienās, mēnešos, gados vai citās laika vienībās (*time to-event-data*). Šādi dati visbiežāk sastopami tādās jomās kā medicīna, bioloģija, sabiedrības veselība, epidemioloģija, inženierzinātne, ekonomika un demogrāfija. Neskatoties uz to, ka tālāk apskatītās metodes var tikt pielietotas jebkurā no minētajām nozarēm, terminoloģija vēsturiski balstās uz medicīnu un bioloģiju, tāpēc tradicionāli subjekta jeb pacienta nāve vai kādas slimības iestāšanās ir notikums, līdz kuram tiek uzņemts laiks, ko sauksim par izdzīvošanas laiku. Savukārt, inženierzinātnēs un ekonomikā attiecīgais notikums varētu būt, piemēram, kādas iekārtas kalpošanas vai darba meklētāja bezdarba ilgums.

Piemērs 1. Īsi minēsim dažas iespējamas izdzīvošanas analīzes problēmas.

- Pētījums, kurā novēro, cik ilgi pacienti izdzīvo pēc sirds transplantācijas.
- Recidīvisma pētījums - tiek uzņemts laiks, cik ilgi ieslodzītie pēc nosacītās atbrīvošanas atkārtoti nonāk policijas redzeslokā.
- Laiks, kurā pacients atgriežas darbā pēc plānveida operācijas (notikums, līdz kuram tiek uzņemts laiks, var būt arī "labs").

1.1 Izdzīvošanas datu sadalījuma raksturlielumi [3]

Pieņemsim, ka X ir nenegatīvs gadījuma lielums, kas raksturo laiku līdz kādam mūs interesējošam notikumam. Apskatīsim četras dažādas funkcijas, kas raksturo X sadalījumu, turklāt, zinot

jebkuru no šīm funkcijām, pārējās trīs var tikt viennozīmīgi izteiktas.

Definīcija 1. Par izdzīvošanas funkciju sauc funkciju $S(x) = P(X > x)$.

Izdzīvošanas funkcija $S(x)$ raksturo varbūtību subjektam, kas izdzīvojis līdz laika momentam x, izdzīvot laika momentā x. Ja X ir nepārtraukts gadījuma lielums, $S(x)$ ir nepārtraukta stingri dilstoša funkcija, turklāt tā ir sadalījuma funkcijas $F(x) = P(X \leq x)$ papildinājums, t.i., $S(x) = 1 - F(x)$. Izdzīvošanas funkciju var definēt arī kā integrāli pa varbūtību blīvuma funkciju, $f(x)$,

$$S(x) = P(X > x) = \int_x^\infty f(t)dt,$$

no kurienes

$$f(x) = -\frac{dS(x)}{dx}.$$

Attēlojot dažādas izdzīvošanas funkcijas grafiski, tām novērojamas vairākas kopīgas īpašības:

- tās ir dilstošas līknes;
- laika momentā $x = 0$, $S(x) = S(0) = 1$, jo varbūtība izdzīvot laika momentā 0 ir 1;
- laika momentā $x = \infty$, $S(x) = S(\infty) = 0$, t.i., ja pētījumu būtu iespējams turpināt bezgalīgi ilgi, tad pienāktu brīdis, kad visi subjekti būtu miruši.

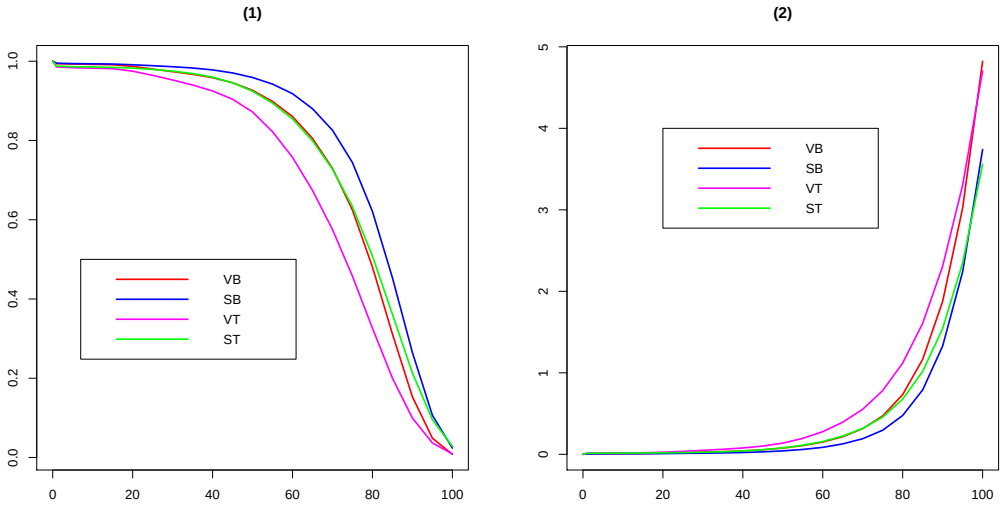
Savukārt, ja X ir diskrēts gadījuma lielums, tad $S(x)$ ir dilstoša pakāpienveida funkcija. Pieņemsim, ka X pieņem vērtības x_i, $i = 1, 2, \ldots$, turklāt $x_1 < x_2 < \ldots$, ar varbūtībām $p(x_i) = P(X = x_i)$, tad $S(x)$ var izteikt

$$S(x) = P(X > x) = \sum_{x_i > x} p(x_i).$$

Piemērs 2. Amerikas Savienotajās valstīs Veselības departaments katru gadu publicē izdzīvošanas varbūtības visiem nāves cēloņiem, iedalot cilvēkus grupās pēc rases un dzimuma. 1.1 attēlā redzamas izdzīvošanas funkcijas pēc attiecīgā iedalījuma ASV 2006.gadam. No tā iespējams secināt, ka vislielākās izdzīvošanas varbūtības ir baltās rases sievietēm, tad baltās rases vīriešiem un tumšās rases sievietēm, savukārt vissliktākās izdzīvošanas iespējas ASV ir tumšās rases vīriešiem.

Definīcija 2. Par riska funkciju sauc funkciju

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x}.$$



Att. 1.1: Izdzīvošanas funkcijas (1) un riska funkcijas (2) visiem nāves cēloņiem ASV 2006.gadam iedalījumā pēc rases (B-baltā, T-tumšā) un dzimuma (V-vīrieši, S-sievietes).

Riska funkcija raksturo tūlītēju potenciālu subjektam, kas pārdzīvojis laika momentu x, piedzīvot interesējošo notikumu nākamajā laika vienībā. Riska funkcijai $h(x)$ nav noteiktas vienotas formas, vienīgais nosacījums, ka $h(x) \geq 0$.

Ja X ir nepārtraukts gadījuma lielums, tad

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d(\ln S(x))}{dx}.$$

Aplūkosim arī kumulatīvo riska funkciju $H(x) = \int_0^x h(u)du = -\ln S(x)$. Izdzīvošanas funkciju nepārtrauktiem izdzīvošanas laikiem varam izteikt ar kumulatīvā riska funkciju sekojoši

$$S(x) = \exp\left(-H(x)\right) = \exp\left(-\int_0^x h(u)du\right).$$

Kad X ir diskrēts gadījuma lielums, riska funkciju var izteikt kā nosacīto varbūtību

$$h(x_i) = P(X = x_i | X \geq x_i) = \frac{p(x_i)}{S(x_{i-1})}, \quad i = 1, 2, \ldots,$$

kur $S(x_0) = 1$. Tā kā $p(x_i) = S(x_{i-1}) - S(x_i)$, tad $h(x_i) = 1 - S(x_i)/S(x_{i-1})$, $i = 1, 2, \ldots$. Ievērosim arī, ka

$$S(x) = \prod_{x_i \leq x} S(x_i)/S(x_{i-1}),$$

no kurienes varam iegūt izdzīvošanas funkcijas saistību ar riska funkciju

$$S(x) = \prod_{x_i \leq x} (1 - h(x_i)).$$

Piemērs 3. 1.1 attēlā redzamas riska funkcijas Piemērā 2 minētajiem datiem pēc attiecīgā iedalījuma. Visām grupām riska funkcija sākumā ir konstanta, bet pie nedaudz atšķirīgiem vecumiem katrai grupai sāk augt, jo, palielinoties cilvēka vecumam, nāves risks palielinās dabīgās novecošanas dēļ.

Definīcija 3. Par sagaidāmo atlikušo dzīves ilgumu (*mean residual life*) laika momentā x sauc

$$mrl(x) = E(X - x | X > x).$$

Subjektiem, kas izdzīvojuši laika momentā x, $mrl(x)$ raksturo sagaidāmo atlikušo izdzīvošanas laiku. Vidējais dzīves ilgums $\mu = mlr(0)$ sakrīt ar laukumu zem izdzīvošanas funkcijas $S(x)$, savukārt $mlr(x)$ sakrīt ar laukuma daļu pa labi no x zem izdzīvošanas funkcijas izdalītu ar $S(x)$.

Nepārtrauktiem gadījuma lielumiem

$$mrl(x) = \frac{\int_x^\infty (t - x) f(t) dt}{S(x)} = \frac{\int_x^\infty S(t) dt}{S(x)}$$

un

$$\mu = E(X) = \int_0^\infty t f(t) dt = \int_0^\infty S(t) dt.$$

Arī X dispersija var tikt izteikta ar izdzīvošanas funkcijas palīdzību

$$D(X) = 2 \int_0^\infty t S(t) dt - \left[\int_0^\infty S(t) \right]^2,$$

kā arī X sadalījuma p-tā kvantile x_p ir mazākā vērtība tāda, ka

$$S(x_p) \leq 1 - p, \text{ t.i., } x_p = \inf\{t : S(t) \leq 1 - p\}.$$

No kurienes varam definēt izdzīvošanas laika mediānu kā X sadalījuma 0.5-to kvantili, nepārtrauktiem gadījuma lielumiem izpildās $S(x_{0.5}) = 0.5$.

Piemērs 4. No 1.1 attēla aptuveni varam atrast izdzīvošanas laika mediānu tumšās rases vīriešiem 2006.gadā, kas ir apmēram 75 gadi.

1.2 Biežāk lietotie sadalījumi

Lai arī darbā lielākā uzmanība ir vērsta tieši uz konkrētas neparametriskās statistikas metodes pielietojumu, izdzīvošanas datu analīzē bieži nākas sastapties ar sadalījumiem, ko plaši pielieto parametriskajos modeļos. Turklāt šie sadalījumi nav izvēlēti tikai pateicoties popularitātei izdzīvošanas datu pētnieku vidū, bet arī tam, ka tie sniedz labāku priekšstatu par iepriekšējā nodaļā apskatītajiem lielumiem un funkcijām, it īpaši, par riska funkciju $h(x)$.

Riska funkciju, izdzīvošanas funkciju, varbūtību blīvuma funkciju un sagaidāmo dzīves ilgumu apkopojums dažiem pazīstamākajiem sadalījumiem attēlots 1.1 tabulā. Vēl dažus no tiem apskatīsim sīkāk.

Eksponenciālais sadalījums ir ne tikai viens no vēsturiski nozīmīgākajiem, bet arī matemātiski vienkāršākajiem sadalījumiem izdzīvošanas analīzē. To raksturo konstanta riska funkcija $h(x) = \lambda$. Eksponenciālajam sadalījumam piemīt īpašība

$$P(X \geq x + z | X \geq x) = P(X \geq z),$$

ko mēdz saukt par atmiņas trūkuma (*lack of memory*) īpašību. Šīs īpašības dēļ arī atlikušais sagaidāmais dzīves ilgums ir konstanta funkcija

$$E(X - x | X > x) = E(X) = \frac{1}{\lambda},$$

jo laiks līdz notikumam nav atkarīgs no pagātnes, ko atspoguļo arī konstantā riska funkcija - novecošanās neietekmē notikuma iestāšanās risku. Šo īpašību dēļ eksponenciālā sadalījuma pielietošana ir diezgan ierobežota gan veselības, gan industriālajā sfērā.

Veibuls 1939. gadā piedāvāja lietot kādu sadalījumu, kas aprakstīja materiālu dzīves ilgumu. Lai gan viņš nebija pirmais, tomēr šo sadalījumu vēlāk nosauca viņa vārdā - par Veibula sadalījumu. Šī sadalījuma izdzīvošanas funkcija ir $S(x) = e^{-\lambda x^{\alpha}}$, $x > 0$. Eksponenciālais sadalījums ir Veibula sadalījuma speciālgadījums, kad $\alpha = 1$. Veibula sadalījums var tikt pielāgots gan augošai ($\alpha > 1$), gan dilstošai ($\alpha < 1$), gan konstantai ($\alpha = 1$) riska funkcijai. Ir skaidrs, ka šī sadalījuma forma ir atkarīga no parametra α, tāpēc to mēdz dēvēt par formas (*shape*) parametru. Augošas Veibula sadalījuma riska funkcijas piemērs varētu būt leikēmijas pacienti, kuru izdzīvošanas izredzes pasliktinās līdz ar laiku, savukārt dilstošas riska funkcijas piemērs ir operāciju pacienti, kuru izdzīvošanas izredzes ir sliktākas uzreiz pēc operācijas, bet ar laiku risks nomirt samazinās.

Tabula 1.1: Riska, izdzīvošanas un varbūtību blīvuma funkcijas un sagaidāmie dzīves ilgumi dažiem biežāk izmantotajiem sadalījumiem

Sadalījums	$h(x)$	$S(x)$	$f(x)$	$E(X)$
Eksponenciālais $\lambda > 0$ $x \geq 0$	λ	$e^{-\lambda x}$	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$
Veibula $\alpha, \lambda > 0$ $x \geq 0$	$\alpha\lambda x^{\alpha-1}$	$e^{-\lambda x^{\alpha}}$	$\alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}}$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma $\beta, \lambda > 0$ $x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - \frac{\int_0^{\lambda x} u^{\beta-1}e^{-u}du}{\Gamma(\beta)}$	$\frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$
Normālais $\sigma > 0$ $-\infty < x < \infty$	$\frac{f(x)}{S(x)}$	$1 - \Phi\left(\frac{x-\mu}{\sigma}\right)$	$\frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{(2\pi)^{1/2}\sigma}$	μ
Pareto $\theta, \lambda > 0$ $x \geq \lambda$	$\frac{\theta}{x}$	$\frac{\lambda^{\theta}}{x^{\theta}}$	$\frac{\theta\lambda^{\theta}}{x^{\theta+1}}$	$\frac{\theta\lambda}{\theta-1}$, ja $\theta > 1$

Līdzīgas īpašības kā Veibula sadalījumam ir Gamma sadalījumam, kura blīvuma funkcija ir izskatā

$$f(x) = \frac{\lambda^{\beta}x^{\beta-1}e^{-\lambda x}}{\Gamma(\beta)},$$

kur $\lambda > 0$, $\beta > 0$, $x > 0$ un $\Gamma(\beta) = \int_0^{\infty} u^{\beta-1}e^{-u}du$ ir tā saucamā gamma funkcija. Tāpat kā Veibula sadalījuma gadījumā, eksponenciālais sadalījums ir Gamma sadalījuma speciālgadījums, kad $\beta = 1$, tuvojas normālajam sadalījumam, kad $\beta \to \infty$, un dod χ^2_{ν} sadalījumu, kad $\nu = 2\beta$ (β vesels skaitlis) un $\lambda = 1/2$.

Ar citu izdzīvošanas datu analīzē pielietoto sadalījumu aprakstu var iepazīties, piemēram, grāmatā [3].

1.3 Novērojumu cenzēšana un nošķelšana

Bieži gadās, ka subjekta precīzu izdzīvošanas laiku nav iespējams novērot. Piemēram, ja līdz pētījuma beigām subjekts tā arī nepiedzīvo interesējošo notikumu vai mirst cita iemesla nevis pētāmās slimības dēļ, vai arī pētījuma laikā maina dzīvesvietu un vairs nevar turpināt dalību, tad novērotais izdzīvošanas laiks tiek uzskatīts par cenzētu. Šādi nepilnīgi dati ir izdzīvošanas analīzes datu kopu iezīme, tāpēc apskatīsim sīkāk dažādas iespējamās cenzēšanas un nošķelšanas shēmas.

