LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

PĀRBĪDES FUNKCIJA DIVU IZLAŠU GADĪJUMĀ

DIPLOMDARBS

Autore: **Lidija Januševa**
Stud. apl. Nr.: lj07012
Darba vadītājs: docents Dr. math. Jānis Valeinis

RĪGA 2012

Anotācija

Darbā ir aplūkota vispārīgā pārbīdes funkcija divu izlašu gadījumā un ranžētu izlašu pielietojums. Šīs izlases ieviesa *McIntyre* 1952. gadā kā metodi, kas būtiski samazina datu apsekojumu izmaksas. Tomēr ranžētas izlases var pielietot arī liela datu apjoma gadījumos. Izrādās, ka pārbīdes funkcija cieši saistīta gan ar statistikā pazīstamo lokācijas modeli, gan ar – kvantiļu-kvantiļu grafikiem. Darbā tiek pārbaudīts lokācijas un lokācijas-skalēšanas modelis, konstruējot vienlaicīgās ticamības joslas vispārīgai pārbīdes funkcijai, vertikālās pārbīdes funkcijai, kvantiļu starpības funkcijai, kvantiļu-kvantiļu (Q-Q) grafikiem un varbūtību-varbūtību (P-P) grafikiem. Darba mērķi ir salīdzināt savā starpā šīs pieejas, kā arī pielietot ranžētās izlases praktiskām datu problēmām.

Atslēgas vārdi: pārbīdes funkcija, strukturālo attiecību modeļi, ranžētas izlases

Abstract

Thesis contains analysis of general shift function in the two-sample case and application of ranked set sampling. Ranked set sampling was introduced by McIntyre in 1952 as a cost-effective survey sampling method, yet ranked set sampling can be applied for data reduction as well. It turns out that the general shift function is closely connected to location model and quantile-quantile plot. The location model and the location-scale model have been tested by constructing simultaneous confidence bands for the general shift function, the vertical shift function, quantile distance function, quantile-quantile plot (Q-Q) and probability-probability plot (P-P). The goal of this study is to compare these methods and apply ranked set sampling on real problems.

Keywords: general shift function, structural relationship models, ranked set sampling

Saturs

Apzīmējumi

F, G	izlašu teorētiskās sadalījuma funkcijas
F_n, G_m	izlašu empīriskās sadalījuma funkcijas
$N(\mu, \sigma^2)$	normāli sadalīts gadījuma lielums ar vidējo vērtību μ un dispersiju σ^2
$\xrightarrow{d}$	konverģence pēc sadalījuma
SRS	vienkārša gadījuma izlase
RSS	ranžēta izlase
BRSS	balansēta ranžēta izlase
$X_{k\times m}$	balansēta ranžēta izlase apjomā $k \cdot m$
$X_{[r]i}$	i-tais izmērītais elements ar rangu r
RE	relatīvā efektivitāte
MSE	vidējā kvadrātiskā kļūda
ARE	asimptotiskā relatīvā efektivitāte
UMVUE	minimālais dispersijas nenovirzīts novērtējums

Ievads

Medicīnā un farmācijā, ieviešot jaunus medikamentus, ir svarīgi noteikt, vai visiem populācijas locekļiem jauno zāļu iedarbība dod uzlabojumus. Šajā gadījumā populārākie testi nesniegs atbildi uz šo jautājumu. Piemēram, t-tests divu izlašu gadījumā pārbauda hipotēzi par vidējo vērtību vienādību, bet nesniedz informāciju par sagaidāmo uzlabojumu. Atbildi par iespējamo uzlabojumu var saņemt, pārbaudot hipotēzi par lokācijas modeli starp kontroles grupu un eksperimentālo grupu, neinteresējoties par konkrētu sadalījuma funkciju veidu.

Pieņemsim, ka dotas divas neatkarīgas izlases $X_1, X_2, \ldots, X_n$ un $Y_1, Y_2, \ldots, Y_m$ ar sadalījuma funkcijām attiecīgi F un G. Starp šīm izlasēm pastāv lokācijas modelis, ja

$$F(x) = G(x + \theta),\ x \in \mathbb{R},$$

kur θ ir lokācijas parametrs. Lokācijas modelis ir *Freitag* un *Munk* [1] definēto strukturālo modeļu speciālgadījums.

Pastāv vairākas iespējas lokācijas modeļa pārbaudei. Viena no tām ir konstruēt vienlaicīgās ticamības joslas pašam parametram θ. *Doksum* un *Sievers* savā publikācijā [2] definē vispārīgo pārbīdes funkciju

$$\Delta(x) = G^{-1}(F(x)) - x,\ x \in \mathbb{R}. \tag{0.1}$$

Konstruējot vienlaicīgās ticamības joslas šai funkcijai, iespējams pārbaudīt ne tikai lokācijas modeli, bet arī – lokācijas-skalēšanas modeli.

Alternatīva pieeja lokācijas modeļa analīzē ir atņemt no otrās izlases lokācijas parametru, iegūstot sadalījuma funkciju $G_\theta(x)$ un veikt divu sadalījuma funkciju vienādības pārbaudi $H_0 : F(x) = G_\theta(x)$. Šajā gadījumā var izmantot klasisko Kolmogorova-Smirnova testu, vai arī – konstruēt vienlaicīgās ticamības joslas varbūtību-varbūtību (P-P) un kvantiļu-kvantiļu (Q-Q) grafikiem. Pirms veikt hipotēžu pārbaudi lokācijas parametrs θ ir jānovērtē. Šo pieeju ir pētījis savā diplomdarbā Cielēns [3].

Cita pieeja ir konstruēt vienlaicīgās ticamības joslas doto izlašu kvantiļu starpības funkcijai. Šo funkciju ir apskatījuši *P. Laake*, *K. Laake* un *R. Aaberge* savā publikācijā [4]. Šai funkcijai ticamības joslas var konstruēt, arī izmantojot empīriskās ticamības (EL) metodi, ko savā disertācijā ir aprakstījis Valeinis [5].

Ghosh un *Tiwari* savā publikācijā [6] apskatīja ne tikai vispārīgo pārbīdes funkciju (0.1), bet arī – vertikālo pārbīdes funkciju $\Lambda(t) = G(F^{-1}(t)) - t,\ t \in [0, 1]$ un piedāvāja

ar tās palīdzību pārbaudīt vertikālo pārbīdi. Hipotēzes pārbaude par lokācijas modeli ir nepieciešama iepriekšējo preparātu un jauno preparātu iedarbības salīdzinājumā, eksperimentos ar dzīvniekiem, jaunu ārstniecības metožu ieviešanā u. c.

McIntyre 1952. gadā savā publikācijā [7] piedāvāja jaunu datu apsekojumu metodi, kas pazīstama kā ranžētu izlašu veidošana. Šī metode ļauj, samazinot apsekojumu izmaksas, iegūt izlasi, kas labi reprezentē populāciju. Pēdējo 20 gadu laikā interese par šo metodi ir pieaugusi. Mūsdienās ranžētas izlases tiek izmantotas arī liela izlases apjoma gadījumos. *Ghosh* un *Tiwari* savā publikācijā [6] apskatīja ranžētu izlašu pielietojumu divu izlašu problēmās.

Diplomdarba mērķi ir veikt grafisko hipotēžu pārbaudi par lokācijas modeli reālām datu problēmām, izmantojot dažādas pieejas – konstruējot vienlaicīgās ticamības joslas pārbīdes funkcijai, vertikālās pārbīdes funkcijai, varbūtību-varbūtību un kvantiļu-kvantiļu grafikiem, kā arī – kvantiļu starpības funkcijai, un salīdzināt šīs pieejas savā starpā, kā arī pielietot ranžētās izlases reālām datu problēmām. Šiem mērķiem tiek izmantota brīvpieejas programma R.

Darbs sastāv no trim nodaļām un pielikuma. Pirmajā nodaļā tiek apskatīta vispārīgā pārbīdes funkcija un sniegts priekšstats par strukturālo modeļiem, hipotēžu pārbaudi fiksēta lokācijas parametra gadījumā, hipotēžu pārbaudi, izmantojot varbūtību-varbūtību, kvantiļu-kvantiļu grafikus, vertikālo pārbīdes funkciju un kvantiļu starpības funkciju. Otrajā nodaļā tiek apskatīts ranžētu izlašu pielietojums hipotēžu pārbaudē par lokācijas modeli un to rašanās vēsture. Trešā nodaļa veltīta grafiskai hipotēžu pārbaudei par lokācijas modeli trim reālām datu problēmām. Divi datu piemēri tika izmantoti no *Doksum* un *Sievers* [2] un *Doksum* [8] publikācijām, viens datu piemērs tika iegūts no ASV Nacionālā Veselības Statistikas centra (`http://www.cdc.gov/nchs`). Pielikumā atrodams programmas R kods.

1. PĀRBĪDES FUNKCIJA

Pieņemsim, ka $X_1, X_2, \ldots, X_n$ ir neatkarīgi vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju F un $Y_1, Y_2, \ldots, Y_m$ ir neatkarīgi vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju G. Savā publikācijā *Doksum* un *Sievers* [2] apskatīja šo izlašu vispārīgo pārbīdes funkciju $\Delta(x)$. Lai ieviestu pārbīdes funkcijas jēdzienu, šajā nodaļā vispirms tiks apskatīti divu izlašu strukturālo attiecību modeļi un to hipotēžu pārbaudes iespējas. Pēc tam tiks definēta pārbīdes funkcija un apskatītas iespējas tās pielietojumam hipotēžu pārbaudē par divu izlašu lokācijas un lokācijas-skalēšanas modeli, kā arī – tiks apskatīta vēl viena alternatīva šīs hipotēžu pārbaudes iespēja.

1.1. Strukturālo attiecību modeļi

Statistikā ļoti svarīga problēma ir divu izlašu salīdzināšana un pārbaude, vai starp tām eksistē kāda sakarība. Jēdziens strukturālie attiecību modeļi parādījās salīdzinoši nesen 2005. gada *Freitag* un *Munk* publikācijā [1]. Šie modeļi ir veidoti kā vispārinājums un sevī ietver divas problēmas – lokācijas skalēšanas modeli un Lēmaņa alternatīvo modeli.

Definīcija 1. Starp divu izlašu $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ sadalījuma funkcijām F un G pastāv lokācijas modelis, ja

$$F(x) = G(x + \theta), \tag{1.1}$$

kur θ ir lokācijas parametrs.

Definīcija 2. Starp divu izlašu $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ sadalījuma funkcijām F un G pastāv lokācijas-skalēšanas modelis, ja

$$F(x) = G\left(\frac{x - \mu}{\sigma}\right), \ x \in \mathbb{R}, \tag{1.2}$$

kur σ ir skalēšanas parametrs un μ – lokācijas parametrs.

Ja $\sigma = 1$, tad saka, ka starp divu izlašu sadalījuma funkcijām pastāv lokācijas modelis, ja $\mu = 0$, tad saka, ka starp izlašu sadalījuma funkcijām pastāv skalēšanas modelis.

Definīcija 3. Par kvantiļu funkciju F^{-1} sauc sadalījuma funkcijas F kreiso inverso funkciju, kas tiek definēta

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}, \ t \in [0, 1], \ x \in \mathbb{R}. \tag{1.3}$$

Definīcija 4. Par empīrisko sadalījuma funkciju sauc funkciju

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}, \tag{1.4}$$

kur X_i ir izlases elementi, n – izlases apjoms un

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

Empīriskā kvantiļu funkcija F_n^{-1} ir attiecīgi empīriskās sadalījuma funkcijas F_n kreisā inversā funkcija:

$$F_n^{-1}(t) = \inf\{x : F_n(x) \ge t\},\ t \in [0,1],\ x \in \mathbb{R}. \tag{1.5}$$

Vienādojumu (1.2) var ekvivalenti izteikt arī ar kvantiļu funkcijām:

$$F^{-1}(t) = \sigma G^{-1}(t) + \mu,\ t \in [0,1]. \tag{1.6}$$

Definīcija 5. Starp divu izlašu $X_1, \dots, X_n$ un $Y_1, \dots, Y_m$ sadalījuma funkcijām F un G pastāv Lēmaņa alternatīvais modelis, ja

$$F(x) = 1 - (1 - G(x))^{1/h},\ x \in \mathbb{R},\ h > 0.$$

Izmantojot kvantiļu funkcijas, iegūst

$$F^{-1} = G^{-1}(1 - (1-t)^h),\ t \in [0,1]. \tag{1.7}$$

Strukturālo attiecību modeļi, kas izteikti ar kvantiļu funkcijām, kā (1.6) un (1.7), var tikt pārveidoti vispārējā formā. Pieņemsim, ka sadalījuma funkcijas F un G ir no funkciju klases

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

Definīcija 6. [1] Pieņemsim, ka $\mathcal{H} \subseteq \mathbb{R}^l$ un $\phi_1 : \mathbb{R} \times \mathcal{H} \to \mathbb{R}$, $\phi_2 : [0,1] \times \mathcal{H} \to [0,1]$. Funkcijas F un G ir saistītas ar strukturālu attiecību, kuru veido ϕ_1 un ϕ_2, ja $(F,G) \in \mathcal{U}_{\phi_1,\phi_2} =: \mathcal{U}$, kur modeļu klase $\mathcal{U}$ tiek definēta kā

$$\mathcal{U} := \left\{(F,G) \in \mathcal{F}^2 \times \mathcal{F}^2 \mid \exists h \in \mathcal{H} : F^{-1}(t) = \phi_1(G^{-1}(\phi_2(t,h)),h),\ t \in [0,1]\right\}.$$

No definīcijas 6 izriet, ka izvēloties $\phi_1(x,(h_1,h_2)^T) = h_1 + h_2 x$ un $\phi_2(t,h) \equiv t$, tiek iegūts lokācijas-skalēšanas modelis, bet izvēloties $\phi_1(x,h) \equiv x$ un $\phi_2(t,h) = 1-(1-t)^h$ tiek iegūts Lēmaņa alternatīvais modelis.

Parametra h novērtējumam tiek izmantots Mallova attālums [9], ar kuru var tikt aprēķināts attālums starp kvantiļu funkcijām. Vispārīgā gadījumā teorētiskais parametrs h_0 strukturālo attiecību modelim tiek aprēķināts

$$h_0 = argmin_{h\in\mathcal{H}} \left\{ \frac{1}{b-a} \int_a^b F^{-1}(t) - \phi_1(G^{-1}(\phi_2(t,h)))^2 dt \right\}.$$

Teorētiskās funkcijas aizvietojot ar to empīriskajām versijām, iegūst parametra novērtējumu.

1.1.1. Hipotēžu pārbaude fiksēta lokācijas parametra gadījumā

Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju F un $Y_1, \ldots, Y_n$ ir neatkarīgi vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju G. Divu izlašu sadalījuma funkciju vienādības pārbaudei var izmantot Kolmogorova-Smirnova testu divu izlašu gadījumam. Testa statistika tiek uzdota šādi:

$$D = \sup_x \sqrt{\frac{nm}{n+m}} |F_n(x) - G_m(x)|,$$

kur n un m ir izlašu apjomi, bet F_n un G_m – izlašu empīriskās sadalījuma funkcijas (skat. (1.4)).

Kolmogorova-Smirnova testu var izmantot hipotēžu pārbaudei par lokācijas modeli. Šajā gadījumā var definēt sadalījuma funkciju $G_\theta(x) = G(x) - \theta$, ko iegūst otrās izlases sadalījuma funkcijai atņemot lokācijas parametru θ, un veikt hipotēžu pārbaudi:

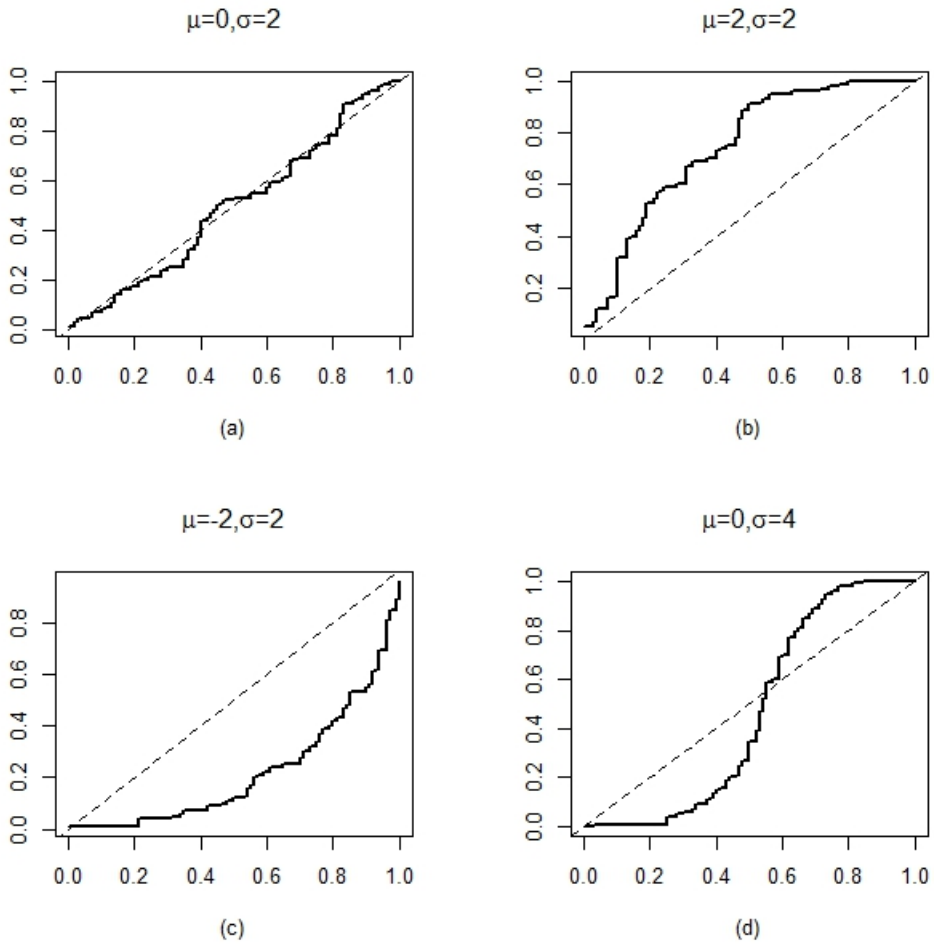
$$\begin{aligned} H_0 &: F(x) = G_\theta(x) \\ H_1 &: F(x) \neq G_\theta(x) \end{aligned} \tag{1.8}$$

1.1.2. Varbūtību-varbūtību un kvantiļu-kvantiļu grafiku pielietojums empīriskajos procesos

Cielēns [3] savā diplomdarbā ir pētījis iespēju pārbaudīt hipotēzi par lokācijas modeļa eksistenci, konstruējot vienlaicīgās ticamības joslas varbūtību-varbūtību (P-P) un kvantiļu-kvantiļu (Q-Q) grafikiem.

