LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKIE PROCESI AR PIELIETOJUMIEM STATISTIKĀ

MAĢISTRA DARBS

Autors: **Juris Cielēns**

Stud. apl. jc05001

Darba vadītājs: docents Dr.math. Jānis Valeinis

RĪGA 2012

Anotācija

Darbā aplūkoti galvenie empīrisko procesu un Vapnika-Červonenkis teorijas rezultāti. Tiek aplūkoti metriskās un Vapnika-Červonenkis entropijas pielietojumi kā arī empīrisko procesu pielietojumi dažādās statistikas apakšnozarēs. Detalizētāk tiek aprakstīti pielietojumi statistiskos testos, modeļu novērtējumos, kā piemērs tiek veikta empīriskā procesa butstrapošana lokācijas-skalēšanas modelim ar dažādām butstrapa metodēm.

Atslēgas vārdi: empīriskie procesi, Vapnika-Červonenkis teorija, gludinātais butstraps

Abstract

This thesis contains an overview of the main results of empirical process and Vapnik-chervonenkis theory. Some applications of metric and Vapnin-chervonenkis entropy are shown. This work contains detailed analysis of emprirical process bootstrapping for location-scale model as well as applications to statistical tests and model estimation.

Keywords: empirical processes, Vapnik-Chervonenkis theory, smoothed bootstrap, location-scale model

Saturs

Apzīmējumi

$F(x)$	Teorētiskā sadalījuma funkcija
$F_n(x)$	Empīriskā sadalījuma funkcija
1_a	Indikatorfunkcija - 1_a=1 ja ir spēkā apgalvojums a un ir 0 pretējā gadījumā
$\rightarrow_{gd}$	Gandrīz droša konverģence
$\rightarrow_d$	Konverģence pēc sadalījuma
$U(a,b)$	Vienmērīgi sadalīts gadījuma lielums intervālā [a,b]
$N(\mu,\sigma^2)$	Normāli sadalīts gadījuma lielums ar vidējo vērtību μ un standartnovirzi σ
$l^\infty(T)$	Telpa, kura satur visas ierobežotas funkcijas no T

Ievads

Empīrisko procesu teorija sāka attīstīties 20. gadsimta 30.-40. gados un viens no svarīgākajiem tā laika rezultātiem ir Glivenko-Kantelli teorēma [1], kura publicēta 1933. gadā. Tajā pašā gadā tika definēta Kormogorova-Smirnova statistika. Sākotnēji tās asimptotiskais sadalījums bija pieejams tabulās, bet vēlāk 1948. gadā Doob [2] piedāvāja jaunu pieeju šīs statistikas asimptotiskā sadalījuma iegūšanai un trīs gadus vēlāk Donskers [3] pierādīja, ka asimptotiskais sadalījums ir Brauna tilts. Šis rezultāts mūsdienās pazīstams kā Donskera teorēma. Šajā laikā radās nepieciešamība pēc vispārināta empīriskā procesa definīcijas kā arī tika meklēti nosacījumi, pie kādiem varētu vispārināt Glivenko-Kantelli un Donskera teorēmas. Šo rezultātu vispārināšanai tika izmantota metriskā entropija un 1971. gadā Vapniks un Červonenkis ieviesa savu entropijas definīciju, ar kuras palīdzību bija iespējams novērtēt empīriskā procesa konverģences ātruma augšējo robežu.

Nozīmīgs pavērsiens empīrisko procesu pielietošanā bija iespēja izmantot butstrapa metodi empīrisko procesu statistikām. Vairāk par šo tēmu var uzzināt publikācijā [4]. 1990. gadā Beirlants un Deheuvels [5] aplūkoja empīrisko procesu varbūtību-varbūtību un kvantiļu-kvantiļu grafikiem tādējādi aizsākot empīrisko procesu pētīšanu divu izlašu gadījumam. Šajā laikā daudzu problēmu risināšanā tiek mēģināts pielietot empīrisko procesu teoriju un 1992. gadā Velners [6] publicē nelielu apkopojumu par svarīgākajiem pielietojumiem. 2008. gadā Horvats [7] publicē empīrisko procesu ROC līknēm, vienlaicīgi piedāvājot gludinātā butstrapa metodi ticamības joslu konstruēšanai.

Šo statistikas virzienu - empīriskos procesus - pētu jau vairākus gadus un svarīgs rezultāts ir publikācija [8], kurā tiek salīdzinātas empīrisko procesu un empīriskās ticamības funkcijas metodes divu izlašu gadījumā. Laikam ejot pētāmās problēmas kļuva sarežģītākas un pētījuma rezultāts ir šis maģistra darbs. Maģistra darba mērķis ir iepazīties ar vispārināto empīrisko procesu teoriju un tās pielietojumiem statistikā. Lai sasniegtu mērķi tika izvirzīti sekojoši uzdevumi

- iepazīties ar vispārināto empīriskā procesa definīciju un kritērijiem, pēc kuriem funkcijas klasificē P-Glivenko-Kantelli un P-Donskera klasēs;
- iepazīties ar Vapnika-Červonenkis teoriju un tās nozīmi empīrisko procesu pielietošanā;
- aplūkot empīrisko procesu pietietojumus dažādās statistika apakšnozarēs;

- veikt butrapa metožu analīzi lokācijas-skalēšanas empīriskajam procesam.

Maģistra darbs sastāv no divām daļām un pielikuma. Pirmajā daļā apkopotas empīriskā procesa definīcijas un svarīgākie rezultāti kā arī sniegts ieskats Vapnika-Červonenkis teorijā, kuru bieži izmanto empīrisko procesu konverģences ātruma mērīšanai. Otrā daļa satur empīrisko procesu pielietojumu piemērus, kuri sadalīti trīs daļās - empīriko procesu pielietojumi testos, empīrisko procesu butstraps un empīrisko procesu pielietojums modeļu novērtēšanā. Pielikums satur darbā izmantoto attēlu un tabulu iegūšanai izmantotos algoritmus valodā R.

1. Empīriskie procesi un to asimptotika

1.1. Empīriskais process

Šajā nodaļā tiks aprakstīti svarīgākie rezultāti un definīcijas saistībā ar empīrisko procesu teoriju. Empīrisko procesu teorija sākās ar izlases sadalījuma funkcijas novērtēšanu un šī novērtējuma pētīšanu.
Pieņemsim, ka $X_1, X_2, ..., X_n$ ir neatkarīgu un vienādi sadalītu gadījuma izlase ar sadalījuma funkciju F. Empīriskā sadalījuma funkcija tiek definēta sekojoši:

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

Kā jau tika minēts ievadā, pirmais svarīgākais rezultāts ir Glivenko-Kantelli teorēma.

Teorēma 1.1. [9, 1. lpp] (Glivenko-Kantelli) Ir spēkā

$$\sup_{-\infty<x<\infty} (|F_n(x) - F(x)|) \to_{gd} 0. \tag{1.1}$$

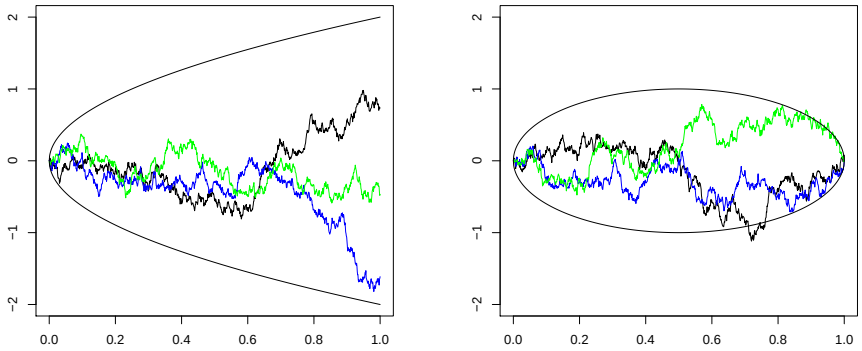
Kā vēlāk izrādījās, ļoti svarīga nozīme ir statistikai, kuru iegūst normējot Glivenko-Kantelli teorēmā minēto ar $\sqrt{n}$. Pēc centrālās robežteorēmas ir spēkā

$$\sqrt{n}(F_n(x) - F(x)) \to_d N(0, F(x)(1 - F(x))). \tag{1.2}$$

Teorēma 1.2. [10, 1901. lpp] (Donskera) Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju F, tad

$$\sup_{-\infty<x<\infty} |\sqrt{n}(F_n(x) - F(x))| \to_d \sup_{-\infty<x<\infty} |B(F(x))|, \tag{1.3}$$

kur B apzīmē Brauna tiltu.



