LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKO PROCESU PIELIETOJUMS STRUKTURĀLO ATTIECĪBU MODEĻOS

DIPLOMDARBS

Autors: **Juris Cielēns**

Stud. apl. jc05001

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2010

Anotācija

Darbā aplūkoti strukturālo attiecību modeļi un to pārbaude, izmantojot empīrisko procesu metodes. Tiek pierādīti kvantiļu-kvantiļu, varbūtību-varbūtību, lokācijas-skalēšanas un Lēmaņa alternatīvu empīrisko procesu asimptotiskie sadalījumi. Ar simulāciju palīdzību programmā R analizēta ticamības intervālu pārklājuma precizitāte, kā arī noteiktas dažādu statistiku kritiskās vērtības.

Atslēgas vārdi: empīriskie procesi, strukturālo attiecību modeļi

Abstract

This thesis contains analysis of structural relationship models using empirical process methods. The asymptotic distribution of empirical quantile-quantile, probability-probability, location-scale and Lehman alternatives ir proved. Using simulations in progamm R, the coverage accuracy of confidence bands is analyzed and the critical values of different statistics are estimated.

Keywords: empirical processes, structural relationship models

Saturs

Apzīmējumi **2**

Ievads **3**

1. Empīriskie procesi vienas izlases gadījumā **4**

1.1. Empīriskā procesa definīcija 4

1.2. P-Donskera klašu piemēri 11

2. Empīriskie procesi divu izlašu gadījumā **13**

2.1. Strukturālo attiecību modeļi 13

2.2. P-P un Q-Q grafiki 14

2.3. Divu sadalījuma funkciju vienādības pārbaude 16

2.4. Lokācijas-skalēšanas modelis 20

2.5. Lēmaņa alternatīvu modelis 23

3. Praktiskā daļa **26**

3.1. Lokācijas-skalēšanas modeļa analīze 26

3.2. Lēmaņa alternatīvu modeļa analīze 27

3.3. Reālu datu analīze 31

Nobeigums **33**

Izmantotā literatūra un avoti **34**

A Pielikums **36**

A1. Blīvuma funkcijas gludināšanas kodolu metode 36

A2. Izmantoto programmu kods 37

Apzīmējumi

F_1, F_2	izlašu teorētiskās sadalījuma funkcijas
F_{1n}, F_{2m}	izlašu empīriskās sadalījuma funkcijas
$W(t)$	Brauna kustība
$B(t)$	Brauna tilts
$I_{\{a\}}$	Indikatorfunkcija - $I = 1$, ka ir spēkā a un 0, ja a nav spēkā
$N(\mu, \sigma^2)$	Normāli sadalīts gadījuma lielums ar vidējo vērtību μ un dispersiju σ^2
$U(a, b)$	Vienmērīgi sadalīts gadījuma lielums intervēlā $[a, b]$
$Exp(\lambda)$	Eksponenciāli sadalīts gadījuma lielums ar parametru λ
χ^2_k	χ^2 sadalīts gadījuma lielums ar k brīvības pakāpēm
$\rightarrow_d$	Konverģence pēc sadalījuma
$\xrightarrow{p}$	Konverģence pēc varbūtības
$\xrightarrow{as}$	Gandrīz droša konverģence
$l^\infty(T)$	Ierobežotas funkcijas telpā T
$C[a, b]$	Nepārtrauktas funkcijas intervālā $[a, b]$

Ievads

Statistikā svarīga loma ir hipotēžu pārbaudei. Šajā darbā tiek aplūkota hipotēžu pārbaude par populācijas sadalījumu veidu vienas izlases, kā arī divu izlašu gadījumā. Tiek padziļināti analizēts Kolmogorova-Smirnova tests vienkāršām un saliktām hipotēzēm no empīrisko procesu viedokļa. Empīrisko procesu teorja labi aprakstīta Billingsley [1], Shorack un Welner [2] un Van der Waart [3] grāmatās.

Divu izlašu gadījumā hipotēze $H_0 : \ F_1(x) = F_2(x), \ \forall x \in \mathbb{R}$ tiek pārbaudīta konstruējot vienlaicīgos ticamības intervālus varbūtību-varbūtību (P-P) un kvantiļu-kvantiļu (Q-Q) grafikiem. P-P un Q-Q procesu asimptotika ir zināma un tiek plaši izmantota, tomēr precīzu pierādījumu atrast neizdevās. Šo procesu asimptotika tika pierādīta izmantojot idejas no Horwath publikācijas [4]. Apkopojumu par teorētiski asimptotiskajiem sadalījumiem ar praktisku pielietojumu vienkāršas hipotēzes gadījumā skatīt publikācijā [5]. Teorētiskās metodes implementētas programmā R, veicot pārklājumu precizitātes simulācijas, kā arī apskatīts reāls datu piemērs.

Visbeidzot tiek analizēti vispārēji strukturāli modeļi divu izlašu gadījumā, kas pirmo reizi definēti publikācijās Freitag un Munk [6, 7]. Visvienkāršākais ir lokācijas modelis, kurš pieņem, ka divas sadalījuma funkcijas atšķiras viena no otras ar zināmu nobīdes parametru, tas ir, $F_1(x) = F_2(x - \theta)$ visiem $x \in \mathbb{R}$. Šādi modeļi ir svarīgi, piemēram, medicīnā, kur ieviešot jaunas zāles ir sagaidāms uzlabojums. Šajā gadījumā tiek pārbaudīta hipotēze par pašu lokācijas modeli, neinteresējoties par konkrēto sadalījuma funkciju veidu. Tādus modeļus pieņemts saukt par semiparametriskiem modeļiem.

Šajā darbā tiek pierādīti lokācijas-skalēšanas un Lēmaņa alternatīvu strukturālo modeļu empīrisko procesu asimptotiskie sadalījumi, kas ir jauns rezultāts. Ar to palīdzību var kontruēt ticamības intervālus un veikt hipotēžu pārbaudi strukturālajiem modeļiem. Ticamības intervāli strukturālajiem modeļiem, izmantojot empīrisko ticamības funkciju, iegūti Valeiņa [8] disertācijā.

Darbs sastāv no trīs nodaļām un pielikuma. Pirmajā nodaļā tiek sniegta empīriskā procesa definīcija klasiskajā un vispārinātā gadījumā, attēlota tā saistība ar Brauna tiltu un sniegti piemēri funkciju klasēm, kurām ir spēkā saistība ar Brauna tiltu. Otrā nodaļa satur strukturālo modeļu apskatu kā arī asimptotisko sadalījumu šo modeļu statistikām. tiek aplūkoti empīriskie procesi divu izlašu gadījumā. Trešā nodaļa satur modeļu konverģences un pārklājumu precizitātes analīzi. Pielikums satur izmantoto programmu kodu.

1. Empīriskie procesi vienas izlases gadījumā

Šajā nodaļā tiks sniegta empīriskā procesa definīcija vispārīgā gadījumā, aplūkotas Brauna tilta īpašības, kā arī svarīgākās empīrisko procesu teorēmas vienas izlases gadījumā. Tiks analizēta Kolmogorova-Smirnova statistika ne tikai vienkāršu, bet arī saliktu hipotēžu gadījumā, kā arī sniegti daži piemēri funkciju klasēm, kuras veido P-Donskera klases.

1.1. Empīriskā procesa definīcija

Vienkāršākajā gadījumā empīriskais process tiek veidots no empīriskās un teorētiskās sadalījuma funkcijas starpības. Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgu, vienādi sadalītu (turpmāk tiks lietots saīsinājums no angļu valodas - i.i.d.) gadījuma lielumu izlase ar sadalījuma funkciju F.

Definīcija 1.1. Par empīrisko sadalījuma funkciju sauc funkciju

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}.$$

Apgalvojums 1.1.

$$\sqrt{n}(F_n(x) - F(x)) \to_d N(0, F(x)(1 - F(x))).$$

Pierādījums. Rezultāts seko no centrālās robežteorēmas, no tā, ka

$$\mathbb{E}(F_n(x)) = F(x) \text{ un } \mathbb{D}(F_n(x)) = \frac{F(x)(1 - F(x))}{n}.$$

□

Teorēma 1.2. [2, 1. lpp] (Glivenko-Kantelli) Ir spēkā

$$\sup_{-\infty<x<\infty}(|F_n(x)-F(x)|)\xrightarrow[n\to\infty]{as}0.$$

Pēc daudzdimensionālās centrālās robežteorēmas [3, Piemērs 2.18] katriem $x_1,...,x_k\in\mathbb{R}$.

$$\sqrt{n}(F_n(x_1)-F(x_1),...,F_n(x_k)-F(x_k))\to_d(G_F(x_1),...,G_F(x_k)),$$

kur labās puses vektors ir ar daudzdimensiju normālo sadalījumu ar vidējo vērtību 0 un kovariāciju

$$\mathbb{E}G_F(x_i)G_F(x_j)=F(x_i\wedge x_j)-F(x_i)F_(x_j),$$

kur $x_i\wedge x_j=min(x_i,x_j)$.

Definīcija 1.2. [3, 257. lpp] Telpu $D[a,b]$, $a,\,b\in\mathbb{R}$ sauc par Skorohoda telpu, ja tā satur visas funkcijas $z:[a,b]\to\mathbb{R}$, kuras ir nepārtrauktas no labās puses un kurām intervālā $[a,b]$ eksistē robeža no kreisās puses. Šādas funkcijas tiek sauktas par 'cadlag' (continue à droite, limites à gauche) funkcijām.

Teorēma 1.3. [3, 266. lpp] Ja $X_1,X_2,...$ ir i.i.d. gadījuma lielumi ar sadalījuma funkciju F, tad empīriskais process $\sqrt{n}(F_n-F)$ konverģē pēc sadalījuma Skorohoda telpā $D[-\infty,\infty]$ uz gadījuma procesu G_F, kura robežsadalījumam ir daudzdimensionāls normālais sadalījums ar vidējo vērtību 0 un kovariāciju $\mathbb{E}G_F(t_i)G_F(t_j)=F(t_i\wedge t_j)-F(t_i)F(t_j)$.

Teorēma 1.4. [4, 1901. lpp] Pieņemsim, ka $X_1,...,X_n$ i.i.d. gadījuma lielumi ar sadalījuma funkciju F, tad

$$\sup_{-\infty<x<\infty}|\sqrt{n}(F_n(x)-F(x))|\xrightarrow{as}\sup_{-\infty<x<\infty}|B(F(x))|,\tag{1.1}$$

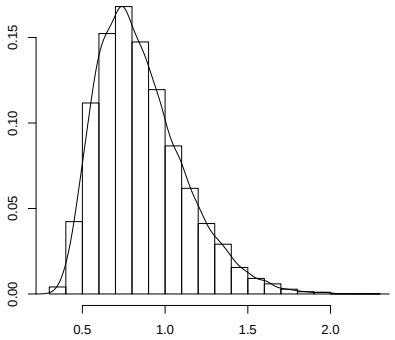
kur B apzīmē Brauna tiltu, kas tiks definēts vēlāk. Kreisās puses izteiksme ir plašāk pazīstama ar nosaukumu - Kolmogorova-Smirnova statistika.

Kolmogorova-Smirnova statistika bieži tiek izmantota, lai pārbaudītu hipotēzi par izlases empīriskās sadalījuma funkcijas atbilstību kādai konkrētai sadalījuma funkcijai. Šīs statistikas sadalījums ir bezgalīga summa [1, 104. lpp]

$$P(\sup_{-\infty<x<\infty}|\sqrt{n}(F_n(x)-F(x))|\le\alpha)=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2\alpha^2}.$$

Grafiski statistikas sadalījumu var, piemēram, iegūt veicot simulācijas un attēlojot tās histogrammā. Tika veiktas 10000 Kolmogorova-Smirnova statistikas simulācijas normāli

sadalītu gadījuma lielumu izlasēm un aprēķināta statistika. Rezultātus var aplūkot attēlā 1.1.



1.1. att.: Kolmogorova-Smirnova statistikas histogramma. 10000 reizes simulētas normāli sadalītas gadījuma izlases ar apjomu 1000. Histogrammai pievienota blīvuma funkcija, kas iegūta izmantojot kodolu gludināšanas metodi. 0.95-tā kvantile ir aptuveni 1.36.

Statistikas teorijā tiek aplūkotas ne tikai vienkāršas hipotēzes, bet arī saliktas. Piemēram, hipotēze par $N(0,1)$ sadalījumu ir vienkārša, bet par $N(\mu,\sigma^2)$ ir salikta. Šādas ir hipotēzes par modeļiem, kuri iekļauj sevī parametru novērtēšanu. Kolmogorova-Smirnova testu ar parametru novērtēšanu analizējis Lilliefors [9] un šādi testi mūsdienās tiek saukti par Lilliefora testiem. Parametru novērtēšanas gadījumā izmainās statistika t.i. iegūst

$$\sup_{-\infty<x<\infty} |\sqrt{n}(F_n(x) - F(x,\hat{\theta}))|, \tag{1.2}$$

kur $\hat{\theta}$ - novērtētie parametri. Piemēram $N(\mu,\sigma^2)$ sadalījumam $\hat{\theta} = \{\overline{x}, s^2\}$, kur $\overline{x}$ ir vidējās vērtības novērtējums un s^2 - dispersijas novērtējums. Izrādās, ka saliktu hipotēžu gadījumā asimptotiskais sadalījums 1.1 vairs nav spēkā. To viegli pamatot ar Teilora rindas palīdzību. Sekojošo izvedumu var skatīt [10, 452. lpp].