Definīcija 7. Par varbūtību-varbūtību grafiku sauc funkciju $F(G^{-1}(t))$, $0 < t < 1$.

Aizstājot abu izlašu sadalījuma funkcijas ar to empīriskajām versijām (1.4) un (1.5), iegūst empīrisko P-P grafiku.



1. att.: Empīrisko P-P grafiku piemēri sadalījuma likumiem $N(0,1)$ un $N(\mu,\sigma)$. Ģenerēto izlašu apjomi $n = 100$.

1. attēlā redzams, ka varbūtību-varbūtību grafiki vienmēr atrodas intervālā $(0,1)$. Ja izlašu sadalījuma likumi ir vienādi, tad empīriskais P-P grafiks ir tuvs taisnei $y = t$ (a). Ja izlašu sadalījuma likumi ir no vienas klases, bet otrās izlases matemātiskā cerība ir lielāka nekā pirmās izlases matemātiskā cerība, tad grafiks noliecas virs diagonāles (b). Ja otrās izlases matemātiskā cerība ir mazāka nekā pirmās izlases matemātiskā cerība, tad grafiks noliecas zem diagonāles (c). Gadījumā, ja izlašu sadalījuma likumi ir no vienas klases, bet atšķiras dispersija, tad grafiks atrodas gan virs diagonāles, gan zem tās (d).

Palielinot izlases apjomu, empīriskais varbūtību-varbūtību grafiks tiecas uz teorētisko. Šo grafiku starpību normē ar $\sqrt{n}$, lai pētītu to pie dažādiem izlašu apjomiem, un iegūst empīrisko procesu divu izlašu gadījumā.

Definīcija 8. [10] Par empīrisko varbūtību-varbūtību procesu sauc

$$\Delta_{nm}(t) = \sqrt{n} \sup_{0<t<1} |F_n(G_m^{-1}(t)) - F(G^{-1}(t))|.$$

Definīcija 9. Par kvantiļu-kvantiļu grafiku sauc funkciju $F^{-1}(G(x))$, $x \in \mathbb{R}$.

Aizstājot abu izlašu sadalījuma funkcijas ar to empīriskajām versijām (1.4) un (1.5), iegūst empīrisko Q-Q grafiku.

Definīcija 10. [10] Par empīrisko kvantiļu-kvantiļu procesu sauc

$$\delta_{nm}(x) = \sqrt{n} \sup_{-\infty<x<+\infty} |f(F^{-1}(G(x)))(F_n^{-1}(G_m(x)) - F^{-1}(G(x)))|. \tag{1.9}$$

Izmantojot šo procesu asimptotiskos sadalījumus, var iegūt kritisko vērtību c pie fiksēta nozīmības līmeņa α (skat. [3]).
Ticamības joslas tiek konstruētas šādi:

1. P-P grafikam:

$$P\left(F_n(G_m^{-1}(t)) - \frac{c}{\sqrt{n}} \leq F(G^{-1}(t)) \leq F_n(G_m^{-1}(t)) + \frac{c}{\sqrt{n}}\right) = 1-\alpha,\ \forall t \in (0,1);$$

2. Q-Q grafikam:

$$P\left(F_n^{-1}(G_m(x)) - \frac{c}{\sqrt{n}f(F^{-1}(G(x)))} \leq F^{-1}(G(x)) \leq F_n^{-1}(G_m(x)) + \frac{c}{\sqrt{n}f(F^{-1}(G(x)))}\right) = 1-\alpha,\ \forall x \in \mathbb{R}.$$

Apgalvojums 1. *Ja nulles hipotēze* (1.8) *ir spēkā, tad P-P procesam* Δ_{nm} *izpildās*

$$\forall \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F(G^{-1}(t)) - t) - (B^{(n)}(t) + \sqrt{\frac{n}{m}}B^{(m)}(t))| > \epsilon) = 0 \text{ } g.d.$$

Apgalvojums 2. *Ja nulles hipotēze* (1.8) *ir spēkā, tad Q-Q procesam* δ_{nm} *izpildās*

$$\forall \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{-\infty<x<+\infty} |\sqrt{n}f(x)(F_n^{-1}(G_m(x)) - x) - (B^{(n)}(G(x)) + \sqrt{\frac{n}{m}}B^{(m)}(G(x)))| > \epsilon) = 0 \text{ } g.d.$$

Definīcija 11. Stohastisku procesu $\{X(t) : t \in T\}$ sauc par Gausa procesu, ja $\forall t_1, \ldots, t_n \in T$ vektoram $(X(t_1), X(t_2), \ldots, X(t_n))$ ir daudzdimensionālais normālais sadalījums.

Definīcija 12. Stohastisku procesu $\{W(t) : t \geq 0\}$ sauc par Brauna kustību ar sākumu punktā $x \in \mathbb{R}$, ja

1. $W(0) = x$;

2. pieaugumi $W(t_i) - W(t_{i-1})$ ir neatkarīgi $\forall t_i$, $i = 1, 2 \ldots, n$;

3. $\forall t \geq 0$ un $\forall h > 0 : W(t+h) - W(t) \sim N(0, h)$;

4. funkcija $t \mapsto W(t)$ ir nepārtraukta gandrīz droši.

Ja $x = 0$, tad $\{W(t) : t \geq 0\}$ sauc par standarta Brauna kustību.

Definīcija 13. Par Brauna tiltu sauc Gausa procesu $\{B(t) : t \in [0, 1]\}$, kura kovariāciju struktūra ir $cov(B(s), B(t)) = s(1-t),\ s < t$. Šī procesa asimptotiskais sadalījums ir tāds pats, kā $W(t) - tW(1)$ sadalījums, kur $W(t)$ – standarta Brauna kustība.

Redzams, ja H_0 ir spēkā, abu procesu Δ_{nm} un δ_{nm} asimptotiskais sadalījums ir divu neatkarīgu Brauna tiltu summa. Šādam sadalījumam var aprēķināt kritisko vērtību un konstruēt vienlaicīgās ticamības joslas P-P un Q-Q grafikiem. Veicot simulācijas pie $\alpha = 0.05$, kritiskā vērtība iznāca 1.88.

Hipotēze H_0 par divu izlašu sadalījumu vienādību tiek noraidīta, ja P-P grafikam konstruētās ticamības joslas neiekļauj taisni $y = t,\ t \in [0, 1]$, kā arī gadījumā, kad Q-Q grafikam konstruētās ticamības joslas neiekļauj taisni $y = x,\ x \in \mathbb{R}$.

Lai konstruētu vienlaicīgās ticamības joslas kvantiļu-kvantiļu grafikam, nepieciešams novērtēt pirmās izlases blīvuma funkciju. To ir iespējams izdarīt, izmantojot blīvuma funkcijas neparametrisko kodolu gludināšanas metodi.

Blīvuma funkcijas neparametriskā kodolu gludināšanas metode

Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi, vienādi sadalīti gadījuma lielumi ar nezināmu sadalījuma funkciju $F(x)$ un blīvuma funkciju $f(x) = F'(x)$. Zināms, ka histogramma aproksimē teorētisko blīvuma funkciju, neatkarīgi no sadalījuma veida. Pēc definīcijas

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}.$$

Aizstājot sadalījuma funkciju F ar tās novērtējumu F_n (1.4), iegūst

$$\hat{f}(x) = \lim_{h \to 0} \frac{F_n(x+h) - F_n(x)}{h}.$$

Atmetot robežu, iegūst blīvuma funkcijas novērtējumu

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I_{\{x < X_i \leq x+h\}},$$

kur h ir joslas platums, kas ir atkarīgs no izlases apjoma n, un

$$I_{\{x<X_i\leq x+h\}} = \begin{cases} 1, & X_i \in (x, x+h] \\ 0, & X_i \notin (x, x+h] \end{cases}.$$

Izvēloties intervālu simetriski pret punktu x, iegūst

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} I_{\{|x-X_i|<h\}} = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{2} I_{\{|\frac{x-X_i}{h}|<1\}}.$$

Definīcija 14. Pieņemsim, ka K ir kodols un h ir joslas platums, tad par gludināto blīvuma funkciju sauc funkciju

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{1.10}$$

kur K ir kāda sadalījuma funkcija, saukta par kodolu.

Šīs idejas autors ir *Rosenblatt* [11]. Kodolu funkcijas piemēri:

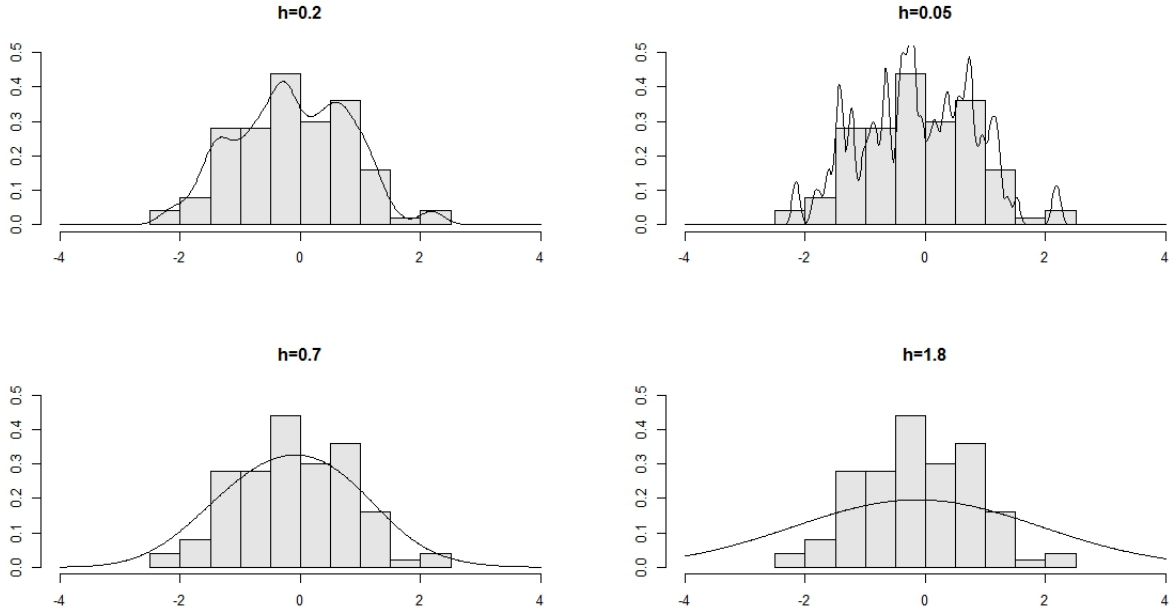
1. Vienmērīgais kodols $K(u) = \frac{1}{2} I_{\{-1\leq u\leq 1\}}$;
2. Normālais jeb Gausa kodols $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$;
3. Bi-kvadrātiskais kodols $K(u) = \frac{15}{16}(u^2-1)^2 I_{\{|u|\leq 1\}}$.

Vispārīgā gadījumā būtiskākie pieņēmumi: $h \to 0$, $nh \to \infty$, kad $n \to \infty$,

1. $\int K(u)du = 1$;
2. $\int uK(u)du = 0$;
3. $\int u^2 K(u)du < \infty$.

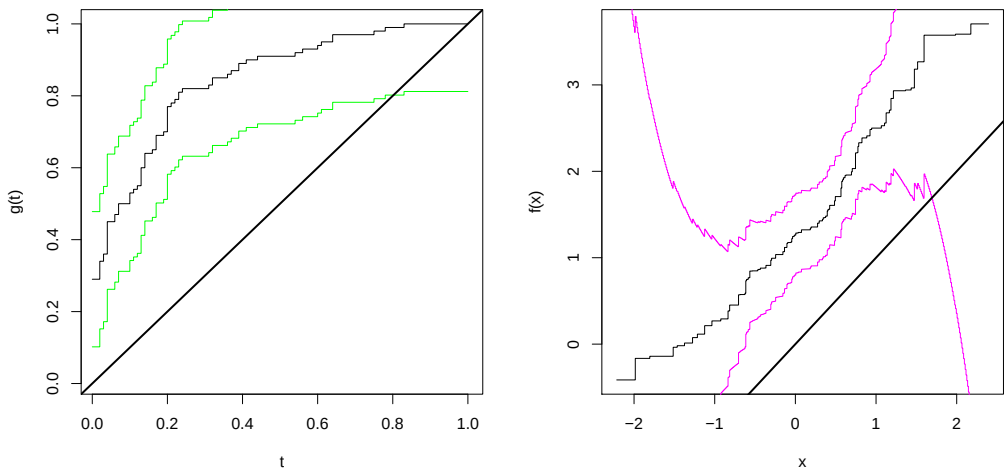
Praksē pielietojot blīvuma funkcijas neparametrisko kodolu gludināšanas metodi pastāv divas problēmas – kodola K un joslas platuma h izvēle. Izrādās, ka svarīgāka ir joslas platuma h izvēle, ko parāda 2. attēls.

Tika simulēti dati no normālā sadalījuma ar parametriem 0 un 1 un pielietota neparametriskā kodolu gludināšanas metode blīvuma funkcijas novērtējumam, izvēloties dažādus joslas platumus h. Redzams, ka izvēloties pārāk mazu vai pārāk lielu joslas platumu, blīvuma funkcijas novērtējums nav pietiekami labs. Literatūrā sastopamas dažādas h novērtēšanas metodes, no kurām pazīstamākās ir krosvalidācijas un ievietošanas metodes. h izvēlei skat. [12].



2. att.: Histogramma un tās neparametriskais novērtējums izmantojot kodolu gludināšanu ar Gausa kodolu un dažādām joslas platuma h izvēlēm.

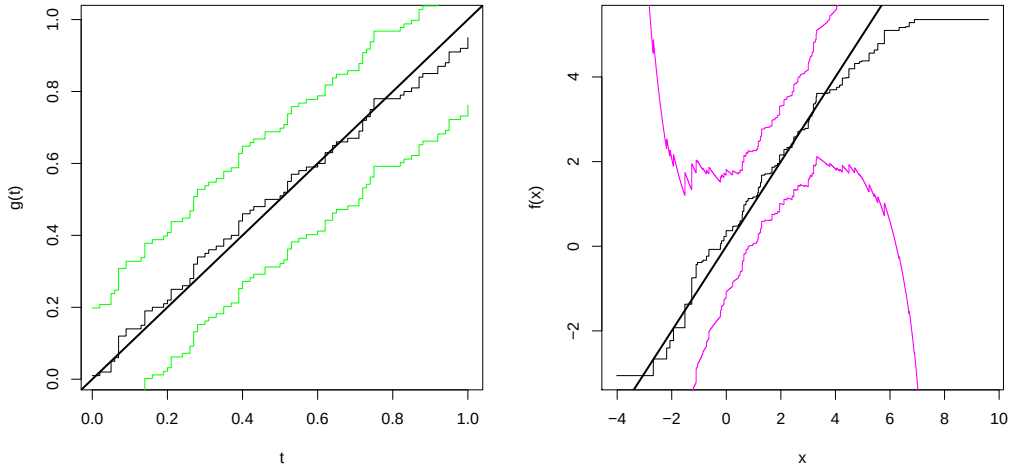
Pārbaudīsim hipotēzi par divu izlašu sadalījumu funkciju vienādību, izmantojot varbūtību-varbūtību un kvantiļu-kvantiļu grafikus. Izlases $X \sim N(0,1)$ un $Y \sim N(1.5,1)$ satur neatkarīgi ģenerētus gadījuma lielumus apjomā $n = 100$. 3. attēlā redzami izlašu empīriskie varbūtību-varbūtību (pa labi) un kvantiļu-kvantiļu (pa kreisi) grafiki ar vienlaicīgās ticamības joslām. Tā kā diagonāles neiekļaujas ticamības joslās, tad hipotēze par abu sadalījumu funkciju vienādību tiek noraidīta.



3. att. Noraidītas hipotēzes piemērs, $X \sim N(0,1)$, $Y \sim N(1.5,1)$, $n = 100$, $\alpha = 0.05$.

4. attēlā redzami izlašu $X \sim N(2,2)$ un $Y \sim N(2,2)$ empīriskie varbūtību-varbūtību

(pa labi) un kvantiļu-kvantiļu (pa kreisi) grafiki ar vienlaicīgās ticamības joslām. Tā kā diagonāles iekļaujas ticamības joslās, tad hipotēze par abu sadalījumu funkciju vienādību netiek noraidīta.



4. att. Nenoraidītas hipotēzes piemērs, $X \sim N(2,2)$, $Y \sim N(2,2)$, $n = 100$, $\alpha = 0.05$.

1.2. Pārbīdes funkcijas vienlaicīgās ticamības joslas

Doksum un *Sievers* savā publikācijā [2] lokācijas modeļa $F(x) = G(x+\theta)$, $\forall x$ vietā aplūkoja vispārīgu gadījumu $F(x) = G(x+\Delta(x))$ kādai funkcijai $\Delta(x)$, kuru nosauca par vispārīgu pārbīdes funkciju.

Definīcija 15. Par divu izlašu $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ ar sadalījuma funkcijām attiecīgi F un G vispārīgo pārbīdes funkciju sauc funkciju

$$\Delta(x) = G^{-1}(F(x)) - x, \ x \in \mathbb{R}. \tag{1.11}$$

Izlašu sadalījuma funkciju F un G vietā ievietojot to empīriskās versijas F_n un G_m, iegūst empīrisko pārbīdes funkciju

$$\hat{\Delta}(x) = G_m^{-1}(F_n(x)) - x, \ x \in \mathbb{R}.$$

Gadījumā, kad pastāv lokācijas modelis, $\Delta(x) = \theta$.

Teorēma 3. [8] *Pieņemsim, ka sadalījuma funkcijai $G(x)$ ir nepārtraukts atvasinājums $g(x)$, kurš apmierina nevienādību $0 < g(x) < \infty$. Tad*

$$\sqrt{n+m}\left[\hat{\Delta}(x) - \Delta(x)\right] \xrightarrow{d} \frac{B(F(x))}{g(G^{-1}(F(x)))\sqrt{\lambda(1-\lambda)}}, \tag{1.12}$$

kur $\lambda \in [0,1]$, $\lambda_N = \dfrac{n}{N}$ un $\lambda_N \to \lambda$, kad $n \to \infty$, un B ir Brauna tilts (skat. definīciju 13).

Vispārīgo pārbīdes funkciju ļoti bieži pielieto medicīnā, ieviešot jaunas zāles. Tāpēc *Doksum* un *Sievers* savā publikācijā [2] izvirzīja šādus jautājumus:

1. Vai jauno zāļu iedarbība ir pozitīva visiem populācijas locekļiem, t. i., $\forall x : \Delta(x) > 0$?

2. Ja uz pirmo jautājumu atbilde ir noraidoša, tad kādai populācijas daļai zāļu iedarbība ir pozitīva, t. i., $\{x : \Delta(x) > 0\}$?