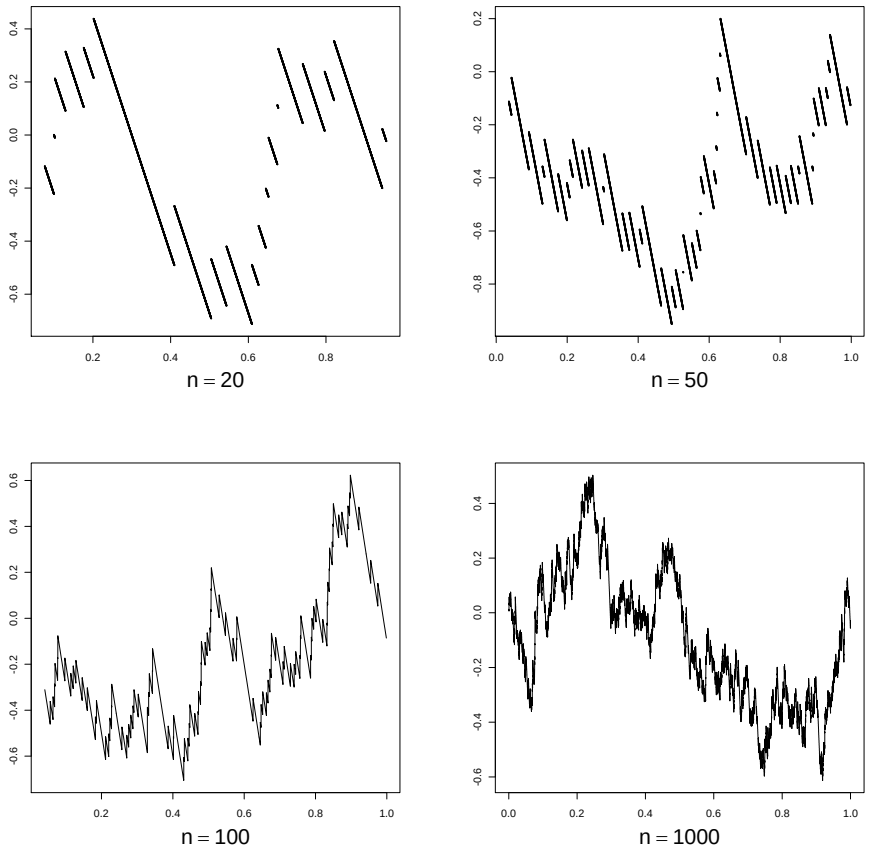
1.1. att.: Brauna kustības (pa kreisi) un Brauna tilta realizācijas piemēri ar 95% punktveida ticamības joslu.

Kreisās puses izteiksmi sauc par Kolmogorova-Smirnova statistiku. Sākotnēji šīs statistikas asimptotiskais sadalījums tika publicēts tabulās un tikai vēlāk Donskers [3] pierādīja, ka tas robežsadalījums ir suprēmums pa moduli no Brauna tilta.

Definīcija 1.1. [11] Par standarta Brauna kustību jeb Vīnera procesu sauc stohastisku procesu $\{W(t), t \geq 0\}$, kuram

- ir stacionāri un neatkarīgi ieaugumi, t.i. $W(t) - W(s), t > s$ sadalīti pēc likuma N(0,t-s) un katram $0 \leq t_1 \leq t_2 \leq ... \leq t_n$ pieaugumi $W(t_2) - W(t_1)$,...,$W(t_n) - W(t_{n-1})$ ir neatkarīgi;
- funkcija $t \to W(t)$ ir gandrīz droši nepārtraukta;
- $cov(W(t), W(s)) = s$, ja $t > s$.

Definīcija 1.2. [11] 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. Svarīga Brauna tilta īpašība ir tā, ka šis process sākuma un beigu punktos pieņem vērtību 0, t.i. $B(0) = B(1) = 0$. Vairāk par Brauna kustību kā stohastisku procesu var uzzināt, piemēram, [11]. Attēlā 1.1. redzami daži Brauna kustības un Brauna tiltu realizācijas piemēri, kā arī šo realizāciju vērību 95% punktveida ticamības josla . Brauna kustības gadījumā ticamības joslu konstruēšanai var izmantot faktu, ka procesa vērtība katrā tā punktā ir gadījuma lielums, kas sadalīts pēc likuma $N(0,t)$. Brauna tilta gadījumā ticamības josla noteikta veicot simulācijas.



1.2. att.: Empīriskā procesa $\sqrt{n}(F_n(x) - F(x))$ konverģence uz Brauna tiltu. Simulētas izlases ar apjomu n un sadalījuma funkciju $U(0,1)$.

Statistikā izteiksme $\sqrt{n}(F_n(x) - F(x))$ tiek saukta par empīrisko procesu un visi rezultāti un pētījumi, kas saistīti ar šo izteiksmi pieder pie empīrisko procesu teorijas. Analītisks pierādījums empīriskā procesa konverģencei uz Brauna tiltu ir ļoti sarežģīts un netiks šajā darbā aprakstīts, tomēr šo rezultātu var arī novērot konstruējot empīrisko procesu simulētām gadījuma izlasēm, pakāpeniski palielinot izlases apjomu. Attēlā 1.2. var aplūkot, kā izskatās empīriskā procesa grafiks, ja tiek simulēta gadījuma izlase ar sadalījumu U(0,1). Minētajā attēlā redzami grafiki pie izlases apjoma $n = 20, 50, 100$ un1000.

Līdzīgi rezultāti ir spēkā izlases kvantiļu funkcijai. Kvantiļu funkcijas empīriskais novērtējums tiek definēts sekojoši:

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

Un līdzīgi kā sadalījuma funkcijai, tai ir spēkā [12]

$$\sup_{0<t<1}(|F_n^{-1}(t) - F^{-1}(t)|) \to_{gd} 0, \text{ kad } n \to \infty. \tag{1.4}$$

Kvantiļu funkcijas empīriskā procesa asimptotiskais sadalījums ir sarežģītāks nekā sadalījuma funkcijai, tomēr, nedaudz modificējot pašu kvantiļu procesu, iespējams asimptotisko sadalījumu vienkāršot.

Teorēma 1.3. [12, 31. lpp] Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar kvantiļu funkciju F^{-1} un blīvuma funkciju f, tad

$$\sup_{0<t<1} |\sqrt{n} f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))| \to_d \sup_{0<t<1} |B(t)|. \tag{1.5}$$

Empīrisko procesu, kuru veido no izlases kvantiļu funkcijas, vairākās publikācijās ir aprakstījis Csorgo [12]. Tā kā empīriskā procesa pielietojums kvantiļu funkcijām ir gandrīz tik pat plaši pētīts kā sadalījuma funkcijas gadījums, izteiksme $\sqrt{n}(F_n^{-1}(t) - F^{-1}(t))$ tiek saukta arī par empīrisko kvantiļu procesu. Kvantiļu funkcijas gadījumā iespējams uzzīmēt līdzīgus grafikus kā sadalījuma funkcijai, kuri parāda konverģenci uz Brauna tiltu. Attēlā 1.3. redzams asimptotiskā sadalījuma salīdzinājums izteiksmēm $\sqrt{n}(F_n^{-1}(t) - F^{-1}(t))$ un $\sqrt{n} f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$. To, ka kreisajā pusē redzamie grafiki nekonverģē uz Brauna tiltu, var viegli pateikt pēc vērtībām grafiku gala punktos.