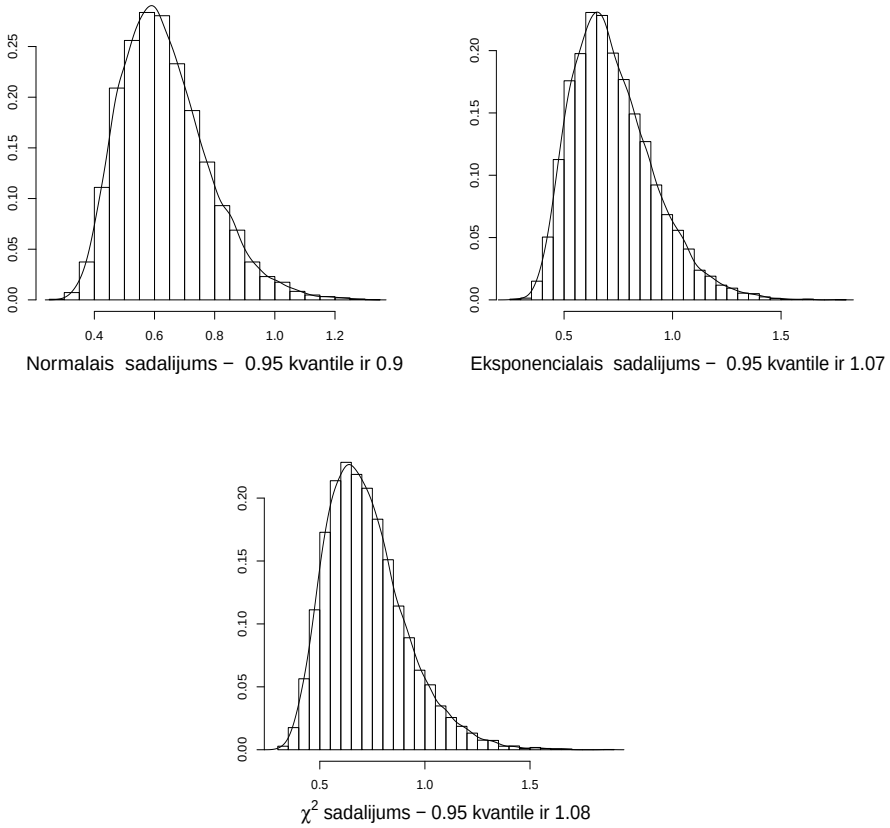
Fiksētam x ir spēkā

$$F_n(x) - F(x,\hat{\theta}) = (F_n(x) - F(x,\theta_0)) - (F(x,\hat{\theta}) - F(x,\theta_0)$$
$$= (F_n(x) - F(x,\theta_0)) - \left((\hat{\theta}-\theta_0)\tfrac{\partial}{\partial\theta}F(x,\theta)|_{\theta=\theta_0} + o_p(\hat{\theta}-\theta_0)\right).$$

No kurienes neko, ka

$\sqrt{n}(F_n(x) - F(x,\hat{\theta})) = \sqrt{n}(F_n(x) - F(x,\theta_0)) - \sqrt{n}(\hat{\theta}-\theta_0)\frac{\partial}{\partial\theta}F(x,\theta)|_{\theta=\theta_0} + o_p(\sqrt{n}(\hat{\theta}-\theta_0))$. Rezultātā var secināt, ka statistikas (1.2) asimptotiskais sadalījums ir atkarīgs no parametru novērtējumiem un pārbaudāmā populācijas sadalījuma veida.

Attēlā 1.2. var aplūkot statistikas (1.2) histogrammas dažādiem sadalījumiem ar 0.95 kvantilēm. Normāli sadalītām gadījuma izlasēm tā ir 0.9, eksponenciāli sadalītām 1.07, un χ^2 sadalītām 1.06. Svarīgākais secinājums ir tāds, ka asimptotiskais sadalījums ir atkarīgs no novērtētajiem parametriem, kuri savukārt ir atkarīgi no izlases.

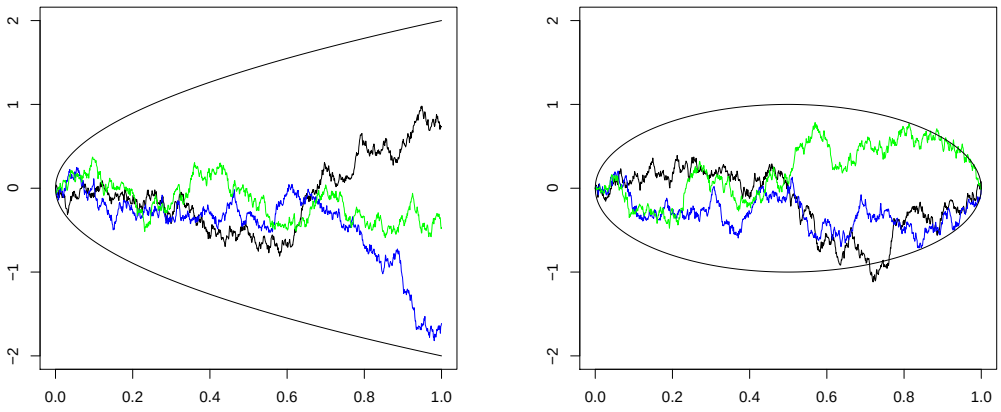


1.2. att.: Lilliefora statistikas hitogramma dažādām sadalījuma funkcijām. Redzams, ka Kritiskās vērtības konkrētam ticamības līmenim ir atkarīgas no dotās sadalījuma funkcijas. Simulācijas veiktas 10000 reizes izlasēm ar apjomu 1000.

Iepriekš Teorēmā 1.4 tika minēts, ka Kolmogorova-Smirnova statistikas sadalījums tiecas uz suprēmumu pa Brauna tiltu.

Definīcija 1.3. Par Brauna tiltu sauc Gausa procesu $\{B(t) : t \in [0,1]\}$, kura kovariāciju struktūra ir $\text{cov}(B(s), B(t)) = s(1-t)$, kur $s < t$. Šī procesa sadalījums ir tāds pats kā ar $W(t) - tW(1)$ sadalījums, kur W - standarta Brauna kustība.

Lai noteiktu iespējamo Brauna kustības un Brauna tilta attīstību, iespējams konstruēt punktveida ticamības joslas. Ir zināms, ka gadījuma lielumam, kura sadalījums ir $N(0, \sigma^2)$ 95% ticamības intervāls ir aptuveni $[-2\sigma, 2\sigma]$, piemērus skatīt attēlā 1.3.



1.3. att.: Pa kreisi - Brauna kustības realizācijas piemēri, pa labi - Brauna tilta realizācijas piemēri. Gludā līnija atdala 95% punktveida ticamības joslu.

Aplūkosim empīriskā procesa vispārēju definīciju. Pieņemsim, ka P ir varbūtību mērs mērojamā telpā $(\mathcal{X}, \mathcal{F})$, kur $\mathcal{X}$ ir kaut kāda kopa un $\mathcal{F}$ ir minimāla σ-algebra uz $\mathcal{X}$.

Definīcija 1.4. [3, 269. lpp] Par empīrisko varbūtību mēru sauc

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

kur δ_x ir varbūtību sadalījums, kurš deģenerēts punktā x.

Definīcija 1.5. [3, 269. lpp] Pieņemsim, ka $f : \mathcal{X} \to \mathbb{R}$ ir mērojama funkcija. Funkcijas f sagaidāmā vērtība pie empīriskā mēra tiek apzīmēta ar $\mathbb{P}_n f$, pie mēra P - ar Pf, un tiek definēta:

$$\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), \qquad Pf = \int f dP.$$

No lielo skaitļu likuma seko, ka $\mathbb{P}_n f$ konverģē gandrīz droši uz Pf, katram f, kuram Pf ir definēts.

Definīcija 1.6. [3, 269. lpp] Mērojamu funkciju $f : \mathcal{X} \to \mathbb{R}$ klase $\mathcal{A}$ tiek saukta par P-Glivenko-Kantelli klasi, ja

$$||\mathbb{P}_n f - Pf||_{\mathcal{A}} = \sup_{f \in \mathcal{A}} |\mathbb{P}_n f - Pf| \to 0 \ g.d.$$

Definīcija 1.7. [3, 269. lpp] Par empīrisko procesu funkcijai f sauc

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf).$$

No daudzdimensionālās centrālās robežteorēmas seko, ka katrai galīgai mērojamu funkciju f_i kopai, kur $Pf_i^2 < \infty$ ir spēkā

$$(\mathbb{G}_n f_1, \mathbb{G}_n f_2, ..., \mathbb{G}_n f_k) \to_p (\mathbb{G}_p f_1, \mathbb{G}_p f_2, ..., \mathbb{G}_p f_k),$$

kur labās puses vektoram ir daudzdimensionāls normālais sadalījums ar kovariācijām

$$\mathbb{E}\mathbb{G}_p f \mathbb{G}_p g = Pfg - PfPg. \tag{1.3}$$

Definīcija 1.8. [3, 269. lpp] Mērojamu funkciju klase $\mathcal{A}$ tiek saukta par P-Donskera klasi, ja procesu virkne $\mathbb{G}_n f : f \in \mathcal{A}$ konverģē uz robežprocesu $\mathbb{G}_p$ telpā $l^\infty(\mathcal{A})$, kur robežprocess $\mathbb{G}_p$ ir Gausa process ar vidējo vērtību 0 un kovariācijām 1.3 un tiek saukts par P-Brauna tiltu. Speciālgadījumā, ja $\mathcal{A}$ satur indikatorfunkcijas, tad Teorēma 1.2 ir speciālgadījums P-Glivenko-Kantelli klasei un Teorēma 1.4 ir speciālgadījums P-Donskera klasei.

Ļoti uzskatāmi empīriskā proccesa konverģenci uz Brauna tiltu var parādīt grafiski. Pētījumā tika ģenerētas trīs izlases ar sadalījuma likumu $U[0,1]$, ar apjomu $n = 20$, $n = 100$ un $n = 1000$ un uzzīmēti empīriskā varbūtību procesa attēli. Attēlā 1.4. redzami minētie posmi, kas grafiski parāda procesa konverģenci.

Definīcija 1.9. Par empīrisko kvantiļu funkciju sauc funkciju

$$F_n^{-1}(t) := \inf\{x : F_n(x) \geq t\},\ 0 < t < 1.$$

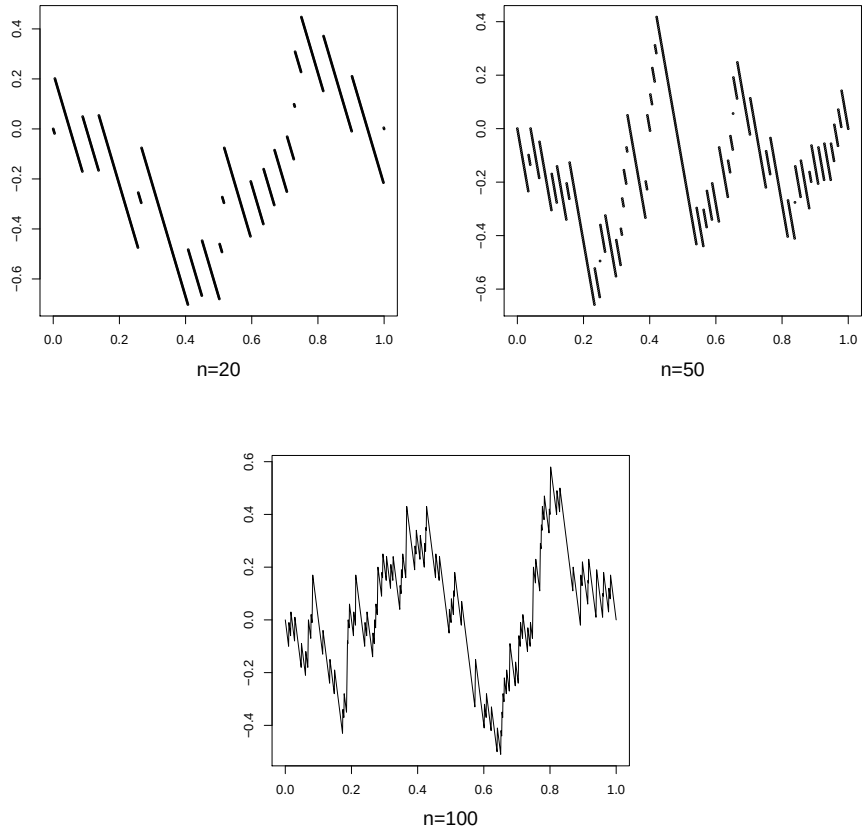
Teorēma 1.5. *Līdzīgi kā sadalījuma funkcijām, kvantiļu funkcijām ir spēkā*

$$\sup_{0<t<1} (|F_n^{-1}(t) - F^{-1}(t)|) \xrightarrow[n\to\infty]{as} 0.$$

Teorēma 1.6. [11, 31. lpp] Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar kvantiļu funkciju $F^{-1}(x)$ un empīrisko kvantiļu funkciju $F_n^{-1}(x)$, tad

$$\forall\ \epsilon > 0 \lim_{n\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_n^{-1}(t) - F^{-1}(t)) - \frac{1}{f(F^{-1}(t))} B(t)| \geq \epsilon) = 0\ g.d.$$

šādu pierakstu kā Teorēmā 1.6, lieto Horwath [4] un tas tiks izmantots turpmāk pierādījumos. Tā kā vienai izlasei var konstruēt gan empīrisko, gan kvantiļu procesu, nepieciešams sasprast, vai šie procesi ir kaut kā saistīti, t.i., vai eksistē kāda sakarība starp tiem. Attēlojot abus procesus vienā koordinātu plaknē, redzam, ka tie ir saistīti. Attēlā 1.5. attēloti



1.4. att.: Empīriskā procesa $\sqrt{n}(F_n(x) - F(x))$ konverģence un Brauna tiltu. Izlases ar apjomu n.