Cenzēšana no labās puses

Šis cenzēšanas veids ir visbiežāk sastopams izdzīvošanas analīzes datu kopās. Ieviesīsim sekojošus apzīmējumus. n subjektu izdzīvošanas laikus apzīmēsim ar X_1,...,X_n, kas ir neatkarīgi un vienādi sadalīti (turpmāk lietosim apzīmējumu iid) gadījuma lielumi ar kādu sadalījuma funkciju $F(x)$. Ar Y_1,...,Y_n apzīmēsim fiksētus cenzēšanas laikus katram subjektam. Precīzs i-tā subjekta izdzīvošanas laiks cenzēšanas no labās puses gadījumā ir zināms tikai tad, ja X_i ir mazāks vai vienāds ar Y_i, pretējā gadījumā pacients ir "izdzīvotājs" un par viņa izdzīvošanas laiku tiek uzskatīts cenzēšanas laiks Y_i. Datus ērti reprezentēt ar gadījuma lielumu pāriem (Z_i, δ_i), kur δ_i ir cenzēšanas indikators ($\delta_i = 1$, ja Z_i atbilst īstajam izdzīvošanas laikam, savukārt $\delta_i = 0$, ja izdzīvošanas laiks ir cenzēts), un $Z_i = \min(X_i, Y_i)$.

Cenzēšanu no labās puses var iedalīt vairākos tipos.

Definīcija 4. Izdzīvošanas laiks tiek saukts par *pirmā tipa* cenzētu no labās puses, ja interesējošais notikums neiestājas līdz kādam iepriekš norādītam laikam, kas var būt atšķirīgs katram subjektam.

Definīcija 5. Ja pētījums tiek turpināts, kamēr pirmie r subjekti piedzīvo interesējošo notikumu, tad pārējo subjektu izdzīvošanas laiki tiek saukti par *otrā tipa* cenzētiem no labās puses. $r < n$ ir kāds pētījuma sākumā noteikts vesels skaitlis.

Definīcija 6. Izdzīvošanas laiks tiek saukts par *gadījuma* cenzētu no labās puses, ja subjekts, kas vēl nav piedzīvojis notikumu, nevar turpināt dalību pētījumā no pētāmā notikuma neatkarīgu iemeslu dēļ.

Arī katru no šiem tipiem var iedalīt sīkāk. Šis iedalījums smalki aprakstīts Klein un Moeschberger grāmatā [3]. Komentēsim nedaudz katru no cenzēšanas tipiem. Pirmā tipa cenzēšana parādās, kad pētījumam ir noteikts ierobežots laiks, piemēram, pacientiem tiek dotas zāles tieši vienu mēnesi, vai arī, kad, novērojot zāļu iedarbību uz pelēm, pētījumam beidzoties, tās nepieciešams nogalināt. Otrā tipa cenzēšana ļauj ietaupīt ne tikai laiku, bet arī līdzekļus, piemēram, novērojot iekārtas darbības ilgumu, pētījums tiek pārtraukts pēc noteikta skaita iekārtu salūšanas. Gadījuma cenzēšana parādās, kad vēlamies novērtēt kāda notikuma robežsadalījumu, bet daži subjekti saskaras ar apstākļiem, kas liedz turpināt piedalīties pētījumā. Šos apstākļus sauc par konkurējošiem riskiem (*competing risks*).

Cenzēšana no kreisās puses un intervālu cenzēšana

Cenzēšana no kreisās puses notiek, ja subjekts ir piedzīvojis notikumu jau pirms iestāšanās pētījumā, tāpēc precīzs izdzīvošanas laiks nav zināms. Piemēram, pētījumā pāraudzināšanas iestādes jauniešiem tiek jautāts "Kad jūs pirmo reizi lietojāt marihuānu?" Ir iespējams izvēlēties atbildi "Esmu lietojis, bet neatceros, cik man bija gadu". Tātad šajā gadījumā respondents notikumu ir piedzīvojis un par cenzēto izdzīvošanas laiku tiek uzskatīts viņa pašreizējais vecums.

Izmantosim tādus pašus apzīmējumus kā cenzēšanas no labās puses gadījumā. i-tā subjekta izdzīvošanas laiks X_i tiek uzskatīts par cenzētu no kreisās puses, ja tas ir mazāks nekā cenzēšanas laiks Y_i. Datus ar šādu cenzēšanas shēmu attēlosim līdzīgi kā iepriekš ar gadījumu lielumu pāriem (Z_i, ε_i), kur ε_i ir cenzēšanas indikators ($\varepsilon_i = 1$, ja Z_i atbilst īstajam izdzīvošanas laikam, $\varepsilon_i = 0$, ja izdzīvošanas laiks ir cenzēts), bet $Z_i = \max(X_i, Y_i)$.

Intervālu cenzēšana ir vispārīgāks cenzēšanas veids, kas rodas, ja izdzīvošanas laiks ietilpst kādā noteiktā intervālā. Tas var notikt, ja pētījuma gaitā tiek veikti periodiski subjektu apsekojumi, piemēram, pirmajā apsekojumā subjekts bija vesels, bet nākamajā jau bija slims ar pētāmo slimību.

Ir skaidrs, ka pētījuma gaitā var sastapties ar jebkuru cenzēšanas veidu kombināciju, tomēr intervālu cenzēšanu var uztvert kā divu iepriekšminēto cenzēšanas veidu vispārinājumu. Ja kreisais intervāla galapunkts ir 0, bet labais galapunkts ir Y_i, tad mums ir gadījums ar cenzēšanu no kreisās puses, savukārt, ja kreisais galapunkts ir Y_i, bet labais ir bezgalība, tad pastāv cenzēšana no labās puses.

Novērojumu nošķelšana

Novērojumu nošķelšanu ir viegli sajaukt ar cenzēšanu, taču nošķelšanas gadījumā pētījumā tiek iekļauti tikai tie subjekti, kuru izdzīvošanas laiks iekļaujas kādā noteiktā laika intervālā (I_1, I_2). Par subjektiem, kuru izdzīvošanas laiks sniedzas ārpus šī intervāla, pētniekam nav nekādas informācijas, cenzēšanas gadījumā vismaz ir pieejama daļēja informācija. Kad I_2 ir bezgalība, mēs saskaramies ar nošķelšanu no kreisās puses, savukārt nošķelšana no labās puses ir gadījumā, kad I_1 ir 0.

1.4 Neparametriskie novērtējumi

Šajā nodaļā apskatīsim dažus novērtējumus bāzētus uz izlasi ar cenzēšanu no labās puses, kas ļauj izdarīt secinājumus par laika X līdz kādam notikumam sadalījumu. Pieņemsim, ka izdzīvošanas un cenzēšanas laiki ir neatkarīgi, kā arī mums ir k atšķirīgi novērojumi $t_1 < t_2 < \ldots < t_k$, un laikā t_i ir novēroti d_i notikumi. Savukārt ar r_i apzīmēsim subjektu skaitu, kas laikā t_i atrodas riska grupā, t.i., vēl nav piedzīvojuši notikumu līdz laikam t_i.

Definīcija 7. Izdzīvošanas funkcijas $S(t)$ Kaplāna-Meijera (1958) novērtējums ir

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{r_i}\right), \quad 0 \leq t \leq t_k. \tag{1.41}$$

Apgalvojums 1. *Kaplāna-Meijera novērtējumu var pārrakstīt alternatīvā formā*

$$1 - \hat{F}(t) = \prod_{i=1}^{n} \left[\frac{n-i}{n-i+1}\right]^{I(Z_{(i)} \leq t, \delta_{(i)}=1)}, \quad \textit{visiem} \quad t \leq Z_{(n)}, \tag{1.42}$$

kur $Z_{(1)} \leq ... \leq Z_{(n)}$ ir sakārtoti dati Z_i, bet $\delta_{(i)}$ ir cenzēšanas indikators, kas asociēts ar novērojumu $Z_{(i)}$. Turklāt definējam $1 - \hat{F}(t) = 0$, kad $t > Z_{(n)}$.

Formu (1.42) izmantosim 3.2 nodaļā, bet šajā nodaļā turpmāk izmantosim formulējumu (1.41). Kā redzams, Kaplāna-Meijera novērtējums ir pakāpienveida funkcija ar lēcieniem novērotajos laikos. Lēcienu lielums atkarīgs ne tikai no novēroto notikumu skaita laikā t_i, bet arī no cenzēto novērojumu skaita pirms t_i. Gadījumā, kad dati nesatur cenzētus novērojumus, Kaplāna-Meijera novērtējums reducējas uz empīrisko izdzīvošanas funkciju.

Apgalvojums 2. *Grīnvuda formula Kaplāna-Meijera novērtējuma dispersijas novērtēšanai ir*

$$\hat{V}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i:t_i \leq t} \frac{d_i}{r_i(r_i - d_i)},$$

no kurienes standartnovirze Kaplāna-Meijera novērtējumam ir $(\hat{V}(\hat{S})(t))^{1/2}$.

Kaplāna-Meijera novērtējumu var pielietot arī kumulatīvās riska funkcijas $H(t)$ novērtēšanai, izmantojot saistību ar izdzīvošanas funkciju $\hat{H}(t) = -\ln(\hat{S}(t))$. Cits kumulatīvās riska funkcijas novērtējums ir tā saucamais Nelsona-Ālena (1972, 1978) novērtējums, kas labāk piemērots mazām izlasēm.

Definīcija 8. Kumulatīvās riska funkcijas $H(t)$ Nelsona-Ālena novērtējums ir

$$\tilde{H}(t) = \sum_{i:t_i \le t} \frac{d_i}{r_i}, \quad 0 \le t \le t_k. \tag{1.43}$$

Apgalvojums 3. *Nelsona-Ālena novērtējuma dispersijas novērtējums ir*

$$\sigma_H^2(t) = \sum_{i:t_i \le t} \frac{d_i}{r_i^2}.$$

Balstoties uz $\tilde{H}(t)$, var iegūt arī izdzīvošanas funkcijas novērtējumu formā $\tilde{S}(t) = \exp(-\tilde{H}(t))$.

Nelsona-Ālena novērtējumam ir divi galvenie pielietojumi, analizējot datus. Pirmkārt, tas palīdz izvēlēties piemērotāko parametrisko modeli, kas apraksta datus, piemēram, attēlojot $\tilde{H}(t)$ pret t būs aptuveni lineārs grafiks, ja datiem atbilst eksponenciālais sadalījums ar riska funkciju λ. Otrkārt, Nelsona-Ālena novērtējums ļauj iegūt "rupjus" riska funkcijas $h(t)$ novērtējumus. Labākus $h(t)$ novērtējumus var iegūt, nogludinot Nelsona-Ālena novērtējuma lēcienus ar kodolu metodi (skatīt, piemēram, [3]).

Punktveida ticamības intervāli izdzīvošanas funkcijai $S(t)$

Izmantosim iepriekš aprakstīto Kaplāna-Meijera izdzīvošanas funkcijas novērtējumu un tā standartnovirzi, lai konstruētu ticamības intervālu ar nozīmības līmeni $1-\alpha$ izdzīvošanas funkcijai $S(t)$ fiksētā laikā t_0.

Apzīmēsim ar $\sigma_S^2 = \hat{V}(\hat{S}(t))/\hat{S}^2(t)$. Biežāk lietotie $100(1-\alpha)\%$ ticamības intervāli izdzīvošanas funkcijai fiksētā laikā t_0 tiek saukti par lineārajiem ticamības intervāliem un ir definēti formā

$$\left[\hat{S}(t_0) - Z_{1-\alpha/2}\sigma_S(t_0)\hat{S}(t_0),\ \hat{S}(t_0) + Z_{1-\alpha/2}\sigma_S(t_0)\hat{S}(t_0)\right],$$

kur $Z_{1-\alpha/2}$ ir $1-\alpha/2$ standartnormālā sadalījuma kvantile.

Labākus ticamības intervālus iespējams konstruēt, vispirms transformējot $S(t_0)$. Viena no iespējām ir logaritmiskā transformācija, kas dod $100(1-\alpha)\%$ ticamības intervālus izdzīvošanas funkcijai laikā t_0 formā

$$\left[\hat{S}(t_0)^{1/\theta},\ \hat{S}(t_0)^{\theta}\right], \quad \text{kur} \quad \theta = \exp\left\{\frac{Z_{1-\alpha/2}\sigma_S(t_0)}{\ln(\hat{S}(t_0))}\right\}.$$

Var ievērot, ka logaritmiskā transformācija dod intervālus, kas nav simetriski ap izdzīvošanas

funkcijas novērtējumu $\hat{S}(t_0)$.

Vienlaicīgās ticamības joslas izdzīvošanas funkcijai $S(t)$

Punktveida ticamības intervāli ir noderīgi, ja interesējamies tikai par kādu fiksētu laiku t_0, taču bieži vien praktiskos pielietojumos pētniekus interesē atrast augšējo un apakšējo robežu, kurā iekļaujas īstā izdzīvošanas funkcija ar zināmu nozīmības līmeni visos laika momentos t, t.i., mēs gribam atrast divas funkcijas $L(t)$ un $U(t)$ tā, ka $1-\alpha = P(L(t) \leq S(t) \leq U(t),$ katram $t_L \leq t \leq t_U)$. $[L(t), U(t)]$ sauksim par $100(1-\alpha)\%$ vienlaicīgo ticamības joslu izdzīvošanas funkcijai $S(t)$.

Vispirms izvēlēsimies $t_L < t_U$ tā, lai t_L ir lielāks vai vienāds ar t_1, savukārt t_U ir mazāks vai vienāds ar t_k. Pieņemsim, ka n ir izlases apjoms, kurai vēlamies konstruēt ticamības joslas izdzīvošanas funkcijai, tad definēsim

$$a_L = \frac{n\sigma_S^2(t_L)}{1+n\sigma_S^2(t_L)}$$

un

$$a_U = \frac{n\sigma_S^2(t_U)}{1+n\sigma_S^2(t_U)},$$

turklāt tiek prasīts, lai $0 < a_L < a_U < 1$.

Lai konstruētu $100(1-\alpha)\%$ ticamības joslu izdzīvošanas funkcijai $S(t)$ apgabalā $[t_L, t_U]$, vispirms atrodam ticamības koeficientu $c_\alpha(a_L, a_U)$ no tabulas C.3 pielikumā C grāmatā [3]. Tālāk līdzīgi kā punktveida ticamības intervālu gadījumā lineārās ticamības joslas tiek definētas formā

$$[\hat{S}(t) - c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t),\ \hat{S}(t) + c_\alpha(a_L, a_U)\sigma_S(t)\hat{S}(t)]$$

un logaritmiskās transformācijas joslas tiek definētas formā

$$[\hat{S}(t)^{1/\theta},\ \hat{S}(t)^{\theta}], \quad \text{kur} \quad \theta = \exp\left\{\frac{c_\alpha(a_L, a_U)\sigma_S(t)}{\ln(\hat{S}(t))}\right\}.$$

Vidējā izdzīvošanas ilguma un izdzīvošanas laika mediānas novērtējumi

Vidējais izdzīvošanas ilgums un izdzīvošanas laika mediāna var tikt izteikti kā funkcijas no izdzīvošanas funkcijas $S(t)$. Šo lielumu neparametriskie novērtējumi var tikt iegūti vienkāršā veidā - aizstājot nezināmo izdzīvošanas funkciju ar tās Kaplāna-Meijera novērtējumu attiecīgajā formulā.