3. Vai pastāv lokācijas modelis, t. i., vai $\exists\, \theta$ tāda, ka $\forall x : \Delta(x) = \theta$?

4. Ja uz iepriekšējo jautājumu atbilde ir noraidoša, vai eksistē α un β tādi, ka $\forall x : \Delta(x) = \alpha + \beta x$?

Uz šiem jautājumiem atbildes var sniegt pārbīdes funkcijas $\Delta(x)$ vienlaicīgās ticamības joslas $(\Delta_*(x), \Delta^*(x))$.

1. Jauno zāļu iedarbība ir pozitīva visiem populācijas locekļiem, ja $\forall x : \Delta_*(x) > 0$.

2. Jauno zāļu iedarbība ir pozitīva tai populācijas daļai, kurai $\{x : \Delta_*(x) > 0\}$.

3. Lokācijas modelis nepastāv, t. i., neeksistē šāda θ, ja nav tādas horizontālas taisnes, kura ietilpst ticamības joslās.

4. Šādi α un β neeksistē, ja nav tādas taisnes, kura ietilpst ticamības joslās.

Ievieš apzīmējumus $N = n + m$ un $M = nm/N$. *Doksum* un *Sievers* [2] apskatīja divas ticamības joslu konstruēšanas iespējas. Pirmā no tām balstīta uz Kolmogorova-Smirnova statistiku

$$T(F_n, G_m) = D_N = \sqrt{M} \sup_x |F_n(x) - G_m(x)|.$$

Teorēma 4. [2] *Pārbīdes funkcijas $\Delta(x)$ $1-\alpha$ vienlaicīgās ticamības joslas tiek uzdotas*

$$\left(G_m^{-1}\left(F_n(x) - \frac{K_{S,\alpha}}{\sqrt{M}}\right) - x, G_m^{-1}\left(F_n(x) + \frac{K_{S,\alpha}}{\sqrt{M}}\right) - x \right),\ x \in \mathbb{R}, \tag{1.13}$$

kur $K_{S,\alpha}$ ir izvēlēts no Kolmogorova-Smirnova tabulām tā, lai $P(D_n \le K_{S,\alpha}) = 1 - \alpha$.

Šīs ticamības joslas sauc par S ticamības joslām un apzīmē ar $(S_*(x), S^*(x))$. Ievieš apzīmējumus:

- $[t]$ – skaitļa t veselā daļa;

- $\langle t \rangle$ – vismazākais veselais skaitlis, kurš ir lielāks vai vienāds ar t;

- $X_1 < \ldots < X_n$ un $Y_1 < \ldots < Y_m$ – izlašu X un Y statistikas;
- $Y(j) = -\infty (j < 0)$, $Y(j) = \infty (j \geq m+1)$.

Izmantojot ieviestos apzīmējumus S ticamības joslas (1.13) var pārrakstīt:

$$(S_*(x), S^*(x)) = \left(Y\left\{ \left\langle m\left(\frac{i}{n} - \frac{K_{S,\alpha}}{\sqrt{M}} \right) \right\rangle \right\} - x, Y\left\{ \left[m\left(\frac{i}{n} + \frac{K_{S,\alpha}}{\sqrt{M}} \right) \right] + 1 \right\} - x \right),$$

$\forall x \in [X(i), X(i+1)),\ i = 0, 1, \ldots, n$ un $X(0) = -\infty,\ X(n+1) = \infty$.

Doksum un *Sievers* publikācijā [2] apskatīja arī vienlaicīgās ticamības joslas, kas balstītas uz svērto suprēma normas statistiku:

$$W_N = W_N(F_n, G_m) = \sqrt{M} \sup_{\{x: a \leq F_n(x) \leq b\}} \frac{|F_n(x) - G_m(x)|}{\Psi(H_N(x))},$$

kur $H_N(x) = \lambda F_n(x) + (1-\lambda) G_m(x),\ \lambda = n/N$ un $0 \leq a < b \leq 1$.
Izvēloties $\Psi(t) = \sqrt{t(1-t)}$, katram x iegūst aptuveni vienādu svaru tādā nozīmē, ka

$$\frac{\sqrt{M}(F_n(x) - G_m(x))}{\Psi(H_N(x))}$$

ir asimptotiskā dispersija neatkarīgi no x. Risinot nevienādību $|W_N(F_n, G_m)| \leq K$ attiecībā pret G_m un ievietojot $\Psi(t) = \sqrt{t(1-t)}$, iegūst

$$(G_m(x) - F_n(x))^2 \leq K^2(\lambda F_n(x) + (1-\lambda) G_m(x)) \frac{1 - (\lambda F_n(x) + (1-\lambda) G_m(x))}{M}, \quad (1.14)$$

$\forall x \in \{x : a \leq F_n(x) \leq b\}$.
Apzīmē $c = K^2/M,\ u = F_n(x)$ un $v = G_m(x)$. Tad (1.14) var pārrakstīt kā $d(v) \leq 0$, kur

$$d(v) = (1 + c(1-\lambda)^2)v^2 - (2u - c(1-\lambda)(2\lambda u - 1))v + u^2 - c\lambda u + c\lambda^2 u^2.$$

Tā kā $v^2 > 0$, $d(v) \leq 0$ tad un tikai tad, ja v atrodas starp divām reālām vienādojuma $d(v) = 0$ saknēm.

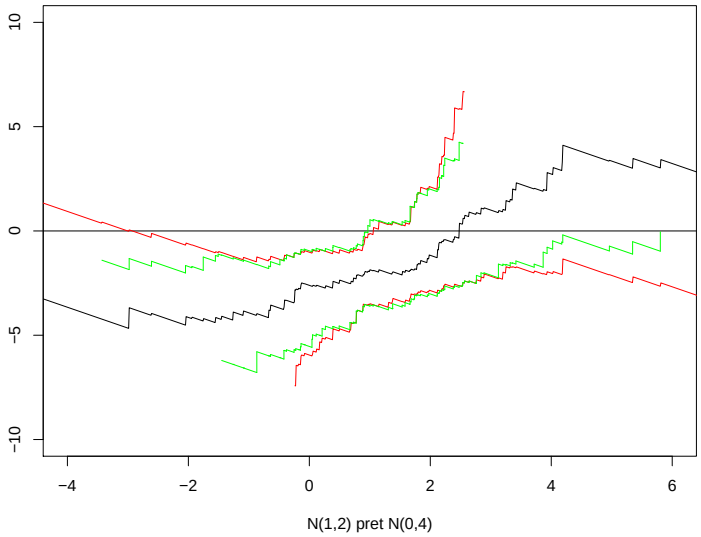
Teorēma 5. [2] *Ja* $P_{F=G}(W_N \leq K) = 1 - \alpha$, *tad uz* W_N *ar* $\Psi(t) = \sqrt{t(1-t)}$ *balstītas pārbīdes funkcijas* $\Delta(x)$ *vienlaicīgās ticamības joslas tiek uzdotas*

$$(G_m^{-1}(h^-(F_n(x))) - x, G_m^{-1}(h^+(F_n(x))) - x),\ x \in \{x : a \leq F_n(x) \leq b\},$$

kur

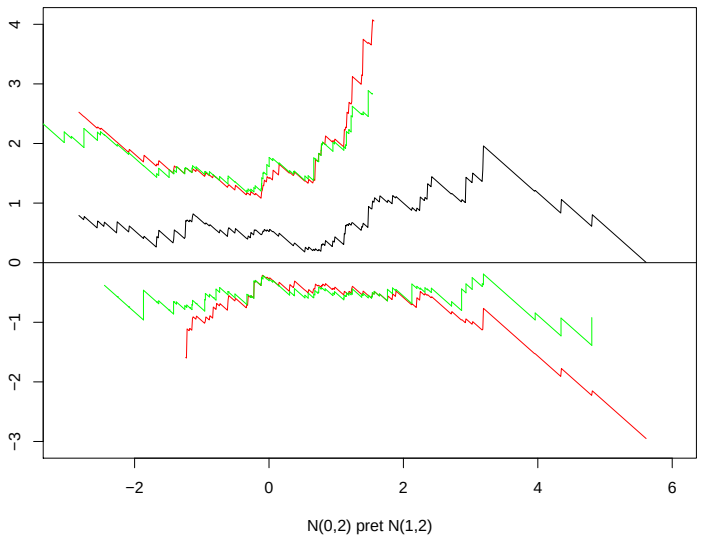
$$h^{\pm}(u) = \frac{u + \frac{1}{2}c(1-\lambda)(1-2\lambda u) \pm \frac{1}{2}\sqrt{c^2(1-\lambda)^2 + 4cu(1-u)}}{1 + c(1-\lambda)^2}.$$

un $c = \dfrac{K^2}{M}$.
Šīs ticamības joslas sauc par W ticamības joslām un apzīmē ar $(W_*(x), W^*(x))$.



5. att. Vienlaicīgās ticamības joslas pārbīdes funkcijai.

5. attēlā redzams izlašu $X \sim N(1,2)$ un $Y \sim N(0,4)$ pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām. Sarkanajā krāsā attēlotas $(S_*(x), S^*(x))$ vienlaicīgās ticamības joslas un zaļā krāsā – $(W_*(x), W^*(x))$ vienlaicīgās ticamības joslas. Redzams, ka $(W_*(x), W^*(x))$ joslas galos ir šaurākas par $(S_*(x), S^*(x))$ joslām. Tā kā neviena horizontāla taisne neietilpst ticamības joslās, tad var izdarīt secinājumu, ka starp izlasēm X un Y nepastāv lokācijas modelis. Starp šīm izlasēm var pastāvēt lokācijas-skalēšanas modelis, jo var atrast taisni, kas iekļaujas šajās ticamības joslās.



6. att. Vienlaicīgās ticamības joslas pārbīdes funkcijai.

6. attēlā redzams izlašu $X \sim N(0,2)$ un $Y \sim N(1,2)$ pārbīdes funkcijas novērtējums

ar vienlaicīgās ticamības joslām. Sarkanajā krāsā attēlotas $(S_*(x), S^*(x))$ vienlaicīgās ticamības joslas un zaļā krāsā – $(W_*(x), W^*(x))$ vienlaicīgās ticamības joslas. Šajā gadījumā nevaram noraidīt hipotēzi par lokācijas modeli starp šīm izlasēm, jo var atrast horizontālu taisni, kas ietilpst ticamības joslās. Tāpat redzams, ka pārbīdes funkcija $\Delta(x)$ ir tuva taisnei $y = 1$, tātad šo izlašu lokācijas parametrs $\theta = 1$, kas atbilst patiesībai.

1.3. Kvantiļu starpības funkcija

Izrādās, ka pārbaudīt hipotēzi par lokācijas modeli (1.8) var, konstruējot vienlaicīgās ticamības joslas divu kvantiļu funkciju starpībai. *P. Laake*, *K. Laake* un *R. Aaberge* 1985. gadā savā publikācijā [4], analizējot saikni starp hospitalizāciju un mirstību, definēja kvantiļu starpības funkciju:

$$\Lambda(t) = F^{-1}(t) - G^{-1}(t),\ 0 \le t \le 1. \tag{1.15}$$

Atzīmēsim, ka

$$\int_0^1 \Lambda(t)dt = E(X) - E(Y).$$

Atcerēsimies, ka *Doksum* un *Sievers* savā publikācijā [2] definēja pārbīdes funkciju $\Delta(x) = G^{-1}(F(x)) - x$, kas patiesībā ir horizontālais attālums starp F un G attiecīgajā punktā x. Acīmredzami $\Lambda(F(x)) = -\Delta(x)$.

Kvantiļu funkcijas aizvietojot ar to empīriskajām versijām (1.5), iegūstam $\Lambda(t)$ novērtējumu

$$\hat{\Lambda}(t) = F_n^{-1}(t) - G_m^{-1}(t),\ 0 \le t \le 1.$$

Tiek izmantots, ka $\sqrt{n}|F_n^{-1}(t) - F^{-1}(t)|$ un $\sqrt{m}|G_m^{-1}(t) - G^{-1}(t)|$ konverģē uz Gausa procesu (skat. definīciju 11).

Tiek pieņemts, ka $f(x)$ un $g(y)$ ir $F(x)$ un $G(y)$ nepārtraukti atvasinājumi, kas apmierina $0 < f(x) < \infty$ un $0 < g(y) < \infty$. Ar $B_1(t)$ un $B_2(t)$ tiek apzīmēti Brauna tilti (skat. definīciju 13).

Apgalvojums 6.

$$\sqrt{n}|F_n^{-1}(t) - F^{-1}(t)| \xrightarrow{d} \frac{B_1(t)}{f(F^{-1}(t))}.$$

Apgalvojums 7.

$$\sqrt{m}|G_m^{-1}(t) - G^{-1}(t)| \xrightarrow{d} \frac{B_2(t)}{g(G^{-1}(t))}.$$

Ja $n, m \to \infty$ tā, ka $\dfrac{n}{N} \to \theta$, kur $N = n + m$, tad

$$\sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)| = \sqrt{\frac{N}{n}}\sqrt{n}|F_n^{-1}(t) - F^{-1}(t)| - \sqrt{\frac{N}{m}}\sqrt{m}|G_m^{-1}(t) - G^{-1}(t)|,$$

kas nozīmē

$$\sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)| \xrightarrow{d} \frac{1}{\sqrt{\theta}} \frac{B_1(t)}{f(F^{-1}(t))} - \frac{1}{\sqrt{1-\theta}} \frac{B_2(t)}{g(G^{-1}(t))}. \tag{1.16}$$

Tas, savukārt, nozīmē, ka $\sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)|$ asimptotiskā dispersija ir

$$k^2(t) = \left\{ \frac{1}{\theta f^2(F^{-1}(t))} + \frac{1}{(1-\theta) g^2(G^{-1}(t))} \right\} t(1-t).$$

Turpmāk, pētot šo problēmu, publikācijā [4] tiek izdarīts pieņēmums, ka otrās izlases sadalījuma funkcija G ir zināma. Līdz ar to tiek salīdzināta izlase, kuras sadalījuma funkcija ir zināma, ar izlasi, kuras sadalījuma funkcija nav zināma. Šajā gadījuma iegūst

$$\sqrt{n}|\hat{\Lambda}(t) - \Lambda(t)| = \sqrt{n}|F_n^{-1}(t) - F^{-1}(t)|$$

un

$$\sqrt{n}|\hat{\Lambda}(t) - \Lambda(t)| \xrightarrow{d} \frac{B_1(t)}{f(F^{-1}(t))}.$$

Gadījumā, kad G ir zināms, $\sqrt{n}|\hat{\Lambda}(t) - \Lambda(t)|$ asimptotiskā dispersija ir

$$\gamma^2(t) = \frac{1}{f^2(F^{-1}(t))} t(1-t).$$

Tās nezināmā daļa ir $f^2(F^{-1}(t))$. Blīvuma funkcijas novērtējumam var pielietot (1.10). Sadalījuma funkciju F var novērtēt, integrējot gludināto blīvuma funkciju.

Izmanto Kolmogorova-Smirnova statistiku $\sqrt{n} \sup_t |F_n(F^{-1}(t)) - t|$ un iegūst

$$\sqrt{n} \sup_t |F_n(F^{-1}(t)) - t| = \sqrt{n} \sup_t |F_n(\Lambda(t) + G^{-1}(t)) - t|.$$

No tā seko

$$P(\sqrt{n} \sup_t |F_n(\Lambda(t) + G^{-1}(t)) - t| \leq k) = 1 - \alpha,$$

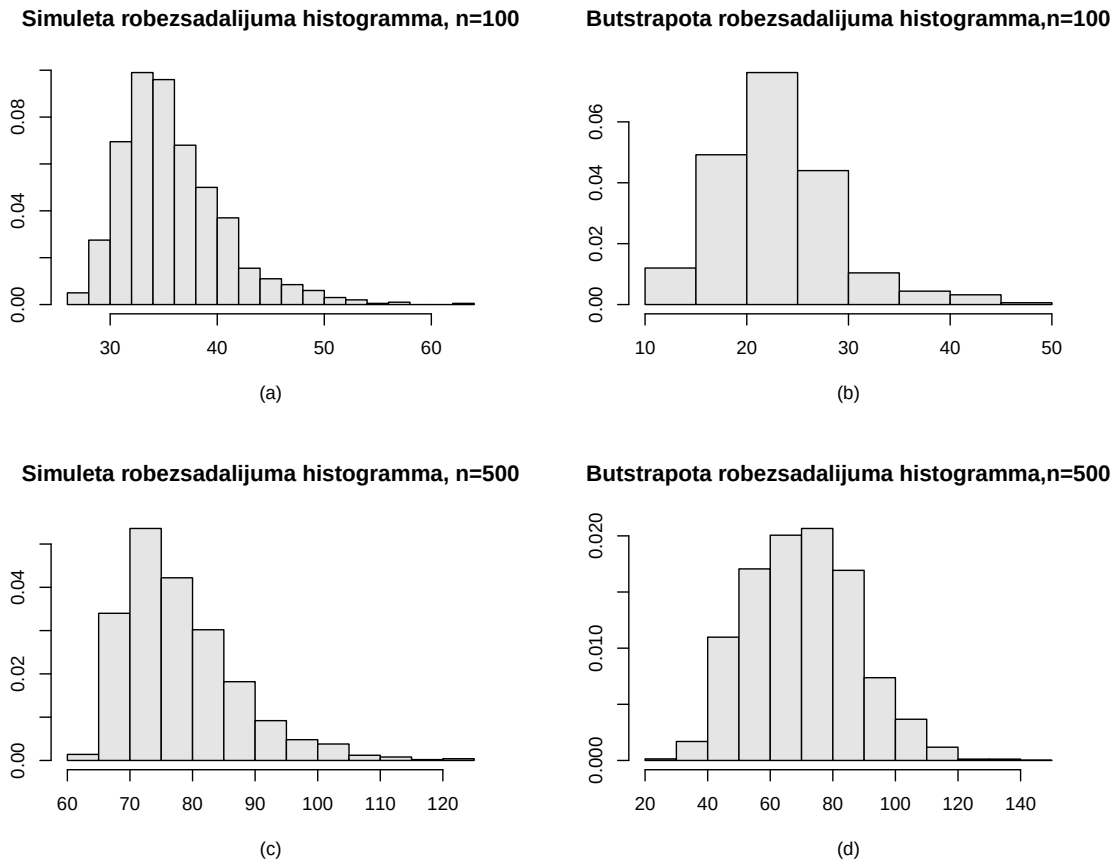
un kvaniļu starpības funkcijas $\Lambda(t)$ vienlaicīgās ticamības joslas tiek uzdotas

$$\left(F_n^{-1}\left(t - \frac{k}{\sqrt{n}}\right) - G^{-1}(t), F_n^{-1}\left(t + \frac{k}{\sqrt{n}}\right) - G^{-1}(t) \right),$$

kur k ir Kolmogorova-Smirnova statistikas $(1-\alpha)$ kvantile.