Aplūkosim vispārināta empīriskā procesa 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 1.3. [13, 269. lpp] Par empīrisko varbūtību mēru sauc

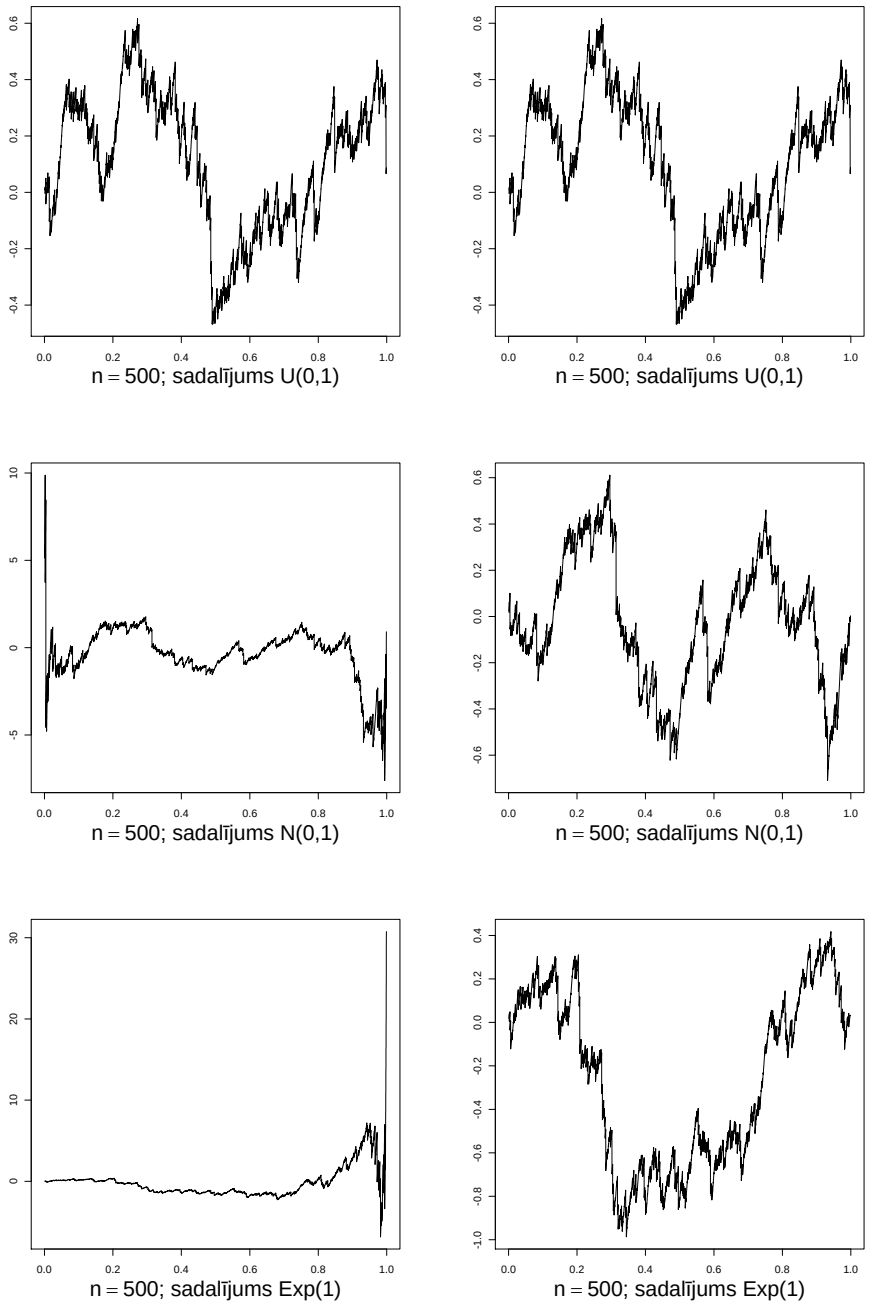
$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

kur δ_x ir varbūtību sadalījums, kurš deģenerēts punktā x.

Definīcija 1.4. [13, 269. lpp] Pieņemsim, ka $g : \mathcal{X} \to \mathbb{R}$ ir mērojama funkcija. Funkcijas g sagaidāmā vērtība pie empīriskā mēra tiek apzīmēta ar $\mathbb{P}_n g$, pie mēra P - ar Pg, un tiek definēta:

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

No lielo skaitļu likuma seko, ka $\mathbb{P}_n g$ konverģē gandrīz droši uz Pg, katram g, kuram Pg ir definēts. Saistībā ar parasto empīrisko procesu tika minēti divi svarīgi rezultāti - Glivenko-Kantelli Teorēma 1.1 un Donskera Teorēma 1.3. Vispārinātā gadījumā funkciju klases var sagrupēt aktarībā no tā, kura teorēma ir spēkā.



1.3. att.: Empīriskā kvantiļu procesa konverģence uz Brauna tiltu. Kreisajā pusē redzami grafiki izteiksmei $\sqrt{n}(F_n^{-1}(t) - F^{-1}(t))$, labajā pusē izteiksmei $\sqrt{n}f(F^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$. Procesu realizācijas simulētas ņemot izlases apjomu $n = 500$.

Definīcija 1.5. [13, 269. lpp] Mērojamu funkciju $g : \mathcal{X} \to \mathbb{R}$ klase $\mathcal{A}$ tiek saukta par P-Glivenko-Kantelli klasi, ja

$$||\mathbb{P}_n g - Pg||_{\mathcal{A}} = \sup_{g \in \mathcal{A}} |\mathbb{P}_n g - Pg| \to_{gd} 0. \tag{1.6}$$

Bieži publikācijās lieto arī pierakstu

$$\sup_{g \in \mathcal{A}} \left| \int g d(P_n - P) \right| \to_{gd} 0.$$

Definīcija 1.6. [13, 269. lpp] Par empīrisko procesu funkcijai g sauc

$$\mathbb{G}_n g = \sqrt{n}(\mathbb{P}_n g - Pg). \tag{1.7}$$

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

$$(\mathbb{G}_n g_1, \mathbb{G}_n g_2, ..., \mathbb{G}_n g_k) \to_d (\mathbb{G}_p g_1, \mathbb{G}_p g_2, ..., \mathbb{G}_p g_k), \tag{1.8}$$

kur labās puses vektoram ir daudzdimensionāls normālais sadalījums ar kovariācijām.

$$\mathbb{E}\mathbb{G}_p f \mathbb{G}_p g = Pfg - PfPg. \tag{1.9}$$

Definīcija 1.7. [13, 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 (1.9) un tiek saukts par P-Brauna tiltu. Ja $\mathcal{A}$ satur indikatorfunkcijas, tad Teorēma 1.1 ir speciālgadījums P-Glivenko-Kantelli klasei un Teorēma 1.2 ir speciālgadījums P-Donskera klasei.

1.2. Metriskā entropija

Iepriekšējā nodaļā tika definētas P-Glivenko-Kantelli un P-Donskera klases. Šajā nodaļā tiks sniegti nepieciešamie nosacījumi, lai funkcijas piederētu kādai no minētajām klasēm. Lai formulētu nepieciešamos nosacījumus, tiek izmantota tā sauktā metriskā entropija. Pietiekoši plaša informācija par metrisko entropiju pieejama Van Der Waart [13] un van de Geer [14] grāmatās, kuras tika izmantotas apkopojot tālāk uzskaitītos rezultātus.