Atcerēsimies, ka vidējais izdzīvošanas laiks ir $\mu = \int_0^\infty S(t)dt$, un novērtējumu iegūsim, aizstājot $S(t)$ ar $\hat{S}(t)$. Jāatzīmē, ka šāds novērtējums ir piemērots tikai gadījumā, kad lielākais novērotais izdzīvošanas laiks nav cenzēts, pretējā gadījumā to var mākslīgi pārveidot par izdzīvošanas laiku vai arī izvēlēties intervālu $[0, \tau]$, kur τ ir pētnieka izvēlēts maksimālais iespējamais izdzīvošanas laiks. Tad novērtētais vidējais izdzīvošanas laiks ir

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(t)dt,$$

kur τ ir maksimālais novērojums vai arī pētnieka noteiktā maksimālā robeža. Šī novērtējuma dispersija ir

$$\hat{V}(\hat{\mu}_\tau) = \sum_{i=1}^{k} \left[\int_{t_i}^{\tau} \hat{S}(t)dt \right]^2 \frac{d_i}{r_i(r_i - d_i)}.$$

$100(1-\alpha)\%$ ticamības intervāls vidējam izdzīvošanas laikam μ ir formā

$$\left[\hat{\mu}_\tau - Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\mu}_\tau)},\ \hat{\mu}_\tau + Z_{1-\alpha/2}\sqrt{\hat{V}(\hat{\mu}_\tau)} \right].$$

Piezīme 4. *Gadījumā, kad nav cenzēšanas, vidējā izdzīvošanas ilguma novērtējums reducējas uz izlases vidējo vērtību.*

Kaplāna-Meijera novērtējums var tikt izmantots arī sadalījuma kvantiļu x_p novērtējumu iegūšanai. Atgādināsim, ka $x_p = \inf\{t : S(t) \leq 1-p\}$. Kad $p = 1/2$, x_p ir izdzīvošanas laika mediāna, kuras novērtējums ir $\hat{x}_p = \inf\{t : \hat{S}(t) \leq 1-p\}$. $100(1-\alpha)\%$ ticamības intervāls kvantilei x_p, bāzēts uz izdzīvošanas funkcijas lineāro ticamības intervālu, ir visu laika punktu t kopa, kas apmierina nosacījumu

$$-Z_{1-\alpha/2} \leq \frac{\hat{S}(t) - (1-p)}{\hat{V}^{1/2}(\hat{S}(t))} \leq Z_{1-\alpha/2}.$$

Savukārt, $100(1-\alpha)\%$ ticamības intervāls kvantilei x_p balstīts uz logaritmisko transformāciju ir visu punktu t kopa, kas apmierina nosacījumu

$$-Z_{1-\alpha/2} \leq \frac{\Big(\ln\{-\ln[\hat{S}(t)]\} - \ln\{-\ln[1-p]\}\Big)\Big(\hat{S}(t)\ln[\hat{S}(t)]\Big)}{\hat{V}^{1/2}(\hat{S}(t))} \leq Z_{1-\alpha/2}.$$

2 Empīriskās ticamības funkcijas metode

Šajā nodaļā īsi apskatīsim empīriskās ticamības funkcijas metodi, par kuras pamatlicēju tiek uzskatīts Ovens ([8],[9]) un metodes vispārinājumu vienas izlases gadījumā, kas pieļauj traucējošo (*nuisance*) parametru klātbūtni izmantotajās novērtējošajās funkcijās.

Lai definētu empīriskās ticamības funkcijas metodi, pieņemsim, ka X_1,...,X_n ir iid gadījuma lieluma izlase ar sadalījuma funkciju F.

Definīcija 9. Funkcijas F empīriskā (neparametriskā) ticamības funkcija ir

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

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

Acīmredzami, ka $L(F)$ nav 0 tikai tādiem sadalījumiem ar atomiem punktos X_i, un $L(F)$ ir varbūtība iegūt tieši dotās izlases $X_1, \ldots, X_n$ novērojumu vērtības.

Zināms, ka neparametrisko ticamības funkciju maksimizē empīriskā sadalījuma funkcija $F_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i < x)$ [9], tātad F_n ir sadalījuma funkcijas F neparametriskais vislielākās ticamības novērtējums. Empīriskās ticamības funkcijas attiecība var tikt definēta sekojoši

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

Pieņemsim, ka mēs interesējamies par kādu p-dimensionālu parametru θ asociētu 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 $E_F m_j(x, \theta) = 0$.

Tālāk definēsim profila empīrisko ticamības funkciju

$$\mathrm{EL}(\theta) = \max \left\{ \prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m(X_i, \theta) = 0 \right\}, \tag{2.01}$$

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 (2.01) 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

$$l_E(\theta) = \ln \mathrm{EL}(\theta) = \sum_{i=1}^{n} \ln\{1 + \lambda^T m(X_i,\theta)\}, \tag{2.02}$$

kur λ ir Lagranža reizinātāju vektors.

Qin un Lawless savā darbā [10] apskata nosacījumus, kuriem izpildoties, var pierādīt, ka empīriskās ticamības attiecības statistika hipotēzei $H_0 : \theta = \theta_0$ ir

$$W_E(\theta_0) = 2l_E(\theta_0) - 2l_E(\widehat{\theta}) \to_d \chi^2_p, \quad \text{kad } n \to \infty,$$

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

Piemērs 5. Lai konstruētu ticamības intervālus ar empīriskās ticamības funkcijas metodi sadalījuma funkcijas F vidējai vērtībai $\mu = EX = \int_{-\infty}^{+\infty} x dF(x)$, izmantosim profila empīriskās ticamības attiecības funkciju

$$\mathrm{EL}(\mu) = \max\left\{ \prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu \right\}. \tag{2.03}$$

Pielietojot Lagranža reizinātāju metodi, iegūstam

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

Pie zināmiem nosacījumiem Ovens [8] pierādīja, ka $-2\ln \mathrm{EL}(\mu_0) \to_d \chi^2_1$, kad $n \to \infty$ un $\mu = \mu_0$.

Lai atļautu traucējošo (*nuisance*) parametru klātbūtni novērtējošajās funkcijās, nepieciešams vispārināt EL metodi. 2009. gadā to izdarīja Hjort, McKeague un Van Keilegom ([6]), ieviešot 4 nosacījumus, kuriem izpildoties, var noteikt statistikas robežsadalījumu. Traucējošo parametru aizvietošana ar novērtējumiem pirms tam jau tika izmantota vairākos izdzīvošanas datu analīzes kontekstos (Qin un Jing [11], Wang un Jing [2]), ko apskatīsim vēlāk.

Šajā gadījumā novērtējošās funkcijas dotas formā $m(X,\theta,h)$, kur h ir traucējošais parametrs

ar nezināmu īsto vērtību h_0, savukārt tā novērtējumu apzīmēsim ar $\hat{h}$.

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

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

un atrast ticamības intervālus parametram θ formā $\{\theta : \mathrm{EL}_n(\theta, h_0) > c\}$, kur statistikas robežsadalījums apskatīts [10].

Lietosim 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 (parasti $a_n = 1$) un U - nedeģenerēts p-dimensionāls gadījuma vektors. V_2 - $p \times p$ pozitīvi definīta kovariāciju matrica.

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

(A1) $n^{-1/2} \sum_{i=1}^{n} m(X_i, \theta_0, \hat{h}) \to_d U$.

(A2) $a_n n^{-1} \sum_{i=1}^{n} m^{\otimes 2}(X_i, \theta_0, \hat{h}) \to_p V_2$.

(A3) $a_n \max_{1 \le i \le n} ||n^{-1/2} m(X_i, \theta_0, \hat{h})|| \to_p 0$.

Kad nezināmais h_0 tiek aizstāts ar novērtējumu $\hat{h}$, izpildoties nosacījumiem $(A0)$ - $(A3)$, iespējams vispārināt Qin un Lawless [10] iegūtos rezultātus.

Teorēma 5. *[6] Ja* $(A0)$ - $(A3)$ *ir spēkā, tad* $-2a_n^{-1} \log EL_n(\theta_0, \widehat{h}) \to_d U^T V_2^{-1} U$.

Piemērs 6. Apskatīsim nosacījumus $(A0)$ - $(A3)$ vienkāršākajā gadījumā – vidējai vērtībai, t.i., kad $m(X_i, \theta, h) = X_i - \mu$. Pieņemsim, ka $a_n = 1$, U ir normāli sadalīts gadījuma lielums ar vidējo vērtību 0 un dispersiju σ^2, un $V_2 = \sigma^2$, jo nav nekādu traucējošo parametru, kurus būtu nepieciešams novērtēt.

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

$(A1)$ seko no Centrālās robežteorēmas [12]:

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

Kā redzams, arī $(A2)$ izpildās, izmantojot Lielo skaitļu likumu [12]:

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

Savukārt nosacījums $(A3)$ ir spēkā, jo

$$\max_{1 \le i \le n} |m(X_i, \theta_0, \hat{h})/\sqrt{n}| = \max_{1 \le i \le n} |(X_i - \mu)/\sqrt{n}|, \tag{2.04}$$

un tā kā $\max_{1 \le i \le n} |X_i| = o_{pr}(\sqrt{n})$ [8], tad (2.04) pēc varbūtības tiecas uz 0.

Redzams, ka Teorēma 5 dod rezultātu formā $-2 \log EL_n(\theta_0, \widehat{h}) \to_d U^2/\sigma^2$, kas sakrīt ar Ovena [8] rezultātu, tā kā $U^2 \sim \chi_1^2$, jo $U \sim N(0, 1)$.

Piemērs 7. [6] Pieņemsim, ka X_1, …, X_n iid ar nezināmu sadalījuma blīvuma funkciju f_0, kas ir vienmērīgi nepārtraukta, bet nav vienmērīgā sadalījuma. Lielums $\theta_0 = \int f_0^2 dx$ ir bieži pielietots interesējošais parametrs dažādās problēmās, kas saistītas ar neparametrisko blīvuma funkcijas novērtēšanu, piemēram, Hodges-Lehmann lokācijas novērtējuma asimptotiskā sadalījuma dispersija ir proporcionāla lielumam $1/\theta_0^2$.

Par 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.

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

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

Definēsim

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

kas ir $\sum_{i=1}^{n} m(X_i, \theta_0, f_0)/\sqrt{n}$ asimptotiskā dispersija. (A2) ir spēkā, kad $V_2 = V$, savukārt (A1) ir spēkā ar $U \sim N(0, 4V)$, pie nosacījumiem $\sqrt{n}b \to \infty$ un $\sqrt{n}b^2 \to 0$. (A0) un (A3) apskatīti [6]. Rezultātā varam secināt, ka

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

kas ir piemērs tam, ka traucējošo parametru novērtēšana var mainīt statistikas robežsadalījumu.

3 EL metode izdzīvošanas datiem

Kopš Ovens 1988. gadā pirmoreiz aprakstīja empīriskās ticamības funkcijas metodi, tā kļuvusi par ļoti populāru un plaši pielietotu neparametriskās statistikas metodi iid datiem, lai gan jāatzīmē, ka pirmoreiz tā minēta Thomas un Grunkemeier 1975. gada publikācijā [1] tieši izdzīvošanas analīzes kontekstā. Ilgu laiku šī metode tika atstāta novārtā, jo cenzēšanas klātbūtne radīja sarežģījumus tās pielietošanā.

Pieņemsim, ka ir dots iid gadījuma lielums X_1,...,X_n ar nezināmu sadalījuma funkciju F un izdzīvošanas funkciju S, kas satur izdzīvošanas laikus, savukārt Y_1,...,Y_n arī iid gadījuma lielums ar sadalījuma funkciju G, kas satur cenzēšanas laikus. Tiek pieņemts, ka X_i un Y_i ir neatkarīgi. Gadījuma cenzēšanas modelī patiesie izdzīvošanas laiki X_i,...,X_n nav zināmi, tā vietā mums ir datu pāri $Z_i = \min(X_i, Y_i)$ un $\delta_i = I(X_i < Y_i)$, $i = 1, \ldots, n$. Pieņemsim tāpat kā iepriekš, ka ir novēroti k atšķirīgi izdzīvošanas laiki $0 = t_0 < t_1 < \ldots < t_k < t_{k+1} = \infty$.

3.1 Tradicionālā EL pieeja izdzīvošanas datiem

Jau 1958. gadā [13] Kaplan un Meier parādīja, ka Kaplāna-Meijera izdzīvošanas funkcijas novērtējums formā (1.41) vai (1.42) ir neparametriskais vislielākās ticamības novērtējums, kas maksimizē neparametrisko ticamības funkciju

$$L(S) = \prod_{\delta_i=1} [S(Z_i-) - S(Z_i)] \prod_{\delta_i=0} S(Z_i)$$

pār bezgalīgi dimensionālu parametru telpu $\Theta \equiv \{$*visas izdzīvošanas funkcijas intervālā* $[0, \infty]\}$. Tika pierādīts, ka Θ vietā pietiek izvēlēties $\Theta_t \equiv \{$*visas diskrētās izdzīvošanas funkcijas punktos* $\{t_1, \ldots, t_k\}\}$, turklāt katru $S \in \Theta_t$ var pārrakstīt formā $S(t) = \prod_{j:t_j \le t}(1 - h_j)$, $t \ge 0$, kur $h_j = 1 - S(t_j)/S(t_{j-1})$ visiem $1 \le j \le k$, un tādām S ticamības funkcija $L(S)$ reducējas uz

vienkāršāku formu

$$L(S) = \prod_{j=1}^{k} h_j^{d_j}(1-h_j)^{r_j-d_j},$$

kuru maksimizē $h_j = d_j/r_j$ visiem j. Tā kā mērķis bija atrast ticamības intervāla novērtējumu $S(a)$, kur $a > 0$ fiksēts, tad tika parādīts, ka novērtējums $\hat{S}(a)$ ir asimptotiski normāls un tā dispersijas būtisku novērtējumu dod jau minētā Grīnvuda formula.

Lai gan ar Grīnvuda formulu iegūtie intervālu novērtējumi labi strādā lielām izlasēm, tomēr tika novērots, ka rezultāti nav apmierinoši mazām izlasēm, jo var iekļaut vērtības ārpus intervāla $[0,1]$. 1975. gadā kā alternatīvu hipotēzei $H_0 : S(a) = p$ Thomas un Grunkemeier [1] piedāvāja empīriskās ticamības funkcijas attiecību

$$R(p) = \frac{\sup\left\{ \prod_{j=1}^{k} h_j^{d_j}(1-h_j)^{r_j-d_j} \Big| \prod_{j:t_j \le a} (1-h_j) = p, \text{ un } 0<h_j\le 1,\ 1\le j\le k \right\}}{\prod_{j=1}^{k} \left(\frac{d_j}{r_j}\right)^{d_j} \left(1-\frac{d_j}{r_j}\right)^{r_j-d_j}},$$

un $100(1-\alpha)\%$ ticamības intervālu $S(a)$ formā $\{p : -2\ln R(p) \le \chi_1^2(\alpha)\}$, kur $\chi_1^2(\alpha)$ ir $1-\alpha$ kvantile sadalījumam χ_1^2, turklāt intervāli vienmēr ir $[0,1]$ apakšintervāli.

Pateicoties Thomas un Grunkemeier rezultātiem, sākās neparametriskās ticamības funkcijas metodes strauja attīstība, tomēr ne cenzētiem, bet gan iid datiem. Tikai 1995. gadā Li [5] sniedza stingru metodes pierādījumu ticamības intervālu konstruēšanai izdzīvošanas varbūtībām, kā arī parādīja, ka Thomas un Grunkemeier piedāvātais rezultāts nav tiešas sekas Kaplan un Meier rezultātiem.

Li [5] arī pierādīja, ka

$$R(p_1,\ldots,p_J) = \frac{\sup\{L(S)|S(a_1)=p_1,\ldots,S(a_J)=p_J \text{ un } S\in\Theta\}}{\sup\{L(S)|S\in\Theta\}}$$

un $-2\ln R(p_1,\ldots,p_J \to^d \chi_J^2)$, kad $n\to\infty$, kas ļauj izdarīt secinājumus par izvēlētās izdzīvošanas funkcijas atbilstību datiem.