Tomēr praktiskajos pētījumos izlasēm sadalījuma funkcija G bieži vien nav zināma, tāpēc šī pieeja vairākumos gadījumu nebūs derīga. Tā kā nav zināms robežsadalījums, jāizmanto suprēma statistika

$$\sup_{0<t<1} \sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)|. \tag{1.17}$$



7. att.: Statistikas (1.17) robežsadalījuma histogrammas. (a) un (c) gadījumā robežsadalījumi iegūti, simulējot izlases apjomā 100 un 500 ar normālo sadalījumu $N(0,1)$ un $N(1,1)$, bet (b) un (d) gadījumā robežsadalījumi iegūti, butstrapojot izlases $X \sim N(0,1)$ un $Y \sim N(1,1)$.

7. attēls parāda robežsadalījuma problemātiku – palielinot n, tas konverģē aizvien tālāk. Tāpēc šo statistiku pielietot nevar, jāmeklē cita pieeja, lai konstruētu vienlaicīgās ticamības joslas kvantiļu starpības funkcijai.

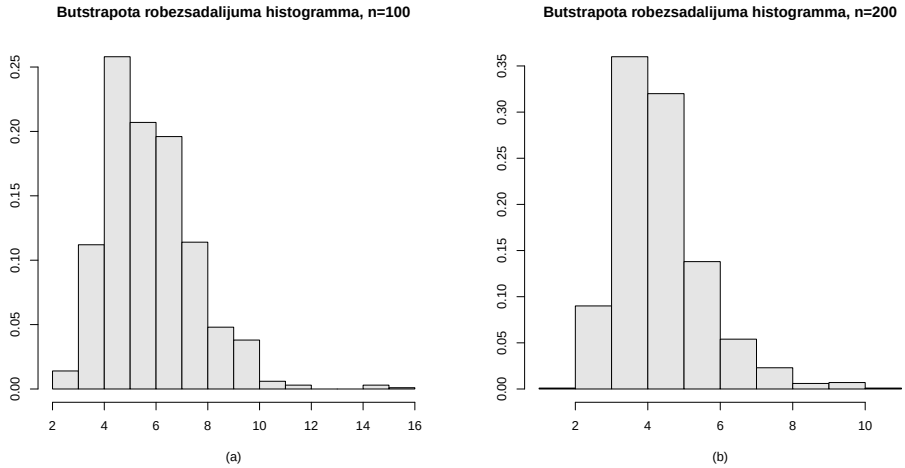
Radās ideja statistiku (1.17) reizināt ar $f(F^{-1}(t))$ līdzīgi kā Q-Q procesa gadījumā (1.9):

$$\sup_{0<t<1} f(F^{-1}(t))\sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)|. \tag{1.18}$$

Līdz ar to no (1.16) iegūst

$$f(F^{-1}(t))\sqrt{N}|\hat{\Lambda}(t) - \Lambda(t)| \xrightarrow{d} \frac{1}{\sqrt{\theta}} B_1(t) - \frac{B_2(t)}{\sqrt{1-\theta}} \frac{f(F^{-1}(t))}{g(G^{-1}(t))}.$$

Ja H_0 (1.8) ir spēkā, tad $\dfrac{f(F^{-1}(t))}{g(G^{-1}(t))} = const$. Šajā gadījumā empīriskā procesa robežsadalījums būs divu Brauna tiltu summa (skat. definīciju 13). Reālās datu problēmās pirmās izlases blīvuma funkcija nav zināma, tāpēc tā ir jānovērtē, izmantojot (1.10).



8. att. Statistikas (1.18) histogramma.

8. attēlā redzamas statistikas (1.18) robežsadalījuma histogrammas, kas iegūtas butstrapojot izlases $X \sim N(0,1)$ un $Y \sim N(1,1)$. Redzam, ka šis empīriskais process nekonverģē aizvien tālāk, palielinot izlašu apjomu n.

2. PĀRBĪDES FUNKCIJA RANŽĒTĀM IZLASĒM

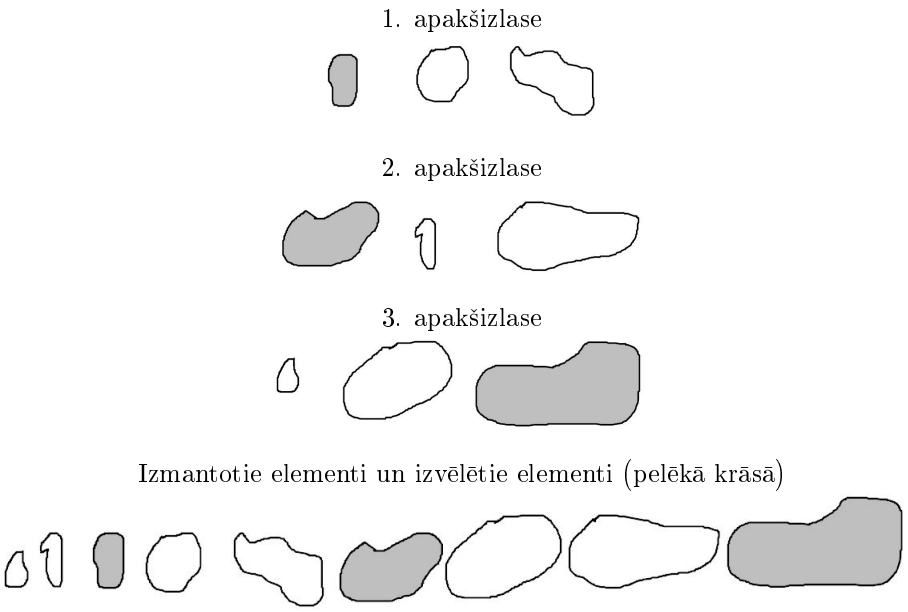
Statistikas jomā nopietnas problēmas rada gadījumi, kad datu apsekojumi ir dārgi un/vai laikietilpīgi. 1952. gadā *McIntyre* [7] piedāvāja izlases metodi, kas kļuva pazīstama kā ranžētu izlašu (RSS) veidošana. RSS veidošanas procedūra piedāvā efektīvu veidu, kā samazinot izmaksas, iegūt reprezentatīvu izlasi. Ilgus gadus šī metode netika plaši pielietota, interese par tās efektivitāti pieauga tikai pēdējo 20 gadu laikā. Kopš tā laika oriģinālā *McIntyre* ideja tikusi pilnveidota. *Ghosh* un *Tiwari* savā 2007. gada publikācijā [6] piedāvā izmantot ranžētas izlases hipotēzes par divu izlašu lokācijas modeli pārbaudei. Šajā nodaļā tiks apskatītas ranžētas izlases, to rašanās vēsture un veidošanas procedūra, kā arī – pielietojums divu izlašu problēmās. Tiks izmantota *Chen*, *Bai* un *Sinha* grāmata [13].

2.1. Ranžētu izlašu veidošana

Ranžētu izlašu veidošanas priekšnosacījums ir pieņēmums, ka pētāmās izlases apakšizlases elementi var tikt sakārtoti augošā secībā bez reāliem mērījumiem, izmantojot lētas metodes.

Apskatīsim *McIntyre* [7] ieviesto ranžētu izlašu veidošanas procedūru. Pirmkārt, tiek izvēlēta vienkārša k-elementu gadījuma izlase. Neveicot reālus mērījumu, šos k elementus sakārto augošā secībā un izvēlas mazāko elementu, lai veiktu mērījumus. Pārējie elementi netiek reāli izmērīti. Nākamā k-elementu vienkārša gadījuma izlase tiek izvēlēta. Tieši tāpat, neveicot reālus mērījumus, k elementus sakārto augošā secībā, tikai šoreiz izvēlas otru mazāko elementu, lai veiktu reālus mērījumus. Process tiek turpināts, līdz no pēdējās k-elementu apakšizlases tiek izvēlēts un reāli izmērīts k-tais elements. Šis process tiek saukts par ciklu. Lai gan vienā ciklā tiek izmantoti k^2 elementi, tikai k no tiem tiek reāli izmērīti. Atkārtojot ciklu m reizes, iegūstam balansētu ranžētu izlasi (BRSS) apjomā $N = mk$.

Mode, *Conquest* un *Marker* [14] apskatīja piemēru, kad ASV Meža dienests veica Klusā okeāna Ziemeļrietumu straumju platības mērīšanu. Šo platību var novērtēt vizuāli, vai arī – veikt precīzu mērīšanu. *Mode*, *Conquest* un *Marker* [14] uzskatāmi attēloja ranžētu izlašu veidošanas lietderību gadījumā, kad pētījuma budžets ļauj veikt precīzus mērījumus tikai trim apgabaliem. 9. attēlā redzama ranžētu izlašu veidošanas procedūra gadījumā, kad $k = 3$ un $m = 1$. Pelēkā krāsā iekrāsoti mērīšanai izvēlētie apgabali. Ranžētu izlašu veidošanas procedūra vispārīgā gadījumā ilustrēta 10. attēlā.



9. att. Ranžētu izlašu veidošanas procedūra, $k = 3,\ m = 1$.

Ranžētu izlašu veidošanas procedūrā kopumā tiek izmantoti $N \cdot k$ elementi, bet tikai N no tiem tiek reāli izmērīti. Procedūras rezultātā tiek iegūta balansēta ranžēta izlase (2.1). Ja katras kolonas elementi ir sakārtoti, tad procedūra tiek saukta par perfektu, pretējā gadījumā – par nepilnīgu. Pirmajā gadījumā lieto apzīmējumu $X_{(r)i}$, otrajā gadījumā attiecīgi – $X_{[r]i}$. Ar $X_{[r]i}$ ($X_{(r)i}$) apzīmē i-to izmērīto elementu ar rangu r.

Ranžētu izlašu veidošana ir piemērota gadījumos, kad pētāmo objektu sakārtošana augošā secībā ir lētāka, nekā katra objekta reāla izmērīšana.

Piemērs 1. Ranžētas izlases tika izmantotas ganību kapacitātes novērtēšanai fermās. Šajā gadījumā izlases elementi ir ganību kvadrāti. Katra kvadrāta mērīšana sevī ietver zāles nopļaušanu, žāvēšanu un barības nosvēršanu, kas ir laikietilpīgi un postoši. Taču ganību kvadrātus var novērtēt pieredzējis speciālists. Šādā situācijā ranžētu izlašu veidošana var būtiski palielināt efektivitāti.

Piemērs 2. Ranžētas izlases tika izmantotas krūmu fitomasas novērtēšanai Apalaču ozolu mežā. Šajā gadījumā izlase sevī ietvēra katra veģetatīvā tipa skaitu nejauši izvēlētos mežaudzes blokos. Neliela daudzuma bloku vizuāla sarindošana augošā kārtībā bija viegli izdarāma.

1. cikls

$$
\begin{array}{ccccccccc}
\boxed{X_{(1)11}} & \leq & X_{(2)11} & \leq & \dots & \leq & X_{(k)11} & \Rightarrow & X_{(1)1} \\
X_{(1)21} & \leq & \boxed{X_{(2)21}} & \leq & \dots & \leq & X_{(k)21} & \Rightarrow & X_{(2)1} \\
 & & & & \vdots & & & & \\
X_{(1)k1} & \leq & X_{(2)k1} & \leq & \dots & \leq & \boxed{X_{(k)k1}} & \Rightarrow & X_{(k)1}
\end{array}
$$

2. cikls

$$
\begin{array}{ccccccccc}
\boxed{X_{(1)12}} & \leq & X_{(2)12} & \leq & \dots & \leq & X_{(k)12} & \Rightarrow & X_{(1)2} \\
X_{(1)22} & \leq & \boxed{X_{(2)22}} & \leq & \dots & \leq & X_{(k)22} & \Rightarrow & X_{(2)2} \\
 & & & & \vdots & & & & \\
X_{(1)k2} & \leq & X_{(2)k2} & \leq & \dots & \leq & \boxed{X_{(k)k2}} & \Rightarrow & X_{(k)2} \\
 & & & & \dots & & & &
\end{array}
$$

m. cikls

$$
\begin{array}{ccccccccc}
\boxed{X_{(1)1m}} & \leq & X_{(2)1m} & \leq & \dots & \leq & X_{(k)1m} & \Rightarrow & X_{(1)m} \\
X_{(1)2m} & \leq & \boxed{X_{(2)2m}} & \leq & \dots & \leq & X_{(k)2m} & \Rightarrow & X_{(2)m} \\
 & & & & \vdots & & & & \\
X_{(1)km} & \leq & X_{(2)km} & \leq & \dots & \leq & \boxed{X_{(k)km}} & \Rightarrow & X_{(k)m}
\end{array}
$$

10. att. Ranžētu izlašu veidošanas procedūra

$$
\begin{array}{cccc}
X_{[1]1} & X_{[1]2} & \dots & X_{[1]m} \\
X_{[2]1} & X_{[2]2} & \dots & X_{[2]m} \\
\dots & \dots & \dots & \dots \\
X_{[k]1} & X_{[k]2} & \dots & X_{[k]m}
\end{array}
\tag{2.1}
$$

Piemērs 3. Ranžētas izlases var tikt izmantotas atsevišķos medicīnas pētījumos. Piemēram, nosakot noteiktu medicīnas mērījumu normas robežas, kas parasti sevī ietver dārgus laboratorijas testus. Ranžētas izlases tika izmantotas bilirubīna normas robežu noteikšanai jaundzimušo asinīs. Lai noteiktu šādas robežas, jāņem asins paraugi no izlasē iekļautajiem jaundzimušajiem un jātestē laboratorijā. No otras puses, nelielam skaitam bērnu bilirubīna līmeņa noteikšana var tikt veikta, apskatot zīdaiņus. Ja viņu seja, krūškurvis un pārējā ķermeņa daļa ir dzeltenīga, bilirubīna līmenis asinīs ir paaugstināts. Ranžētām izlasēm ir potenciāls pielietojums klīniskajos pētījumos. Parasti pacienta piedalīšanās izmaksas klīniskajos pētījumos ir ļoti augstas. Taču pacientus, kuri piedalīsies pētījumos, var izvēlēties, izmantojot ranžētu izlašu tehniku, pamatojoties uz dažādu informāciju, piemēram, vecumu, svaru, augumu, asinsspiediena līmeni, slimības vēsturi u.c., ko var iegūt ar salīdzinoši nenozīmīgām izmaksām.

2.2. Sakārtošanas mehānismi

Ranžētu izlašu veidošanas procedūrā esam ieguvuši BRSS (2.1). Jāatzīmē, ka elementi $X_{[r]i}$ ir savstarpēji neatkarīgi un tie $X_{[r]i}$, kas atrodas vienā rindā, ir vienādi sadalīti. Sākotnējās izlases blīvuma funkcijas un sadalījuma funkcijas apzīmēsim attiecīgi ar f un F.

Iegūstam

$$f_{(r)}(x) = \frac{k!}{(r-1)!(k-r)!} F^{r-1}(x)[1-F(x)]^{k-r} f(x).$$

Viegli pārliecināties, ka visiem x

$$f(x) = \frac{1}{k} \sum_{r=1}^{k} f_{(r)}(x). \tag{2.2}$$

Vienādība (2.2) norāda uz RSS veidošanas procedūras pamatotību.

Sakārtošanas mehānismu sauc par atbilstošu, ja katram x ir spēkā fundamentāls vienādojums:

$$F(x) = \frac{1}{k} \sum_{r=1}^{k} F_{[r]}(x).$$

Apskatīsim citus sakārtošanas mehānismus:

1. Nepilnīgs sakārtošanas mehānisms

   Gadījumā, kad pastāv sakārtošanas mehānisma kļūdas, $f_{(r)}$ nav statistikas ar rangu r blīvuma funkcija. Tomēr var izteikt attiecīgo sadalījuma funkciju $F_{[r]}$:

   $$F_{[r]}(x) = \sum_{s=1}^{k} p_{sr} F_{(s)}(x),$$

   kur p_{sr} apzīmē varbūtību, ar kādu s-tai (skaitliskai) statistikai tiek piešķirts rangs r. Ja šīs (kļūdu) varbūtības viena balansētas ranžētas izlases cikla robežās ir vienādas, tad $\sum_{s=1}^{k} p_{sr} = \sum_{r=1}^{k} p_{sr} = 1$. Tātad

   $$\frac{1}{k} \sum_{r=1}^{k} F_{[r]}(x) = \frac{1}{k} \sum_{r=1}^{k} \sum_{s=1}^{k} p_{sr} F_{(s)}(x) = \frac{1}{k} \sum_{s=1}^{k} \left(\sum_{r=1}^{k} p_{sr} \right) F_{(s)}(x) = F(x).$$

2. Daudzdimensiju izlases, kas iegūtas sakārtojot vienu no mainīgajiem.

   Tiks apskatīts divdimensiju gadījums. Pieņemsim, ka jāveic secinājumi, balstoties uz X un Y kopīgo sadalījumu. Arī šajā gadījumā var tikt pielietota ranžētu izlašu veidošanas shēma. Izlašu elementi tiek sakārtoti, balstoties uz vienu no mainīgajiem, piemēram, Y. Procedūras otrajā solī tiek izmērīti abu mainīgo izvēlētie elementi.

Ar $f(x,y)$ apzīmēsim X un Y kopīgo blīvuma funkciju, ar $f_{[r]}(x,y)$ – $X_{[r]}$ un $Y_{[r]}$ kopīgo blīvuma funkciju. Varam rakstīt

$$f_{[r]}(x,y) = f_{X|Y_{[r]}}(x \mid y) g_{[r]}(y)$$

un

$$f(x,y) = \frac{1}{k} \sum_{r=1}^{k} f_{[r]}(x,y).$$

2.3. Matemātiskās cerības novērtējums

Ar $h(x)$ apzīmēsim jebkuru funkciju no x. Ar μ_h apzīmēsim $h(X)$ matemātisko cerību, t. i., $\mu_h = Eh(X)$. Apskatīsim μ_h novērtējumu, izmantojot ranžētas izlases. Par $h(x)$ var izvēlēties dažādas funkcijas, piemēram:

1. $h(x) = x^l$ kur $l = 1, 2, ...$;
2. $h(x) = I\{x \leq c\}$, kur $I\{\cdot\}$ ir indikātorfunkcija;
3. $h(x) = \frac{1}{\lambda} K\left(\frac{t-x}{\lambda}\right)$, kur K ir dotā funkcija un λ – dotā konstante.