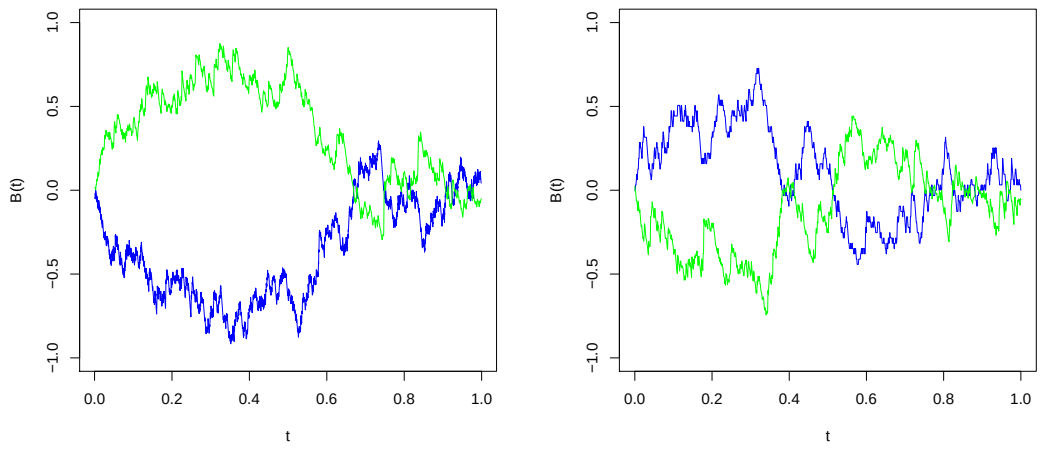
vabrūtību un kvantiļu procesi vienai un tai pašai izlasei. Kā redzams attēlā, procesi ir ar pretējām zīmēm, t.i., abu procesu summa ir 0.

Tas, vai funkciju klase ir Glivenko-Kantelli, vai Donskera ir atkarīgs no klases izmēra. Galīga integrējamu funkciju klase vienmēr ir Glivenko-Kantelli un galīga kvadrātiski integrējamu funkciju klase - Donskera [3, 270. lpp]. Lai izmērītu telpas $\mathcal{A}$ izmēru, tiek izmanoti entropijas termini. Tiek aplūkota tā sauktā iekavu entropija (angliski - bracketing entropy) attiecībā pret $L_r(P)$ normu $||f||_{P,r}$, kur $L_r(P)$ ir mērojamu funkciju klase ar īpašību $f \in L_r(P)$, ja f^r ir integrējama pēc P. Pieņemsim, ka ir dotas divas funkcijas l un u. Ar iekavām $[l, u]$ apzīmē tādu funkciju f kopu, kurām $l \leq f \leq u$ [3, 270. lpp].

Definīcija 1.10. Par $\epsilon - iekavu$ klasē $L_r(P)$ sauc iekavu $[l, u]$ ar īpašību

$$P(u - l)^r < \epsilon.$$

Definīcija 1.11. Par iekavu skaitli $N_{[\,]}(\epsilon, \mathcal{A}, L_r(P))$ sauc mazāko $\epsilon - iekavu$ skaitu, ar kurām var noklāt $\mathcal{A}$.



1.5. att.: Varbūtību un kvantiļu procesi vienai un tai pašai izlasei. Pa kreisi - normāli sadalītu gadījuma lielumu N(0,1) izlase, pa labi - U(0,1) gadījuma lielumu izlase.

Teorēma 1.7. [3, 270. lpp] Mērojamu funkciju klase $\mathcal{A}$ ir P-Glivenko-Kantelli, ja $\forall\, \epsilon > 0$

$$N_{[\,]}(\epsilon, \mathcal{A}, L_1(P)) < \infty.$$

Lielākajai daļai klašu, iekavu numuri $N_{[\,]}(\epsilon, \mathcal{A}, L_1(P))$ tiecas uz bezgalību, kad $\epsilon \downarrow 0$. Pietiekams nosacījums, lai klase būtu Donskera ir, ka tie netiecas uz bezgalību pārāk strauji. Šis ātrums tiek mērīts ar iekavu integrāli

$$J_{[\,]}(\delta, \mathcal{A}, L_2(P)) = \int_0^{\delta} \sqrt{\ln N_{[\,]}(\epsilon, \mathcal{A}, L_1(P))} d\epsilon.$$

Teorēma 1.8. *Mērojamu funkciju klase $\mathcal{A}$ ir P-Donskera, ja*

$$J_{[\,]}(1, \mathcal{A}, L_2(P)) < \infty.$$

Visas minētās definīcijas un teorēmas, spriedumi un teorēmu pierādījumi atrodami avotā $[3, 270.lpp]$.

1.2. P-Donskera klašu piemēri

Turpmāk minētie piemēri atrodami literatūras avotā citea13.

Piemērs 1. (Sadalījuma funkcija). Pieņemsim, ka funkciju klase $\mathcal{A}$ satur visas indikatoru funkcijas formā $f_t = I_{(-\infty,t]}$, kur $t \in \mathbb{R}$, tad empīriskais process $\mathbb{G}f_t$ ir klasiskais empīriskais process $\sqrt{n}(F_n(t) - F(t))$. Tiek aplūkotas iekavas formā $[I_{(-\infty,t_{i-1})}, I_{(-\infty,t_i)}]$ punktu režģim

$-\infty = t_0 < t_1 < ... < t_k = \infty$ ar īpašību $F(t_i-) - F(t_{i-1}) < \epsilon$ katram i. Šo iekavu $L_1(F)$ izmērs ir ϵ un to kopējais skaits k var tikt izvēlēts mazāks par $\epsilon/2$. Tā kā $Ff^2 < Ff$ katram $0 < f < 1$, iekavu $L_2(F)$ izmērs ir ierobežots ar $\sqrt{\epsilon}$, kas nozīmē, ka $N_{[\ ]}(\sqrt{\epsilon}, \mathcal{A}, L_2(F)) \leq 2/\epsilon$. Līdz ar to šī funkciju klase apmierina P-Donskera klases nosacījumus.

Piemērs 2. (Parametriska klase). Pieņemsim, ka $\mathcal{A} = \{f_\theta : \theta \in \Theta\}$ ir mērojamu funkciju klase, $\Theta \in \mathbb{R}^d$ ir indeksu kopa. Pieņemsim ,ka eksistē mērojama funkcija m tāda, ka

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq m||\theta_1 - \theta_2|| \text{ katram} \theta_1, \theta_2.$$

Ja $P|m|^r < \infty$, tad var pierādīt [3, 271. lpp], ka eksistē konstante K, atkarīga tikai no Θ un d tāda, ka iekavu numuri apmerina nevienādību

$$N_{[\ ]}(\epsilon||m||_{P,r}, \mathcal{A}, L_r(F)) \leq K\left(\frac{diam\Theta}{\epsilon}\right)^d, \text{ katram } 0 < \epsilon < diam\Theta.$$

Līdz ar to entropija ir ar kārtu mazāku par $(1/\epsilon)$ un iekavu entropijas integrālis konverģē.

Piemērs 3. (Gludas funkcijas). Pieņemsim, ka $\mathbb{R}^d = \cup_j I_j$ ir kubu ar izmēru 1 daļa un $\mathcal{A}$ ir visu funkciju $f : \mathbb{R}^d \to \mathbb{R}$ klase, kurām eksistē parciālie atvasinājumi līdz kārtai α un kuras ir vienmērīgi ierobežotas ar konstantēm M_j katrā no kubiem I_j. Tad katram var pierādīt, ka $V \geq d/\alpha$ un katram varbūtību mēram P klases $\mathcal{A}$ iekavu numuri apmierina nevienādību

$$\log N_{[\ ]}(\epsilon, \mathcal{A}, L_r(P)) \leq K\left(\frac{1}{\epsilon}\right)^V \left(\sum_{j=1}^{\infty}(M_j^r P(I_j))^{\frac{V}{V+r}}\right)^{\frac{V+r}{r}}.$$

Ja nevienādības labā puse konverģē tad klase $\mathcal{A}$ ir P-Donskera.

Piemērs 4. (Soboļeva klases). Pieņemsim, ka funkciju klase $\mathcal{A}$ satur visas funkcijas $f : [0,1] \to \mathbb{R}$ tādas, ka $||f||_\infty \leq 1$ un kuru (k-1)-ais atvasinājums ir absolūti nepārtraukts ar $\int (f^{(k)})^2(x)dx \leq 1$ kādam fiksētam k. Tad eksistē konstante K tāda, ka katram $\epsilon > 0$

$$\log N_{[\ ]}(\epsilon, \mathcal{A}, ||.||_\infty) \leq K\left(\frac{1}{\epsilon}\right)^{1/k}.$$

Šī klase ir P-Donskera katram $k \geq 1$ un katram P.

Piemērs 5. (Ierobežota variācija). Pieņemsim, ka funkciju klase $\mathcal{A}$ satur visas monotonas funkcijas $f : \mathbb{R} \to [-1,1]$ Tad eksistē konstante K tāda, ka katram varbūtību mēram P

$$\log N_{[\ ]}(\epsilon, \mathcal{A}, L_2(P)) \leq K\left(\frac{1}{\epsilon}\right).$$

Šī klase ir P-Donskera katram P.

2. Empīriskie procesi divu izlašu gadījumā

Iepriekšējā nodaļā tika aplūkoti dažu statistiku asimptotiskie sadalījumi vienas izlases gadījumā. Ļoti svarīga problēma statistikā ir divu izlašu salīdzināšana pārbaude vai starp tām eksistē kāda sakarība. Šajā nodaļā tiks apskatīti strukturālo attiecību modeļi, pierādītas to empīrisko procesu asimptotikas. Pieņemsim, ka dotas divas izlases $X_1, ...X_n$ i.i.d. un $Y_1, ..., Y_m$ i.i.d. ar sadalījuma funkcijām attiecīgi F_1 un F_2.

2.1. Strukturālo attiecību modeļi

Strukturālo attiecību modeļi kā jēdziens parādījās salīdzinoši nesen Freitag un Munk (2005) publikācijā [7]. Paši par sevi šie modeļi ir veidoti kā vispārinājums un iekļauj sevī divas nozīmīgas problēmas - lokācijas-skalēšanas modeli, un proporcionālā riska jeb Lēmaņa alternatīvu modeli.