3.2 EL metode ar novērojumu modifikāciju

Ja pārrakstām empīriskās ticamības attiecības funkciju formā

$$R(\theta_0) = \frac{\sup\{L(S)|\theta(S) = \theta_0\}}{\sup\{L(S)\}},$$

tad ilgu laiku bija zināmas tikai dažas $\theta(S)$ (vai $\theta(H)$) vienkārši pielietojamas formas. Šajā nodaļā apskatīsim Wang un Jing [2] piedāvāto empīriskās ticamības funkcijas metodi, kas ļauj izdarīt secinājumus daudz plašākai parametru klasei.

Mēs interesēsimies par sekojošu funkcionāli no sadalījuma funkcijas $F(\cdot)$ formā

$$\theta(F) = \int_0^\infty \xi(t)dF(t),$$

kur $\xi(t)$ ir kāda (nenegatīva) mērojama funkcija ar $E\xi(X) < \infty$.

Piemērs 8. Ja $\xi(t) = t$, tad $\theta(F)$ ir vidējais izdzīvošanas ilgums.

Piemērs 9. Ja $\xi(t) = I(t \geq t_0)$, tad $\theta(F)$ ir izdzīvošanas varbūtība $1 - F(t_0)$ fiksētā punktā t_0.

Piemērs 10. Ja $\xi(t) = I(t \leq t_0)/[1 - F(t)]$, tad $\theta(F)$ ir kumulatīvā riska funkcija $\int_0^{t_0}(1 - F(t))^{-1}dF(t)$ fiksētā punktā t_0.

Aizstājot $F(t)$ ar tās Kaplāna-Meijera novērtējumu $\hat{F}(t)$ formā (1.42), iegūstam $\theta(F)$ novērtējumu

$$\hat{\theta} = \theta(\hat{F}) = \int_0^{Z_{(n)}} \xi(t)d\hat{F}(t). \tag{3.21}$$

Šī novērtējuma asimptotiskās īpašības jau ir vairākkārt pētītas, piemēram, Stute [14] pierādīja, ka, uzliekot noteiktus nosacījumus uz $\xi(\cdot)$, $\sqrt{n}(\theta(\hat{F}) - \theta(F))$ sadalījums ir asimptotiski normāls ar vidējo vērtību 0 un kādu dispersiju σ^2. Lai aprēķinātu ticamības intervālus parametram $\theta(F)$, nepieciešams novērtēt asimptotisko dispersiju σ^2, kas ir matemātiski sarežģītā izskatā (skatīt Lemmu 7). Wang un Jing savā publikācijā [2] piedāvā σ^2 novērtēšanai izmantot džeknaifa metodi, kas dod būtisku asimptotiskās dispersijas novērtējumu [15].

Lemma 6.

$$\theta(F) = E(\xi(X)) = E\left(\frac{\xi(Z)\delta}{1 - G(Z)}\right).$$

Pierādījums.

$$\begin{aligned} E\left(\frac{\xi(Z_i)\delta_i}{1-G(Z_i)}\right) &= E\left[\frac{\xi(\min(X_i,Y_i))I(X_i<Y_i)}{1-G(\min(X_i,Y_i))}\right] \\ &= \int\int_{x<y} \frac{\xi(x)}{1-G(x)} dF(x)dG(y) \\ &= \int_0^\infty \frac{\xi(x)}{1-G(x)} \int_x^\infty dG(y)dF(x) \\ &= \int_0^\infty \frac{\xi(x)}{1-G(x)}(1-G(x))dF(x) \\ &= \int_0^\infty \xi(x)dF(x) \\ &= \theta(F), \quad i=1,\ldots,n. \end{aligned}$$

□

Tātad, problēma pārbaudīt, vai θ_0 ir $\theta(F)$ īstā vērtība, ir ekvivalenta problēmai pārbaudīt, vai $E\left(\frac{\xi(Z_i)\delta_i}{1-G(Z_i)}\right) = \theta_0$, $i=1,\ldots,n$. To var izdarīt, pielietojot empīriskās ticamības funkcijas metodi, ko ieviesa Owen ([8]). Pieņemsim, ka F_p ir sadalījuma funkcija, kas uzliek varbūtību p_i datu punktam $\frac{\xi(Z_i)\delta_i}{1-G(Z_i)}$, tad

$$\theta(F_p) = \sum_{i=1}^n p_i \left(\frac{\xi(Z_i)\delta_i}{1-G(Z_i)}\right).$$

Tā kā $G(\cdot)$ ir nezināma $\theta(F_p)$ definīcijā, aizstāsim to ar tās Kaplan-Meier novērtējumu $\hat{G}_n(t)$ formā

$$1-\hat{G}_n(t) = \prod_{i=1}^n \left[\frac{n-i}{n-i+1}\right]^{I(Z_{(i)}\le t, \delta_{(i)}=0)}.$$

Tālāk varam definēt empīriskās ticamības funkciju punktā θ_0

$$L(\theta_0) = \max \prod_{i=1}^n p_i$$

pie ierobežojumiem

$$p_i > 0, \quad i=1,\ldots,n, \quad \sum_{i=1}^n p_i = 1, \quad \text{un} \quad \sum_{i=1}^n p_i \left(\frac{\xi(Z_i)\delta_i}{1-\hat{G}_n(Z_i)}\right) = \theta_0.$$

Savukārt, pielietojot Lagranža reizinātāju metodi tāpat kā iepriekšējā nodaļā, empīriskās tica-

mības attiecības funkcija tiek definēta sekojoši

$$R(\theta_0) = \prod_{i=1}^{n} np_i = \prod_{i=1}^{n} \left(1 + \lambda(V_{ni} - \theta_0)\right),$$

kur $V_{ni} = \frac{\xi(Z_i)\delta_i}{1-\hat{G}_n(Z_i)}$ ir modificētie novērojumi un λ ir atrisinājums vienādojumam

$$\frac{1}{n}\sum_{i=1}^{n} \frac{(V_{ni} - \theta_0)}{1 + \lambda(V_{ni} - \theta_0)} = 0.$$

Attiecīgā empīriskās ticamības funkcijas metodes statistika ir

$$l(\theta_0) = -2\ln R = 2\sum_{i=1}^{n} \ln\{1 + \lambda(V_{ni} - \theta_0)\}.$$

Apzīmēsim $W_{ni} = V_{ni} - \theta_0$ un $\bar{W}_n = n^{-1}\sum_{i=1}^{n} W_{ni}$, kā arī

$$\gamma_1(x) = \frac{1}{\bar{H}(x)} \int I(x < s)\xi(s)\gamma_0(s)d\tilde{H}_1(s),$$
$$\gamma_2(x) = \iint \frac{I(s < x, s < t)\xi(t)\gamma_0(t)}{\bar{H}^2(s)} d\tilde{H}_0(s)d\tilde{H}_1(t)$$

Definēsim

$H(s) = P(Z \le s), \quad \bar{H}(s) = P(Z_1 > s), \quad \tilde{H}_0(s) = P(Z_1 > s, \delta_1 = 0),$

$\tilde{H}_1(s) = P(Z_1 > s, \delta_1 = 1), \quad \gamma_0(x) = \exp\left\{\int_0^{x-} \frac{d\tilde{H}_0(s)}{\bar{H}(s)}\right\}$, kur $x-$ nozīmē "neieskaitot x",

$C(x) = \int_0^{x-} \frac{dG(s)}{(1-H(s))(1-G(s))}, \quad \tau_H = \inf\{t : H(t) = 1\}.$

Lemma 7. *Pieņemsim, ka izpildās nosacījumi*

(C1) $\int_0^{\tau_H} \xi^2(x)\gamma_0^2(x)d\tilde{H}_1(x) < \infty$,

(C2) $\int_0^{\tau_H} \xi(x)C^{1/2}(x)dF(x) < \infty$,

(C3) $\int_0^{\tau_H} \frac{\xi^2(x)dF(x)}{1-G(x)} < \infty$,

(C4) $\tau_F = \tau_H$ *un* $F(\tau_F) = F(\tau_F-)$.

Tad

$$\sqrt{n}\bar{W}_n \to^d N(0, \sigma^2),$$

kur

$$\sigma^2 = D(\xi(Z_1)\gamma_0(Z_1)\delta_1 + \gamma_1(Z_1)(1-\delta_1) - \gamma_2(Z_1)).$$

Wang un Jing savā publikācijā [2] pierāda, ka pie zināmiem nosacījumiem $rl(\theta_0)$, kur $r = (n^{-1}\sum_{i=1}^{n} W_{ni}^2)/\sigma^2$, ir asimptotiski sadalīts kā χ_1^2. Lai konstruētu ticamības intervālu θ_0, vispirms nepieciešams novērtēt koeficientu r.

Tā kā parametra θ asimptotiskā dispersija ir ļoti sarežģītā teorētiskā formā kā redzams Lemmā 7, arī funkciju F un G aizstāšana ar novērtējumiem dotu teorētiski sarežģītu σ^2 novērtējumu. Alternatīvi σ^2 var novērtēt ar džeknaifa metodi, turklāt Stute [15] parādīja, ka šādā veidā iespējams iegūt būtisku σ^2 novērtējumu.

Džeknaifa procedūra σ^2 novērtēšanai [15]

Integrāli $\int \xi(t)d\hat{F}_n(t)$, kuram vēlamies novērtēt asimptotisko dispersiju, sauksim par Kaplāna-Meijera integrāli. Jāatzīmē, ka Kaplāna-Meijera integrāļi nav iid gadījuma lielumu svērtās summas kā tas būtu,aizstājot sadalījuma funkciju F ar empīrisko sadalījuma funkciju. Apzīmēsim

$$T_i = \frac{\delta_{(i)}}{n-i+1}\prod_{j=1}^{i-1}\left[\frac{n-j}{n-j+1}\right]^{\delta_{(j)}},$$

no kurienes

$$S_n \equiv \int \xi(t)d\hat{F}_n(t) = \sum_{i=1}^{n} T_i\xi(Z_{(i)}).$$

Stute [14] pierādīja centrālo robežteorēmu cenzētiem datiem

$$\sqrt{n}\left(\int \xi(t)d\hat{F}_n(t) - S\right) \to^d N(0,\sigma^2)$$

kur $S \equiv \lim_{n\to\infty}\int \xi(t)d\hat{F}_n(t)$. Taču jāatcerās, ka cenzētiem datiem σ^2 var nesakrist ar dispersijas zināmo formu

$$\sigma_0^2 = \int \xi^2(t)dF(t) - \left(\int \xi(t)dF(t)\right)^2.$$

Protams, iid datu gadījumā $\sigma^2 = \sigma_0^2$.

Apzīmēsim ar $\hat{F}_n^{(k)}$ Kaplāna-Meijera novērtējumu dotai izlasei, no kuras izdzēsts novērojums $(Z_{(k)}, \delta_{(k)})$, tad $S_n^{(k)} \equiv S(\hat{F}_n^{(k)})$ ir tā saucamās pseidovērtības (ar džeknaifa metodi iid

datiem sīkāk var iepazīties, piemēram, [16]). Pseidovērtību vidējā vērtība ir sekojošā formā

$$\bar{S_n} = S_n - \xi(Z_{(n)})\frac{\delta_{(n)}(1-\delta_{(n-1)})}{n}\prod_{i=1}^{n-2}\left[\frac{n-i-1}{n-i}\right]^{\delta_{(i)}}$$

un džeknaifa dispersijas novērtējums

$$n\widehat{D}(jack) = (n-1)\sum_{k=1}^{n}\left(S_n^{(k)} - \bar{S}_n\right)^2.$$

Ievērosim, ka $\bar{S}_n$ nesakrīt ar S_n, ja

$$\delta_{(n-1)} = 0 \text{ un } \delta_{(n)} = 1, \tag{3.22}$$

tad (3.22) gadījumā mākslīgi aizstāsim $\delta_{(n)}$ ar 0, un iegūsim modificēto džeknaifa dispersijas novērtējumu $\widehat{D}^*(jack)$.

Teorēma 8. *[15] Pieņemsim, ka izpildās nosacījums (C1), tad $n\widehat{D}^*(jack) \to \sigma^2$ gandrīz droši.*

r novērtējums un neparametriskā Wilks tipa teorēma cenzētiem datiem

Atgādināsim, ka vēlamies novērtēt koeficientu $r = (n^{-1}\sum_{i=1}^{n} W_{ni}^2)/\sigma^2$, lai konstruētu ticamības intervālu parametram θ. No Stute un Wang [17] zināms, ka

$$\bar{V}_n = \frac{1}{n}\sum_{i=1}^{n} V_{ni} \to \theta_0 \text{ gandrīz droši,}$$

no kurienes iegūstam, ka

$$n^{-1}\sum_{i=1}^{n}(V_{ni} - \bar{V}_n)^2 - n^{-1}\sum_{i=1}^{n} W_{ni}^2 \to 0 \text{ gandrīz droši.}$$

Un tā kā $n\widehat{D}^*(jack) \to \sigma^2$ gandrīz droši, tad koeficienta r novērtējumu iegūstam formā

$$\hat{r} = \frac{n^{-1}\sum_{i=1}^{n}(V_{ni} - \bar{V}_n)^2}{n\widehat{D}^*(jack)}.$$

Sekojošā teorēma ir Wilks teorēmas neparametriskā versija cenzētiem datiem pielāgotai empīriskās ticamības funkcijas metodei. Teorēmas pierādījums atrodams [2].

Teorēma 9. *[2] Ja izpildās nosacījumi (C1)–(C4) un θ_0 ir parametra θ īstā vērtība, tad*

$$\hat{r}l(\theta_0) \to^d \chi^2_1. \tag{3.23}$$

Vienkārša pieeja konstruēt ticamības intervālu parametram θ ar nozīmības līmeni α, balstoties uz (3.23), ir $I_\alpha = \{\theta : \hat{r}l(\theta) \leq c_\alpha\}$, kur $P(\chi^2_1 \leq c_\alpha) \leq 1 - \alpha$.

Piezīme 10. *Logaritmiskā empīriskā ticamības attiecības funkcija $l(\theta_0)$ izskatās līdzīgi kā gadījumā bez datu cenzēšanas, taču jāņem vērā, ka V_{ni}, $i = 1, \ldots, n$ vairs nav neatkarīgi un vienādi sadalīti gadījuma lielumi sadalījuma funkcijas $G(z)$ novērtēšanas dēļ.*

Piezīme 11. *Wang un Jing pierādīja Teorēmu 9 jau 2001. gadā, balstoties uz Ovena 1990. gada publikāciju, taču šī problemātika var tikt uztverta arī kā Hjort, McKeague un Van Keilegom ([6]) empīriskās ticamības funkcijas metodes vispārinājuma speciālgadījumu. Nosacījums (A1) izpildās ar $U \sim N(0, \sigma^2)$, savukārt nosacījums (A2) izpildās ar $V_2 = n^{-1} \sum_{i=1}^n W_{ni}^2$, un Teorēma 5 dod rezultātu, kas sakrīt ar Wang un Jing [2] rezultātu.*

4 Divu izlašu problēmas

Izdzīvošanas datu analīzē bieži ir svarīgi salīdzināt tieši divas izlases, lai noteiktu kādu zāļu vai procedūru ietekmi uz subjektu izdzīvošanas laikiem. Kopš Ovens [8] ieviesa empīriskās ticamības funkcijas metodi, ir bijuši vairāki mēģinājumi to vispārināt arī divu izlašu gadījumam. Vieni no pirmajiem bija Qin un Zhao [7], kas 2000. gadā iepazīstināja ar empīriskās ticamības funkcijas metodi divu izlašu vidējo vērtību starpībai un sadalījuma funkciju starpībai kādā fiksētā punktā. Nedaudz vēlāk Valeinis savā doktora disertācijā [18] un iesniegtajā publikācijā [19] 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 [18].