Pieņemsim, ka Q ir mērs telpā $(\mathcal{X}, \mathcal{A})$, $\mathcal{G}$ ir funkciju klase uz $\mathcal{X}$ un $L_p(P) = \{g : \mathcal{X} \to \mathbf{R} : \int |g|^p dP < \infty\}, 1 \leq p < \infty$. Apzīmēsim

$$||g||_{p,Q}^p = \int |g|^p dQ.$$

Definīcija 1.8. [14, 16. lpp] Katram $\delta > 0$ aplūkojam tādu funkciju kopumu $g_1, ..., g_N$, ka katram $g \in \mathcal{G}$ eksistē $j = j(g) \in \{1, ..., N\}$, ka

$$||g - g_j||_{p,Q} \leq \delta.$$

Ar $N_p(\delta, \mathcal{G}, Q)$ apzīmē mazāko N vērtību, kurai eksistē šāds pārklājums ar lodēm ar rādiusu δ un centriem $g_1, g, ..., g_N$ ($N_p(\delta, \mathcal{G}, Q) = \infty$, ja šāds pārklājums neeksistē). Tad

$N_p(\delta, \mathcal{G}, Q)$ tiek saukts par δ-pārklājuma skaitli un $H_p(\delta, \mathcal{G}, Q) = \log N_p(\delta, \mathcal{G}, Q)$ sauc par δ-entropiju no $\mathcal{G}$ $L_p(Q)$ metrikai. Klasi $\mathcal{G}$ sauc par pilnīgi pierobežotu, ja $H_p(\delta, \mathcal{G}, Q) < \infty$ katram $\delta > 0$.

Definīcija 1.9. [14, 16. lpp] Ar $N_{p,B}(\delta, \mathcal{G}, Q)$ apzīmē mazāko N vērtību, kurai eksistē funkciju pāri $\left\{[g_j^L, g_j^U]\right\}_{j=1}^N$ tādi, ka $||g_j^U - g_j^L||_{p,Q} \leq \delta$ katram $j = 1, .., N$ un katram $g \in \mathcal{G}$ eksistē $j = j(g) \in \{1, ..., N\}$ tādi, ka

$$g_j^L \leq g \leq g_j^U.$$

$N_{p,B}(\delta, \mathcal{G}, Q) = \infty$, ja neeksistē galīga šādu pāru (iekavu) kopa. Tad $H_{p,B}(\delta, \mathcal{G}, Q) = \log N_{p,B}(\delta, \mathcal{G}, Q)$ sauc par δ-entropiju ar iekavām no $\mathcal{G}$.

Definīcija 1.10. [14, 17. lpp] Ar $N_\infty(\delta, \mathcal{G})$ apzīmē mazāko N vērtību, kurai eksistē $\{g_j\}_{j=1}^N$ tādi, ka

$$\sup_{g \in \mathcal{G}} \min_{j=1,...,N} |g - g_j|_\infty \leq \delta.$$

$N_\infty(\delta, \mathcal{G}) = \infty$, ja neeksistē galīga kopa ar šādu īpašību. Tad $H_\infty(\delta, \mathcal{G}) = \log N_\infty(\delta, \mathcal{G})$ sauc par δ-entropiju no $\mathcal{G}$ suprēma normai. Ievērosim, ka šajā definīcijā suprēms nav atkarīgs no mēra Q.

Šīs entropiju definīcijas saista sekojošs rezultāts:

Teorēma 1.4. [14, 17. lpp] Katram $1 \leq p < \infty$

$$H_p(\delta, \mathcal{G}, Q) \leq H_{p,B}(\delta, \mathcal{G}, Q), \text{visiem } \delta > 0.$$

Ja Q ir varbūtību mērs, tad

$$H_{p,B}(\delta, \mathcal{G}, Q) \leq H_\infty\left(\frac{\delta}{2}, \mathcal{G}\right), \text{visiem } \delta > 0.$$

Teorēma 1.5. [14, 23. lpp] Pieņemsim, ka

$$H_{1,B}(\delta, \mathcal{A}, P) < \infty \text{ katram } \delta > 0,$$

tad $\mathcal{G}$ ir P-Glivenko-Kantelli klase.

Pierādījums. Izvēlamies patvaļīgu $\delta > 0$ un pieņemsim, ka $\left\{[g_j^L, g_j^U]\right\}$ ir δ-iekavu kopa no $\mathcal{G}$, t.i. $||g_j^U - g_j^L||_{1,P} \leq \delta, j = 1, ..., N$ un katrs $g \in \mathcal{G}$ atrodas starp kādu no pāriem $[g_j^L, g_j^U]$.

Tad

$$\int gd(P_n - P) = \int gdP_n - \int gdP \tag{1.10}$$

$$\leq \int g_j^U dP_n - \int gdP = \int g_j^U d(P_n - P) + \int (g_j^U - g)dP \tag{1.11}$$

$$\leq \int g_j^U d(P_n - P) + \delta.$$

Līdzīgi

$$\int gd(P_n - P) \geq \int g_j^L d(P_n - P) - \int (g - g_j^L)dP \geq \int g_j^L d(P_n - P) - \delta.$$

Tā ka $\left\{[g_j^L, g_j^U]\right\}_{j=1}^N$ ir galīga, rezultātā iegūst, ka

$$\max_{j=1,\ldots,N} \left| \int g_j^U d(P_n - P) \right| \leq \delta \ g.d.$$

kā arī

$$\max_{j=1,\ldots,N} \left| \int g_j^L d(P_n - P) \right| \leq \delta \ g.d.$$

Tātad rezultātā

$$\sup_{g \in \mathcal{G}} \left| \int gd(P_n - P) \right| \leq 2\delta \ g.d.$$

□

Teorēma 1.6. [14, 90. lpp] Pieņemsim, ka

$$\int_0^1 H_{2,B}^{1/2}(u, \mathcal{G}, P)du < \infty,$$

tad $\mathcal{G}$ ir P-Donskera klase. Nodaļas nobeigumā apskatīsim dažus piemērus no [13].

Piemērs 1.1. (Sadalījuma funkcija). Pieņemsim, ka funkciju klase $\mathcal{G}$ 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 $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ 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 $N_{2,B}(\sqrt{\epsilon}, \mathcal{G}, L_2(F)) \leq 2/\epsilon$. Līdz ar to šī funkciju klase apmierina P-Donskera klases nosacījumus.

Piemērs 1.2. (Parametriska klase). Pieņemsim, ka $\mathcal{G} = \{f_\theta : \theta \in \Theta\}$ 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 [13, 271. lpp], ka eksistē konstante K, atkarīga tikai no Θ un d tāda, ka iekavu numuri apmerina nevienādību

$$N_{2,B}(\epsilon||m||_{P,r}, \mathcal{G}, L_r(F)) \leq K\left(\frac{diam\Theta}{\epsilon}\right)^d, \text{ katram } 0 < \epsilon < diam\Theta.$$

Līdz ar to entropija ir ar kārtu mazāku par $(1/\epsilon)$ un iekavu entropijas integrālis konverģē. Seko, ka klase $\mathcal{G}$ ir P-Donskera.

Piemērs 1.3. (Gludas funkcijas). Pieņemsim, ka $\mathbb{R}^d = \cup_j I_j$ ir kubu ar izmēru 1 daļa un $\mathcal{G}$ 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 var pierādīt, ka $V \geq d/\alpha$ un katram varbūtību mēram P klases $\mathcal{G}$ iekavu entropija apmierina evienādību

$$H_{2,B}(\epsilon, \mathcal{G}, L_r(P)) \leq K\left(\frac{1}{\epsilon}\right)^V \left(\sum_{j=1}^{\infty}(M_j^r P(I_j))^{\frac{V}{V+r}}\right)^{\frac{V+r}{r}}.$$

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

Piemērs 1.4. (Soboļeva klases). Pieņemsim, ka funkciju klase $\mathcal{G}$ 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 $\epsilon > 0$

$$H_{[2,B]}(\epsilon, \mathcal{G}, ||.||_\infty) \leq K\left(\frac{1}{\epsilon}\right)^{1/k}.$$

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

Piemērs 1.5. (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

$$H_{[2,B]}(\epsilon, \mathcal{G}, L_2(P)) \leq K\left(\frac{1}{\epsilon}\right).$$

Šī klase ir P-Donskera katram P.