Pieņemsim, ka eksistē $h(X)$ dispersija. μ_h novērtējumu definē sekojoši:

$$\hat{\mu}_{h\cdot\mathrm{RSS}} = \frac{1}{mk} \sum_{r=1}^{k} \sum_{i=1}^{m} h(X_{[r]i}).$$

Teorēma 8. [13] *Pieņemsim, ka ranžētas izlases sakārtošanas mehānisms ir atbilstošs. Tad,*

1. *Matemātiskās cerības novērtējums ir nenovirzīts, t. i.,* $E\hat{\mu}_{h\cdot\mathrm{RSS}} = \mu_h$.
2. $Var(\hat{\mu}_{h\cdot\mathrm{RSS}}) \leq \frac{\sigma_h^2}{mk}$, *kur* σ_h^2 *apzīmē* $h(X)$ *dispersiju, un nevienādība ir stingra, ja sakārtošanas mehānisms ir pilnībā nejaušs.*
3. *Kad* $m \to \infty$,

$$\sqrt{mk}(\hat{\mu}_{h\cdot\mathrm{RSS}} - \mu_h) \to N(0, \sigma^2_{h\cdot\mathrm{RSS}}),$$

kur

$$\sigma^2_{h\cdot\mathrm{RSS}} = \frac{1}{k} \sum_{r=1}^{k} \sigma^2_{h[r]}$$

un $\sigma^2_{h[r]}$ *ir* $h(X_{[r]i})$ *dispersija.*

Teorēmas 8 pierādījums atrodams *Chen*, *Bai* un *Sinha* grāmatā [13].

Zināms, ka $\frac{\sigma_h^2}{mk}$ ir μ_h novērtējuma, kas balstīts uz vienkāršu gadījuma izlasi ar apjomu mk, dispersija. Teorēma 8 parāda, ka μ_h novērtējumam, kas balstīts uz ranžētu izlasi, vienmēr būs mazāka dispersija nekā novērtējumam, kas balstīts uz vienkāršu gadījuma izlasi.

Automātiski tiek pieņemts, ka ranžētu izlašu veidošanas procedūras izmaksas ir niecīgas. Salīdzināsim statistikas procedūras, kas balstīta uz ranžētu izlasi ar apjomu mk, efektivitāti ar statistikas procedūras, kas balstīta uz vienkāršu gadījuma izlasi ar apjomu mk, efektivitāti. Ar $\hat{\mu}_{h\cdot\mathrm{SRS}}$ apzīmēsim vienkāršas gadījuma izlases ar apjomu mk vidējo vērtību. Ranžētas izlases μ_h novērtējuma relatīvo efektivitāti attiecībā pret vienkāršu gadījuma izlasi definē šādi:

$$\mathrm{RE}(\hat{\mu}_{h\cdot\mathrm{RSS}}, \hat{\mu}_{h\cdot\mathrm{SRS}}) = \frac{\mathrm{Var}(\hat{\mu}_{h\cdot\mathrm{SRS}})}{\mathrm{Var}(\hat{\mu}_{h\cdot\mathrm{RSS}})}. \tag{2.3}$$

Teorēma 8 parāda, ka $\mathrm{RE}(\hat{\mu}_{h\cdot\mathrm{RSS}}, \hat{\mu}_{h\cdot\mathrm{SRS}}) \geq 1$.

Pētot relatīvo efektivitāti, iegūstam:

$$\begin{aligned}
\sigma^2_{h\cdot\mathrm{RSS}} &= \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r]} \\
&= \frac{1}{k}\sum_{r=1}^{k}(E\left[h(X_{[r]})\right]^2 - \left[Eh(X_{[r]})\right]^2) \\
&= \frac{1}{k}\sum_{r=1}^{k}E\left[h(X_{[r]})\right]^2 - \mu_h^2 + \mu_h^2 - \frac{1}{k}\sum_{r=1}^{k}\left[Eh(X_{[r]})\right]^2 \\
&= \sigma_h^2 - \frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2.
\end{aligned} \tag{2.4}$$

Līdz ar to relatīvo efektivitāti var izteikt sekojoši:

$$\mathrm{RE}(\hat{\mu}_{h\cdot\mathrm{RSS}}, \hat{\mu}_{h\cdot\mathrm{SRS}}) = \frac{\sigma_h^2}{\sigma^2_{h\cdot\mathrm{RSS}}} = \left[1 - \frac{\frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2}{\sigma_h^2}\right]^{-1}.$$

No pēdējās izteiksmes izriet – kamēr pastāv vismaz viens r tāds, ka $\mu_{h[r]} \neq \mu_h$, relatīvā efektivitāte ir lielāka par 1.

McIntyre [7] veica sekojošu pieņēmumu: ranžētas izlases relatīvā efektivitāte, kas balstīta uz vienkāršu gadījuma izlasi, populācijas matemātiskās cerības novērtējumam atrodas starp 1 un $(k+1)/2$, kur k ir izlases apjoms. Tomēr pamatā esošajam sadalījumam kļūstot asimetriskam, relatīvā efektivitāte samazinās. *Takashi* un *Wakimoto* [15] parādīja, ka gadījumā, kad sakārtošana ir perfekta, $\frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r]}$, kā funkcija no k, dilst, ja k pieaug, kas nozīmē, ka pieaugot izlases apjomam k, relatīvā efektivitāte palielinās.

2.4. Ranžētu izlašu pielietojums pārbīdes funkcijai

Lai gan sākotnēji *McIntyre* [7] ieviesa ranžēšanas procedūru ar nolūku samazināt izmaksas bezgalīgu populāciju izlašu apsekojumos, ar laiku šo metodi sāka pielietot arī gadījumos, kad statistikas programmām ir grūti apstrādāt apsekotās izlases to lielo apjomu dēļ.

Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgu un vienādi sadalītu gadījuma lielumu izlase ar sadalījuma funkciju F un $Y_1, \ldots, Y_m$ – neatkarīgu un vienādi sadalītu gadījuma lielumu izlase ar sadalījuma funkciju G.

Savā publikācijā [6] *Ghosh* un *Tiwari* apskatīja ne tikai horizontālo pārbīdes funkciju (1.11), bet arī vertikālo pārbīdes funkciju.

Definīcija 16. Par divu izlašu $X_1, \ldots, X_n$ un $Y_1, \ldots, Y_m$ vertikālo pārbīdes funkciju sauc funkciju

$$\Lambda(t) = G(F^{-1}(t)) - t,\ 0 \leq t \leq 1. \tag{2.5}$$

Šeit $F^{-1}(t)$ ir $X_1, \ldots, X_n$ kvantiļu funkcija (skat. (1.3)).

Atšķirībā no vispārīgās pārbīdes funkcijas, kas definē horizontālo attālumu starp izlašu sadalījuma funkcijām, vertikālā pārbīdes funkcija definē vertikālo attālumu starp izlašu sadalījuma funkcijām.

Pieņemsim, ka ranžēšanas procedūras rezultātā no sākotnējās izlases $X_1, \ldots, X_n$ tika iegūta balansēta ranžēta izlase

$$X_{k_1 \times m_1} = \begin{Bmatrix} X_{(1)1} & X_{(1)2} & \ldots & X_{(1)m_1} \\ X_{(2)1} & X_{(2)2} & \ldots & X_{(2)m_1} \\ \ldots & \ldots & \ldots & \ldots \\ X_{(k_1)1} & X_{(k_1)2} & \ldots & X_{(k_1)m_1} \end{Bmatrix}, \tag{2.6}$$

un no $Y_1, \ldots, Y_m$ tika iegūta balansēta ranžēta izlase

$$Y_{k_2 \times m_2} = \begin{Bmatrix} Y_{(1)1} & Y_{(1)2} & \ldots & Y_{(1)m_2} \\ Y_{(2)1} & Y_{(2)2} & \ldots & Y_{(2)m_2} \\ \ldots & \ldots & \ldots & \ldots \\ Y_{(k_2)1} & Y_{(k_2)2} & \ldots & Y_{(k_2)m_2} \end{Bmatrix}. \tag{2.7}$$

Ievieš apzīmējumus $M_1 = k_1 \cdot m_1$, $M_2 = k_2 \cdot m_2$, $M = M_1 + M_2$, $q_1 = \dfrac{m_1}{M_1}$, $q_2 = \dfrac{m_2}{M_2}$.

Chen, *Bai* un *Sinha* [13] definēja izlases sadalījuma funkciju

$$F(x) = \frac{1}{k_1} \sum_{i=1}^{k_1} F_{(i)}(x), \tag{2.8}$$

kur $F_{(i)}(x)$ ir uz F balstīta $X_{(i)}$ empīriskā sadalījuma funkcija. Šīs funkcijas empīriskā versija ir

$$\hat{F}_{(i)}(x) = \frac{1}{m_1} \sum_{j=1}^{m_1} I_{X_{(i)j} \leq x}.$$

Līdz ar to var definēt (2.8) empīrisko versiju

$$\hat{F}(x) = \frac{1}{k_1} \sum_{i=1}^{k_1} \hat{F}_{(i)}(x).$$

Tāpat definē otrai izlasei:

$$G(x) = \frac{1}{k_2} \sum_{i=1}^{k_2} G_{(i)}(x), \tag{2.9}$$

kur $G_{(i)}(x)$ ir uz G balstīta $Y_{(i)}$ empīriskā sadalījuma funkcija. Šīs funkcijas empīriskā versija ir

$$\hat{G}_{(i)}(x) = \frac{1}{m_2} \sum_{j=1}^{m_2} I_{Y_{(i)j} \leq x}.$$

Līdz ar to var definēt (2.9) empīrisko versiju

$$\hat{G}(x) = \frac{1}{k_2} \sum_{i=1}^{k_2} \hat{G}_{(i)}(x).$$

Teorēma 9. [6] *Neatkarīgām balansētām ranžētām izlasēm $X_{k_1 \times m_1}$ un $Y_{k_2 \times m_2}$ no sadalījumiem attiecīgi F un G, kad* $\min(m_1, m_2) \to \infty$,

$$\sqrt{k_1 m_1 + k_2 m_2}(\hat{\Delta} - \Delta) \xrightarrow{d} \frac{\mathbb{Z}_\Delta}{g(G^{-1}(F))},$$

kur $\mathbb{Z}_\Delta$ ir Gausa process ar šādu kovariāciju struktūru

$$K_\Delta(x, y) = \frac{1}{\lambda} \left(F(\min(x, y)) - \frac{1}{k_1} \sum_{i=1}^{k_1} F_{(i)}(x) F_{(i)}(y) \right) + \frac{1}{1 - \lambda} \left(G(\min(x, y)) - \frac{1}{k_2} \sum_{i=1}^{k_2} G_{(i)}(x) G_{(i)}(y) \right)$$

un $\dfrac{k_1 m_1}{k_1 m_1 + k_2 m_2} \to \lambda$.

Doksum (1974) [8] teorēma 3 ir *Ghosh* teorēmas 9 speciālgadījums, kad $k_1 = k_2 = 1$, $m_1 = n$ un $m_2 = m$.

Piezīme 10. *Tā kā*

$$\frac{1}{k} \sum_{i=1}^{k} F_{(i)}^2(x) \geq \left(\frac{1}{k} \sum_{i=1}^{k} F_{(i)}(x) \right)^2 = F^2(x), \ \forall x,$$

tad

$$K_{\Delta,\mathrm{BRSS}}(x, x) \leq K_{\Delta,\mathrm{SRS}}(x, x).$$

Tātad horizontālās pārbīdes funkcijas $\Delta(x)$ punktveida ticamības joslas, kas konstruētas, izmantojot balansētu ranžētu izlasi (BRSS), būs šaurākas nekā tās, kas konstruētas, izmantojot vienkāršu gadījuma izlasi (SRS).

Teorēma 11. [6] *Neatkarīgām balansētām ranžētām izlasēm* $X_{k_1\times m_1}$ *un* $Y_{k_2\times m_2}$ *no sadalījumiem attiecīgi* F *un* G, *kad* $\min(m_1, m_2) \to \infty$,

$$\sqrt{k_1 m_1 + k_2 m_2}(\hat{\Lambda} - \Lambda) \xrightarrow{d} \mathbb{Z}_\Lambda,$$

kur $\mathbb{Z}_\Lambda$ *ir Gausa process ar kovariāciju struktūru*

$$\begin{aligned} K_\Lambda(x,y) = \frac{g(F^{-1}(x))\cdot g(F^{-1}(y))}{f(F^{-1}(x))\cdot f(F^{-1}(y))}\cdot\frac{1}{\lambda}\cdot \\ \left(\min(x,y) - \frac{1}{k_1}\sum_{i=1}^{k_1} F_{(i)}(F^{-1}(x))\cdot F_{(i)}(F^{-1}(y))\right) \\ + \frac{1}{1-\lambda}\left(G(F^{-1}(\min(x,y))) - \frac{1}{k_2}\sum_{i=1}^{k_2} G_{(i)}(F^{-1}(x))\cdot G_{(i)}(F^{-1}(y))\right), \end{aligned}$$

kur $x, y \in (0,1)$.

Arī vertikālās pārbīdes funkcijas $\Lambda(t)$ punktveida ticamības joslas, kas konstruētas, izmantojot balansētu ranžētu izlasi (BRSS), būs šaurākas nekā tās, kas konstruētas, izmantojot vienkāršu gadījuma izlasi (SRS).

3. PRAKTISKAIS PIELIETOJUMS

Šajā nodaļā tiks apskatīts teorijas pielietojums praktisku datu problēmām, salīdzinātas dažādas hipotēzes par lokācijas modeli pārbaudes iespējas, kā arī tiks pielietotas ranžētas izlases reāliem datiem.

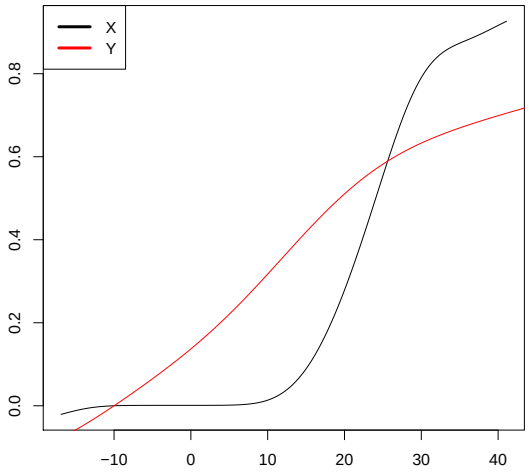
3.1. Ozona ietekme uz svara pieaugumu

Kā pirmie tiks izmantoti dati no [2] publikācijas. Tika veikts eksperiments, ar mērķi noteikt ozona nevēlamo iedarbību uz dzīvu organismu. Pirmā izlase X satur informāciju par 23 žurku svara pieaugumu pēc septiņu dienu uzturēšanās bezozona vidē (kontroles grupa).

```
41.1  38.4  24.4  25.9  21.9  18.3  13.1  27.3  28.5  -16.9  26.0
17.4  21.8  15.4  27.4  19.2  22.4  17.7  26.0  29.4  21.4  26.6  22.7
```

Otrā izlase Y satur datus par 22 žurku svara pieaugumu pēc septiņu dienu uzturēšanās ozona vidē (eksperimentālā grupa).

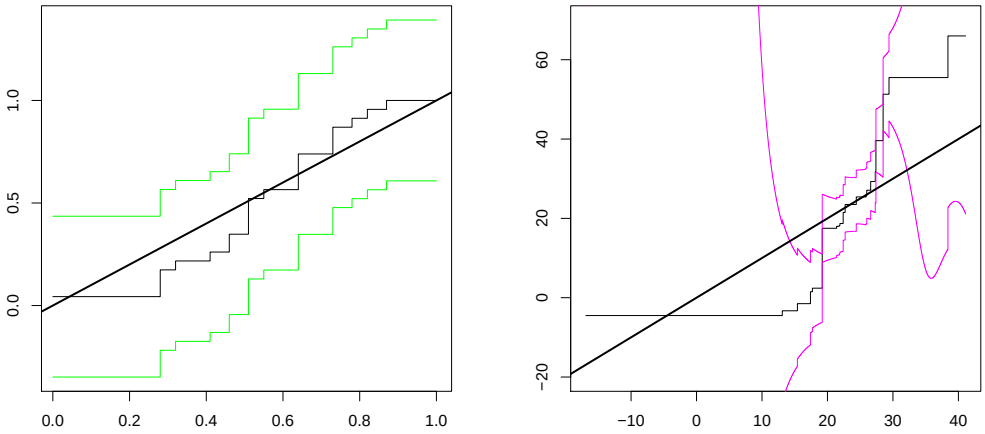
```
10.1  6.1  20.4  7.3  14.3  15.5  -9.9  6.8  28.2  17.9  -9.0,
-12.9  14.0  6.6  12.1  15.7  39.9  -15.9  54.6  -14.7  44.1  -9.0
```



11. att. X un Y sadalījuma funkcijas.

11. attēlā melnā krāsā attēlota kontroles grupas X gludinātā empīriskā sadalījuma funkcija un sarkanā krāsā attēlota eksperimentālās grupas Y gludinātā empīriskā sadalījuma funkcija. Šīs funkcijas iegūtas, pielietojot blīvuma funkcijas neparametrisko kodolu gludināšanas metodi, izmantojot Gausa kodolu.

Pārbaudīsim hipotēzi (1.8) par lokācijas modeli, izmantojot varbūtību-varbūtību un kvantiļu-kvantiļu grafikus ar vienlaicīgās ticamības joslām pie nozīmības līmeņa $\alpha = 0.05$. $\hat{\theta} = -11.4$ tika novērtēta kā abu izlašu empīrisko vidējo vērtību starpība. Rezultāti redzami 12. attēlā. Redzams, ka varbūtību-varbūtību grafika vienlaicīgās ticamības joslas hipotēzi par lokācijas modeli nenoraida, jo tajās ietilpst taisne $y = t,\ t \in [0, 1]$, bet kvantiļu-kvantiļu grafika vienlaicīgās ticamības joslas hipotēzi noraida, jo $y = x,\ x \in \mathbb{R}$ iziet ārpus tām. Q-Q grafika vienlaicīgās ticamības joslas ietekmē pirmās izlases blīvuma funkcija.



12. att. X un Y empīriskie P-P un Q-Q grafiki ar vienlaicīgās ticamības joslām.