Definīcija 2.1. citea16 Starp divu izlašu sadalījuma funkcijām pastāv lokācijas-lkalēšanas modelis, ja

$$F_1(x) = F_2\left(\frac{x-\mu}{\sigma}\right) := F_2(x,h), x \in \mathbb{R}.$$

Šo attiecību var izteikt arī ar kvantiļu funkcijām

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

Definīcija 2.2. [6] Starp divu izlašu sadalījuma funkcijām pastāv Lēmaņa alternatīvu modelis, ja

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

Izmantojot kvantiļu funkcijas, iegūst

$$F_1^{-1}(t) = F_2^{-1}(1-(1-t)^h), t \in [0,1].$$

Pieņemsim, ka sadalījuma funkcijas F_1 un F_2 ir elementi no funkciju klases

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

Definīcija 2.3. [7] Pieņemsim, ka $\mathcal{H} \subseteq \mathbb{R}^l$ un $\phi_1 : \mathbb{R} \times \mathcal{H} \to \mathbb{R}, \phi_2 : [0,1] \times \mathcal{H} \to [0,1]$. Funkcijas F_1 un F_2 ir saistītas ar strukturālu attiecību, kuru veido ϕ_1 un ϕ_2, ja $(F_1, F_2) \in \mathcal{U}_{\phi_1,\phi_2} =: \mathcal{U}$, kur modeļu klase $\mathcal{U}$ tiek definēta kā

$$\mathcal{U} := \left\{ (F_1, F_2) \in \mathcal{F}^2 \times \mathcal{F}^2 |\ h \in \mathcal{H} \text{ tāds, ka } F_1^{-1}(t) = \phi_1(F_2^{-1}(\phi_2(t,h))), t \in [0,1] \right\}.$$

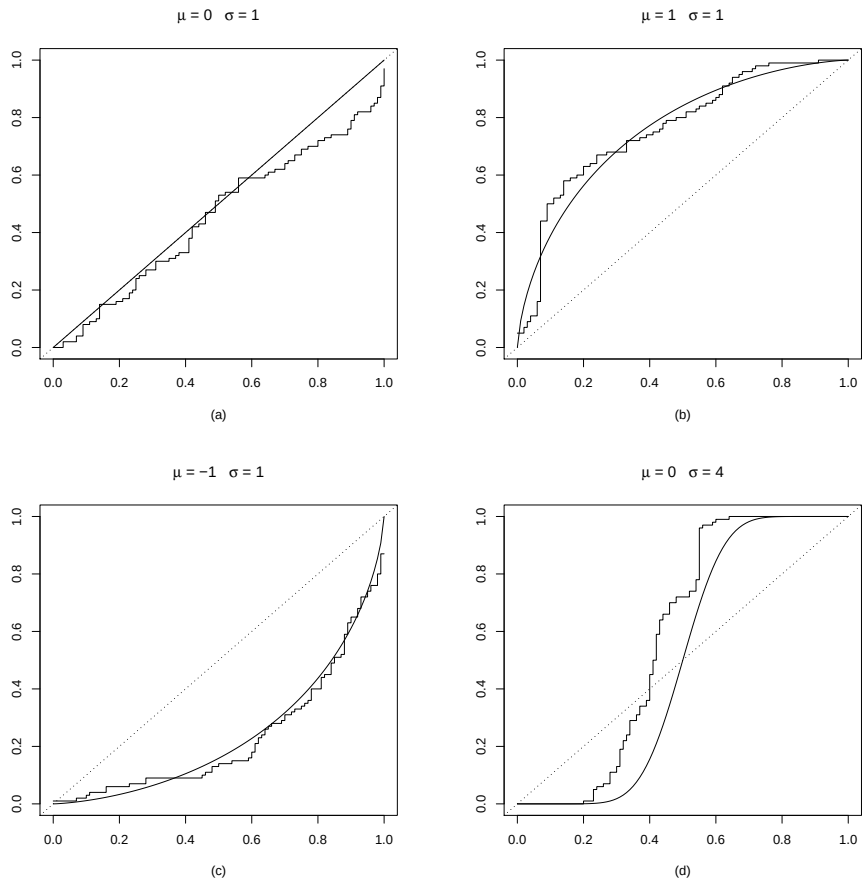
No definīcijas izriet, ka izvēloties $\phi_2(t,h) \equiv t$ un $\phi_1(x,(h_1,h_2)^T) = h_1 + h_2 x$, tiek iegūts Lokācijas-Skalēšanas modelis un ņemot $\phi_1(x,h) \equiv x$ un $\phi_2(t,h) = 1-(1-t)^h$ iegūst Lēmaņa alternatīvu modeli. Turpmāk darbā šiem modeļiem tiks analizētas gan vienkāršās, gan saliktās hipotēzes. Lai novērtētu parametrus h, tiek izmantots Mallova attālums [12, 511. lpp], ar kuru var tikt aprēķināts attālums starp kvantiļu funkcijām. Vispārīgā gadījumā strukturālo attiecību modelim teorētiskie parametri h_0 tiek aprēķināti kā

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

aizvietojot teorētiskās sadalījuma funkcijas ar empīriskajām, iegūst parametra novērtējumu.

2.2. P-P un Q-Q grafiki

Definīcija 2.4. [13, 242. lpp.] Par varbūtību-varbūtību (P-P) grafiku sauc funkciju $F_1(F_2^{-1}(t))$, kur $0 < t < 1$. Aizstājot F_1 un F_2 ar to empīriskajām versijām, iegūst empīrisko P-P grafiku.



2.1. att.: Empīrisko un teorētiskko P-P grafiku piemēri, sadalījuma likumiem $N(0,1)$ pret $N(\mu,\sigma^2)$. Ģenerēto izlašu apjomi n=100.

Attēlā 2.2. redzami daži empīrisko un teorētisko P-P grafiku salīdzinājumu piemēri. Empīriskais P-P grafiks tiks iegūts ģenerējot gadījuma izlases ar apjomu 100. Ja izlašu sadalījuma likumi ir vienādi, tad empīriskais P-P grafiks tuvs taisnei $y = x$. Ja izlašu sadalījuma likumi ir no vienas klases, bet atšķiras matemātiskā cerība, tad grafiks noliecas virs vai zem diagonāles, atkarībā no tā, vai otrās izlases matemātiskā cerība ir lielāka vai mazāka par pirmās izlases matemātisko cerību (b,c). Ja izlašu sadalījuma likumi ir no vienas klases, bet atšķiras dispersija, grafiks tiek saspiests uz vidu (d). Palielinoties izlases apjomam empīriskais P-P grafiks tiecas uz teorētisko, lai pētītu šo grafiku starpību pie dažādiem izlašu apjomiem, to normē ar $\sqrt{n}$ un iegūst empīrisko procesu divu izlašu gadījumā.

Definīcija 2.5. [14, 28. lpp.] Par empīrisko P-P procesu sauc

$$PP(t) = \sqrt{n}(F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t))).$$

Šo procesu var lietot gan divu izlašu sadalījuma funkciju vienādības pārbaudei, gan

arī lai noteiktu doto izlašu empīriskā P-P grafika ticamības joslu. Līdzīgi var aplūkot kvantiļu-kvantiļu grafiku.

Definīcija 2.6. [13, 242. lpp.] Par kvantiļu-kvantiļu (Q-Q) grafiku sauc $F_1^{-1}(F_2(t))$ kur $0 < t < 1$. Līdzīgi kā P-P grafikam, aizvietojot teorētiskās funkcijas ar empīriskajām, iegūst empīrisko grafiku.

Definīcija 2.7. [14, 28. lpp.] Par empīrisko kvantiļu - kvantiļu procesu sauc

$$QQ(x) = \sqrt{n}(F_1^{-1}(F_2(x)))(F_{1n}^{-1}(F_{2m}(x)) - F_1^{-1}(F_2(x)))$$

Šo procesu arī var izmantot divu sadalījuma funkciju pārbaudei. Nākošajā nodaļā tiks sīkāk pētīti abi procesi.

2.3. Divu sadalījuma funkciju vienādības pārbaude

Pieņemsim, ka ir dotas divas neatkarīgas izlases $X_1, X_2, ..., X_n$ un $Y_1, Y_2, ..., Y_m$, kuras abas satur neatkarīgus un vienādi sadalītus gadījuma lielumus, n un m ir izlašu apjomi. Pieņemsim, ka izlašu teorētiskās sadalījuma funkcijas ir F_1 un F_2, izlašu empīriskās sadalījuma funkcijas attiecīgi F_{1n} un F_{2m}. Aplūkosim empīrisko procesu pielietošanu divu izlašu sadalījuma funkciju vienādības pārbaudes hipotēzei 2.1. Šīs hipotēzes pārbaudei bieži tiek lietots Kolmogorova-Smirnova tests divu izlašu gadījumā. Testa statistika tiek definēta kā

$$D := \sup_x \sqrt{\frac{nm}{n+m}} |F_{1n}(x) - F_{2m}(x)|,$$

kur n un m - izlašu apjomi un F_{1n} un F_{2m} - izlašu empīriskās sadalījuma funkcijas. Aplūkosim šīs problēmas alternatīvu risinājumu. Pārbaudam hipotēzi:

$$H_0 : \ F_1(x) = F_2(x) \text{ pret } H_1 : F_1(x) \neq F_2(x). \tag{2.1}$$

Pie pieņēmuma, ka ir spēkā H_0, teorētiskais varbūtību-varbūtību grafiks kļūst daudz vienkāršāks: $F_1(F_2^{-1}(t)) = t$. Tātad, lai pārbaudītu hipotēzi 2.1, nepieciešams konstruēt ticamības joslas PP grafikam $F_1(F_2^{-1}(t))$. Ja diagonāle pilnīgi visos punktos būs iekšā joslā, tad nevarēs noraidīt H_0. Aplūkosim statistiku

$$\sup_{0<t<1} |\sqrt{n}(F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t)))|. \tag{2.2}$$

Šīs statistikas asimptotiskais sadalījums dots publikācijā [14, 29. lpp] un līdzīgas problēmas risinājuma daži soļi doti [4, 1900. lpp]. Lai gan detalizēts pierādījums netika atrasts, tomēr iegūtā informācija bija pietiekoša, lai varētu rekonstruēt pierādījuma gaitu. Turpmākos pierādījumos svarīgs rīks būs teorēma par vidējo vērtību.

Teorēma 2.1. [15, 311. lpp](videjā vērtība) Pieņemsim, ka funkcija $f : [a,b] \to \mathbf{R}$ ir nepārtraukta un diferencējama vaļējā intervālā (a,b). Tad $\exists\ c \in (a,b)$ tāds, ka $f'(c) = \frac{f(b)-f(a)}{b-a}$.

Teorēma 2.2. [11, 13. lpp](Levī nepārtrauktības modulis)
Pieņemsim, ka $B(t)$ ir Brauna tilts intervālā [0,1], tad

$$\limsup_{h\downarrow 0} \sup_{0\le t\le 1-h} \frac{|B(t+h)-B(t)|}{\sqrt{2h\log(1/h)}} = 1 \text{ gandrīz droši.} \tag{2.3}$$

Teorēma 2.3. *Pieņemsim, ka dotas divas izlases $X_1,...,X_n$ un $Y_1,...,Y_m$ katrā no tām elementi ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar sadalījuma funkcijām attiecīgi F_1 un F_2, tad ir spēkā*

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_2^{-1}(t))$$
$$-(B^{(n)}(F_1F_2^{-1}(t))) + \sqrt{\frac{n}{m}}\frac{f_1(F_2^{-1}(t))}{f_2(F_2^{-1}(t))}B^{(m)}(t))| > \epsilon) = 0\ g.d.,$$

kur $B^{(n)}$ un $B^{(m)}$ ir divi neatkarīgi brauna tilti, f_1 un f_2 ir attiecīgi izlašu X un Y teorētiskās blīvuma funkcijas.

Pierādījums. Veicot ekvivalentus pārveidojumus, sadalam doto izteiksmi divos saskaitāmajos.

$$\sqrt{n}(F_{1n}F_{2m}^{-1} - F_1F_2^{-1}(t)) = \sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_{2m}^{-1}(t)) + \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(F_1F_{2m}^{-1}(t) - F_1F_2^{-1}(t)).$$

Pielietojot Teorēmu 1.4 pirmajam saskaitāmajam, iegūst

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_{2m}^{-1}(t)) - B^{(n)}(F_1F_{2m}^{-1}(t))| > \epsilon) = 0\ g.d.$$

Tā kā empīriskā kvantiļu funkcija tiecas uz teorētisko kvantiļu funkciju gandrīz droši, tad Teorāmas 2.2 iegūst, ka

$$\forall\ \epsilon > 0 \lim_{m\to\infty} P(\sup_{0<t<1} |B^{(n)}(F_1F_2^{-1}(t)) - B^{(n)}(F_1F_{2m}^{-1}(t))| > \epsilon) = 0\ g.d.$$

no kurienes seko, ka

$$\forall\, \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_{2m}^{-1}(t)) - B^{(n)}(F_1F_2^{-1}(t))| > \epsilon) = 0 \ g.d.$$

Pielietojot Teorēmu 2.1 otrajam saskaitāmajam, iegūst

$$\sqrt{m}(F_1F_{2m}^{-1}(t) - F_1F_2^{-1}(t)) = f_1(\xi)\sqrt{m}(F_{2m}^{-1}(t) - F_2^{-1}(t)),$$

kur $\xi = \xi_m(t)$ atrodas starp F_{2m}^{-1} un F_2^{-1}. Tā kā, pieaugot izlases apjomam, empīriskā kvantiļu funkcija tiecas uz teorētisko kvantiļu funkciju, tad

$$\forall\, \epsilon > 0 \lim_{n\to\infty} P(\sup_{0<t<1} |\xi_m(t) - F_2^{-1}(t)| > \epsilon) = 0 \ g.d. \ un$$

$$\forall\, \epsilon > 0 \lim_{n\to\infty} P(\sup_{0<t<1} |f_1(\xi_m(t)) - f_1(F_2^{-1}(t))| > \epsilon) = 0 \ g.d.$$