1.3. Vapnika-Červonenkis entropija

Novērtējot nezināmu procesu ir svarīgi gan tas, cik precīzi novērtētais process raksturo patieso procesu, gan arī tā konverģences ātrums. Konverģences ātrums mēdz būt svarīgs, ja nav pieejami daudz datu vai ja process ir tik komplicēts, ka liels datu apjoms būtiski ietekmē tā novērtēšanas laiku. Praksē pastāv iespēja novērtēt empīriskā procesa konverģences ātrumu, izmantojot Vapnika-Červonenkis entropiju. Šajā nodaļā tiks aprakstītas Vapnika-Červonenkis entropijas definīcijas un galvenie rezultāti empīriskā procesa konverģences ātruma augšējās robežas novērtēšanai.

Pieņemsim, ka Λ ir parametru kopa un ka $Q(z,\alpha), \alpha \in \Lambda$ ir indikatora funkciju kopa. Aplūkojam izlasi $Z = z_1, z_2, ..., z_n$. Ar lielumu $N^\Lambda(z_1, z_2, ..., z_n)$, kurš parāda, cik dažādos veidos var sadalīt izlasi Z ar funkcijām no indikatora funkciju klases, raksturosim kopas $Q(z,\alpha)$ dažādību. Formāls pieraksts būtu sekojošs. Aplūkojam n-dimensionālu vektoru

$$q(\alpha) = (Q(z_1,\alpha), Q(z_2,\alpha), ..., Q(z_n,\alpha)), \alpha \in \Lambda,$$

kuru iegūst ņemot dažādas vērtības no Λ. Tad $N^\Lambda(z_1, z_2, ..., z_n)$ ir n-dimensionāla kuba, kurš iegūts no izlases $z_1, z_2, ..., z_n$ un kopas $Q(z,\alpha)$, dažādu virsotņu skaits.

Definīcija 1.11. [15, 42. lpp] Vērtību

$$H^\Lambda(z_1, z_2, ..., z_n) = log N^\Lambda(z_1, z_2, ..., z_n) \tag{1.12}$$

sauc par gadījuma entropiju. Gadījuma entropija raksturo funkciju kopas dažādību uz dotajiem datiem. $H^\Lambda(z_1, z_2, ..., z_n)$ ir gadījuma lielums, jo tas atkarīgs no gadījuma izlases.

Definīcija 1.12. [15, 42. lpp] Vērtību

$$H^\Lambda(n) = \mathbb{E}(H^\Lambda(z_1, z_2, ..., z_n)) \tag{1.13}$$

sauc par indikatora funkciju klases $Q(z,\alpha)$ entropiju pie izlases apjoma n. Šī vērtība ir atkarīga no funkciju klases $Q(z,\alpha)$, varbūtību mēra un izlases apjoma n un raksturo sagaidāmo dažādību pie minētajiem lielumiem.

Virpārināsim entropijas definīciju uz kopu, kura satur ierobežotas funkcijas. Pieņemsim, ka $A \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$ ie ierobežotu funkciju kopa. Izmantojot šo funkciju kopu un izlasi $z_1, z_2, ..., z_n$ var konstruēt sekojošo n-dimensionālo vektoru kopu:

$$q(\alpha) = (Q(z_1,\alpha), Q(z_2,\alpha), ..., Q(z_n,\alpha))$$

šī vektoru kopa pieder n-dimensionālam kubam un satur galīgu minimālo ϵ-tīklu metrikā C (vai L_p). Ar $N = N^{\Lambda}(\epsilon, z_1, z_2, ..., z_n)$ apzīmējam $q(\alpha)$ minimālā ϵ-tīkla elementu skaitu. N arī šoreiz ir gadījuma lielums.

Definīcija 1.13. [15, 44. lpp] Vērtību

$$H^{\Lambda}(\epsilon, z_1, z_2, ..., z_n) = log N^{\Lambda}(\epsilon, z_1, z_2, ..., z_n) \tag{1.14}$$

sauc par gadījuma Vapnika-Červonenkis entropiju jeb saīsināti par VC entropiju funkciju kopai $A \leq Q(z, \alpha) \leq B$.

Definīcija 1.14. [15, 44. lpp] Gadījuma entropijas sagaidāmo vērtību

$$H^{\Lambda}(\epsilon, n) = \mathbb{E}(H^{\Lambda}(\epsilon, z_1, z_2, ..., z_n)) \tag{1.15}$$

sauc par funkciju kopas $A \leq Q(z, \alpha) \leq B$ VC entropiju pie izlases apjoma n. Šāda entropijas definīcija reālām funkcijām ir vispārinājums indikator funkciju entropijai. Indikator funkciju kopai minimālais ϵ-tīkls pie $\epsilon < 1$ nav atkarīgs no ϵ un ir n-dimensiju vienības kuba virsotņu apakškopa. Tādējādi pie $\epsilon < 1$ ir spēkā sekojošas sakarības:

$$N^{\Lambda}(\epsilon, z_1, z_2, ..., z_n) = N^{\Lambda}(z_1, z_2, ..., z_n),$$

$$H^{\Lambda}(\epsilon, z_1, z_2, ..., z_n) = H^{\Lambda}(z_1, z_2, ..., z_n),$$

$$H^{\Lambda}(\epsilon, n) = H^{\Lambda}(n).$$

Papildus jau minētajai VC entropijai, VC teorijā apskatīta arī nedaudz savādāk definēta entropija.

Definīcija 1.15. [15, 55. lpp] Par rūdīto (no angļu valodas - *annealed*) VC entropiju sauc vērtību

$$H^{\Lambda}_{ann}(n) = log \mathbb{E}(N^{\Lambda}(z_1, z_2, ..., z_n)). \tag{1.16}$$

Definīcija 1.16. [15, 55. lpp] Par pieauguma funkciju sauc

$$G^{\Lambda}(n) = ln \sup_{z_1, z_2, ..., z_n} (N^{\Lambda}(z_1, z_2, ..., z_n)). \tag{1.17}$$

Teorēma 1.7. [15, 55. lpp] Katram n ir spēkā nevienādības

$$H^{\Lambda}(n) \leq H^{\Lambda}_{ann}(n) \leq G^{\Lambda}(n). \tag{1.18}$$

Tālāk tiks minēti svarīgākie rezultāti empīriskā procesa konverģences augšējās robežas noteikšanai.

Teorēma 1.8. [15, 70. lpp] Pieņemsim, ka $Q(z,\alpha)$, $\alpha \in \Lambda$ ir indikator funkciju klase un $H^{\Lambda}_{ann}(n)$ tās atbilstošā rūdītā VC entropija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \epsilon\right\} \leq 4e^{\left(\frac{H^{\Lambda}_{ann}(2n)}{n} - \epsilon^2\right)n}. \tag{1.19}$$

Teorēma 1.9. [15, 70. lpp] Pieņemsim, ka $Q(z,\alpha)$, $\alpha \in \Lambda$ ir indikator funkciju klase un $H^{\Lambda}_{ann}(n)$ tās atbilstošā rūdītā VC entropija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \epsilon\right\} \leq 4e^{\left(\frac{H^{\Lambda}_{ann}(2n)}{n} - \frac{\epsilon^2}{4}\right)n}. \tag{1.20}$$

Šīs robežas ir netriviālas (t.i. katram $\epsilon > 0$ nevienādības labā puse tiecas uz 0, ja n tiecas uz bezgalību) ja

$$\lim_{n\to\infty}\frac{H^{\Lambda}_{ann}(n)}{n} = 0.$$

Tā kā tehniski ir ļoti sarežģīti novērtēt rūdīto VC entropiju un līdz ar to arī visu nevienādības labās puses izteiksmi, tad var pielietot nevienādību starp rūdīto VC entropiju un augšanas funkciju. Rezultātā iegūst sekojošo.