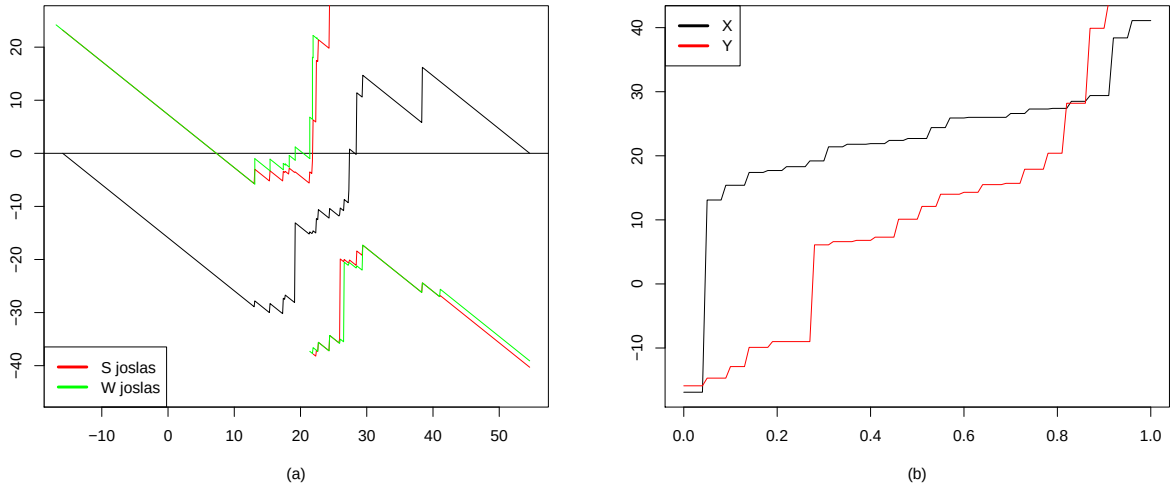
Veicot hipotēžu pārbaudi par lokācijas modeli (1.8) pie nozīmības līmeņa $\alpha = 0.05$, izmantojot Kolmogorova-Smirnova testu, p-vērtība iznāca 0.5957. Tā kā p-vērtība ir lielāka par nozīmības līmeni α, tad hipotēzi (1.8) nevar noraidīt.

Lai pārbaudītu hipotēzi par lokācijas modeli un lokācijas-skalēšanas modeli, tika konstruēts izlašu X un Y pārbīdes funkcijas novērtējums $\hat{\Delta}(x) = G_m^{-1}(F_n(x)) - x$ ar vienlaicīgās ticamības joslām pie nozīmības līmeņa $\alpha = 0.05$ pēc *Doksum* un *Sievers* [2]. Rezultāti redzami 13. attēlā (a). Sarkanā krāsā attēlotas $(S_*(x), S^*(x))$ joslas, un zaļā krāsā – $(W_*(x), W^*(x))$ joslas. Hipotēze par lokācijas modeli netiek noraidīta, jo varam atrast horizontālu taisni, kas ietilpst gan $(S_*(x), S^*(x))$ joslās, gan $(W_*(x), W^*(x))$ joslās. Pārbīdes funkcijas novērtējums $\hat{\Delta}(x)$ rāda, ka ozona ietekmē žurkas vidējais svara pieaugums tiek samazināts, ko apstiprina arī 14. attēlā redzamais kastu grafiks, tomēr liels svara pieaugums kontroles grupā eksperimenta grupā kļūst vēl lielāks, ko apstiprina 13. attēlā (b) redzamie abu izlašu empīriskie kvantiļu grafiki.

Ghosh un *Tiwari* savā publikācijā [6] piedāvā konstruēt vienlaicīgās ticamības joslas, butstrapojot kritisko vērtību. Tomēr šajā gadījumā rodas tādas pašas problēmas kā ar

kvantiļu starpības funkcijas robežsadalījumu – palielinot n, tas konverģē aizvien tālāk.

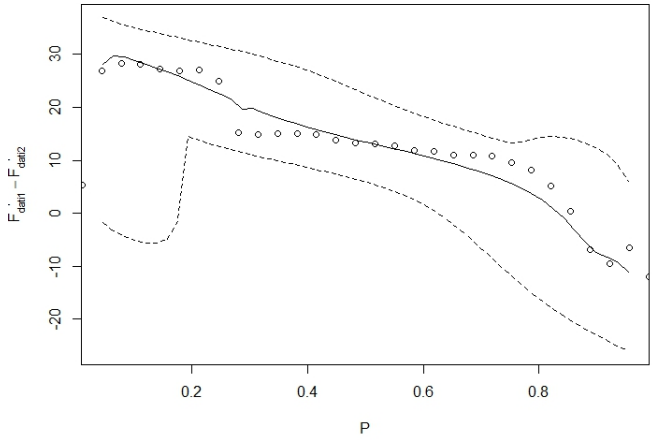
Šiem datiem nav nepieciešamības pielietot ranžētas izlases, jo datu apjoms nav liels un statistikas programmām nesagādā grūtības.



13. att.: Pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām un abu izlašu empīriskie kvantiļu grafiki.

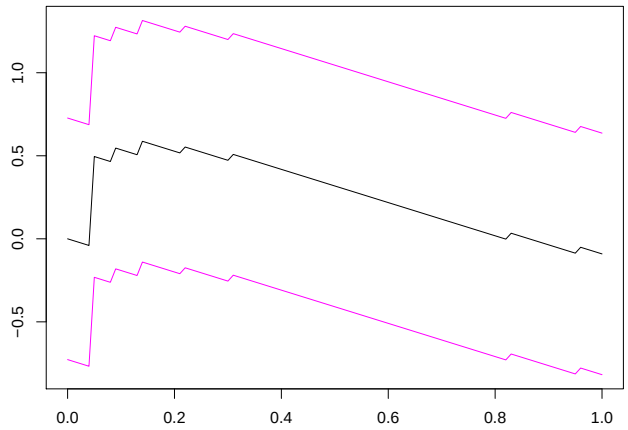
14. att. Kastu grafiki izlasēm X un Y.

Pārbaudīt hipotēzi par lokācijas modeli var arī izmantojot kvantiļu starpības funkciju (1.15). 15. attēlā redzams kvantiļu starpības funkcijas novērtējums ar vienlaicīgās ticamības joslām, kas konstruētas, izmantojot empīriskās ticamības (EL) metodi. Vienlaicīgās ticamības joslas noraida hipotēzi par lokācijas modeli, jo nav tādas horizontālas taisnes, kas ietilpst šajās joslās. Mēģinot konstruēt ticamības joslas, izmantojot statistiku (1.18), joslas iznāca pārāk platas, līdz ar to izmantot tās hipotēžu pārbaudei nebija jēgas.



15. att. Ticamības joslas kvantiļu starpības funkcijas novērtējumam.

Ghosh un *Tiwari* [6] apskatīja ne tikai horizontālo pārbīdes funkciju (1.11), bet arī vertikālo pārbīdes funkciju (2.5). Vertikālās pārbīdes funkcijas novērtējums X un Y izlasēm apskatāms 16. attēlā. Redzams, ka hipotēze par vertikālo pārbīdi netiek noraidīta, jo var atrast horizontālu taisni, kas ietilpst ticamības joslās.



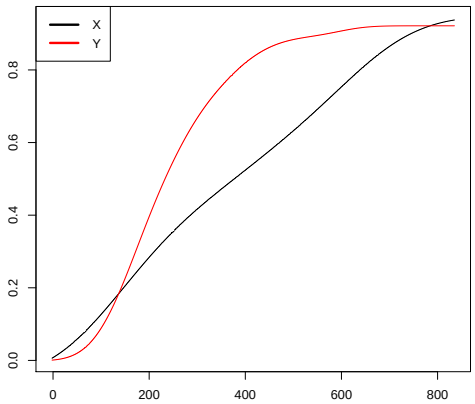
16. att. Vertikālās pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām.

Pārbaudot hipotēzi par lokācijas modeli izlasēm X un Y, kvantiļu-kvantiļu grafika un kvantiļu starpības funkcijas gadījumā hipotēze tiek noraidīta, tomēr kvantiļu starpības funkcija nenoraida hipotēzi par lokācijas-skalēšanas modeli. Varbūtību-varbūtību grafika un vispārīgās pārbīdes funkcijas gadījumā hipotēze par lokācijas modeli netiek noraidīta. Arī Kolmogorova-Smirnova tests nenoraida hipotēzi par lokācijas modeli. Tāpat netiek noraidīta vertikālā pārbīde starp izlasēm X un Y.

3.2. Tuberkulozes nūjiņu ietekme uz organismu

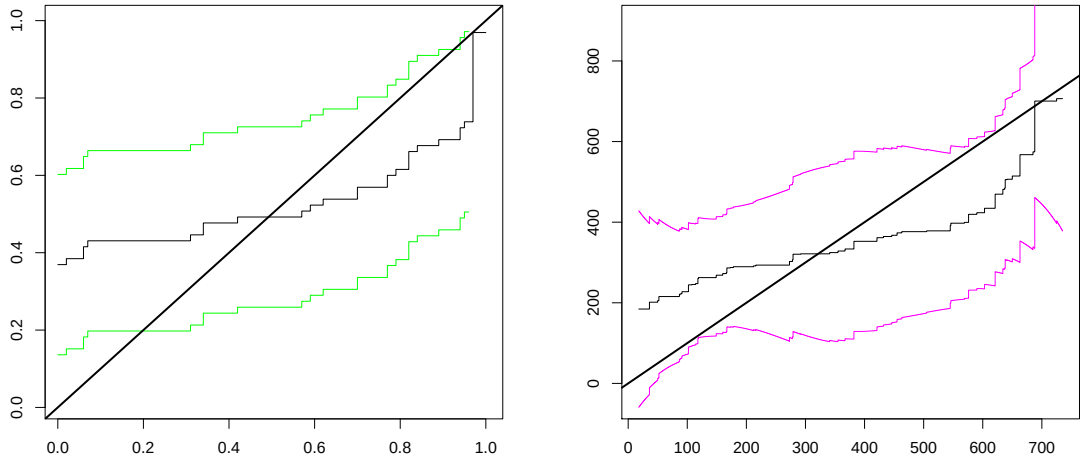
Nākamie tiks izmantoti dati no [8] publikācijas. Eksperimenta laikā 60 jūrascūciņas saņēma tuberkulozes nūjiņu devu. Izlase X satur datus par 65 jūrascūciņu dzīves ilgumu dienās (kontroles grupa), un izlase Y satur datus par 60 jūrascūciņu dzīves ilgumu dienās pēc tuberkulozes nūjiņu saņemšanas (eksperimenta grupa).

17. attēlā melnā krāsā attēlota kontroles grupas X gludinātā empīriskā sadalījuma funkcija un sarkanā krāsā attēlota eksperimentālās grupas Y gludinātā empīriskā sadalījuma funkcija. Tāpat kā iepriekš, šīs funkcijas iegūtas, pielietojot blīvuma funkcijas neparametrisko kodolu gludināšanas metodi, izmantojot Gausa kodolu.



17. att. X un Y sadalījuma funkcijas.

Lai pārbaudītu hipotēzi (1.8) par sadalījumu $F(x)$ un $G_\theta(x)$ vienādību var izmantot P-P un Q-Q grafikus un konstruēt tiem vienlaicīgās ticamības joslas, kā arī – veikt Kolmogorova-Smirnova testu. Šajā gadījumā $\hat{\theta} = -108$ tika novērtēta kā abu izlašu vidējo vērtību starpība. 18. attēlā redzami varbūtību-varbūtību un kvantiļu-kvantiļu grafiki ar vienlaicīgās ticamības joslām. Redzams, ka P-P grafiks hipotēzi (1.8) noraida, jo taisne $y = t,\ t \in [0,1]$ iziet ārpus ticamības joslām, bet Q-Q grafiks šo hipotēzi nenoraida, jo taisne $y = x,\ x \in \mathbb{R}$ ietilpst vienlaicīgās ticamības joslās.

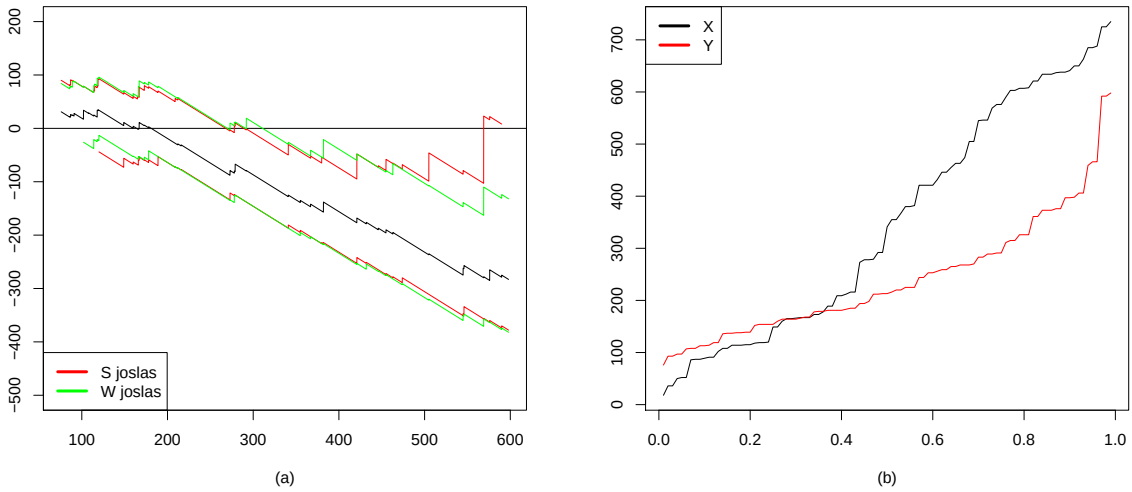


18. att. X un Y empīriskie P-P un Q-Q ar vienlaicīgās ticamības joslām.

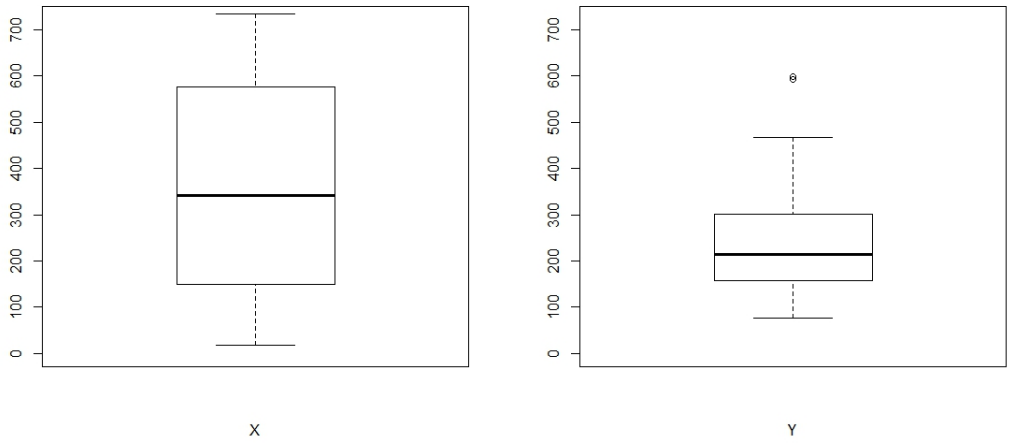
Pārbaudot hipotēzi par lokācijas modeli (1.8) pie nozīmības līmeņa $\alpha = 0.05$, izmantojot Kolmogorova-Smirnova testu, p-vērtība iznāca 0.0004. Tā kā p-vērtība ir mazāka par nozīmības līmeni α, tad hipotēze (1.8) tiek noraidīta.

Ar mērķi pārbaudīt hipotēzi par lokācijas modeli un lokācijas-skalēšanas modeli, tika konstruēts izlašu X un Y pārbīdes funkcijas novērtējums $\hat{\Delta}(x) = G_m^{-1}(F_n(x)) - x$ ar vienlaicīgās ticamības joslām pie nozīmības līmeņa $\alpha = 0.05$ pēc *Doksum* un *Sievers* [2]. Rezultāti redzami 19. attēlā (a). Sarkanā krāsā attēlotas $(S_*(x), S^*(x))$ joslas, un zaļā krāsā – $(W_*(x), W^*(x))$ joslas. Šajā gadījumā hipotēze par lokācijas modeli tiek noraidīta, nav tādas horizontālas taisnes, kas ietilpst vienlaicīgās ticamības joslās. Tomēr lokācijas-skalēšanas modelis netiek noraidīts, jo var atrast taisni, kas ietilpst ticamības joslās. Tā kā $W_*(x) < 0$, $\forall x$, tad pēc *Doksum* tuberkulozes nūjiņas negatīvi ietekmē jūrascūciņu dzīves ilgumu, ko apstiprina arī 20. attēlā redzamie izlašu kastu grafiki.

Šiem datiem nav nepieciešamības pielietot ranžētas izlases, jo datu apjoms nav liels un statistikas programmām nesagādā grūtības.

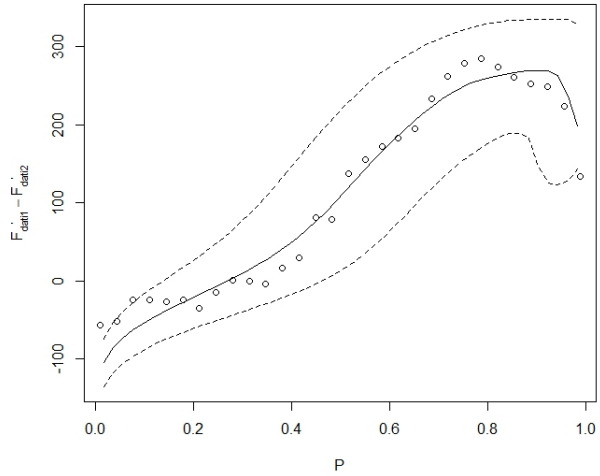


19. att.: Pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām un empīriskie kvantiļu grafiki.



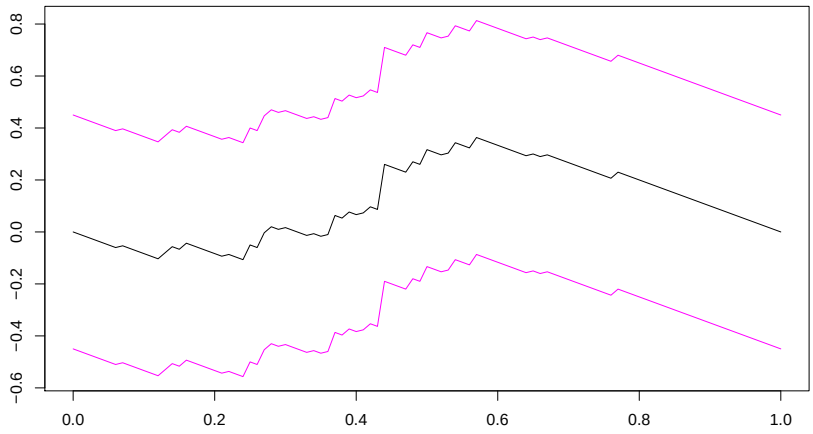
20. att. Kastu grafiki izlasēm X un Y.

Hipotēzes pārbaudei par lokācijas modeli var izmantot arī kvantiļu starpības funkciju (1.15). Tiks izmantotas ticamības joslas, pielietojot EL metodi. Mēģinot konstruēt ticamības joslas, izmantojot statistiku (1.18), līdzīgi kā ar datiem par ozona ietekmi uz organismu, joslas iznāca pārāk platas, līdz ar to izmantot tās hipotēžu pārbaudei nebija jēgas. Kā redzams 21. attēlā, hipotēze par lokācijas modeli tiek noraidīta.