4.1 Empīriskā ticamības funkcijas metode divām izlasēm

Pieņemsim, ka ir dotas divas neatkarīgas 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 interesējošam parametram Δ, turklāt θ_0 ir traucējošais 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 formā, t.i.,

$$E_{F_1} w_1(X, \theta_0, \Delta) = 0, \tag{4.11}$$

$$E_{F_2} w_2(Y, \theta_0, \Delta) = 0. \tag{4.12}$$

Piemērs 11. Apskatīsim divu izlašu vidējo vērtību starpību. Apzīmēsim $\theta_0 = \int x dF_1(x)$ un $\Delta = \int y dF_2(y) - \int x dF_1(x)$. Lai (4.11) un (4.12) būtu spēkā, izvēlamies

$$w_1(X, \theta_0, \Delta) = X - \theta_0, \quad w_2(Y, \theta_0, \Delta) = Y - \theta_0 - \Delta$$

kā novērtējošās funkcijas.

Piemērs 12. Ja interesējamies par kvantiļu – kvantiļu jeb Q-Q grafiku, tad $\theta_0 = F_2(t)$ un $\Delta = F_1^{-1}(F_2(t))$. Novērtējošās funkcijas ir formā

$$w_1(X, \theta_0, \Delta) = I_{\{X \leq \Delta\}} - \theta_0, \quad w_2(Y, \theta_0, \Delta) = I_{\{Y \leq t\}} - \theta_0.$$

Definīcija 10. $X_1, \ldots, X_n \sim F_1$ iid un $Y_1, \ldots, Y_m \sim F_2$ iid, empīriskā (neparametriskā) 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_{2m} [7]. 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

$$\mathrm{EL}(\Delta, \theta) = \sup_{\theta, p, q} \prod_{i=1}^{n}(np_i) \prod_{j=1}^{m}(mq_j), \tag{4.13}$$

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

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i w_1(X_i, \theta, \Delta) = 0,$$

$$q_i \geq 0, \quad \sum_{i=1}^{m} q_j = 1, \quad \sum_{i=1}^{m} q_j w_2(Y_j, \theta, \Delta) = 0.$$

Ja θ ir dots, tad eksistē viens vienīgs (4.13) atrisinājums, ja 0 pieder $w_1(X_i, \theta, \Delta)$ un $w_2(Y_j, \theta, \Delta)$ izliektajai čaulai.

Maksimizācijas problēmu var atrisināt ar Lagranža reizinātāju metodi. To sīkāk šeit neizrakstīsim, bet apskatīsim publikācijā [7] iegūtos rezultātus. Varam iegūt profila empīriskās

ticamības attiecības funkciju parametram Δ izskatā

$$\mathrm{EL}(\Delta,\theta)=\prod_{i=1}^{n}\left\{\frac{1}{1+\lambda_1(\theta)w_1(X_i,\theta,\Delta)}\right\}\prod_{j=1}^{m}\left\{\frac{1}{1+\lambda_2(\theta)w_2(Y_j,\theta,\Delta)}\right\},$$

kur λ_1 un λ_2 ir Lagranža reizinātāji un atrodami no vienādojumiem

$$\frac{1}{n}\sum_{i=1}^{n}\frac{w_1(X_i,\theta,\Delta)}{1+\lambda_1(\theta)w_1(X_i,\theta,\Delta)}=0, \tag{4.14}$$

$$\frac{1}{m}\sum_{j=1}^{m}\frac{w_2(Y_j,\theta,\Delta)}{1+\lambda_2(\theta)w_2(Y_j,\theta,\Delta)}=0. \tag{4.15}$$

Savukārt logaritmiskā empīriskā ticamības attiecības statistika ir

$$-2\ln\mathrm{EL}(\Delta,\theta)=2\sum_{i=1}^{n}\ln(1+\lambda_1(\theta)w_1(X_i,\theta,\Delta))+2\sum_{j=1}^{m}\ln(1+\lambda_2(\theta)w_2(Y_j,\theta,\Delta)).$$

Pieņemot, ka $\partial w_1(X_i,\theta,\Delta)/\partial\theta$ un $\partial w_2(Y_j,\theta,\Delta)/\partial\theta$ eksistē, apzīmēsim

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

Lai atrastu $\theta=\hat{\theta}$, kas maksimizē $\mathrm{EL}(\Delta)$, pielīdzinām nullei $\mathrm{EL}(\Delta,\theta)$ atvasinājumu pēc θ un iegūstam vienādojumu

$$\frac{1}{n}\lambda_1(\theta)\sum_{i=1}^{n}\frac{\alpha_1(X_i,\theta,\Delta)}{1+\lambda_1(\theta)w_1(X_i,\theta,\Delta)}+\frac{1}{n}\lambda_2(\theta)\sum_{j=1}^{m}\frac{\alpha_2(Y_j,\theta,\Delta)}{1+\lambda_2(\theta)w_2(Y_j,\theta,\Delta)}=0. \tag{4.16}$$

Autori Qin un Zhao [7] ievieš sekojošus pieņēmumus:

(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 integrējamu funkciju $G_1(x)$, savukārt $\alpha_2(y,\theta,\Delta)$ un $w_2^3(y,\theta,\Delta)$ šajā apkārtnē ir ierobežotas ar integrējamu funkciju $G_2(y)$, un $E_{F_1}\alpha_1(x,\theta,\Delta)$ un $E_{f_2}\alpha_2(y,\theta,\Delta)$ nav 0.

(iii) $m/n\to k$, kad $n,m\to\infty$, un $0<k<\infty$.

Teorēma 12. *(Qin un Zhao, [7]) Izpildoties pieņēmumiem (i), (ii), (iii),*

$$-2\ln EL(\Delta_0, \hat{\theta}) \to_d \chi_1^2,$$

kur $\hat{\theta}$ ir parametra θ_0 būtisks novērtējums, kas maksimizē (4.16).

Izmantojot šo rezultātu, varam konstruēt uz EL metodi balstītus ticamības intervālus interesējošajam parametram Δ formā $\{\Delta : -2\ln \mathrm{EL}(\Delta, \hat{\theta}) \leq c_\alpha\}$, kur $P(\chi_1^2 \leq c_\alpha) \leq 1 - \alpha$.

4.2 EL metode ar novērtētiem parametriem divām izlasēm

Līdzīga situācija kā vienas izlases gadījumā ir arī ar empīriskās ticamības funkcijas metodi divu izlašu problēmām izdzīvošanas datiem. Līdz šim ir atrasti robežsadalījumi nedaudziem interesējošiem parametriem, piemēram, divu izdzīvošanas funkciju starpībai un attiecībai, divu riska funkciju starpībai un attiecībai, kā arī kvantiļu-kvantiļu grafikiem gadījuma cenzētiem datiem. Tāpēc šī darba galvenais mērķis ir vispārināt 3.2 nodaļā apskatīto empīriskās ticamības funkcijas metodi divu izlašu problēmām, apvienojot Wang un Jing [2] rezultātu vienas izlases un Qin un Zhao [7] rezultātu divu izlašu gadījumā, turklāt, pielietojot Hjort, McKeague un Van Keilegom [6] ieviestos nosacījumus.

Jāatzīmē, ka Valeinis savā doktora disertācijā [18] jau mēģināja vispārināt empīriskās ticamības funkcijas metodi ar novērtētiem parametriem divām izlasēm, taču galvenais uzdevums bija konstruēt ticamības joslas strukturālajiem attiecību modeļiem, savukārt, mēs gribam vērst uzmanību uz metodes pielietojumu izdzīvošanas analīzē. Cieši sekosim minētās disertācijas notācijai, bet no jauna caurskatīsim galvenos pierādījumus, izlabojot dažas neprecizitātes.

Pieņemsim, ka mums dota līdzīga situācija, kā aprakstīts 4.1 nodaļas sākumā, galvenās atšķirības ir, ka izlases $X_1,\ldots,X_n$ un $Y_1,\ldots,Y_m$ var nebūt iid un novērtējošās funkcijas

$$w_1(X, \theta_0, \Delta, h) \quad \text{un} \quad w_2(Y, \theta_0, \Delta, h)$$

var saturēt traucējošo parametru h ar nezināmu īsto vērtību h_0. Ar $\hat{h}$ apzīmēsim parametra h *plug-in* novērtējumu.

Pieņemsim, ka abām izlasēm izpildās nosacījumi (A0)–(A3) no 2 nodaļas, un nosacījumi (i)–(iii) no 4.1 nodaļas, t.i.,

(A0) Maksimizācijas problēmai (4.16) eksistē atrisinājums.

(A1) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_1(X_i, \theta_0, \Delta, \hat{h}) \to_d U_1$, kur $U_1 \sim N(0, V_1)$;
$\frac{1}{\sqrt{m}} \sum_{j=1}^{m} w_2(Y_j, \theta_0, \Delta, \hat{h}) \to_d U_2$, kur $U_2 \sim N(0, M_1)$.

(A2) $n^{-1} \sum_{i=1}^{n} w_1^2(X_i, \theta_0, \Delta, \hat{h}) \to_p V_2$;
$m^{-1} \sum_{j=1}^{m} w_2^2(Y_j, \theta_0, \Delta, \hat{h}) \to_p M_2$.

(A3) $\max_{1 \le i \le n} ||n^{-1/2} w_1(X_i, \theta_0, \Delta, \hat{h})|| \to_p 0$;
$\max_{1 \le j \le m} ||m^{-1/2} w_2(Y_j, \theta_0, \Delta, \hat{h})|| \to_p 0$.

Papildus pieņemsim, ka

(A4) $n^{-1} \sum_{i=1}^{n} \alpha_1(X_i, \theta_0, \Delta, \hat{h}) \to_p V_3$;
$m^{-1} \sum_{j=1}^{m} \alpha_2(Y_j, \theta_0, \Delta, \hat{h}) \to_p M_3$.

Teorēma 13. *Pieņemsim, ka izpildās nosacījumi (A0)–(A4) un (i)–(iii), tad vienādojumam (4.16) eksistē sakne $\hat{\theta}$, kas ir θ_0 būtisks novērtējums, un*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, \frac{V_1 M_2^2 V_3^2 + k M_1 V_2^2 M_3^2}{(M_2 V_3^2 + k V_2 M_3^2)^2}\right), \tag{4.21}$$

$$-2 \ln EL(\Delta, \hat{\theta}, \hat{h}) \to^d \frac{M_1 V_3^2 + k V_1 M_3^2}{M_2 V_3^2 + k V_2 M_3^2} \chi_1^2, \quad \textit{kad} \quad n \to \infty, \tag{4.22}$$

vai

$$-2 \ln EL(\Delta, \hat{\theta}, \hat{h}) * r \to^d \chi_1^2, \quad \textit{kad} \quad n \to \infty, \quad \textit{kur} \quad r = \frac{M_2 V_3^2 + k V_2 M_3^2}{M_1 V_3^2 + k V_1 M_3^2}. \tag{4.23}$$

Piezīme 14. *Kad $V_1 = V_2$ un $M_1 = M_2$ statistika tiecas uz standarta χ_1^2 sadalījumu. Tāpat kā vienas izlases gadījumā konstanti $r = \frac{M_2 V_3^2 + k V_2 M_3^2}{M_1 V_3^2 + k V_1 M_3^2}$ ir grūti vispārēji komentēt, jo tā var atšķirties katrā pielietojumā. Savukārt, kad ir teorētiski sarežģīti iegūt V_1 un M_1 formas, būtisku novērtējumu iegūšanai var izmantot butstrapa metodi [6].*

Teorēmas 13 pierādījumam nepieciešamas dažas lemmas.

Lemma 15. *Pieņemsim, ka $1/3 < \eta < 1/2$ un Teorēmas 13 nosacījumi ir spēkā, tad*

$$\lambda_1(\theta) = O_p(n^{-\eta}), \tag{4.24}$$

$$\lambda_2(\theta) = O_p(n^{-\eta}) \tag{4.25}$$

vienmērīgi kādā apkārtnē $\theta \in \{\theta : |\theta - \theta_0| \leq cn^{-\eta}\}$*, kur* c *ir kāda pozitīva konstante.*

Lemma 16. *Pieņemsim, ka Teorēmas 13 nosacījumi ir spēkā un dots* η *kā Lemmā 15. Tad ar varbūtību, kas tiecas uz 1, vienādojumam (4.16) eksistē sakne* $\hat{\theta}$ *tāda, ka*

$$|\hat{\theta} - \theta_0| = O_p(n^{-\eta}).$$

Lemmas 15 un Lemmas 16 pierādījumu skatīt [7] vai [18].

Lemma 17. *Pieņemsim, ka Teorēmas 13 nosacījumi izpildās, tad ar* $\hat{\theta}$ *kā Lemmā 16 ir spēkā*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N\left(0, \frac{V_1 M_2 V_3^2 + k M_1 V_2 M_3^2}{c_1^2}\right), \tag{4.26}$$

$$\lambda_1(\hat{\theta}) = -k\left(\frac{M_3}{V_3}\right)\lambda_2(\hat{\theta}) + o_p(n^{-1/2}), \tag{4.27}$$

$$\sqrt{n}\lambda_2(\hat{\theta}) \to^d N\left(0, \frac{V_3^2(M_1 V_3^2 + k V_1 M_3^2)}{k c_1^2}\right), \tag{4.28}$$

kur $c_1 = M_2 V_3^2 + k V_2 M_3^2$.

Pierādījums. Apzīmēsim $\lambda_1(\theta) = \lambda_1$, $\hat{\lambda}_1 = \lambda_1(\hat{\theta})$, $\lambda_2(\theta) = \lambda_2$, $\hat{\lambda}_2 = \lambda_2(\hat{\theta})$, un

$$Q_{1n}(\theta, \lambda_1, \lambda_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{w_1(X_i, \theta, \Delta)}{1 + \lambda_1 w_1(X_i, \theta, \Delta)},$$

$$Q_{2n}(\theta, \lambda_1, \lambda_2) = \frac{1}{m}\sum_{j=1}^{m} \frac{w_2(Y_j, \theta, \Delta)}{1 + \lambda_2 w_2(Y_j, \theta, \Delta)},$$

$$Q_{3n}(\theta, \lambda_1, \lambda_2) = \lambda_1 \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_1(X_i, \theta, \Delta)}{1 + \lambda_1 w_1(X_i, \theta, \Delta)} + \lambda_2 \frac{1}{n}\sum_{j=1}^{m} \frac{\alpha_2(Y_j, \theta, \Delta)}{1 + \lambda_2 w_2(Y_j, \theta, \Delta)}.$$

No Lemmas 16 ir spēkā

$$Q_{in}(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2) = 0 \quad \text{katram} \quad i = 1, 2, 3.$$

No Teilora izvirzījuma, Lemmas 15 un Lemmas 16 ir spēkā

$$0 = Q_{in}(\hat{\theta}, \hat{\lambda}_1, \hat{\lambda}_2) = Q_{in}(\theta_0, 0, 0) + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \theta}(\hat{\theta} - \theta_0) + \\ + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \lambda_1}\hat{\lambda}_1 + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \lambda_2}\hat{\lambda}_2 + O_p(n^{-2\eta}), \quad i = 1, 2, 3.$$

Tātad

$$Q_{in}(\theta_0, 0, 0) + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \theta}(\hat{\theta} - \theta_0) + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \lambda_1}\hat{\lambda}_1 + \frac{\partial Q_{in}(\theta_0, 0, 0)}{\partial \lambda_2}\hat{\lambda}_2 = \\ = o_p(n^{-1/2}), \quad i = 1, 2, 3.$$