Teorēma 1.10. [15, 72. lpp] Pieņemsim, ka $Q(z,\alpha)$, $\alpha \in \Lambda$ ir indikator funkciju klase un $G^{\Lambda}(n)$ tās atbilstošā augšanas funkcija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \epsilon\right\} \leq 4e^{\left(\frac{G^{\Lambda}(2n)}{n} - \epsilon^2\right)n}. \tag{1.21}$$

Teorēma 1.11. [15, 72. lpp] Pieņemsim, ka $Q(z,\alpha)$, $\alpha \in \Lambda$ ir indikator funkciju klase un $G^{\Lambda}(n)$ tās atbilstošā augšanas funkcija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \epsilon\right\} \leq 4e^{\left(\frac{G^{\Lambda}(2n)}{n} - \frac{\epsilon^2}{4}\right)n}. \tag{1.22}$$

Šīs robežas ir netriviālas, ja

$$\lim_{n\to\infty}\frac{G^{\Lambda}(n)}{n} = 0.$$

Jāatzīmē, ka ja šis nosacījums nav spēkā, tad eksistē varbūtību mēri $F(z)$, kuriem vienādība

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \epsilon\right\} = 0$$

nav spēkā.

Lai aptvertu ne tikai indikator funkciju klases, bet arī reālu funkciju klases, nepieciešams vispārināt minētos rezultātus. Lai pārnestu rezultātus uz reālu funkciju klasēm, Vapnik [15] iesaka sekojošu pieeju.

Pieņemsim, ka $Q(z,\alpha), \alpha \in \Lambda$ ir reālu funkciju kopa, kurai ir spēkā

$$A = \inf_{\alpha,z} Q(z,\alpha) \leq Q(z,\alpha) \leq \sup_{\alpha,z} Q(z,\alpha) = B,$$

kur A un B var būt bezgalība. Vaļēju intervālu (A,B) apzīmēsim ar $\mathcal{B}$. No reālo funkciju kopas $Q(z,\alpha)$ konstruē indikatoru kopu:

$$I(z,\alpha,\beta) = 1_{Q(z,\alpha)\geq\beta}, \alpha \in \Lambda, \beta \in \mathcal{B},$$

Dotai funkcijai $Q(z,\alpha^*)$ un dotai vērtībai β^*, indikators $I(z,\alpha^*,\beta^*)$ pieņem vērtību 1 tājā apgabalā, kur $Q(z,\alpha^*) \geq \beta^*$ un 0 tur, kur $Q(z,\alpha^*) < \beta^*$.

Ja $Q(z,\alpha)$ ir indikator funkciju kopa, tad tā sakrīt ar indikatoru kopu $I(z,\alpha,\beta), \beta \in (0,1)$. Balstoties uz šo vispārinājumu, tiek apskatīti trīs atsevišķi reālu funkciju klašu gadījumi.

Teorēma 1.12. [15, 74. lpp] Pieņemsim, ka $A \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$ ir reālu, ierobežotu funkciju klase un $H_{ann}^{\Lambda,B}(n)$ tās atbilstošā rūdītā entropija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)\right| > \epsilon\right\} \leq 4e^{\left(\frac{H_{ann}^{\Lambda,B}(2n)}{n} - \frac{\epsilon^2}{(B-A)^2}\right)n}. \tag{1.23}$$

Teorēma 1.13. [15, 74. lpp] Pieņemsim, ka $0 \leq Q(z,\alpha) \leq B, \alpha \in \Lambda$ ir reālu, ierobežotu, nenegatīvu funkciju klase un $H_{ann}^{\Lambda,B}(n)$ tās atbilstošā rūdītā VC entropija. Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda} \frac{\int Q(z,\alpha)dF(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}} > \epsilon\right\} \leq 4e^{\left(\frac{H_{ann}^{\Lambda,B}(2n)}{n} - \frac{\epsilon^2}{4B}\right)n}. \tag{1.24}$$

Ievērosim, ka šie gadījumi ir tiešs vispārinājums no rezultātiem indikator funkciju klasēm. Trešais gadījums ir nedaudz sarežģītāks.

Teorēma 1.14. [15, 74. lpp] Pieņemsim, ka $0 \leq Q(z,\alpha), \alpha \in \Lambda$ ir reālu, nenegatīvu funkciju klase un $H_{ann}^{\Lambda,B}(n)$ tās atbilstošā rūdītā VC entropija un eksistē $p > 2$ tāds ka gadījuma lieluma $\xi_\alpha = Q(z,\alpha)$ p-tais normalizētais moments eksistē

$$m_p(\alpha) = \sqrt[p]{\int Q^p(z,\alpha)dF(z)}.$$

Tad ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)dF(z)}}>a(p)\epsilon\right\}\leq 4e^{\left(\frac{H_{ann}^{\Lambda,B}(2n)}{n}-\frac{\epsilon^2}{4}\right)n}, \tag{1.25}$$

kur

$$a(p)=\sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}} \tag{1.26}$$

Visbeidzot pielietojot šiem rezultātiem sakarību starp rūdīto VC entropiju un pieauguma funkciju iegūst, ka reālu, ierobežotu funkciju klasei $A\leq Q(z,\alpha)\leq B$ ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}|\int Q(z,\alpha)dF(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)|>\epsilon\right\}\leq 4e^{\left(\frac{G^{\Lambda,B}(2n)}{n}-\frac{\epsilon^2}{(B-A)^2}\right)n}, \tag{1.27}$$

reālu, ierobežotu, nenegatīvu funkciju klasei $0\leq Q(z,\alpha)\leq B$ ir spēkā nevienādība

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)dF(z)}}>\epsilon\right\}\leq 4e^{\left(\frac{G^{\Lambda,B}(2n)}{n}-\frac{\epsilon^2}{4B}\right)n} \tag{1.28}$$

un reālu, nenegatīvu funkciju klasei $0\leq Q(z,\alpha)$, kurai eksistē p-tais nomalizētais moments un $p>2$ ir spēkā

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)dF(z)-\frac{1}{n}\sum_{i=1}^{n}Q(z_i,\alpha)}{\sqrt[p]{\int Q^p(z,\alpha)dF(z)}}>a(p)\epsilon\right\}\leq 4e^{\left(\frac{G^{\Lambda,B}(2n)}{n}-\frac{\epsilon^2}{4}\right)n}, \tag{1.29}$$

kur $G^{\Lambda,B}(n)$ ir atbilstošās klases agušanas funkcija.

1.4. Vapnika-Červonenkis dimensija

Pēc pēdējo gadu notikumiem ekonomikā zinātnieki aktīvāk sākuši meklēt iespējas precīzāk prognozēt ekonomisko procesu laikrindas. Papildus centieniem veikt prognozes pēc iespējas precīzāk, pēdējos gados kā svarīgs kritērijs veiksmīgai darbībai tiek virzīts modeļa sarežģītības novērtējums. Tas nozīmē, ka pat ja modelis strādā ļoti precīzi, tas nav derīgs ja ir pārāk sarežģīts un prognozēšana prasa tik ilgu laiku, ka notikums jau ir bijis, bet modeļa novērtēšana vēl tikai notiek. McDonald, Shalizi un Shervish [16] piedāvā jaunu metodi, kurā tiek mēģināts atrast efektīvāko modeli gan precizitātes, gan sarežģītības ziņā vienlaicīgi. Šajā gadījumā modeļa sarežģītības noteikšanai tiek izmantota Vapnika-Červonenkis dimensija, kā arī tiek izmantoti iepriekšējā nodaļā apskatītie rezultāti modeļa konverģences ātruma noteikšanai.

Lai novērtētu iepriekšējā nodaļā minētās nevienādības, jāprot novērtēt vai nu rūdītā VC entropija, vai funkciju klases Q augšanas funkcija. Tomēr pastāv iespēja konverģences ātrumu aproksimēt ar Vapnika-Červonenkis dimensiju. Šajā nodaļā tiks sniegta šī jēdziena definīcija, daži rezultāti par to, kā tā saistās ar funkciju klases augšanas funkciju $G(n)$, kā arī algoritms Vapnika-Červonenkis dimensijas (turpmāk tekstā VC dimensija) analītiskai novērtēšanai.