21. att. Ticamības joslas kvantiļu starpības funkcijas novērtējumam.

Vertikālās pārbīdes funkcijas (2.5) novērtējums ar ticamības joslām redzams 22. attēlā. Kritiskā vērtība iegūta ar butstrapa metodi. Šajā gadījumā hipotēze par to, ka starp izlasēm pastāv vertikālā pārbīde, netiek noraidīta, jo var atrast horizontālu taisni, kas ietilpst ticamības joslās.

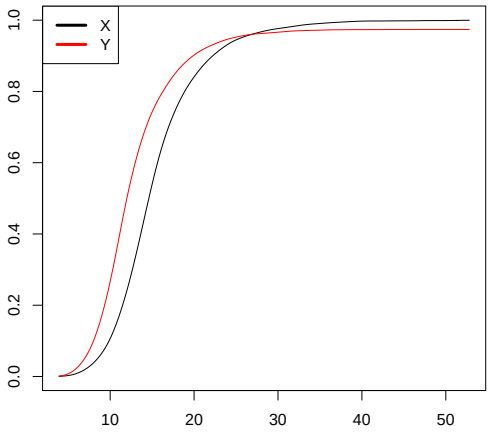


22. att. Vertikālās pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām.

Pārbaudot hipotēzi par lokācijas modeli izlasēm X un Y, kvantiļu-kvantiļu grafika vienlaicīgās ticamības joslas hipotēzi nenoraida, visos pārējos gadījumos hipotēze tika noraidīta. Vertikālās pārbīdes funkcijas ticamības joslas nenoraida hipotēzi, ka starp izlasēm X un Y pastāv vertikālā pārbīde.

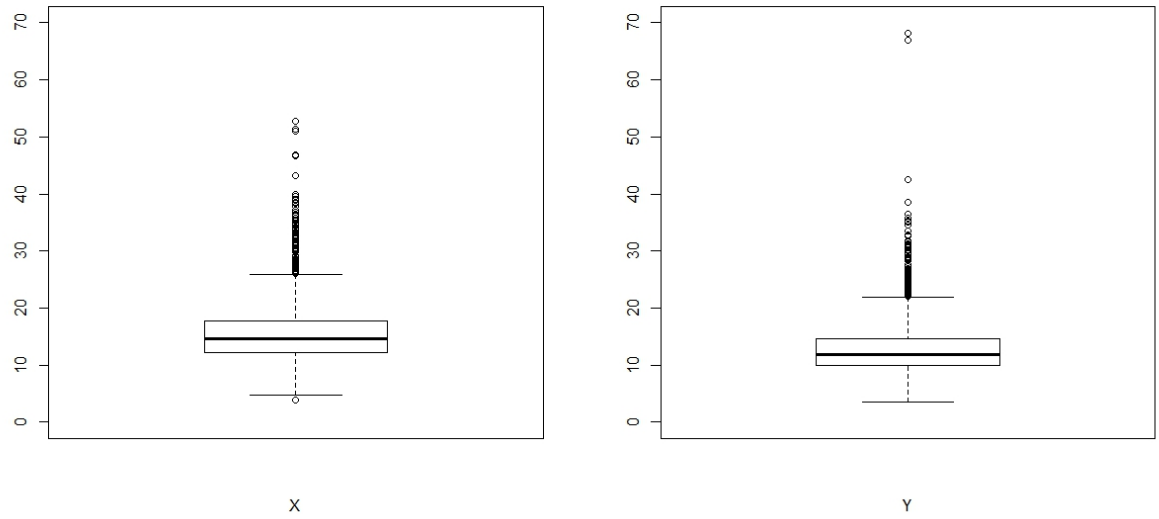
3.3. Prostatas specifiskā antigēna pārbaudes testa ieviešanas ietekme uz mirstību no prostatas vēža

Savā publikācijā [6] *Ghosh* un *Tiwari* salīdzināja mirstību ASV no prostatas vēža pirms un pēc prostatas specifiskā antigēna (PSA) pārbaudes testa ieviešanas. Šis tests tika ieviests 90. gadu vidū. Tā efektivitāti var pārbaudīt, salīdzinot mirstības no prostatas vēža koeficientus pirms un pēc pārbaudes testa ieviešanas. Tika iegūti dati no ASV Nacionālā Veselības Statistikas Centra (`http://www.cdc.gov/nchs`), kas satur informāciju pa apgabaliem par mirstību no prostatas vēža uz 100000 iedzīvotājiem. Datu iegūšanai nepieciešamā programma pieejama `http://seer.cancer.gov/seerstat/`. Apgabali ar neesošu koeficientu vai nulles koeficientu tika ignorēti. Rezultātā X satur datus par 2687 apgabaliem laika posmā no 1990. līdz 1992. gadam, un Y satur datus par 2619 apgabaliem laika posmā no 1999. līdz 2001. gadam. Izlašu empīriskās sadalījuma funkcijas attēlotas 23. attēlā. Aplūkojot sadalījuma funkcijas var izdarīt pieņēmumu, ka starp šīm izlasēm varētu pastāvēt lokācijas modelis.



23. att. X un Y sadalījuma funkcijas.

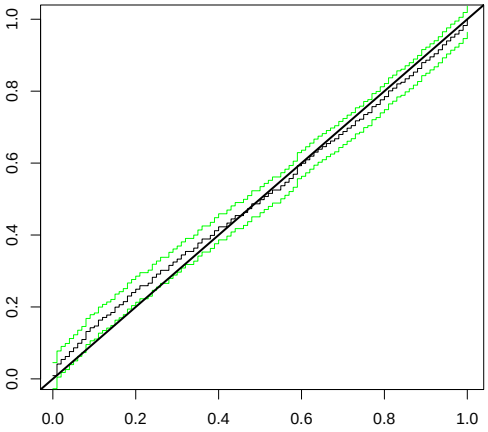
24. attēlā redzami X un Y kastu grafiki. Šajos grafikos ir redzams, ka gan X, gan Y satur daudz izlecējus.



24. att. X un Y kastu grafiki.

Lai pārbaudītu hipotēzi (1.8) par abu sadalījumu vienādību var izmantot P-P un Q-Q grafikus un konstruēt tiem vienlaicīgās ticamības joslas. Šajā gadījumā $\hat{\theta} = -2.7$ tika novērtēta kā abu izlašu vidējo vērtību starpība. 25. attēlā redzams empīriskais varbūtību-varbūtību grafiks ar vienlaicīgās ticamības joslām. Redzams, ka joslas ir pārāk šauras un hipotēžu pārbaudi veikt nav iespējams. Tam par iemeslu ir lielie izlašu apjomi. Tā kā empīriskais P-P grafiks ir ļoti tuvs diagonālei, tad var izteikt pieņēmumu, ka $F(x)$ un $G_\theta(x)$ sadalījuma likumi ir vienādi, tas nozīmē, ka starp izlašu X un Y sadalījuma funkcijām F un G pastāv lokācijas modelis.

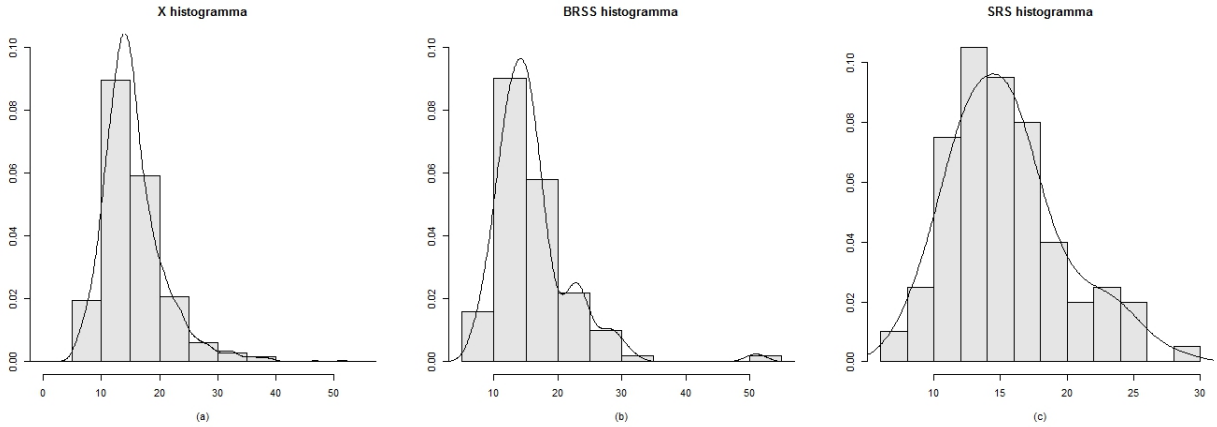
Izlašu X un Y Q-Q grafikam vienlaicīgās ticamības konstruēt neizdevās, jo tās ir atkarīgas no pirmās izlases blīvuma funkcijas, bet izlases lielā apjoma dēļ neizdevās iegūt kvantiļu funkcijas novērtējumu, izmantojot blīvuma funkcijas neparametrisko gludināšanas metodi.



25. att. X un Y empīriskais P-P grafiks ar vienlaicīgās ticamības joslām.

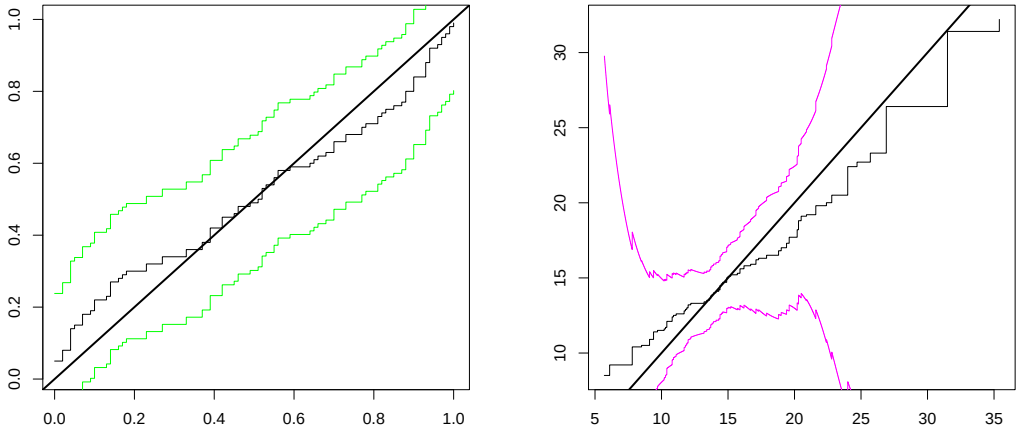
Veicot hipotēžu pārbaudi par lokācijas modeli (1.8) pie nozīmības līmeņa $\alpha = 0.05$, izmantojot Kolmogorova-Smirnova testu, p-vērtība iznāca 0.0007. Tā kā p-vērtība ir mazāka par nozīmības līmeni α, tad hipotēze (1.8) tiek noraidīta. Iespējams, pie vainas lielais izlašu apjoms.

Pielietosim ranžētās izlases, kā to iesaka darīt *Ghosh* un *Tiwari* [6]. Ranžētā izlase reprezentē sāknotnējo izlasi daudz labāk nekā vienkārša gadījuma izlase. To apstiprina 26. attēlā redzamās histogrammas ar blīvuma funkciju novērtējumiem. (a) gadījumā redzama sākotnējās izlases X histogramma, (b) gadījumā redzama no X iegūtās balansētās ranžētās izlases (BRSS) ar apjomu $k \cdot m = 100$ histogramma un (c) gadījumā – no X iegūtās vienkāršās gadījuma izlases (SRS) ar apjomu 100 histogramma.



26. att.: Sākotnējās izlases, ranžētas izlases un vienkāršas gadījuma izlases histogrammas ar blīvuma funkcijas novērtējumu.

Ranžēšanas procedūras rezultātā no X tika iegūta izlase $X_{10\times10}$ apjomā $k\cdot m = 100$ un no Y tika iegūta izlase $Y_{10\times10}$ apjomā $k\cdot m = 100$. Pārbaudīsim hipotēzi (1.8) par šo izlašu lokācijas modeli. Līdzīgi kā gadījumā ar sākotnējiem datiem no otrās izlases empīriskās sadalījuma funkcijas atņem $\hat{\theta} = -2.7$. Ranžēto izlašu empīriskie varbūtību-varbūtību un kvantiļu-kvantiļu grafiki ar vienlaicīgās ticamības joslām redzami 27. attēlā. Redzams, ka gan taisne $y = x,\ x \in \mathbb{R}$, Q-Q grafika gadījumā, gan taisne $y = t,\ t \in [0, 1]$, P-P grafika gadījumā, ietilpst ticamības joslās, tātad hipotēze par lokācijas modeli netiek noraidīta. Tā kā ranžētās izlases labi reprezentē sākotnējās izlases, tad var apgalvot, ka arī hipotēze par X un Y lokācijas modeli netiek noraidīta.



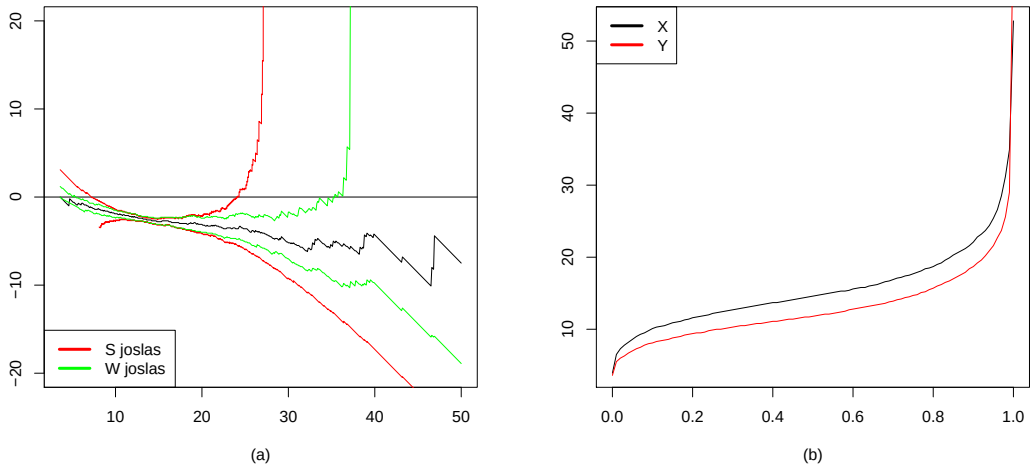
27. att.: Ranžēto izlašu $X_{10\times10}$ un $Y_{10\times10}$ empīriskie P-P un Q-Q grafiki ar vienlaicīgās ticamības joslām.

Veicot hipotēžu pārbaudi par lokācijas modeli (1.8) pie nozīmības līmeņa $\alpha = 0.05$ ranžētām izlasēm $X_{10\times10}$ un $Y_{10\times10}$, izmantojot Kolmogorova-Smirnova testu, p-vērtība iznāca 0.9671. Tā kā p-vērtība ir lielāka par nozīmības līmeni α, tad hipotēze (1.8) netiek noraidīta. Zināms, ka ranžētās izlases labi reprezentē sākotnējo izlasi, līdz ar to var apgalvot, ka hipotēze par X un Y lokācijas modeli netiek noraidīta.

Nākamais solis – pārbaudīt hipotēzi par lokācijas un lokācijas-skalēšanas modeli. 28. attēlā (a) redzams pārbīdes funkcijas novērtējums $\hat{\Delta}(x) = G_m^{-1}(F_n(x)) - x$ ar $(S_*(x), S^*(x))$ (sarkanā krāsā) un $(W_*(x), W^*(x))$ (zaļā krāsā) vienlaicīgās ticamības joslām pēc *Doksum* un *Sievers* [2]. Atkal sastopamies ar lielā datu apjoma problēmu – gan S, gan W joslas pārāk tuvu pietuvojas pašam pārbīdes funkcijas novērtējumam, līdz ar to hipotēžu pārbaudi veikt nav iespējams. Redzams, ka funkcija pie $x > 30$ sāk dīvaini uzvesties un ticamības joslas sāk tiekties uz bezgalību. Pie vainas varētu būt 24. attēlā

redzamie izlecēji. 28. attēlā (b) redzami izlašu X un Y empīriskie kvantiļu grafiki.

Arī šajā gadījumā pielietosim ranžētās izlases, kā to iesaka darīt *Ghosh* un *Tiwari* [6]. Ranžēšanas procedūras rezultātā no X tika iegūta izlase $X_{10\times30}$ apjomā $k \cdot m = 300$ un no Y tika iegūta izlase $Y_{10\times30}$ apjomā $k \cdot m = 300$. Ranžēto izlašu $X_{10\times30}$ un $Y_{10\times30}$ pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām redzams 29. attēlā. Tā kā var atrast horizontālu taisni, kas iekļaujas vienlaicīgās ticamības joslās, tad hipotēze par lokācijas modeli izlasēm $X_{10\times30}$ un $Y_{10\times30}$ netiek noraidīta. Tā kā ranžētās izlases labi reprezentē sākotnējās izlases, tad var apgalvot, ka arī hipotēze par lokācijas modeli starp izlasēm X un Y arī netiek noraidīta. Pie tam, tā kā pārbīdes funkcijas novērtējums $\forall x$ ir negatīvs, var izdarīt secinājumu, ka visiem x ir novērojams mirstības no prostatas vēža samazinājums.

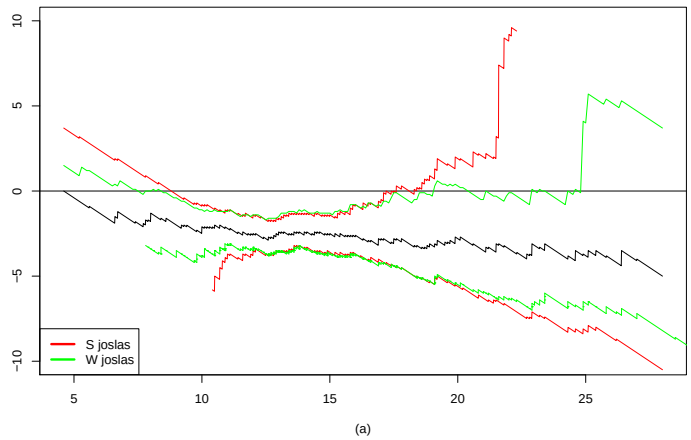


28. att.: Pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām un empīriskie kvantiļu grafiki.