No pieņēmumiem (A2) un (A4) viegli parādīt, ka

$$\frac{\partial Q_{1n}(\theta_0, 0, 0)}{\partial \theta} \to V_3 \text{ g.d.}, \quad \frac{\partial Q_{1n}(\theta_0, 0, 0)}{\partial \lambda_1} \to -V_2 \text{ g.d.}, \quad \frac{\partial Q_{1n}(\theta_0, 0, 0)}{\partial \lambda_2} = 0,$$

$$\frac{\partial Q_{2n}(\theta_0, 0, 0)}{\partial \theta} \to M_3 \text{ g.d.}, \quad \frac{\partial Q_{2n}(\theta_0, 0, 0)}{\partial \lambda_1} = 0 \quad \frac{\partial Q_{2n}(\theta_0, 0, 0)}{\partial \lambda_2} \to -M_2 \text{ g.d.},$$

$$\frac{\partial Q_{3n}(\theta_0, 0, 0)}{\partial \theta} = 0, \quad \frac{\partial Q_{3n}(\theta_0, 0, 0)}{\partial \lambda_1} \to V_3 \text{ g.d.}, \quad \frac{\partial Q_{3n}(\theta_0, 0, 0)}{\partial \lambda_2} \to kM_3 \text{ g.d.}$$

Tad

$$\begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\lambda}_1 \\ \hat{\lambda}_2 \end{pmatrix} = -S^{-1} \begin{pmatrix} Q_{1n}(\theta_0, 0, 0) \\ Q_{2n}(\theta_0, 0, 0) \\ 0 \end{pmatrix} + o_p(n^{-1/2}),$$

kur

$$S = \begin{pmatrix} V_3 & -V_2 & 0 \\ M_3 & 0 & -M_2 \\ 0 & V_3 & kM_3 \end{pmatrix}.$$

Tā kā

$$S^{-1} = \frac{1}{c_1} \begin{pmatrix} M_2V_3 & kV_2M_3 & V_2M_2 \\ -kM_3^2 & kV_3M_3 & M_2V_3 \\ V_3M_3 & -V_3^2 & V_2M_3 \end{pmatrix},$$

tad

$$\hat{\theta} - \theta_0 = \frac{1}{c_1}\left(M_2V_3Q_{1n}(\theta_0, 0, 0) + kV_2M_3Q_{2n}(\theta_0, 0, 0)\right) + o_p(n^{-1/2}), \tag{4.29}$$

$$\lambda_1 = -\left(\frac{kM_3}{c_1}\right)(M_3 Q_{1n}(\theta_0,0,0) - V_3 Q_{2n}(\theta_0,0,0)) + o_p(n^{-1/2}), \tag{4.210}$$

$$\lambda_2 = \left(\frac{V_3}{c_1}\right)(M_3 Q_{1n}(\theta_0,0,0) - V_3 Q_{2n}(\theta_0,0,0)) + o_p(n^{-1/2}). \tag{4.211}$$

No nosacījuma (A1) ir spēkā

$$\sqrt{n}\begin{pmatrix} Q_{1n}(\theta_0,0,0) \\ Q_{2n}(\theta_0,0,0) \end{pmatrix} \to^d N\left(0, \begin{bmatrix} V_1 & 0 \\ 0 & k^{-1}M_1 \end{bmatrix}\right),$$

ko pielietojot vienādojumiem (4.29), (4.210) un (4.211), iegūstam Lemmas 17 pierādījumu. □

Teorēmas 13 pierādījums

No Lemmas 17 un Teilora izvirzījuma mums ir

$$\begin{aligned}\ln \mathrm{EL}(\Delta,\hat{\theta},\hat{h}) = &- n\lambda_1(\hat{\theta})w_{1x}(\hat{\theta}) + \frac{n}{2}\lambda_1^2(\hat{\theta})w_{2x}(\hat{\theta}) - \\ &- m\lambda_2(\hat{\theta})w_{1y}(\hat{\theta}) + \frac{m}{2}\lambda_2^2(\hat{\theta})w_{2y}(\hat{\theta}) + o_p(1),\end{aligned}$$

kur

$$w_{1x}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_1(X_i,\theta,\Delta,\hat{h}), \quad w_{2x}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_1^2(X_i,\theta,\Delta,\hat{h}),$$

$$w_{1y}(\theta) = \frac{1}{m}\sum_{j=1}^{m} w_2(Y_j,\theta,\Delta,\hat{h}), \quad w_{2y}(\theta) = \frac{1}{m}\sum_{j=1}^{m} w_2^2(Y_j,\theta,\Delta,\hat{h}).$$

No vienādojumiem (4.14) un (4.15) ir spēkā

$$w_{1x}(\hat{\theta}) = \lambda_1(\hat{\theta})w_{2x}(\hat{\theta}) + o_p(n^{-1/2}),$$

$$w_{1y}(\hat{\theta}) = \lambda_2(\hat{\theta})w_{2y}(\hat{\theta}) + o_p(n^{-1/2}).$$

Izmantojot Lemmu 17 un faktu, ka

$$w_{2x}(\hat{\theta}) = V_2 + o_p(1), \quad w_{2y}(\hat{\theta}) = M_2 + o_p(1),$$

seko

$$
\begin{aligned}
-2\ln \mathrm{EL}(\Delta,\hat{\theta},\hat{h}) &= n\lambda_1^2(\hat{\theta})w_{2x}(\hat{\theta}) + m\lambda_2^2(\hat{\theta})w_{2y}(\hat{\theta}) + o_p(1) = \\
&= nk^2\frac{M_3^2}{V_3^2}\lambda_2^2(\hat{\theta})V_2 + m\lambda_2^2(\hat{\theta})M_2 + o_p(1) = \\
&= \frac{k(M_2V_3^2 + kV_2M_3^2)}{V_3^2}[\sqrt{n}\lambda_2(\hat{\theta})]^2 + o_p(1) \to^d \\
&\to^d \frac{M_1V_3^2 + kV_1M_3^2}{M_2V_3^2 + kV_2M_3^2}\chi_1^2.
\end{aligned}
$$

Tad

$$
-2\ln \mathrm{EL}(\Delta,\hat{\theta},\hat{h}) * r = -2\ln \mathrm{EL}(\Delta,\hat{\theta},\hat{h})\frac{M_2V_3^2 + kV_2M_3^2}{M_1V_3^2 + kV_1M_3^2} \to^d \chi_1^2.
$$

EL metode strukturālajiem attiecību modeļiem

Jau tika minēts, ka Valeiņa doktora disertācijā [18] galvenais mērķis bija konstruēt ticamības intervālus strukturālajiem attiecību modeļiem, kurus vispārējā formā ieviesa vācu matemātiķi Freitag un Munk [20]. Strukturālo attiecību modeļi sevī iekļauj labāk zināmos lokācijas-skalēšanas modeļus (Hettmansperger, [21]) un Lēmaņa alternatīvu modeļus (Lehmann, [22]), kuri bieži sastopami arī izdzīvošanas datu analīzē.

Tā kā darba gaitā tika izlabots disertācijā [18] atrodamais robežsadalījums, tad kā piemēru īsi apskatīsim EL metodi ar novērtētiem parametriem lokācijas-skalēšanas modelim.

Definīcija 11. Klasisks lokācijas-skalēšanas modelis divām neatkarīgām izlasēm $X_1,\dots,X_n$ un $Y_1,\dots,Y_m$ ar sadalījuma funkcijām attiecīgi F_1 un F_2 tiek definēts

$$
F_1(t) = F_2\left(\frac{t-\mu}{\sigma}\right) =: F_2(t,h), \quad t \in \mathbb{R} \tag{4.212}
$$

kādam parametram $h = (\mu,\sigma)$ un $\sigma > 0$. Šo pašu attiecību var izteikt arī ar kvantiļu funkcijām

$$
F_1^{-1}(u) = F_2^{-1}(u)\sigma + \mu, \quad u \in [0,1]. \tag{4.213}
$$

Vienkāršības labad pieņemsim, ka $\sigma \equiv 1$, t.i., starp izlasēm pastāv *šifta* jeb lokācijas modelis. Tad interesējošais parametrs ir formā

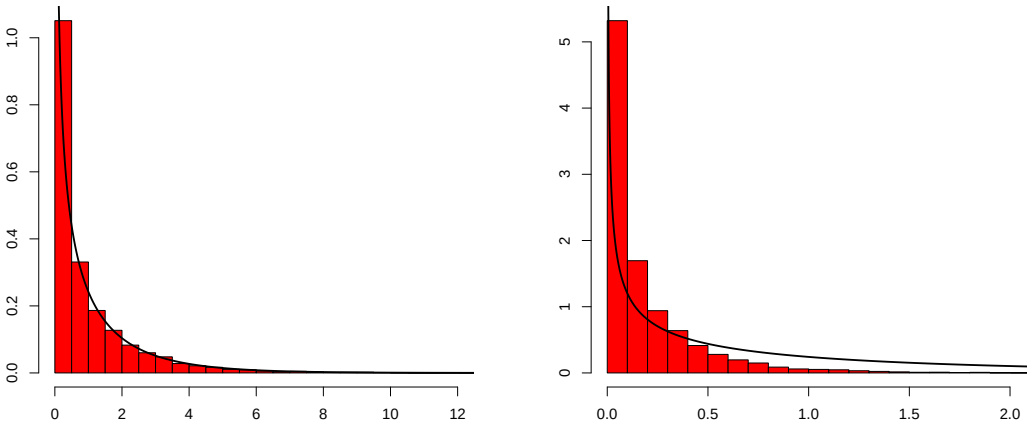
$$
\Delta = F_1(F_2^{-1}(t) + \mu)
$$

un traucējošais parametrs $\theta_0 = F_2^{-1}(t) + \mu_0$. Novērtējošās funkcijas ir formā

$$w_1(X, \theta_0, \Delta, t, \hat{\mu}) = I(X \leq \theta_0) - \Delta,$$

$$w_2(Y, \theta_0, \Delta, t, \hat{\mu}) = I(Y \leq \theta_0 - \hat{\mu}) - t,$$

kur parametra μ novērtēšanai izmantosim abu izlašu vidējo vērtību starpību $\hat{\mu} = \bar{X} - \bar{Y}$, bet sarežģītākos gadījumos var izmantot Mallova attālumu (skatīt [23]).



Att. 4.1: Histogrammas 10 000 reižu simulētai statistikai $-2\ln \mathrm{EL}(\Delta, \hat{\theta}, t, \hat{\mu})$ lokācijas modeļa pārbaudei gadījuma lielumu $X \sim N(0,1)$ un $Y \sim N(1,1)$ realizācijām apjomā $n = 100$. Pa kreisi $\hat{\mu} = -1$, pa labi $\hat{\mu} = \bar{X} - \bar{Y}$.

Valeinis [18] pierāda, ka pieņēmumi $(A0)$ - $(A4)$ ir spēkā lokācijas modeļiem ar

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

un

$$M_1 = M_2 = t(1 - t),$$

kā rezultātā parametra μ novērtēšana neizbojā robežsadalījum, t.i., $-2\ln \mathrm{EL}(\Delta, \hat{\theta}, t, \hat{\mu}) \to_d \chi_1^2$. Tomēr empīriski ar simulāciju palīdzību varam pārliecināties par pretējo.

10 000 reižu simulējam divas izlases ar $F_1 = N(0,1)$ un $F_2 = N(1,1)$ apjomā $n = 100$. 4.1 attēlā pa kreisi redzama histogramma simulētai statistikai, kad $\hat{\mu} = -1$, t.i., netiek novērtēts, savukārt pa labi redzama histogramma statistikai ar novērtēto lokācijas parametru $\hat{\mu}$ katram izlašu pārim. Abām histogrammām pievienota χ_1^2 sadalījuma blīvuma funkcija, un redzams, ka tā labi apraksta tikai histogrammu, kurā traucējošais parametrs netiek aizstāts ar novērtējumu.

Tikai novērtējošā funkcija w_2 satur μ novērtējumu, tāpēc M_1 nav vienāds ar M_2, bet M_1 teorētiskā forma nav triviāli atrodama. Lai noteiktu statistikas $-2\ln \mathrm{EL}(\Delta, \hat{\theta}, t, \hat{\mu})$ robežsadalījumu, var izmantot butstrapa datu pārkārtošanas metodi M_1 novērtēšanai, kas varētu tikt izdarīts pētījuma turpinājumā, attiecīgi paplašinot tematiku.

4.3 Metodes pielietojums izdzīvošanas datiem

Apskatīsim standarta divu izlašu izlašu problēmu ar cenzētiem datiem no labās puses. Pieņemsim, ka mums ir divas neatkarīgas izlases ar iid izdzīvošanas laikiem X_{ji} un cenzēšanas laikiem Y_{ji}, kur $j = 1, 2$, bet $i = 1, \ldots, n_j$, turklāt $n_2/n_1 \to k$, kad $n_1, n_2 \to \infty$. Apzīmēsim $n_1 = n$ un $n_2 = m$. X_{ji} un Y_{ji} sadalījuma funkcijas apzīmēsim attiecīgi ar F_j un G_j. Mūsu rīcībā ir sekojoši novērojumi

$$(Z_{ji}, \delta_{ji}),$$

kur, līdzīgi kā vienas izlases gadījumā, $Z_{ji} = \min(X_{ji}, Y_{ji})$ un $\delta_{ji} = I(X_{ji} \leq Y_{ji})$. Izdzīvošanas funkcijas apzīmēsim ar $S_j = 1 - F_j$, $j = 1, 2$.

Interesējošais parametrs Δ ir

$$\Delta = \int \xi(y) dF_2(y) - \int \xi(x) dF_1(x)$$

un novērtējošās funkcijas šajā gadījumā ir formā

$$w_1(x, \theta_0, \Delta, h) = w_1(x, \theta_0, \Delta, G_1) = \frac{\xi(x)\delta_1}{1 - G_1(x)} - \theta_0,$$

$$w_2(y, \theta_0, \Delta, h) = w_2(y, \theta_0, \Delta, G_2) = \frac{\xi(y)\delta_2}{1 - G_2(x)} - \theta_0 - \Delta,$$

kur $\theta_0 = \int \xi(x) dF_1(x)$. Līdzīgi kā vienas izlases gadījumā Δ ir, piemēram, divu vidējo izdzīvošanas ilgumu starpība, ja $\xi(t) = t$.

Apzīmēsim

$$V_{ni} = \frac{\xi(Z_{1i})\delta_{1i}}{1 - \hat{G}_1(Z_{1i})}, \quad \bar{V}_n = \frac{1}{n}\sum_{i=1}^{n} V_{ni},$$

$$V_{mi} = \frac{\xi(Z_{2i})\delta_{2i}}{1 - \hat{G}_2(Z_{2i})}, \quad \bar{V}_m = \frac{1}{m}\sum_{i=1}^{m} V_{mi}.$$

Teorēma 18. *Pieņemsim, ka* $\hat{h} = (\hat{G}_1, \hat{G}_2)^T$, *kur* $\hat{G}_j$ *ir* G_j *Kaplāna-Meijera novērtējums,* $j =$

1, 2, *tad*

$$-2\ln EL(\Delta,\hat{\theta},\hat{h}) * \frac{\hat{M}_2 V_3^2 + k\hat{V}_2 M_3^2}{\hat{M}_1 V_3^2 + k\hat{V}_1 M_3^2} \to^d \chi_1^2,$$

kur

$$\hat{V}_1 = n\widehat{D}^*(jack), \quad \hat{M}_1 = m\widehat{D}^*(jack),$$

$$\hat{V}_2 = n^{-1}\sum_{i=1}^{n}(V_{ni} - \bar{V}_n)^2, \quad \hat{M}_2 = m^{-1}\sum_{i=1}^{m}(V_{mj} - \bar{V}_m)^2,$$

un $V_3 = M_3 = -1$.