Pēc Teorēmas 1.6

$$\forall\, \epsilon > 0 \lim_{m\to\infty} P(\sup_{0<t<1} |\sqrt{m}(F_{2m}^{-1}(t) - F_2^{-1}(t)) - \frac{1}{f(F_2^{-1}(t))}B^{(m)}(t)| \geq \epsilon) = 0 \text{ gandrīz droši.}$$

Līdz ar to, otro saskaitāmo var novērtēt šādi:

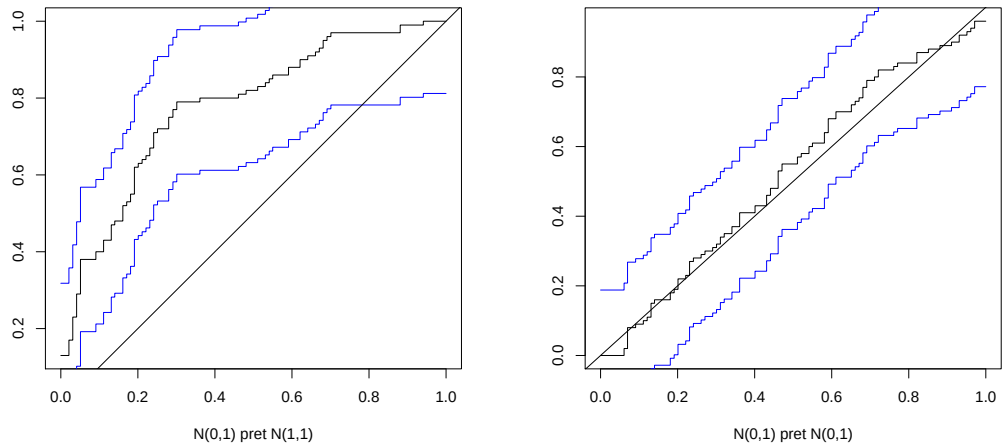
$$\forall\, \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{\tfrac{n}{m}}\sqrt{m}(F_1F_{2m}^{-1}(t) - F_1F_2^{-1}(t)) - \sqrt{\tfrac{n}{m}}\frac{f_1(F_2^{-1}(t))}{f_2(F_2^{-1}(t))}B^{(m)}(t)| > \epsilon) = 0 \ g.d.$$

□

Pie pieņēmuma, ka nulles hipotēze 2.1 ir spēkā, asimptotiskais sadalījums vienkāršojas, t.i.

$$\forall\, \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - t) - (B^{(n)}(t) + \sqrt{\frac{n}{m}}B^{(m)}(t))| > \epsilon) = 0 \ g.d.$$

Šādam sadalījumam var aprēķināt kritiskās vērtības un, konstruēt vienlaicīgās ticamības joslas empīriskajam P-P grafikam. HipotēzeH_0 tiek noraidīta, ja kontruētās joslas neiekļauj taisni $y = t$, kur $t \in [0, 1]$ kaut vienā punktā. Noenoraidītas un noraidītas hipotēzes piemērus var aplūkot attēlā 2.2.



2.2. att.: Nenoraidītas un noraidītas hipotēzes piemēri, konstruējot vienlaicīgās ticamības joslas P-P grafikiem

Šīs pašas problēmas risināšanai var izmantot arī Q-Q grafiku un tā empīrisko procesu. Šajā gadījumā nedaudz tiek lietota nedaudz savādāka statistika un asimptotiskais sadalījums ir vienkāršāks. Izmantojam statistiku

$$\sup_{0<t<1} |\sqrt{n} f_1(F_1^{-1}F_2(x))(F_{1n}^{-1}F_{2m}(x) - F_1^{-1}F_2(x))|,$$

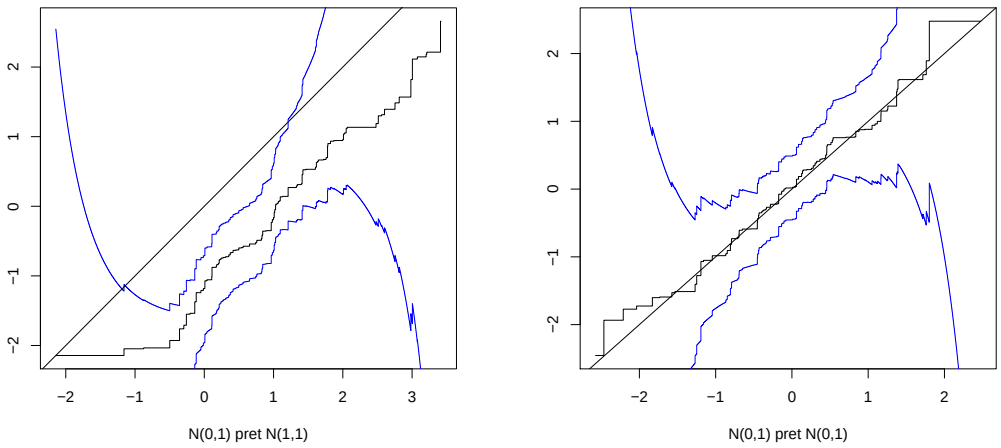
kur f_1 ir pirmās izlases blīvuma funkcija. Rīkojoties kā teorēmas 2.3 pierādījumā, var pierādīt teorēmu empīriskā kvantiļu-kvantiļu procesa asimptotikai.

Teorēma 2.4. *Pieņemsim, ka dotas divas izlases* $X_1, ..., X_n$ *un* $Y_1, ..., Y_m$ *katrā no tām elementi ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar sadalījuma funkcijām attiecīgi* F_1 *un* F_2*, tad ir spēkā*

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{-\infty<x<\infty} |\sqrt{n} f_1(F_1^{-1}F_2(x))(F_{1n}^{-1}F_{2m}(x) - F_1^{-1}F_2(x)) - (B^{(n)}(F_2(x)) + \frac{\sqrt{n}}{\sqrt{m}} B^{(m)}(F_2(x)))| > \epsilon) = 0 \ g.d.,$$

kur $B^{(n)}$ *un* $B^{(m)}$ *ir divi neatkarīgi brauna tilti,* f_1 *un* f_2 *ir attiecīgi izlašu* X *un* Y *teorētiskās blīvuma funkcijas.*

Kā redzams teorēmā, ja divām kvantiļu funkcijām izpildās hipotēze H_0, tad empīriskā kvantiļu procesa asimptotika ir divu neatkarīgu brauna tiltu summa. Šajā gadījumā blīvuma funkcijas iekļaušana izmantotajā statistikā nodrošina ērtāku rezultātu. Tomēr var rasties problēmas konstruējot ticamības joslas, jo tās iekļauj izlases blīvuma funkciju, kas jānovērtē, piemēram, ar kodolu metodi. Attēlā 2.3. redzami nenoraidītas un noraidītas hipotēzes piemēri.



2.3. att.: Nenoraidītas un noraidītas hipotēzes piemēri, konstruējot vienlaicīgās ticamības joslas.

2.4. Lokācijas-skalēšanas modelis

Šajā nodaļā tiks aplūkots iepriekš aprakstītais lokācijas-skalēšanas modelis. Divām dotām izlasēm var pārbaudīt gan vienkāršu hipotēzi par konkrētām parametru μ un σ vērtībām, gan saliktu hipotēzi par modeļa eksistenci. Lai veiktu otro uzdevumu nepieciešams novērtēt parametrus. Pārbaudāmā hipotēze ir sekojoša:

$$H_0 : \ F_1(x) = F_2\left(\frac{x-\mu}{\sigma}\right) \text{ pret } H_1 : F_1(x) \neq F_2\left(\frac{x-\mu}{\sigma}\right), \tag{2.4}$$

kur μ - lokācijas parametrs, σ - skalēšanas parametrs.

Šo modeli var arī uzrakstīt, izmantojot kvantiļu funkcijas

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

Mallova attālums, skatīt [6, 125. lpp], parametru novērtēšanai ir šāds:

$$M(F_1, F_2) := \int_0^1 (F_1^{-1}(t) - \sigma F_2^{-1}(t) - \mu)^2 dt.$$

Lai iegūtu μ un σ novērtējumus, aizvietojam teorētisko kvantiļu funkcijas ar to atbilstošajām empīriskajām versijām. Veicot ekvivalentus pārveidojumus iegūst

$$\hat{\sigma} = \frac{\int_0^1 F_{1n}^{-1}(t) F_{2m}^{-1}(t) dt - \int_0^1 F_{1n}^{-1}(t) dt \int_0^1 F_{2m}^{-1}(t) dt}{\int_0^1 (F_{2m}^{-1}(t))^2 dt - (\int_0^1 (F_{2m}^{-1}(t) dt)^2},$$

$$\hat{\mu} = \int_0^1 F_{1n}^{-1}(t) dt - \hat{\sigma} \int_0^1 F_{2m}^{-1}(t) dt.$$

Ievērosim, ka, ja sakrīt izlašu apjomi, tad tiek iegūti lineārās regresijas modeļa parametru novērtējumi. Tad novētrētos parametrus var pierakstīt vienkāršāk

$$\hat{\sigma} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_{(i)}Y_{(i)} - \frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\frac{1}{n}\sum_{i=1}^{n} Y_i)^2},$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i - \hat{\sigma}\frac{1}{n}\sum_{i=1}^{n} Y_i,$$

kur $X_{(i)}$ un $Y_{(i)}$ - augošā secībā sakārtoti doto izlašu elementi.

Kad novērtēti parametri, tos var izmantot modeļa pārbaudē. Izmantosim empīrisko procesu

$$PP_{LS}(t) := \sqrt{n}(F_{1n}(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}) - F_1(\sigma F_2^{-1}(t) + \mu)).$$

Teorēma 2.5. *Ir spēkā*

$$\forall \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |PP_{LS}(t) - (B^{(n)}(F_1(\sigma F_2^{-1}(t) + \mu)) + \sigma\sqrt{\frac{n}{m}}\frac{f_1(\sigma F_2^{-1}(t) + \mu)}{f_2(F_2^{-1}(t))}B^{(m)}(t)$$
$$+\sqrt{n}f_1(\sigma F_2^{-1}(t) + \mu)((\hat{\mu} - \mu) + F_2^{-1}(t)(\hat{\sigma} - \sigma)))| > \epsilon) = 0 \ g.d.$$

Pierādījums. Doto izteiksmi sadalam divos saskaitāmos līdzīgi kā Teorēmas 2.3 pierādījumā.

$$\begin{aligned} PP_{LS}(t) := \sqrt{n}((&F_{1n}(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}) - F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu})) \\ &+(F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}) - F_1(\sigma F_2^{-1}(t) + \mu))). \end{aligned} \tag{2.5}$$

Pirmajam saskaitāmajam no 2.5 pēc teorēmas 1.4 ir spēkā

$$\forall \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}((F_{1n}(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}) - F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}))$$
$$-B^n(F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu})))| > \epsilon) = 0 \ g.d.$$

No Teorēmas 2.2 ir spēkā

$$\forall \ \epsilon \ P(\sup_{0<t<1} |B^{(n)}(F_1(\sigma F_2^{-1}(t) + \mu)) - B^{(n)}((F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}))| > \epsilon) = 0 \ g.d.$$

Aplūkosim otru saskaitāmo no 2.5 un izmantosim Teorēmu 2.1, pēc kuras ir spēkā

$$\sqrt{n}(F_1(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}) - F_1(\sigma F_2^{-1}(t) + \mu)) = \sqrt{n}f_1(\xi_m(t))(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu} - \sigma F_2^{-1}(t) - \mu),$$

kur $\xi_m(t)$ atrodas starp $\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu}$ un $\sigma F_2^{-1}(t) + \mu$. Tā kā $\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu} \to \sigma F_2^{-1}(t) + \mu$, kad $m \to \infty$, tad arī $\xi_m(t) \to \sigma F_2^{-1}(t) + \mu$, kad $m \to \infty$. Aplūkosim $\sqrt{n}(\hat{\sigma}F_{2m}^{-1}(t) + \hat{\mu} - \sigma F_2^{-1}(t) - \mu)$. Šo starpību var pārveidot par

$$\sqrt{\tfrac{n}{m}}\sqrt{m}(\hat{\sigma}F_{2m}^{-1}(t)+\hat{\mu}-\hat{\sigma}F_2^{-1}(t)-\hat{\mu})+\sqrt{n}(\hat{\sigma}F_2^{-1}(t)+\hat{\mu}-\sigma F_2^{-1}(t)-\hat{\mu})$$

$$+\sqrt{n}(\sigma F_2^{-1}(t)+\hat{\mu}-\sigma F_2^{-1}(t)-\mu) \tag{2.6}$$

Aplūkosim katru saskaitāmo atsevišķi. Pirmais saskaitāmais no 2.6 vienkāršojas uz

$$\sqrt{\frac{n}{m}}\sqrt{m}\hat{\sigma}(F_{2m}^{-1}(t)-F_2^{-1}(t)).$$

Pēc Teorēmas 1.6 un fakta, ka $\hat{\sigma}\to\sigma$, kad $m\to\infty$ ir spēkā

$$\forall\epsilon>0\ \lim_{n,m\to\infty}P(\sup_{0<t<1}|\sqrt{m}\hat{\sigma}(F_{2m}^{-1}(t)-F_2^{-1}(t))-\frac{\sigma}{f_2(F_2^{-1}(t))}B^{(m)}(t)|>\epsilon)=0\ g.d.$$