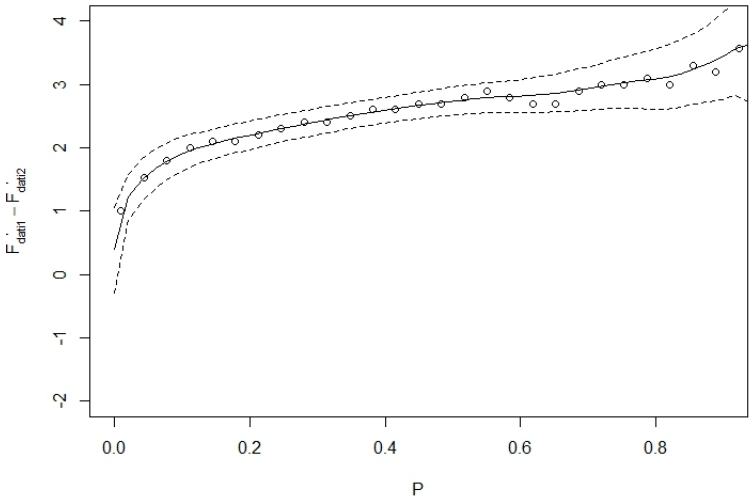
Pārbaudīsim hipotēzi par lokācijas modeli, izmantojot kvantiļu starpības funkciju. 30. attēlā redzams kvantiļu starpības funkcijas novērtējums izlasēm X un Y ar vienlaicīgās ticamības joslām, izmantojot empīriskās ticamības metodi. Redzams, ka vienlaicīgās ticamības joslas ir pārāk šauras hipotēžu pārbaudes veikšanai. Arī šoreiz tam par iemeslu kalpo lielais datu apjoms.

Izmantojot sakārtošanas procedūru, tika izveidotas balansētas ranžētas izlases $X_{10\times10}$ un $Y_{10\times10}$. Šo izlašu kvantiļu starpības funkcijas novērtējums ar vienlaicīgās ticamības joslām redzams 31. attēlā. Tā kā var atrast horizontālu taisni, kas ietilpst šajās joslās, tad hipotēze par lokācijas modeli izlasēm $X_{10\times10}$ un $Y_{10\times10}$ netiek noraidīta. Zināms, ka ranžētās izlases labi reprezentē sākotnējos datus, tātad netiek noraidīta hipotēze par

lokācijas modeli izlasēm X un Y.



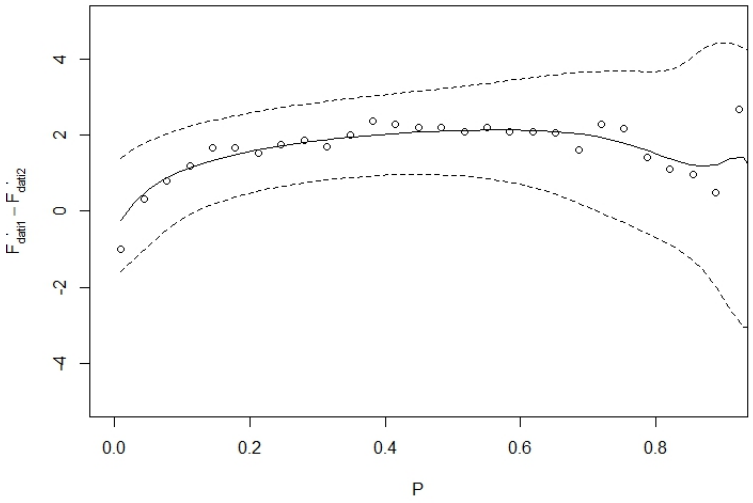
29. att. Ranžētu izlašu pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām.



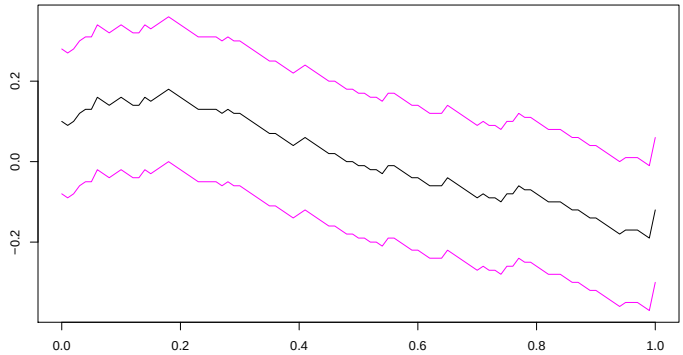
30. att.: Sākotnējo izlašu kvantiļu starpības funkcijas novērtējums ar vienlaicīgās ticamības joslām.

Konstruējot *Ghosh* un *Tiwari* piedāvātās vertikālās pārbīdes funkcijas novērtējumu ar vienlaicīgās ticamības joslām, redzams, ka hipotēze par vertikālo pārbīdi tiek noraidīta (skat. 32. attēlu).

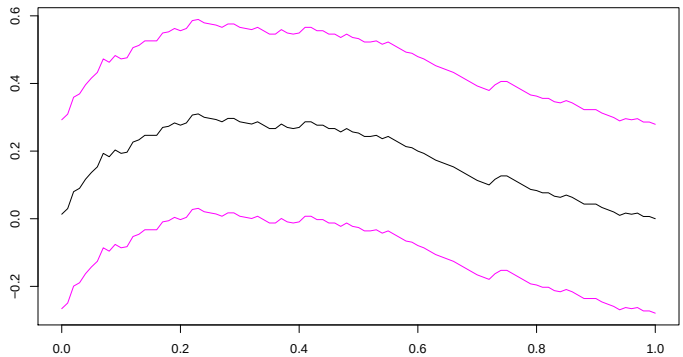
Ranžētu izlašu $X_{10\times10}$ un $Y_{10\times10}$ vertikālās pārbīdes funkcijas novērtējums redzams 33. attēlā. Acīmredzami šajā gadījumā hipotēze par vertikālo pārbīdi netiek noraidīta.



31. att.: Ranžētu izlašu kvantiļu starpības funkcijas novērtējums ar vienlaicīgās ticamības joslām.



32. att.: Sākotnējo datu vertikālās pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām.



33. att.: Ranžētu izlašu vertikālās pārbīdes funkcijas novērtējums ar vienlaicīgās ticamības joslām.

Šo datu analīze uzskatāmi parāda ranžētu izlašu metodes lietderību liela datu apjoma gadījumos. Pielietojot Kolmogorova-Smirnova testu, P-P grafiku, vispārīgo pārbīdes funkciju, kvantiļu starpības funkciju un to vienlaicīgās ticamības joslas lokācijas modeļa hipotēžu pārbaudē, hipotēze tika noraidīta, kaut gan aplūkojot 23. attēlā redzamās X un Y empīriskās sadalījuma funkcijas, kā arī 28. attēlā (b) redzamās empīriskās kvantiļu funkcijas, var izdarīt pieņēmumu, ka starp šīm izlasēm tomēr pastāv lokācijas modelis. Pielietojot ranžētās izlases, nevienā no augstāk minētajiem gadījumiem hipotēze par lokācijas modeli netika noraidīta.

Secinājumi

Darbā tika apskatīta vispārīgā pārbīdes funkcija divu izlašu gadījumā, ranžētas izlases un to pielietojums divu izlašu problēmās. Izmantojot reālu datu piemērus, tika salīdzinātas dažādas lokācijas modeļa grafiskās hipotēžu pārbaudes metodes – vienlaicīgo ticamības joslu konstruēšana vispārīgai pārbīdes funkcijai pēc *Doksum* un *Sievers* [2], kvantiļu-kvantiļu un varbūtību-varbūtību grafikiem pēc Cielēna [3], kvantiļu starpības funkcijai, izmantojot empīriskās ticamības (EL) metodi, pēc Valeiņa [5] un vertikālās pārbīdes funkcijai pēc *Ghosh* un *Tiwari* [6].

Izmantojot statistiku (1.17), radās problēmas ar robežsadalījumu – palielinot izlašu apjomu, tas konverģēja aizvien tālāk, tāpēc radās ideja pielietot statistiku (1.18), līdzīgi kā kvantiļu-kvantiļu procesa gadījumā. Mēģinot konstruēt vienlaicīgās ticamības joslas kvantiļu starpības funkcijai, izmantojot šo statistiku, tās iznāca pārāk platas, līdz ar to hipotēžu pārbaudi veikt neizdevās.

Pielietojot P-P un Q-Q grafiku vienlaicīgās ticamības joslas hipotēžu pārbaudē reālām datu problēmām, redzams, ka šai metodei ir savi trūkumi. Pirmkārt, sākumā ir jānovērtē lokācijas parametrs θ, otrkārt, tā neļauj pārbaudīt hipotēzi par lokācijas-skalēšanas modeli. Šajā darbā θ iespējamie novērtējumi netika apskatīti, hipotēžu pārbaudē izvēloties lokācijas parametru kā izlašu vidējo vērtību starpību.

Doksum un *Sievers* [2] piedāvātām pārbīdes funkcijas vienlaicīgās ticamības joslām S un W un empīriskās ticamības metodei kvantiļu starpības funkcijai [5] ir savas priekšrocības. Tās dod iespēju pārbaudīt hipotēzi par gan par lokācijas modeli, gan par lokācijas-skalēšanas modeli, kā arī iepriekš nav jānovērtē lokācijas parametrs. Verikālā pārbīdes funkcija ļauj pārbaudīt hipotēzi par divu izlašu sadalījumu funkciju vertikālo pārbīdi.

Ranžētu izlašu pielietojums reālām datu problēmām sevi pilnībā attaisnoja. Trešajā datu piemērā no empīriskās sadalījuma funkcijas un empīriskās kvantiļu funkcijas redzams, ka starp izlasēm pastāv lokācijas modelis, tomēr veicot hipotēžu pārbaudes, radās problēmas lielā datu apjoma dēļ. Vairākos gadījumu vienlaicīgās ticamības joslas iznāca pārāk šauras un hipotēžu pārbaudi veikt neizdevās, gadījumā ar varbūtību-varbūtību grafika vienlaicīgajām ticamības joslām radās grūtības novērtēt blīvuma funkciju. Tomēr iegūstot balansētās ranžētās izlases, augstāk minētās problēmas neradās un tika veiktas hipotēžu pārbaudes, kas nenoraidīja hipotēzi par lokācijas modeli.

Darbu turpinot, var analizēt ranžētas izlases un to pielietojumu statistikā un zinātnē, to kā arī turpināt pētīt kvantiļu starpības funkcijas vienlaicīgo ticamības joslu konstruē-

šanu, meklēt pieejas grafiskai hipotēžu pārbaudei un salīdzināt tās savā starpā.

Izmantotā literatūra un avoti

[1] 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*, 94:123–158, 2005.

[2] K. A. Doksum and G. L. Sievers. Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63:421–434, 1976.

[3] J. Cielens. Empīrisko procesu pielietojums strukturālo attiecību modeļos. *Diplomdarbs*, 2010.

[4] P. Laake, K. Laake, and R. Aaberge. On the problem of measuring the distance between distribution functions: Analysis of hospitalization versus mortality. *Biometrika*, 41:515–523, 1985.

[5] J. Valeinis. Confidence bands for structural relationship models. *PhD thesis, Goettingen*, 2007.

[6] K. Ghosh and Tiwari R. C. Empirical process approach to some two-sample problems based on ranked set samples. *Annals of the Institute of Statistical Mathematics*, 59:757–787, 2007.

[7] G. A. McIntyre. A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3:385–390, 1952.

[8] K. A. Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2:267–277, 1974.

[9] C. L. Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43:508–515, 1972.

[10] F. Hsieh and B. W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24:25–40, 1996.

[11] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[12] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Willey, New York, 1980.

[13] Z. Chen, Z. Bai, and B. K. Sinha. *Ranked Set Sampling Theory and Appilcations*. Springer, 2003.

[14] N. A. Mode, L. L Conquest, and Marker D. A. Ranked set sampling for ecological research: Accounting for the total costs of sampling. *Environmetrics*, 10:179–194, 1999.

[15] K. Takahasi and K. Wakimoto. On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20:1–31, 1968.

Pielikums

Izveidoto programmu kods

```
########Ranžētu izlašu veidošana############
a<-scan(file="Mortality rates 1990-1992.txt")
b<-scan(file="Mortality rates 1999-2001.txt")
k<-10
m<-10
z<-c()
y<-c()
BRSS1<-c()
BRSS2<-c()
for (i in 1:m){
for (j in 1:k){
x<-sort(sample(a,k,replace=F))
y[j]<-x[j]}
BRSS1<-c(BRSS1,y)}
for (i in 1:m){
for (j in 1:k){
u<-sort(sample(b,k,replace=F))
z[j]<-z[j]}
BRSS2<-c(BRSS2,u)}

###########Hipotēžu pārbaude, izmantojot P-P un Q-Q grafikus############
par(mfrow=c(1,2))
theta<-mean(dati2)-mean(dati1)
dati2<-dati2-theta
#PP grafiks ar vienlaicigas ticamibas joslam
c<-1.88
Fn1<-ecdf(dati1)
xx<-seq(0,1,by=0.01)
plot(xx,Fn1(quantile(dati2,probs=xx,type=1)),"s",xlab="",ylab="",main="",
ylim=c(0.01,1),xlim=c(0.01,1))
points(xx,Fn1(quantile(dati2,probs=xx,type=1))+c/sqrt(n),type="s",col="green")
points(xx,Fn1(quantile(dati2,probs=xx,type=1))-c/sqrt(n),type="s",col="green")
abline(0,1,lwd=2)
```

```
h<-bw.nrd(dati1)
dens<-function(x){1/(n*h)*sum(dnorm((x-dati1)/h))}
dens<-Vectorize(dens)
sad.fun<-function(x){integrate(dens,-10,x)$value}
sad.fun<-Vectorize(sad.fun)
quant.fun<-function(x){
uniroot(function(z) sad.fun(z)-x,c(-1000,1000))$root}
quant.fun<-Vectorize(quant.fun)
x<-seq(min(dati1),max(dati1),by=0.01)
konst<-dens(quant.fun(sad.fun(x)))
#QQ grafiks ar vienlaicigas ticamibas joslam
plot(x,quantile(dati2,ecdf(dati1)(x),type=1),type="s",xlab="",ylab="",main="")
points(x,quantile(dati2,ecdf(dati1)(x),type=1)+c/(sqrt(n)*konst),type="s",
col="magenta")
points(x,quantile(dati2,ecdf(dati1)(x),type=1)-c/(sqrt(n)*konst),type="s",
col="magenta")
abline(0,1,lwd=2)


##########Doksuma ticamības joslas#############
######S joslas##########
n<-length(dati1)
m<-length(dati2)
N<-m+n
M<-m*n/N
K<-1.36
x<-seq(min(dati2),max(dati2),by=0.01)
punkti.1<-function(x) ecdf(dati1)(x)-K/sqrt(M)
punkti.1<-Vectorize(punkti.1)
punkti1<-punkti.1(x)
punkti11<-punkti1[!punkti1<0]
x1<-x[punkti1>=0]
punkti.2<-function(x) ecdf(dati1)(x)+K/sqrt(M)
punkti.2<-Vectorize(punkti.2)
punkti2<-punkti.2(x)
x2<-x[punkti2<=1]
```

```
punkti22<-punkti2[!punkti2>1]
par(mfrow=c(1,2))
plot(x,quantile(dati2,ecdf(dati1)(x),type=1)-x,type="l",xlim=c(-4,6),
ylim=c(-10,10))
abline(0,0)
lines(x1,quantile(dati2,punkti11,type=1)-x1,col="red")
lines(x2,quantile(dati2,punkti22,type=1)-x2,col="red")


######W joslas#########
K.W<-3.02
c<-K.W^2/M
lambda<-m/N
x<-seq(min(dati2),max(dati2),by=0.01)
h.pluss<-function(u) (u+0.5*c*(1-lambda)*(1-2*lambda*u)+0.5*(c^2*
(1-lambda)^2+4*c*u*(1-u))^0.5)/(1+c*(1-lambda)^2)
h.pluss<-Vectorize(h.pluss)
h.minuss<-function(u) (u+0.5*c*(1-lambda)*(1-2*lambda*u)-0.5*(c^2*
(1-lambda)^2+4*c*u*(1-u))^0.5)/(1+c*(1-lambda)^2)
h.minuss<-Vectorize(h.minuss)
punktiW.1<-function(y) h.minuss(ecdf(dati1)(y))
punktiW.2<-function(z) h.pluss(ecdf(dati1)(z))
punkti1.W<-punktiW.1(x)
punktiW1<-punkti1.W[!punkti1.W<0]
x11<-x[punkti1.W>=0]
punkti2.W<-punktiW.2(x)
punktiW2<-punkti2.W[!punkti2.W>1]
x22<-x[punkti2.W<=1]
lines(x11,quantile(dati2,punktiW1,type=1)-x11,col="green")
lines(x22,quantile(dati2,punktiW2,type=1)-x22,col="green")

###########EL metode#################
library(EL)
EL.plot("qdiff",dati2,dati1,main="",conf.level=0.95,xlim=c(0,0.9),ylim=c(-5,5))
tt<-seq(0.01, 0.99, length=30)
ee<-quantile(dati1,tt)-quantile(dati2,tt)
```

```
points(tt,ee)

##########Vertikālā pārbīdes funkcija##################
B.sak1<-c()
apjoms1<-100
N<-n+m
alpha<-0.05
FF1<-ecdf(dati2)
for (kk in 1:1000){
dati1.b<-sort(sample(dati1,replace=T))
dati2.b<-sort(sample(dati2,replace=T))

FF<-ecdf(dati2)
FF.xx<-ecdf(dati2.b)
xx<-seq(0,1,by=0.001)
B.sak1[kk]<-max(sqrt(N)*abs(FF.xx(quantile(dati1.b,probs=xx,type=1))-
FF(quantile(dati1,probs=xx,type=1))))
}
B.sak1<-sort(B.sak1)
c<-B.sak1[1000*(1-alpha)]
xx<-seq(0,1,by=0.01)
gal1<-min(FF(quantile(dati1,probs=xx,type=1))-xx-c/sqrt(N))
gal2<-max(FF(quantile(dati1,probs=xx,type=1))-xx+c/sqrt(N))
plot(xx,FF(quantile(dati1,probs=xx,type=1))-xx,type="l",ylim=c(gal1,gal2),
xlab="",ylab="")
lines(xx,FF(quantile(dati1,probs=xx,type=1))-xx-c/sqrt(N),col="magenta")
lines(xx,FF(quantile(dati1,probs=xx,type=1))-xx+c/sqrt(N),col="magenta")
```

Diplomdarbs “Pārbīdes funkcija divu izlašu gadījumā” 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.

Autore: Lidija Januševa

(paraksts)

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

(paraksts) *(datums)*

Recenzents: lektors Mag. math Jānis Smotrovs

(paraksts)

Darbs iesniegts Matemātikas nodaļā

(datums)

Dekāna pilnvarotā persona: vecākā metodiķe Dzintra Holsta

(paraksts)

Darbs aizstāvēts Valsts pārbaudījuma komisijas sēdē

_____ prot. Nr. _____

(datums)

Komisijas sekretāre: lektore Baiba Āboltiņa

(paraksts)