Pierādījums. No Lemmas 7 ir spēkā nosacījums (A1)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_1(Z_{1i},\theta_0,\Delta,\hat{G}_1) \to_d N(0,V_1),$$

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m} w_2(Z_{2i},\theta_0,\Delta,\hat{G}_2) \to_d N(0,M_1),$$

kur V_1 un M_1 būtisku novērtējumu iegūšanai varam izmantot aprakstīto dženaifa procedūru, t.i.,

$$\hat{V}_1 = n\widehat{D}^*(jack) \to V_1 \text{ g.d.} \quad \text{un} \quad \hat{M}_1 = m\widehat{D}^*(jack) \to M_1 \text{ g.d.}$$

Savukārt no Stute un Wang [17] ir spēkā nosacījums (A2) ar V_2 un M_2 novērtējumiem

$$\hat{V}_2 = n^{-1}\sum_{i=1}^{n}(V_{ni} - \bar{V}_n)^2 \to^{\text{g.d.}} n^{-1}\sum_{i=1}^{n} w_1^2(Z_{1i},\theta_0,\Delta,\hat{G}_1) \to_p V_2,$$

$$\hat{M}_2 = m^{-1}\sum_{i=1}^{m}(V_{mj} - \bar{V}_m)^2 \to^{\text{g.d.}} m^{-1}\sum_{i=1}^{m} w_2^2(Z_{2j},\theta_0,\Delta,\hat{G}_2) \to_p M_2.$$

No Ovena [8] seko nosacījums (A3), un viegli pārliecināties, ka nosacījums (A4) ir spēkā ar $V_3 = n^{-1}\sum_{i=1}^{n}\alpha_1(Z_{1i},\theta_0,\Delta,\hat{G}_1) = -1$ un $M_3 = m^{-1}\sum_{i=1}^{m}\alpha_2(Z_{2i},\theta_0,\Delta,\hat{G}_2) = -1$. □

5 Simulācijas un datu piemēri

Šajā nodaļā apskatīsim simulāciju rezultātus gan vienas, gan divu izlašu gadījumā, kā arī pielietosim metodi reālu datu piemēram.

Simulācijas

Lai iegūtu cenzētus datus programmā R, vispirms ģenerējam izdzīvošanas laikus un cenzēšanas laikus no eksponenciālā sadalījuma ar parametriem attiecīgi 1 un c, t.i., $F(x) = 1 - \exp(-x)$ un $G(x) = 1 - \exp(-cx)$ visiem $x \geq 0$. $c > 0$ izvēlamies tā, lai cenzētie dati saturētu vēlamo cenzēšanas proporciju.

Vispirms vēlamies pārbaudīt, vai iepriekšējās nodaļās teorētiski pamatotie rezultātā tikpat labi strādā praktiski gan vienas, gan divu izlašu problēmām. Pierādījām, ka statistika (3.23) un (4.23) tiecas uz χ^2_1 sadalījumu, kad $n \to \infty$. Lai to pārbaudītu, simulējam statistiku 10 000 reižu vienas izlases gadījumā un 1000 reižu divu izlašu gadījumā, izvēlamies $\xi(t) = t$ un $c = 0.1$. Tātad vienas izlases gadījumā interesējošais parametrs θ ir vidējais izdzīvošanas ilgums, bet divu izlašu gadījumā - Δ ir šo lielumu starpība. 5.1 un 5.2 attēlā redzamas histogrammas simulētai statistikai dažādiem izlašu apjomiem ar pievienotu χ^2_1 sadalījuma blīvuma funkciju attiecīgi vienas un divu izlašu gadījumā.

Grafiska sadalījuma pārbaude liecina, ka simulētās statistikas histogrammu labi apraksta χ^2_1 sadalījuma blīvuma funkcija, tomēr vēlams salīdzināt arī simulētā un teorētiskā sadalījuma kvantiles. Tabulā 5.1 redzamas 0.95-tās kvantiles, kas vienas un divu izlašu gadījumā iegūtas, izmantojot gan džeknaifa procedūru σ^2 novērtēšanai, gan teorētisko rezultātu $\sqrt{n}\bar{W}_n \to^d N(0, \sigma^2)$ (skatīt Lemmu 7), no kurienes $\hat{\sigma}^2 = 1.14$. Redzams, ka ar novērtēto $\hat{\sigma}^2$ simulētā sadalījuma kvantiles daudz labāk apraksta teorētisko kvantili 3.84, savukārt, izmantojot džeknaifa novērtējumu, kvantile svārstās vai lēni konverģē uz teorētisko kvantili. Tas ir izskaidrojams ar džeknaifa dispersijas novērtējuma nestabilitāti, piemēram, 100 reizes ģenerētām izlasēm ar apjomu $n = 1000$ džeknaifa dispersijas novērtējums svārstās no 0.9 līdz pat 4.1, bet izlasēm ar apjomu

n =10 000 tas svārstās no 1.09 līdz 1.15, tomēr šāda apjoma izlases praktiskos pielietojumos ir reti sastopamas.

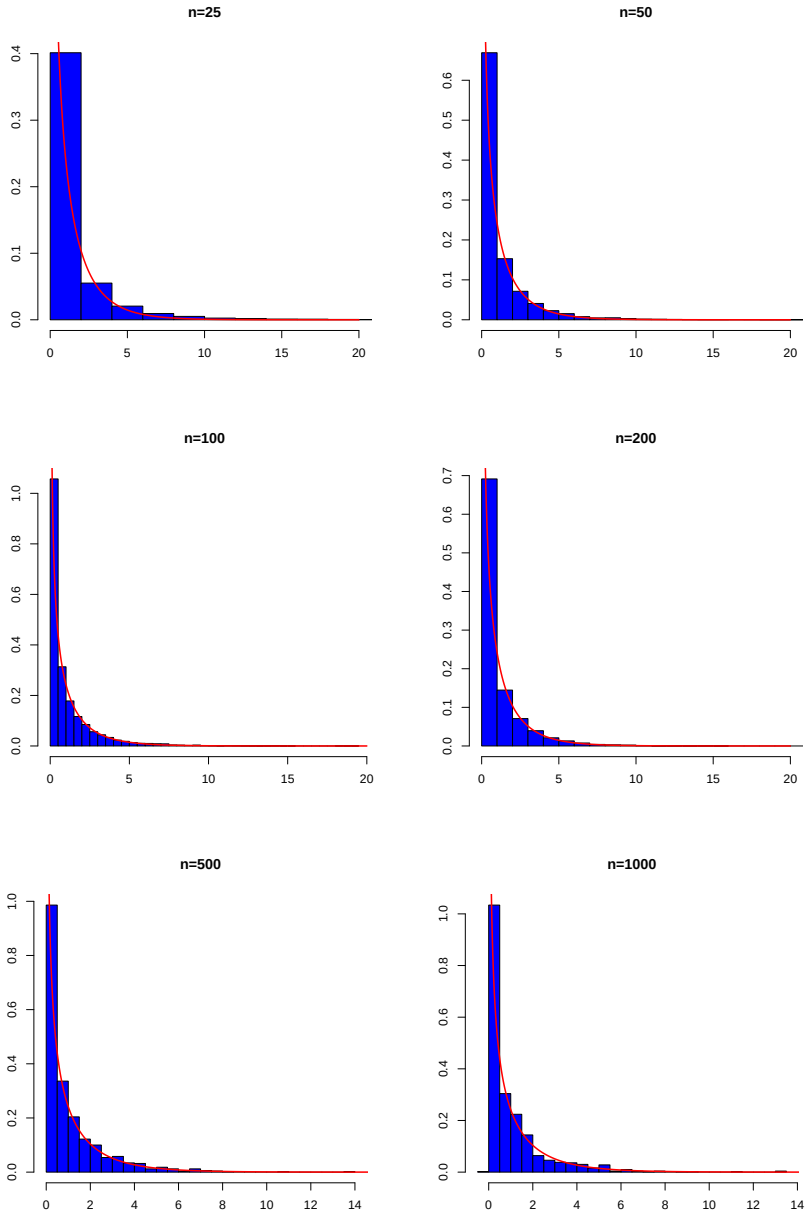
Tabula 5.1: Simulēto sadalījumu 0.95 kvantiles dažādiem n

n	Vienai izlasei		Divām izlasēm	
	$\hat{\sigma}^2 = n\widehat{D}^*(jack)$	$\hat{\sigma}^2 = 1.14$	$\hat{\sigma}_j^2 = n_j\widehat{D}^*(jack)$	$\hat{\sigma}_j^2 = 1.14$
25	5.738	3.903	2.696	4.723
50	4.622	3.645	2.878	4.496
100	4.310	3.783	3.198	3.901
200	4.099	3.721	4.080	3.598
500	4.078	3.817	2.983	4.011
1000	3.981	3.716	3.598	3.843

Tabulā 5.2 redzamas 90 un 95% ticamības intervālu pārklājumu precizitātes vienas izlases gadījumā vidējam izdzīvošanas ilgumam, t.i., $\xi(t) = t$, dažādiem izlašu apjomiem n un cenzēšanas proporcijām c. Salīdzināti ar empīriskās ticamības funkcijas metodi iegūtie un ticamības intervāli, kas balstīti uz $\sqrt{n}(\theta(\hat{F}) - \theta(F))$ asimptotisko normalitāti. Empīriskās ticamības funkcijas metode dod vienmērīgi labākas pārklājumu precizitātes, bet īpaši labākus rezultātus tā dod maziem izlašu apjomiem ar salīdzinošu lielu cenzēšanas proporciju, kad $c = 0.4$. Tomēr, jo lielāka cenzēšanas proporcija, jo sliktākas ir pārklājumu precizitātes ticamības intervāliem iegūtiem, izmantojot abas metodes.

Tabula 5.2: Ticamības intervālu pārklājumu precizitāte vienas izlases gadījumā, $\xi(t) = t$

c	n	$1-\alpha = 0.90$		$1-\alpha = 0.95$	
		Normālie	EL	Normālie	EL
0.10	10	0.770	0.805	0.810	0.864
	20	0.830	0.837	0.870	0.891
	50	0.866	0.892	0.904	0.928
	100	0.892	0.923	0.936	0.951
0.25	10	0.710	0.741	0.730	0.776
	20	0.790	0.785	0.800	0.843
	50	0.786	0.837	0.848	0.896
	100	0.804	0.879	0.856	0.916
0.40	10	0.510	0.675	0.520	0.713
	20	0.600	0.754	0.700	0.792
	50	0.632	0.801	0.674	0.831
	100	0.676	0.821	0.736	0.862

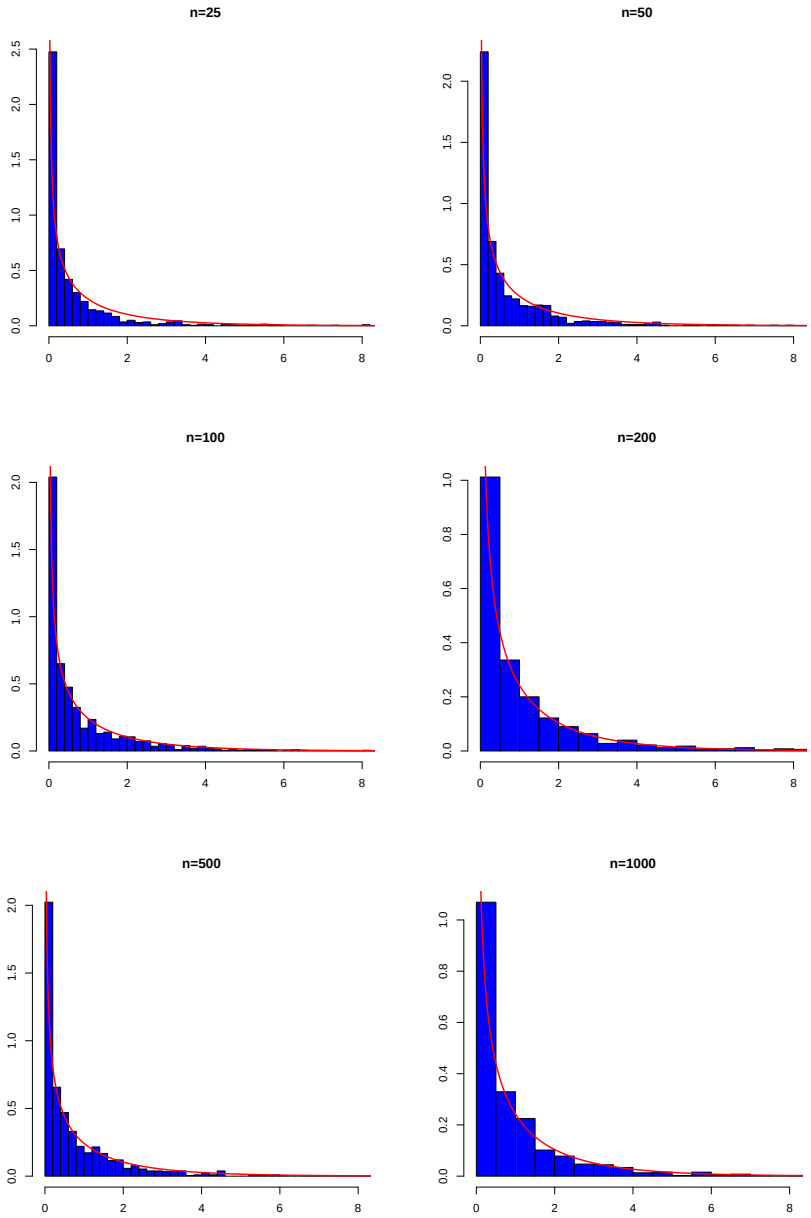


Att. 5.1: Simulētais robežsadalījums vienas izlases gadījumā, $\xi(t) = t$, $c = 0.1$

Primārās biliārās cirozes dati

Dati par primārās biliārās aknu cirozes (pbc) pacientiem iegūti Mayo klīnikas pētījumā, tie atrodami, piemēram, [24]. Pētījumā kopā piedalījās $n = 312$ pacienti, no kuriem 158 saņēma ārstēšanu ar D-penicilamīnu, savukārt 154 saņēma tā saucamās placebo zāles. Dati apraksta izdzīvošanas laiku dienās, cenzēšanas proporcija ir ļoti liela - 187 no 312 novērojumiem ir cenzēti. Mūs interesē salīdzināt divas izlases, pacientu, kas saņēma ārstēšanu, izdzīvošanas laikus un pacientu izdzīvošanas laikus, kas saņēma placebo.

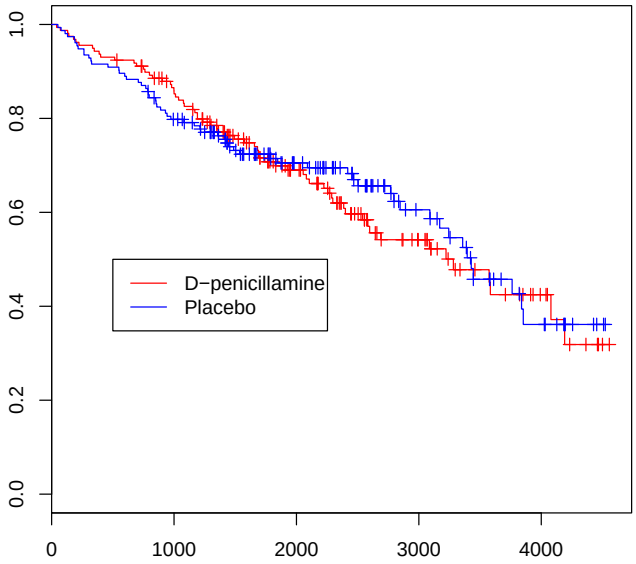
Apskatot abu izlašu izdzīvošanas funkciju novērtējumus 5.3 attēlā, grūti secināt, kurai no



Att. 5.2: Simulētais robežsadalījums divu izlašu gadījumā, $\xi(t) = t$, $c = 0.1$

izlasēm ir labākas izdzīvošanas varbūtības, tādēļ izvirzām hipotēzi, ka gan ārstēšanu, gan placebo saņēmušo pacientu vidējie izdzīvošanas laiki ir vienādi, t.i., $H_0 : \Delta = 0$. Ar empīriskās ticamības funkcijas metodi iegūstam Δ novērtējumu $\hat{\Delta} = 244$ dienas un 95% ticamības intervālu [-506, 1182], un tā kā 0 pieder šim intervālam, tad nevaram noraidīt hipotēzi par vidējo izdzīvošanas ilgumu vienādību. Lai gan intuitīvi šķiet, ka 244 dienas tomēr ir liela atšķirība vidējos izdzīvošanas laikos, pētījums veikts apmēram 10 gadu laikā, kas salīdzinājumā ar 244 dienām ir ilgs laika periods.