Iepriekšējā nodaļā minētajai augšanas funkcijai iespējams noteikt augšējo robežu.

Teorēma 1.15. [15, 79. lpp] Jebkurai augšanas funkcijai ir spēkā vienādība

$$G^{\Lambda}(n) = n\ ln\ 2$$

vai arī tā apmierina nevienādību

$$G^{\Lambda}(n) \leq h\left(ln\ \frac{n}{h} + 1\right),$$

kur h ir tāds naturāls skaitlis, ka pie $n = h$

$$G^{\Lambda}(h) = h\ ln\ 2,$$

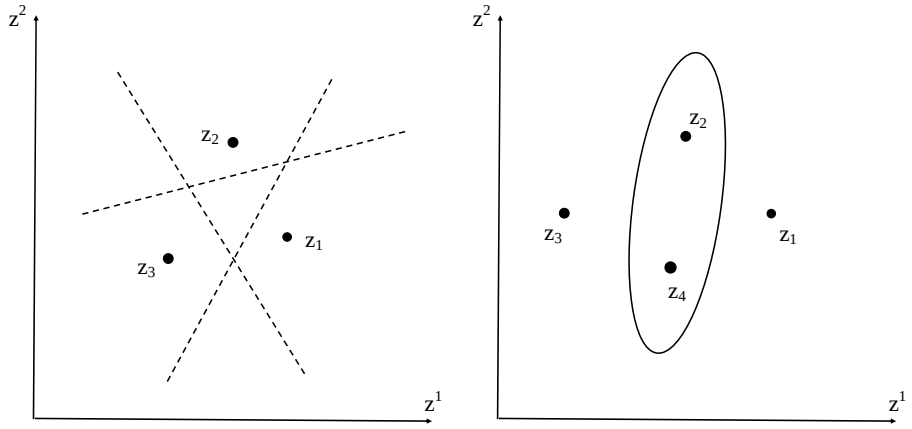
$$G^{\Lambda}(h+1) < (h+1)\ ln\ 2.$$

Izmantojot šos rezultātus, tiek definēta VC dimensija funkciju klasei $Q(z, \alpha)$.

Definīcija 1.17. [15, 79. lpp] Indikatora funkciju kopas $Q(z, \alpha), \alpha \in \Lambda$ VC dimensija ir galīga un vienāda ar skaitli h, ja tās atbilstošā augšanas funkcija ir ierobežota ar logaritmisku funkciju ar koeficientu h. VC-dimensija ir bezgalīga, ja kopas augšanas funkcija ir lineāra.
Kopā ar šo VC dimensijas definīciju Vapniks piedāvā arī alternatīvu definīciju, kuru vieglāk praktiski pielietot.

Definīcija 1.18. [15, 80. lpp] Indikatora funkciju kopas $Q(z, \alpha), \alpha \in \Lambda$ VC dimensija ir maksimālais vektoru $z_1, ..., z_h$ skaits, kuri var tikt sadalīti divās apakškopās visos 2^h veidos. Ja to var izdarīt katram naturālam n, tad kopas $Q(z, \alpha)$ VC dimensija ir bezgalīga.
Lai no indikatora funkciju klasēm definīciju vispārinātu uz reālu funkciju klasēm, izmanto līdzīgu pieeju, kā VC entropijas gadījumā.

Definīcija 1.19. [15, 80. lpp] Pieņemsim, ka $A \leq Q(z, \alpha) \leq B, \alpha \in \Lambda$ ir reālu funkciju kopa, kura ierobežota ar konstantēm A un B. Šīs konstantes var pieņemt arī bezgalīgu



1.4. att.: VC dimensija vektoriem plaknē ir vienāda ar 3, jo ir iespējams sašķelt trīs vektorus, bet ne četrus. Piemēram labās puses attēlā vektorus z_2 un z_4 nav iespējams atšķelt no pārējiem ar indikator funkciju.

vērtību. Kopā ar funkcijām $Q(z,\alpha)$ aplūko arī indikatorus

$$I(z,\alpha,\beta) = \theta(Q(z,\alpha)-\beta), \alpha \in \Lambda, \beta \in (A,B), \tag{1.30}$$

kur

$$\theta(z) = \begin{cases} 0 \; ja \; z < 0 \\ 1 \; ja \; z \geq 0 \end{cases} . \tag{1.31}$$

Līdz ar to kopas $Q(z,\alpha)$ VC dimensija tiek definēta kā atbilstošās indikatoru kopas VC dimensija,

Piemērs 1.6. Aplūkosim lineāru indikator funkciju kopu $Q(z,\alpha)$, n-dimensionālā koordinātu sistēmā, t.i. $z = (z_1,...,z_n), \alpha = (\alpha_1,...,\alpha_n)$,

$$Q(z,\alpha) = \theta\left(\sum_{p=1}^{n} \alpha_p z_p + \alpha_0.\right).$$

Šīs funkciju kopas VC dimensija $h = n+1$, jo tieši tas ir maksimālais vektoru skaits, kurus var sadalīt divās apakškopās 2^h veidos, izmantojot funkcijas no kopas $Q(z,\alpha)$. Mēdz arī teikt, ka funkciju kopa $Q(z,\alpha)$ var sašķelt ne vairāk kā $n+1$ vektoru no n-dimensionālas telpas. Attēlā 1.4. redzams gadījums, kad $n = 2$ - plaknē ar divdimensiju indikator funkcijām iespējams sašķelt ne vairāk kā 3 vektorus.

Piemērs 1.7. Aplūkosim lineāru funkciju kopu $Q(z,\alpha)$, n-dimensionālā koordinātu sistēmā, kur

$$Q(z,\alpha) = \sum_{p=1}^{n} \alpha_p z_p + \alpha_0, \;\; \alpha_1,...,\alpha_n \in (-\infty,\infty).$$

Arī šīs funkciju kopas VC dimensija $h = n + 1$, jo tāda ir atbilstošās indikatoru kopas VC dimensija.

Piemērs 1.8. Aplūkosim lineāru funkciju kopu

$$f(z, \alpha) = \theta(sin\alpha z), \alpha \in \mathbb{R}.$$

Šīs funkciju kopas VC dimensija ir bezgalība. Piemēram, punktus

$$z_1 = 10^{-1}, ..., z_n = 10^{-n}$$

var sašķelt ar funkcijām no šīs kopas. Pieņemot, ka ir dota skaitļu virkne

$$\delta_1, ..., \delta_n, \ \delta_i \in (0, 1),$$

pietiek izvēlēties

$$\alpha = \pi \left(\sum_{i=1}^{n} (1 - \delta_i) 10^i + 1 \right).$$

Pētot pieejamo literatūru nācās secināt, ka lai arī VC dimensija tiek plaši lietota teorētisku rezultātu iegūšanā vispārinātā gadījumā, daudz grūtāk ir to novērtēt konkrētam modelim. Pat mūsdienās tiek meklēti labākie VC dimensijas novērtēšanas algoritmi kā arī to optimālie iestatījumi, kuri tiks pieminēti vēlāk.

Aplūkosim Vapnika, Levina un Cuna piedāvāto VC dimensijas novērtēšanas pieeju [17]. VC dimensijas definīcijā minēts, ka funkcijām no kopas $Q(z, \alpha)$ jāsadala vektori $z_1, ..., z_h$ divās daļās 2^h veidos. Mūsu mērķis ir novērtēt h. Apzīmēsim vienu no daļām ar A un otru ar B. Apzīmēsim $Q(z, \alpha^*) = 0$ ja $z \in A$ un $Q(z, \alpha^*) = 1$ ja $z \in B$. Parametrs α^* šajā gadījumā raksturo konkrētu funkciju no kopas $Q(z, \alpha)$.