Otro saskaitāmo no 2.6 var parveidot kā

$$\sqrt{n}F_2^{-1}(t)(\hat{\sigma}-\sigma)$$

un trešais no 2.6 vienkāršojas uz

$$\sqrt{n}(\hat{\mu}-\mu).$$

Apvienojot veiktos spriedumus, teorēma ir pierādīta. □

Šis ir lokācijas-skalēšanas modeļa asimptotiskais sadalījums vispārīgā gadījumā, ja tiek pārbaudīta saliktā hipotēze. Pie hipotēzes H_0 ir spēkā

$$\forall\epsilon>0\ \lim_{n,m\to\infty}P(\sup_{0<t<1}|PP_{LS}(t)-(B^{(n)}(t)+\sqrt{\frac{n}{m}}B^{(m)}(t)$$
$$+\sqrt{n}f_2(F_2^{-1}(t))((\hat{\mu}-\mu)+F_2^{-1}(t)(\hat{\sigma}-\sigma)))|>\epsilon)=0\ g.d.$$

Līdz ar to pat pie hipotēzes H_0 asimptotiskais sadalījums atkarīgs no nezināmām kvantiļu un blīvuma funkcijām un parametru novērtējumiem. Šim modelim var aplūkot trīs speciālgadījumus:

1. tiek pārbaudīts fiksēts parametrs μ, un novērtēts σ;

2. tiek pārbaudīs fiksēts parametrs σ un novērtēts parametrs μ;

3. tiek pārbaudīti abi fiksēti parametri μ un σ.

   Pēdējā gadījumā asimptotiskais sadalījums sakrīt ar PP empīrisko procesu divu izlašu sadalījuma funkciju pārbaudē iegūto rezultātu, Turklāt ievērosim, ka 2.1 ir speciālgadījums punktam 3) ar $\mu=0$ un $\sigma=1$.

2.5. Lēmaņa alternatīvu modelis

Šajā nodaļā aplūkosim otru no strukturālo attiecību modeļiem - Lēmana alternatīvu modeli.

Definīcija 2.8. Modelis pieder Lēmaņa alternatīvu klasei, ja

$$F_1(x) = 1 - (1 - F_2(x))^{(1/h)},$$

kur $t \in \mathbb{R}$ un $h > 0$. Izmantojot kvantiļu funkcijas, modelis var tiks uzrakstīts kā

$$F_1^{-1}(t) = F_2^{-1}(1 - (1 - t)^h).$$

Lemma 2.6. *Ja funkcijas F_1 un F_2 pieder Lēmaņa alternatīvu klasei, tad ir spēkā proporcionālā riska modelis (proportional hazard model).*

$$\frac{f_2(x)}{1 - F_2(x)} = h\frac{f_1(x)}{1 - F_1(x)}$$

Pierādījums. Atvasinot abas vienādojuma puses pēc x, tiek iegūts

$$f_1(x) = \frac{1}{h} f_2(x)(1 - F_2(x))^{1/h-1}.$$

Veiksim ekvivalentus pārveidojumus.

$$\frac{h f_1(x)}{(1 - F_2(x))^{1/h}} = \frac{f_2(x)}{1 - F_2(x)}.$$

Visbeidzot ievērosim, ka

$$1 - F_1(x) = 1 - (1 - (1 - F_2(x))^{1/h}) = (1 - F_2(x))^{1/h},$$

līdz ar to rezultāts ir pierādīts. □

Mallova attālums šim modelim tiek definēts kā

$$M(F_1, F_2) = \int_0^1 (F_1^{-1} - F_2^{-1}(1 - (1 - t)^h))^2 dt.$$

Salīdzinājumā ar lokācijas-skalēšanas modeli, Lēmana alternatīvu modelis ir daudz sarežģītāks un parametra novērtējumu analītiskās funkcijās uzrakstīt nav iespējams. Kā šajā gadījumā novērtēt parametru h tiks aprakstīts 3.nodaļā. Šobrīd pieņemsim, ka parametrs

ir novērtēts un apzīmēsim to ar $\hat{h}$. Pieņemsim, ka $\hat{h} \xrightarrow{p} h$, kad $n \to \infty$ un aplūkosim empīrisko procesu

$$PP_h(t) := \sqrt{n}(F_{1n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_2^{-1}(1-(1-t)^h)))$$

un pierādīsim atbilstošās statistikas asimptotisko sadalījumu.

Teorēma 2.7. *Ir spēkā*

$$\forall \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |PP_h(t) - (B^{(n)}(F_1(F_2^{-1}(1-(1-t)^h)))$$

$$+\frac{f_1(F_2^{-1}(1-(1-t)^h))}{f_2(F_2^{-1}(1-(1-t)^h))h(1-t)^{h-1}}(\sqrt{\frac{n}{m}}B^{(m)}(1-(1-t)^h) - \sqrt{n}((1-t)^h - (1-t)^{\hat{h}})))| > \epsilon) = 0 \ g.d.$$

Pierādījums. Veicam ekvivalentus pārveidojumus

$$\sqrt{n}(F_{1n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_2^{-1}(1-(1-t)^h)))$$

$$= \sqrt{n}(F_{1n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}}))) \tag{2.7}$$

$$+\sqrt{n}(F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_2^{-1}(1-(1-t)^h)))$$

Aplūkosim pirmo saskaitāmo no 2.7. Pēc teorēmas 1.4 ir spēkā

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{n}(F_{1n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}})))$$

$$-B^{(n)}(F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}})))) = 0 \ g.d.$$

Pēc pieņēmuma, ka $\hat{h} \to h$, kad $n, m \to \infty$ un Levī moduļa Brauna tiltam ir spēkā

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |B^{(n)}(F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}}))) - B^{(n)}(F_1(F_2^{-1}(1-(1-t)^h)))) = 0 \ g.d.$$

Aplūkosim otru saskaitāmo no 2.7. Pielietojot vidējās vērtības teorēmu, iegūst

$$\sqrt{n}(F_1(F_{2m}^{-1}(1-(1-t)^{\hat{h}})) - F_1(F_2^{-1}(1-(1-t)^h)))$$

$$= f_1(\xi_m(t))\sqrt{n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^h)),$$

kur $\xi_m(t)$ ir starp $F_{2m}^{-1}(1-(1-t)^{\hat{h}})$ un $F_2^{-1}(1-(1-t)^h)$. No pieņēmuma, ka $\hat{h} \to h$, kad $n, m \to \infty$ seko, ka $\xi_m(t) \to F_2^{-1}(1-(1-t)^h)$, kad $n, m \to \infty$.

$$\sqrt{n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^h))$$

$$= \sqrt{n}(F_{2m}^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^{\hat{h}}) + F_2^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^h))$$

Pēc Teorēmas 1.6 un Teorēmas 2.2 ir spēkā

$$\forall\ \epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |\sqrt{\frac{n}{m}}\sqrt{m}(F_{2m}^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^{\hat{h}})) \\ -\sqrt{\tfrac{n}{m}}F_2^{-1}(1-(1-t)^h)'B^{(m)}(1-(1-t)^h)| > \epsilon) = 0\ g.d.$$

No vidējās vērtības teorēmas seko

$$\sqrt{n}(F_2^{-1}(1-(1-t)^{\hat{h}}) - F_2^{-1}(1-(1-t)^h)) = (F_2^{-1}(1-(1-t)^h))'((1-t)^h - (1-t)^{\hat{h}})$$

Apkopojot veiktos spriedumus, teorēma ir pierādīta. □

Pie pieņēmuma, ka izpildās saliktā hipotēze par Lēmaņa modeļa izpildīšanos (modelis eksistē), iegūst

$$\forall\epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |PP_h(t) - B^{(n)}(F_1(F_2^{-1}(1-(1-t)^h))) \\ +\sqrt{\frac{n}{m}}B^{(m)}(1-(1-t)^h) - \sqrt{n}((1-t)^h - (1-t)^{\hat{h}}))| > \epsilon) = 0\ g.d.,$$

bet pie pieņēmuma, ka izpildās vienkāršā hipotēze (modelis ir spēkā ar fiksētu parametru h), iegūst vēl vienkāršāku rezultātu.

$$\forall\epsilon > 0 \lim_{n,m\to\infty} P(\sup_{0<t<1} |PP_h(t) - B^{(n)}(F_1(F_2^{-1}(1-(1-t)^h))) + \sqrt{\frac{n}{m}}B^{(m)}(1-(1-t)^h))| > \epsilon) = 0\ g.d.$$

3. Praktiskā daļa

Šajā nodaļā tiks analizēti teorētiskajā daļā apskatītie modeļi, veicot empīriskās pārklājuma precizitātes pārbaudi ar pielietojumu reāliem datiem. Turklāt tiks simulētas dažādas pierādīto asimptotisko sadalījumu kvantiles. Analīzes dažādos posmos tiks izmantota kodolu gludināšana, kas tiks definēta nodaļas sākumā.

3.1. Lokācijas-skalēšanas modeļa analīze

Dažādiem modeļiem tika noteikta statistikas kritiskā vērtība pie līmeņiem $\alpha = 0.10$, $\alpha = 0.05$ un $\alpha = 0.01$. To iespējams izdarīt gan veicot simulācijas ar asimptotisko sadalījumu, gan simulējot izvēlēto statistiku, kas ir vienkāršāks veids. Tā kā lokācijas-skalēšanas modeļiem tikai īpašā gadījumā (ja veic vienkāršas hipotēzes pārbaudi ar fiksētiem parametriem) asimptotiskais sadalījums ir Brauna tiltu summa, tad kritiskās vērtības nepieciešams iegūt simulējot statistiku. Piezīmēsim, ka simulācijas ar Brauna tiltiem minētajā speciālajā gadījumā deva šādus rezultātus: pie $\alpha = 0.10$ kritiskā vērtība ir 1.71, pie $\alpha = 0.05$ - 1.89 un pie $\alpha = 0.01$ - 2.28.

Lai iegūtu korektus rezultātus tika izvēlēti sadalījuma funkciju pāri kuri veido lokācijas-skalēšanas modeli, tika ģenerētas neatkarīgas gadījuma izlases no izvēlētajiem sadalījumiem un aprēķināta testa statistika. Process tika atkārtots 10000 reizes un ģenerēto izlašu apjoms $n = 5000$. Izlases tika ģenerētas fiksējot vienu un to pašu gadījuma lielumu ģenerēšanas sākumpunktu (R komanda - set.seed(10)) Rezultātus var aplūkot tabulā 3.1.

No iegūtajiem rezultātiem var secināt, ka konkrēta modeļa pārbaudes gadījumā kritiskā vērtība nav atkarīga no izlašu sadalījuma likumiem. Pārbaudot hipotēzi par lokācijas-skalēšanas modeļa eksistenci un veicot parametru novērtēšanu, kritiskā vērtība ir atkarīga no sadalījuma funkciju klases, bet nav atkarīga no konkrētiem parametriem.

3.1. tabula: Kritiskās vētības dažādiem lokācijas-skalēšanas modeļa piemēriem. Parametri tiek novērtēti - gadījums, kad tiek pārbaudīta paša modeļa eksistence (saliktā hipotēze). Parametri zināmi - konkrēta modeļa pārbaude (vienkāršā hipotēze), kas sevī ietver divu sadalījuma funkciju modeli kā speciālgadījumu.

		Parametri tiek novērtēti			Parametri zināmi		
1. sad.	2. sad.	90%	95%	99%	90%	95%	99%
$N(0,1)$	$N(2,16)$	1.159655	1.25865	1.45664	1.697056	1.880904	2.234457
$N(1,4)$	$N(-1,9)$	1.159655	1.25865	1.45664	1.697056	1.880904	2.234457
$U[0,1]$	$U[-2,4]$	1.173797	1.286934	1.499066	1.711198	1.895046	2.262742
$U[3,5]$	$U[4,9]$	1.173797	1.286934	1.499066	1.711198	1.895046	2.262742
$Exp(0.5)$	$Exp(0.25)$	2.333452	2.743574	3.521392	1.682914	1.866762	2.276884
$Exp(1)$	$Exp(2)$	2.333452	2.743574	3.521392	1.682914	1.866762	2.276884

Lai pārliecinātos, ka iegūtās vērtības nodrošina vēlamo rezultātu tika veikta pārklājumu precizitātes pārbaude. Šī metode ir sava veida apgriezts process kririskās vērtības simulācijām. Sākotnēji tiek izvēlēta kritiskā vērtība un pārbaudāmās sadalījuma funkcijas. Tad tiek ģenerētas gadījuma lielumu izlases ar izvēlētajām sadalījuma funkcijām un aprēķināts varbūtību-varbūtību grafiks fiksētos punktos. Izmantojot aprēķināto funkciju un kritisko vērtību, tiek konstruētas vienlaicīgās ticamības joslas un ja kaut vienā punktā teorētiskā funkcija iziet ārpus ticamības joslām, modelis tiek noraidīts. Šādas darbības tika atkārtotas 1000 reizes un tika saskaitīts, cik no tām modelis netiek noraidīts. Kritiskā vērtība uzskatāma par pareizu, ja nenoraidīto gadījumu biežums sakrīt ar 1-nozīmības līmeni, kuram atbilst izvēlētā kritiskā vērtība. Tika veiktas pārbaudes gan gadījumam, kad tiek veikta pārbaude ar zināmiem parametriem, gan gadījumam, kad parametrus novētrē.