Konstruēsim arī 95% ticamības intervālu izdzīvošanas varbūtību starpībai $\Delta(t_0) = S_1(t_0) - S_2(t_0)$ punktā $t_0 = 2750$, tātad $\xi(t) = I(t \geq t_0)$, jo no 5.3 attēla redzams, ka apmēram šajā pun-



Att. 5.3: Izdzīvošanas funkciju novērtējums datiem pbc

ktā varētu būt vislielākā izdzīvošanas varbūtību atšķirība abām izlasēm. Ar empīriskās ticamības funkcijas metodi iegūts izdzīvošanas varbūtību starpības novērtējums $\hat{\Delta}(2750) = -0.0441$, un ticamības intervāls [-0.251,0.203], un arī šajā gadījumā nevaram noraidīt hipotēzi par izdzīvošanas varbūtību vienādību abām izlasēm punktā t_0.

Empīriskās ticamības funkcijas metodes viena no priekšrocībām ir tāda, ka tā saglabā apgabalu (*range preserving*), jo balstās uz dotajiem novērojumiem. Tas ir īpaši svarīgi šajā gadījumā, novērtējot izdzīvošanas varbūtības, jo zināms, ka tās pieder intervālam [0,1].

Nobeigums

Maqģistra darbā tika aplūkoti izdzīvošanas analīzes pamatelementi, kurus bija nepieciešams izprast, lai ieviestu empīriskās ticamības funkcijas metodes vispārinājumu izdzīvošanas datiem.

Wang un Jing [2] 2001. gadā pielāgoja empīrisko ticamības funkciju dažādu parametru, kurus var izteikt ar izdzīvošanas funkciju, klasei. Līdz tam literatūrā bija sastopami vairāki atsevišķi rezultāti EL metodes pielietošanai tikai dažiem parametriem, piemēram, ticamības intervālu konstruēšanai izdzīvošanas varbūtībām fiksētā laika momentā. Implementējot Wang un Jing [2] piedāvāto metodi datorprogrammā R, tika veiktas simulācijas, lai praktiski pārliecinātos, kā strādā teorētiskais rezultāts. Konstruējot ticamības intervālus tika noskaidrots, ka EL metode dod augstu ticamības intervālu pārklājumu precizitāti, it īpaši mazu izlašu apjomu un lielas cenzēšanas proporcijas gadījumā.

2009. gadā tika publicēts empīriskās ticamības funkcijas metodes vispārinājums [6], kas pieļāva gan traucējošo parametru novērtēšanu, gan uzlika mazākus nosacījumus sākotnējiem datiem, bet rezultātā tika "izbojāts" robežsadalījums. Wang un Jing [2] rezultāts tika minēts kā EL metodes vispārinājuma piemērs, uz ko balstoties, radās maģistra darba mērķis vispārināt empīrisko ticamības funkciju izdzīvošanas datiem divām izlasēm.

Mērķa sasniegšanai bija nepieciešams labāk izprast Qin un Zhao [7] EL metodes divām izlasēm bez traucējošo parametru klātbūtnes novērtējošajās funkcijās ideju un rezultātu pierādījumu, kas tika izdarīts. Lai arī Valeinis 2007. gadā savā doktora disertācijā [18] aprakstīja empīriskās ticamības funkcijas metodes vispārinājumu divām izlasēm, tomēr darba izstrādāšanas gaitā neatkarīgi tika veikti aprēķini, kuru rezultātā atklājās dažas neprecizitātes minētajā disertācijā, kas arī tika izlabotas.

Darba galvenais mērķis tika sekmīgi sasniegts - tika atrasts robežsadalījums empīriskās ticamības funkcijas metodes statistikai izdzīvošanas datiem divu izlašu gadījumā. Arī šis rezultāts tika pārbaudīts praktiski, veicot simulācijas datorprogrammā R, izmantojot arī Valeiņa un Cera [19] izveidoto datorpaketes R paplašinājumprogrammu EL. Izdzīvošanas analīzē daudz biežāk

ir svarīgi salīdzināt parametrus divām atšķirīgām izlasēm, piemēram, novērojot kādu zāļu iedarbību uz pacientiem salīdzinājumā ar pacientiem, kuri ārstēšanu nesaņem. Līdzīgs reālu datu piemērs tika atrasts, tika pielietoti iegūtie rezultāti, lai konstruētu ticamības intervālus vidējo izdzīvošanas ilgumu starpībai un izdzīvošanas varbūtību starpībai fiksētā punktā. Rezultāti sakrita ar citiem literatūrā sastopamajiem rezultātiem konkrētajiem datiem.

Darba turpinājumā būtu interesanti izpētīt izdzīvošanas datu analīzi sīkāk, lai salīdzinātu empīriskās ticamības funkcijas metodes veikumu ar citām klasiski pielietotām metodēm. Vēl kāds turpmākā pētījuma virziens varētu būt saistīts ar robežsadalījumu atrašanu citiem konkrētiem pielietojumiem, piemēram, darbā minētajiem lokācijas modeļiem. Tā kā darbs bija vairāk teorētisks nevis praktisks, tad būtu vēlams veikt papildus simulācijas un apskatīt datu piemērus it īpaši divu izlašu gadījumā, kas bija apgrūtinoši datorprogrammas salīdzinoši lēnās darbības dēļ. Veicot programmas koda uzlabojumus un optimizāciju, to būtu iespējams piedāvāt lietošanai praksē.

Izmantotā literatūra un avoti

[1] 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.

[2] 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.

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

[4] A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.

[5] G. Li. On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statistics & Probability Letters*, 25(2):95–104, 1995.

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

[7] 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.

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

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

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

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

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

[13] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[14] W. Stute. The central limit theorem under random censorship. *The Annals of Statistics*, pages 422–439, 1995.

[15] W. Stute. The jackknife estimate of variance of a kaplan-meier integral. *The Annals of Statistics*, 24(6):2679–2704, 1996.

[16] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1993.

[17] W. Stute and J. L. Wang. The strong law under random censorship. *The Annals of Statistics*, pages 1591–1607, 1993.

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

[19] J. Valeinis and E. Cers. Extending the two-sample empirical likelihood method. (iesniegta) 2013.

[20] 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.

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

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

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

[24] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*, volume 169. Wiley, 2011.

A Pielikums

Programmas kods cenzētu datu ģenerēšanai

```
library(sm)
library(survival)
library(EL)

#Zimejam teoretiskos sadalijumus
xx<-seq(0,80,by=0.01)
hist(rexp(10000),prob=TRUE,breaks=50)
lines(xx,dexp(xx),type="l")
lines(xx,dexp(xx,0.1),type="l",col="blue")
lines(xx,dexp(xx,0.25),type="l",col="red")
lines(xx,dexp(xx,0.4),type="l",col="darkgreen")

n<-50
#Simulejam cenzetus datus matricaa
cenz<-function(x,c,n)
{
dati<-matrix(0,n,2)
lifetimes <- rexp( n, rate = x)
censtimes <- rexp( n, rate = c)
ztimes <- pmin(lifetimes, censtimes)
status <- as.numeric(censtimes > lifetimes)
#object<-Surv(ztimes,status)
dati[,1]<-ztimes
dati[,2]<-status
return(dati)
}
###dati<-cenz(1,0.1,n)
```

Programma novērojumu modifikācijai

```
#### Funkcija Xi = videja vertiba
xi<-function(x) x
xi1<-Vectorize(xi)

####Survival probability at fixed time t0
#t0<-5
#xi<-function(t)
#{
#if (t<=t0) 1 else 0
```

```
#}
#xi1<-Vectorize(xi)

###Atrast isto theta_0 vertibu
d<-function(x) x*dexp(x)
theta0<-integrate(d,0,20)$value
theta0

####Funkcija g=1-gn Kaplan-Meier novertejums cenzesanas sadalijumam
g<-function(t,z,delta)
{
(ii <- order(x <- z, y <- delta))
mat<-rbind(x,y)[,ii]
z.sort<-mat[1,]
delta.sort<-mat[2,]
vec<-(n-(1:n))/(n-(1:n)+1)
z.sort[delta.sort==1]<-t+1 ### lai iznicinatu, kur 0
prod(vec[z.sort<=t])
}
g<-Vectorize(g,vectorize.args="t")

#xx<-seq(0,6,by=1)              ####Lai uzzimetu g
#plot(xx,1-g(xx,z,delta),type="l")
#lines(xx,g(xx,z,delta),type="l",col="3")
###Iebuveta Kaplan-Meier novertejums
#fit<-survfit(Surv(z,delta==0)~1)
#lines(fit,col="4")

####Izdod modificeto izlasi
fun<-function(dati)
{
z<-dati[,1]
delta<-dati[,2]
gz<-g(z,z,delta)
gz[gz==0]<-1
delta[gz==0]<-0
izl.mod<-xi(z)*delta/gz
return(izl.mod)
}

###Jacknife Sn novertejums un pseidovertibas
sn.jack<-function(dati)
{
z<-dati[,1]
delta<-dati[,2]
(ii <- order(x <- z, y <- delta))
mat<-rbind(x,y)[,ii]
z.sort<-mat[1,]
delta.sort<-mat[2,]
n1<-length(z.sort)
if (delta.sort[n1-1]==0) delta.sort[n1]<-0
vec<-(n1-(1:n1))/(n1-(1:n1)+1)
vec[delta.sort==0]<-1
wn<-c()
wn[1]<-delta.sort[1]/n1
for (i in 2:n1) wn[i]<-prod(vec[1:i-1])/(n1-i+1)
wn[delta.sort==0]<-0
wn1<-c()
```

```
wn1<-wn*xi(z.sort)
sum(wn1)   #xi1-vektorizeta funkcija
}

###Funkcija,kas aprekina jackknife dispersijas novertejumu
n.var.jack<-function(dati)
{
n<-length(dati[,1])
sn.bar<-sn.jack(dati)  ###Jo delta[n]==0 vienmer
snk<-c()
for (i in 1:n) snk[i]<-sn.jack(dati[-i,]) #n vai n-1 ?
(n-1)*sum((snk-sn.bar)^2)
#n.var.jack<-(n-1)*sum(snk^2)-n*(n-1)*sn.bar^2 #Sanaca negativa pec sis f-las
}

###Funkcija,kas aprekina el statistiku
el.fun<-function(data,theta)
{
izl_sort<-c()
izl_sort<-sort(data)

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((data-theta)/(1+lambda*(data-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*(data-theta)))
-2*log(prod(n*p))
}
el.fun<-Vectorize(el.fun,vectorize.args="theta")

#el.fun(fun(cenz(1,0.1,n)),1)
###Parbaudam modifikaciju un atrodam el novertejumu
dati<-cenz(1,0.1,n)
data<-fun(dati)
ee<-seq(min(data)+0.1,max(data)-0.1,by=0.1)
plot(ee,el.fun(data,ee),type="l",xlim=c(min(data),max(data)))

EL<-function(theta) el.fun(data,theta)
EL.value<-optimize(EL,c(min(data),max(data)))$minimum
EL.value

###Funkcija,kas izdod statistiku el*r
rob.sad<-function(dati,theta)
{
mod.izl<-fun(dati)
el.fun(mod.izl,theta)*(sum((mod.izl-mean(mod.izl))^2)/n.var.jack(dati)/n)
}

###Robezsadalijums nesvertais
n<-500
```

```
e<-replicate(1000,el.fun(fun(cenz(1,0.1,n)),1))
hist(e,prob=TRUE,breaks=30,main="",xlab="",ylab="",col="blue",xlim=c(0,20))
sort(e)[950]
###Robezsadalijums svertais
t<-replicate(1000,rob.sad(cenz(1,0.1,n),1))
hist(t, prob=TRUE,breaks=30,main="",xlab="",ylab="",col="blue")
sort(t)[950]
xx<-seq(0,20,by=0.01)
points(xx,dchisq(xx,1),type="l",lwd="2",col="red")
```

Programma σ^2 simulēšanai

```
n<-1500
theta0
wn<-replicate(10000,n^(-1/2)*sum(fun(cenz(1,0.1,n))-theta0))
sigma2<-var(wn)
sigma2

###Robezsadalijums ar sigma^2
n<-1000
sigma2<-1.14
f.sv<-function(dati,theta)
{
mod.izl<-fun(dati)
el.fun(mod.izl,theta)*(sum((mod.izl-mean(mod.izl))^2)/sigma2/n)
}
t.sv<-replicate(10000,f.sv(cenz(1,0.1,n),1))
hist(t.sv, prob=TRUE,breaks=30,main="",xlab="",ylab="",col="blue")
sort(t.sv)[9500]

xx<-seq(0,20,by=0.01)
points(xx,dchisq(xx,1),type="l",lwd="2",col="red")
```

Programma divām izlasēm

```
###Funkcija,kas izdod el*r
r.rob.sad<-function(dati1,dati2)
{
#dati1<-cenz(1,0.1,n)
#dati2<-cenz(1,0.1,n)
mod.izl1<-fun(dati1)
mod.izl2<-fun(dati2)
r<-(sum((mod.izl1-mean(mod.izl1))^2)/n+sum((mod.izl2-mean(mod.izl2))^2)/m)/
(n.var.jack(dati1)+n.var.jack(dati2))
r*EL.means(mod.izl1,mod.izl2)$statistic
}
####Robezsadalijuma vektors
n<-50
m<-n
rob.sad<-c()
n<-25
rob.sad<-replicate(100,r.rob.sad(cenz(1,0.1,n),cenz(1,0.1,n)))
```

Programma datiem pbc

```
####Datu piemers
```

```
attach(pbc)
dati1<-matrix(0,312,3)
dati1[,1]<-time[1:312]
dati1[,2]<-status[1:312]
dati1[,3]<-trt[1:312]
for (i in 1:312) if (dati1[i,2]==1) dati1[i,2]<-0
for (i in 1:312) if (dati1[i,2]==2) dati1[i,2]<-1
#object<-Surv(ztimes,status)
x<-dati1[,1]
y<-dati1[,2]
z<-dati1[,3]
n<-length(z[z==1]) ###D-penicillamine
m<-length(z[z==2]) ###Placebo
###Izlase,kas sanem arstesanu
izl1<-matrix(0,n,2)
izl1[,1]<-x[z==1]
izl1[,2]<-y[z==1]
###Izlase,kas sanem placebo
izl2<-matrix(0,m,2)
izl2[,1]<-x[z==2]
izl2[,2]<-y[z==2]
survdiff(Surv(x,y==1)~z)
plot(survfit(Surv(x,y==1)~z),lty = 1:1,col=c("red","blue"),)
legend(500,0.5,c("D-penicillamine","Placebo"),lty=c(1,1),col=c("red","blue"))
survfit(Surv(x,y==1)~z)
r.rob.sad(izl1,izl2)
```

Programma ticamības intervālu konstruēšanai vidējam izdzīvošanas ilgumam

```
tic.int<-function(dati)
{
data<-fun(dati)
ee<-seq(min(data)+0.001,max(data)-0.001,by=0.1)
EL<-function(theta) rob.sad(dati,theta)
EL.value<-optimize(EL,c(min(data),max(data)))$minimum
R_FF3<-function(theta)
{
R_FF3<-EL(theta)-qchisq(0.95,1)
}
apak.rob<-uniroot(R_FF3,c(min(ee),EL.value))$root-1
aug.rob<-uniroot(R_FF3,c(EL.value+0.001,max(ee)))$root-1
if (apak.rob*aug.rob<0) 1 else 0
}

n<-100
prec.vec<-replicate(1000,tic.int(cenz(1,0.1,n)))
sum(prec.vec)/1000
```

Maģistra darbs “Empīriskās ticamības funkcijas metodes vispārinājums izdzīvošanas datiem” 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: doc. Dr.math. Nadežda Siņenko

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)