Pārklājuma precizitāte tika veikta sadalījuma funkcijām N(0,1) un N(15,9), pārbaudot precīzus modeļa parametrus, sīm pašām funkcijām, novērtējot parametrus un sadalījuma funkcijām Exp(3) un Exp(5) novērtējot parametrus. Tabulā 3.2. attēloti pārbaužu rezultāti.

3.2. Lēmaņa alternatīvu modeļa analīze

Iepriekš, aplūkojot Lēmaņa alternatīvu modeli, tika pieņemts, ka parametrs $\hat{h}$ ir novērtēts. Tagad tiks aprakstīts, kā to izdarīt ar programmas R palīdzību. Līdz šim tika izmantots Mallova attālums, kas definēts strukturālo attiecību modelim pilnā intervālā

3.2. tabula: Pāklājumu precizitāte lokācijas-skalēšanas modelim,pie izlašu apjoma n.

	Konkrēti parametri			$N(0,1)$ un $N(15,9)$			$Exp(3)$ un $Exp(5)$		
n	90%	95%	99%	90%	95%	99%	90%	95%	99%
20	0.924	0.97	0.999	0.97	0.97	0.996	1	1	1
50	0.938	0.968	0.993	0.953	0.953	0.998	0.992	0.998	1
100	0.931	0.961	0.991	0.915	0.96	0.994	0.981	0.995	1
200	0.906	0.948	0.983	0.908	0.942	0.985	0.948	0.981	0.998
500	0.904	0.955	0.997	0.877	0.955	0.984	0.934	0.966	0.997
1000	0.9	0.95	0.989	0.896	0.938	0.989	0.921	0.959	0.994
5000	0.884	0.939	0.989	0.894	0.943	0.987	0.893	0.952	0.994
10000	0.903	0.95	0.992	0.884	0.938	0.991	0.88	0.941	0.987

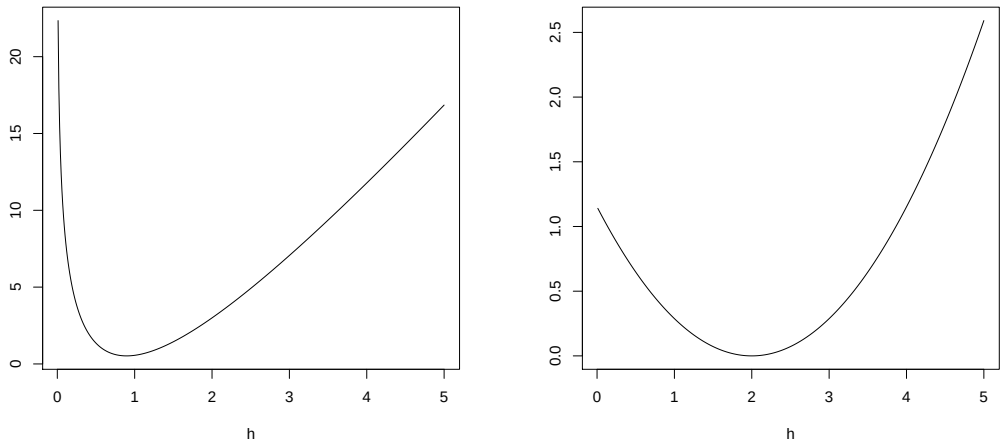
$[0,1]$. Lokācijas - skalēšanas modelim nekādas problēmas neradās, bet var gadīties, ka pie kaut kādiem h Lēmaņa altenatīvu modelim Mallova attālums diverģē. Lai to novērstu tiek lietots tā sauktais nošķeltais Mallova attālums [6, 127. lpp], kas tiek definēts kā

$$\frac{1}{b-a}\int_a^b (F_1^{-1}(t) - F_2^{-1}(1-(1-t)^h))^2 dt,$$

kur $0 < a < b < 1$. Mērķis ir minimizēt šo attālumu, mainot parametru h un par atrisinājumu ņemt to h vērtību, kura dod mazāko attālumu. Līdz ar to sākumā tiek sastādīts potenciālo h vērtību vektors. Pārāk lielu h ņemt nav vērts, jo tad integrālis var diverģēt, darbā tika aplūkotas h vērtības intervālā (0,5] ar atstarpi 0.01. Tad katram h aprēķinam nošķelto Mallova attālumu intervālā [0.05,0.95], tātad tiek izveidots vēl viens vektors. Visbeidzot atrodam jaunā vektora minimālo vērtību, atrodam vietas numuru, kur tā atrodas šajā vektorā un, izvēloties šo pašu numuru sākotnējā h vērtību vektorā, iegūstam rezultātu. Daži detalizēti programmu kodi pieejami pielikumā. Attēlā 3.1. var aplūkot šo realizāciju grafiski.

Šeit aplūkoto izlašu pāri ir tādi, ka zināms vienas izlases sadalījuma likums un otraizlase tiek kontruēta tā, lai būtu spēkā Lēmaņa alternatīvu modelis. Otro izlasi var ģenerēt, piemēram, ņemot izlasi, kas sadalīta U[0,1] un pielietojot inverso transformāciju.

Diezgan precīzi var pārbaudīt procesa asimptotiku vienkāršās hipotēzes gadījumā, jo tā neprasa novērtēt h. Pētījumā tika simulētas izlases ar apjomu $n = 500$ ar dažādiem sadalījuma likumiem un dažādiem h. Simulācijas tika veiktas 10000 reizes. Kritiskās



3.1. att.: Nošķelta Mallova attāluma grafiski piemēri. Attēlā pa kreisi pirmā izlase ir normāli sadalīta, pa labi - eksponenciāli sadalīta. Abos gadījumos otrā izlase konstruēta ar inverso transformāciju no U[0,1] sadalītas izlases tā, ka ir spēkā Lēmaņa alternatīva.

vērtības dažādiem sadalījuma likumiem un dažādiem h vienkāršajai hipotēzei var aplūkot tabulā 3.3. un Pārklājumu precizitāti, kas veikta šim modelim redzama tabulā 3.4. Svarīgākais secinājums ir tāds, ka vienkāršai hipotēzei kritiskā vērtība atkarīga tikai no parametra h un pārklājumu precizitātes pārbaude uzrāda ļoti labus rezultātus.

Vispārīgā gadījumā parametra h vērtība, kas minimizē nošķelto Mallova attālumu nav zināma, tāpēc jāpārbauda plašs intervāls, kas var aizņemt daudz laika. Analizējot modeli un veicot izlašu simulācijas īstā parametra vērtība ir zināma un var izvēlēties šaurāku intervālu. Tomēr šī pieeja metodes pārbaudei neder. Pirmkārt tāpēc, ka pārbaudes process kļūst ļoti ilgs, otrkārt - nevar zināt, vai uzdotā precizitāte ir pietiekoša. Reizēm var gadīties, ka kritiskā vērtība diverģē tikai tāpēc, ka kļūdaini noteikts parametrs h. Saliktas hipotēzes empīriskā procesa simulāciju rezultātus var aplūkot tabulā 3.5. Kā redzams, aprakstītās darbības nav pietiekošas, lai iegūtu konverģējošu asimptotisko sadalījumu.

3.3. tabula: Kritiskās vētrības Lēmaņa alternatīvu modeļiem vienkāršas hipoēzes gadījumā. Simulācijas veiktas 1000 reizes, izlases apjoms n=500.

Sadalījums	h	90%	95%	99%
N(0,1)	0.5	1.945379	2.168986	2.6162
N(0,1)	1.5	1.65469	1.833576	2.213707
N(1,4)	0.5	1.945379	2.168986	2.6162
N(1,4)	1.5	1.65469	1.833576	2.213707
Exp(0.5)	0.5	1.945379	2.146625	2.593839
Exp(1)	1.5	1.65469	1.855936	2.213707
U[-2,3]	0.5	1.923018	2.146625	2.571478
χ^2_4	1.5	1.65469	1.833576	2.213707

3.4. tabula: Pārklājumu precizitāte Lēmaņa alternatīvu modeļiem vienkāršas hipoēzes gadījumā. Simulācijas veiktas 1000 reizes. Vienai no izlasēm dots sadalījuma likums, otra iegūts veicot inverso transformāciju no U[0,1] sadalītu gadījuma lielumu izlases.

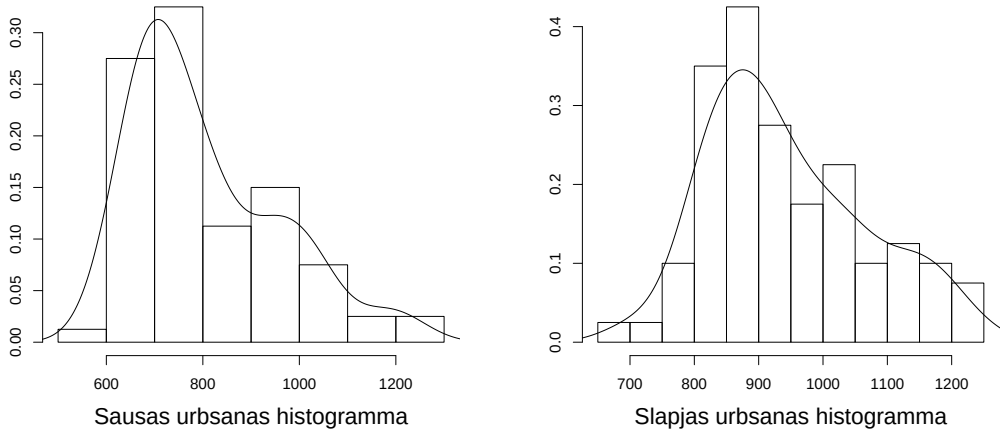
	$N(2,9)$, h=0.5			$Exp(2)$, h=1.5		
n	90%	95%	99%	90%	95%	99%
20	0.923	0.966	0.997	0.905	0.969	0.99
50	0.916	0.966	0.996	0.908	0.954	0.994
100	0.914	0.953	0.989	0.908	0.95	0.986
200	0.907	0.961	0.99	0.89	0.937	0.994
500	0.891	0.945	0.994	0.89	0.951	0.987
1000	0.893	0.947	0.99	0.917	0.95	0.993
5000	0.901	0.95	0.99	0.894	0.942	0.988
10000	0.91	0.956	0.991	0.889	0.952	0.993

3.5. tabula: Kritiskās vētrības Lēmaņa alternatīvu modeļiem saliktās hipoēzes gadījumā. Simulācijas veiktas tikai 100 reizes, jo aizņem daudz laika. No iegūtā rezultāta var secināt, ka aprakstītais novērtēšanas process nav pietiekošs kritiskās vērtības novērtēšanai.

	$N(0,1)$, h=0.5			$Exp(0.5)$, h=3		
n	90%	95%	99%	90%	95%	99%
20	1.699412	1.96774	2.414953	2.414953	2.63856	3.085774
50	1.697056	1.909188	2.404163	2.616295	2.899138	3.323402
100	1.3	1.5	1.8	2.9	3.2	3.8
200	1.343503	1.484924	1.767767	3.181981	3.535534	4.17193
500	1.296919	1.431084	1.744133	3.577709	3.890758	4.740464

3.3. Reālu datu analīze

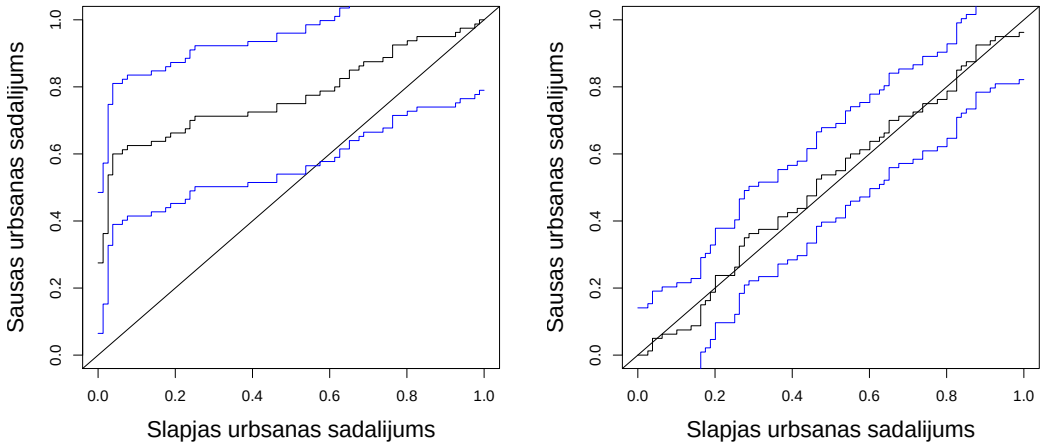
Aplūkosim piemēru no publikācijas [5]. Izlases satur vidējos urbšanas laikus, kuros sasniegts fiksēts dziļums. Tika salīdzināta sausā urbšana un slapjā urbšana. Sākumā aplūkosim abu izlašu histogrammas 3.2..



3.2. att.: Sausās un slapjās urbšanas datu histogrammas, gludinātas ar metodi, kas aprakstīta Pielikumā A1.

Šobrīd mērķis nav noteikt izlašu sadalījumus. Tika veiktas divas hipotēžu pārbudes līmenī $\alpha = 0.05$. Viena par abu izlašu sadalījuma funkciju vienādību, otra par lokācijas-skalēšanas modeļa eksistenci. Pirmā ir vienkārša hipotēze, darbā tika iegūts, ka šādai hipotēzei kritiskā vērtība ir aptuveni 1.88. Otra ir salikta hipotēze. Iepriekš jau tika

salīdzinātas kritiskās vērtības dažādiem sadalījumiem. No aplūkotajiem sadalījumiem, histogrammas vistuvāk ir normālajam sadalījumam, tāpēc tika izvēlēta kritiskā vērtība 1.258. Šāda kritiskā vērtība izvēlēta arī tāpēc, ka tā ir mazākā no aplūkotajām, jo ja hpotēze netiek noraidīta pie šādas kritiskās vērtības, tā netiks noraidīta pie lielākas. Attēlā 3.3. attēloti abu hipotēžu rezultāti. Vienkāršā hipotēze tiek noraidīta, bet salikto hipotēzi nevar noraidīt. Tomēr aplūkoto izlašu sadalījumi nav normālie un iespējams kādam sadalījumam kritiskā vērtība ir vēl mazāka. Šādos gadījumos praksē lieto gludinošo butstrapu, kas aprakstīts Horwath publikācijā [4]. Šī metode novērtē kritisko vērtību no dotatiem datiem.



3.3. att.: Lokācijas skalēšanas modeļu vienkāršaas un saliktas hipotēzes pārbaude. Pa kreisi pārbaude fiksētiem parametriem μ=0 un σ=1. Pa kreisi - parametri tiek novērtēti.

Nobeigums

Darbā tika aplūkoti strukturālo attiecību modeļi un empīrisko procesu pieeja, lai veiktu hipotēžu pārbaudi vai konstruētu ticamības intervālus. Tika pierādīti teorētiskie asimptotiskie sadalījumi ne tikai empīriskajiem varbūtību-varbūtību un kvantiļu-kvantiļu procesiem, bet arī lokācijas-skalēšanas un Lēmaņa alternatīvu modeļos. Pēdējā nodaļā, kas veltīta sadalījumu simulācijām, iegūtie rezultāti liecina, ka vienkāršu hipotēžu gadījumā empīrisko procesu asimptotiskie sadalījumi konverģē ātri un dod labu rezultātu. Tomēr saliktu hipotēžu gadījumā problēmas rada parametru novērtēšana. Situācija ar lokācijas-skalēšanas modeli ir vienkāršāka, jo tā parametru novērtējumus var izteikt ar analītiskām funkcijām. Sliktāk ir ar Lēmaņa alternatīvu modeli, kuram nezinot Mallova attāluma uzvedību, novērtēt parametru ir sarežģīti. Tāpēc būtu tālākā darbā jāanalizē Mallova attāluma asimptotisko uzvedību teorētiski.

Iespējams, ka zem strukturālo modeļu vispārinājuma var iekļaut vēl citus modeļus un analizēt līdzigi kā darbā aplūkotos. Ja parametru novērtējums nav sarežģīts, var veikt simulācijas un izveidot kritisko vērtību tabulas, kuras var izmantot strukturālo modeļu pārbaudei. Tomēr gadījumos, kad izlases sadalījums nav vienkārši nosakāms, tiek izmantots gludinošais butstraps [4]. Tālākā darbā nepieciešams analizēt, vai tas strādā strukturālo attiecību modeļiem vispārējā gadījumā.

Izmantotā literatūra un avoti

[1] P. Billingsley. *Convergence of Probability Measures.* John Wiley & Sons, New York, 1968.

[2] G.R. Shorak and J.A. Wellner. *Empirical Processes with Applications to Statistics.* John Wiley & Sons, New York, 1986.

[3] A.W. Van Der Vaart. *Asymptotic statistics.* Cambridge University Press, 1998.

[4] Z. Horwath, L. Horwath and W. Zhou. Confidence bands for roc curves. *Journal of Statistical Planning and Inference*, 138:1894–1904, 2008.

[5] J. Valeinis, E. Cers, and J. Cielēns. Two-sample problems in statistical data modeling. *Mathematical modelling and analysis*, 15(1):137–151, 2010.

[6] G. Freitag, A. Munk, and M. Vogt. Asessing structural relationships between distributions - a quantile process approach based on mallows distance. *Recent advances and trends in Nonparametric Statistics*, 2008.

[7] 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.

[8] J. Valeinis. Confidence bands for structural relationship models. Dissertation, Goettingen, 2007.

[9] H.W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[10] A DasGupta. *Asymptotic Theory of Statistics and Probability.* Springer, 2008.

[11] M. Csorgo. *Quantile Processes with Statistical Applications*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1983.

[12] Mallows C.L. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43(2):508–515, 1972.

[13] J. Beirlant and P. Deheuvels. On the approximation of p-p and q-q plot processes by brownian bridges. *Statistics and Probability Letters*, 9:241–251, 1990.

[14] F. Hsieh and B. W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1):25–40, 1996.

[15] S.C. Malik and Savita Arora. *Mathematical Analysis*. New Age International (P) Ltd, New Delhi, 1992.

[16] S.J. Sheather and M.C. Jones. A reliable data-based bandwith selection method for kernel density estimation. *Journal of the Royal Statistical Society*, 53:683–690, 1991.

A Pielikums

A1. Blīvuma funkcijas gludināšanas kodolu metode

Pieņemsim, ka ir dota izlase $X_1, ..., X_n$, kur n - izlases apjoms.

Definīcija A1. Funkciju $k : \mathbb{R} \to \mathbb{R}^+$ sauc par kodolu, ja:

1. $\int_{-\infty}^{\infty} k(u)du = 1$;

2. $\forall u\ k(-u) = k(u)$.

Biežāk izmantotās kodolfunkcijas:

1. Vienmērīgais: $k(u) = \frac{1}{2} I_{|u|\leq 1}$, kur I ir indikatorfunkcija;

2. Epanečnikova: $k(u) = \frac{3}{4}(1 - u^2) I_{|u|\leq 1}$;

3. Gausa: $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$.

Definīcija A2. [16, 684. lpp] Pieņemsim, ka k ir kodols un h - joslas platums, tad par gludināto blīvuma funkciju sauc funkciju

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right).$$

Lai vaiktu hipotēžu pārbaudi par strukturālajiem modeļiem, vai konstruētu ticamības joslas, nepieciešams noteikt asimptotisko sadalījumu kvantiles. Tā kā sadalījumi satur nezināmas funkcijas un arī parametrus, kas atkarīgi no datiem, tad šādā gadījumā parasti lieto butstrapa metodes. Jāpiezīmē, ka atšķirībā no P-P procesa, Q-Q empīriskais process satur savā definīcijā pirmās izlases nezināmo blīvuma funkciju, kas jānovērtē. Šim nolūkam tiks lietotas kodolu gludināšanas metodes.

Definīcija A3. Pieņemsim, ka $K(x) = \int_{-\infty}^{x} k(u)du$,kur $-\infty < x < \infty$ tad par gludināto sadalījuma funkciju sauc funkciju

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Definīcija A4. Pieņemsim, ka $\hat{F}_n(x)$ ir gludinātā sadalījuma funkcija, tad par gludināto kvantiļu funkciju sauc

$$\hat{F}_n^{-1}(t) := \inf\{x : \hat{F}_n(x) \geq t\},$$

kur $t \in (0; 1)$.

Praksē pastāv divas ar šo metodi saistītas problēmas - kodola k izvēle un joslas platuma h izvēle. Interesantākais ir tas, ka tieši h novērtēšana rada lielākās problēmas. Ir dažādas metodes (krosvalidācija, tiešās ievietošanas metode, vienādojumu risināšanas metode u.c.), kas novērtē h. Otra problēma ir kodola izvēle, tomēr precīzi novērtējot joslas platumu, atšķirības ir nelielas. Šī iemesla dēļ modeļu analīzē tika izmantots Gausa kodols.

A2. Izmantoto programmu kods

```
n<-1000
t<-seq(0,1,length=n)

##Brauna tilts
plot(t,rbridge(1,n),"l",ylim=c(-2,2),xlab="",ylab="")
points(t,rbridge(1,n),"l",ylim=c(-2,2),col="blue")
points(t,rbridge(1,n),"l",ylim=c(-2,2),col="green")
Up<-2*sqrt(t*(1-t))
Low<--Up
points(t,Up,"l")
points(t,Low,"l")

##Brauna kustība

plot(t,rwiener(1,n),"l",ylim=c(-2,2),xlab="",ylab="")
points(t,rwiener(1,n),"l",ylim=c(-2,2),col="blue")
points(t,rwiener(1,n),"l",ylim=c(-2,2),col="green")
Up<-2*sqrt(t)
Low<--Up
points(t,Up,"l")
```

```
points(t,Low,"l")

## Kolmogorova-Smirnova statistikas histogramma

n<-1000
t<-seq(-3,3,by=0.01)
rez<-c()
for (i in 1:1000){
x<-rexp(n,0.5)
xx<-ecdf(x)
rez[i]<-max(abs(sqrt(n)*(xx(t)-pexp(t,0.5))))
}

hist(rez,breaks=100,main="",xlab="",ylab="Biežums",cex.lab=1.5)

#Empīriskā procesa konstrukcija

n<-20
x<-runif(n,0,1)
T<-seq(0,1,by=0.001)
xx<-ecdf(x)

plot(T,sqrt(n)*(xx(T)-punif(T,0,1)),cex=0.3,ylab="",xlab="n=20",cex.lab=1.5)

#PP grafiku piemēri
n<-100

x<-rnorm(n,0,1)
y<-rnorm(n,0,4)

xx<-ecdf(x)
t<-seq(0,1,by=0.01)
```

```
res<-xx(quantile(y,probs=t,type=1))
plot(t,res,"s",ylim=c(0,1),xlab="(d)",ylab="",main = expression(paste(mu,
 plain(' = 0'),sigma, plain(' = 4'))))
points(t,pnorm(qnorm(t,0,4),0,1),type="l")
abline(0, 1,lty=3)

#Lokācijas-Skalēšanas modeļa simulācijas

set.seed(10)
n<-50000
N<-10000
S<-rep(0,N)
alpha<-c(0.9,0.95,0.99)
for (i in 1:N){

x<-rnorm(n,0,1)
y<-rnorm(n,5,3)

#x<-runif(n,3,5)
#y<-runif(n,4,9)

#x<-rexp(n,1)
#y<-rexp(n,2)

T<-seq(0,1,by=0.001)
s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
Y<-y*s+m
P<-PP(x,Y,T,plot=TRUE)# PP ir Varbūtību-Varbūtību grafika empīriskā versija
#P<-PP(x,Y,T)
S[i]<-max(abs(P-T))
```

```
}
rez<-sqrt(n)*sort(S)[N*alpha]
rez

#Lēmaņa alternatīvu modeļa simulācijas

set.seed(10)
n<-500
h<-0.5
t<-seq(0,1,by=0.001)
alpha<-c(0.9,0.95,0.99)
N<-10000
stat<-c()
for (i in 1:N){
y<-rnorm(n,0,1)
u<-runif(n,0,1)
x<-qnorm(1-(1-u)^h,0,1)

##Empīriskā versija

xx<-ecdf(x)
rez<-xx(quantile(y,probs=1-(1-t)^h,type=1))
stat[i]<-max(abs(rez-t))
}
(sqrt(n)*sort(stat))[N*alpha]
```

Diplomdarbs “Empīrisko procesu pielietojums strukturālo attiecību modeļos” 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: Juris Cielēns

_____ _____

(paraksts) (datums)

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

_____ _____

(paraksts) (datums)

Recenzents: Jevgenijs Carkovs

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts diplomdarbs gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretāre: Ingrīda Uļjane _____

(paraksts)