N reizes tiek atkārtotas tālāk minētās darbības. Tiek definēts

$$X_i^{2n} = z_1^i, \omega_1^i, ; z_2^i, \omega_2^i; ...; z_{2n}^i, \omega_{2n}^i,$$

kur z^i - gadījuma vektoru izlase un ω^i atbilstošā vektoru klase ($\omega_j^i \in \{0; 1\}$). Ar $\nu_1^n(X_i^{2n}, \alpha)$ apzīmē kļūdaino klasifikāciju biežumu pirmajiem n vektoriem, šo biežumu aprēķina

$$\nu_1^n(X_i^{2n}, \alpha) = \frac{1}{n} \sum_{j=1}^{n} |\omega_j^i - f(z_j^i, \alpha)|.$$

Līdzīgi tiek novērtēts klasifikācijas kļūdu biežums otrajiem n vektoriem

$$\nu_2^n(X_i^{2n}, \alpha) = \frac{1}{n} \sum_{j=n+1}^{2n} |\omega_j^i - f(z_j^i, \alpha)|.$$

Mūs interesē starpība starp šiem kļūdu biežumiem, ko apzīmēsim ar

$$\xi(Z_i^{2n}) = \sup_{\alpha\in\Lambda}(\nu_1^n(X_i^{2n},\alpha) - \nu_2^n(X_i^{2n},\alpha)).$$

Balstoties uz šādu N reizes veiktu novērtējumu, tiek aprēķināta vidējā atšķirība starp kļūdu biežumiem izlašu pirmajiem n un otrajien n vektoriem, respektīvi

$$\xi(n) = \frac{1}{N}\sum_{i=1}^{N}\xi(Z_2^{2n}).$$

Atkārtojot šo procedūru dažādiem n, iegūstam novērtējumu izlasi $\xi(n_1), \xi(n_2), ..., \xi(n_k)$. Minētajā publikācijā [17] sniegts šo kļūdu biežumu starpības teorētiskais novērtējums

$$E\{\sup_{\alpha\in\Lambda}(\nu_1^n(X_i^{2n},\alpha) - \nu_2^n(X_i^{2n},\alpha))\} \le \begin{cases} 1, \text{ ja } n/h \le 0.5 \\ C_1\frac{ln(2n/h)+1}{n/h}, \text{ ja } 0.5 < n/h < g, \text{ g ir mazs} \\ C_2\sqrt{\frac{ln(2n/h)+1}{n/h}}, \text{ ja } n/h \text{ ir liels} \end{cases}.$$

Labās puses izteiksmi iespējams aproksimēt ar funkciju

$$\Phi(\tau) = \begin{cases} 1, \text{ ja } \tau \le 0.5 \\ a\frac{ln(2\tau)+1}{\tau-k}\left(\sqrt{1+\frac{b(\tau-k)}{ln(2\tau)+1}}+1\right) \end{cases} \quad \text{ja } \tau \ge 0.5,$$

kur $\tau = n/h$. Funkcijai ir divi brīvi parametri a un b, parametrs k izvēlēts tā, lai $\Phi(0.5) = 1$. Publikācijā [17] šie nezināmie parametri ir novērtēti: $a = 0.16$, $b = 1.2$ un $k = 1.5$.

Līdz ar to VC dimensija funkcijām $Q(z,\alpha)$ tiek meklēta kā mazākais naturālais skaitlis h^*, kurš nodrošina labāko saderību starp funkciju $\Phi(n/h)$ un iegūto izlasi no $\xi(n_i)$

$$h^* = argmin_{h\in\mathbb{N}}\sum_{i=1}^{k}(\xi(n_k) - \Phi(n_i/h))^2.$$

Minētos skaitļus $n_1, ..., n_k$ sauc par dizaina punktiem un sākotnēji var pieņemt, ka katrā no šiem punktiem ģenerē N novērojumus, lai novērtētu $\xi(n_i)$. Tomēr pastāv svarīgs jautājums - vai tiešām visiem n_i jāņem vienādi N? Šo jautājumu ir aprakstījuši Šao, Čerkaskis un Li 2000. gada publikācijā [18].

Tālāko darbību mērķis ir samazināt starpību starp $\xi(n)$ un $\Phi(n/h)$, minimizējot vidējo kvadrātisko kļūdu

$$MSE = E(\xi(n) - \Phi(n/h))^2,$$

tādējādi iegūstot labāku novērtējumu. Sākotnēji ņemam visiem n_i vienādus N_i un rīkojamies saskaņā ar algoritmu

1. Novērtējam MSE aktuālajam modelim (sākotnēji visi N_i sakrīt);

2. Saranžējam dizaina punktus $n_1, ..., n_k$ pēc to ieguldījuma novērtējumā MSE. Dizaina punkta n_i ieguldījums tiek novērtēts

$$Ieguld(n_i) = (MSE_{-i} - MSE)/N_i,$$

   kur MSE ir vidējā kvadrātiskā kļūda visiem dizaina punktiem kopā un MSE_{-i} ir vidējā kvadreātiskā kļūda, ja ņem visus dizaina punktus, izņemot n_i. Kad katram dizaina punktam ieguldījums ir sarēķināts, tos sakārto augošā secīnā pēc ieguldījuma lieluma. Labākais dizaina punkts ir ar vismazāko ieguldījumu.

3. Pārnesam vienu novērojumu no sliktākā dizaina punkta uz labāko un aprēķinam MSE. Ja iegūts lielāks MSE, tad atgriežas pie iepriekšējā varianta un pārceļ novērojumus no otrā sliktākā uz otro labāko dizaina punktu, vēlāk ja nepieciešams no trešā mazākā uz trešo lielāko utt. Procesu atkārto, kamēr iegūst mazāku MSE. Rezultātā iegūto izmainīto novērojumu skaita vektoru $N_1^*, ..., N_k^*$ nosauc par aktuālo modeli.

4. Atkārto soļus 1.-3., līdz tiek sasniegts kāds no kritērijiem:

   - visi dizaina punkti dod nepozitīvu ieguldījumu
   - visi dizaina punkti, kuri dod pozitīvu ieguldījumu ir piesātināti, t.i. satur vismaz 25% no kopējiem novērojumiem.

2. Empīrisko procesu pielietojums

2.1. Empīrisko procesu pielietošana testos

Empīriskajiem procesiem un empīrisko procesu teorijai ir dažādi pielietojumi. Vistiešāk empīrisko procesu var pielietot virzienā, no kura tas ir cēlies un tie ir testi jeb hipotēžu pārbaudes. Vispazīstamākie ir testi par datu sadalījuma funkcijas atbilstību kādai teorētiskai sadalījuma funkcijai. Šāds tests ir piemēram Krāmera-fon Mises kritērijs.

$$H_0 : \ X \sim F(x) \text{ pret } H_1 : H_0 \text{ nav spēkā.}$$

Pieņemsim, ka dota izlase neatkarīgu un vienādi sadalītu gadījuma lielumu $X_1, ..., X_n$ izlase ar teorētisko sadalījuma funkciju $F(x)$. Veicam hipotēžu pārbaudi Krāmera-fon Mises statistika definēta sekojoši:

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x).$$

Testa statistika $T = n\omega^2$. Pie nosacījuma, ka izpildās hipotēze H_0,

$$T \to_d \int_{-\infty}^{\infty} (B(F(x))^2) dF(x), \text{ kad } n \to \infty.$$

Šeit ar B apzīmē Brauna tiltu. Ja hipotēze H_0 nav spēkā, tad $T \to \infty$, kad $n \to \infty$. Krāmera-fon Mises kritēriju var izmantot kā alternatīvu plaši pazīstamajam Kolmogorova-Smirnova statistikai (1.3)

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Šajā gadījumā hipotēžu pārbaudē izmanto statistiku $\sqrt{n}D_n$. Kā redzams, testa statistika ir ļoti līdzīga vienkāršākajam empīriskajam procesam un ja hipotēze ir spēkā, tad $\sqrt{n}D_n \to \sup_x |B(F(x))|$, kad $n \to \infty$ un uz bezgalību ja H_0 nav spēkā.

Abas minētās testa statistikas var vispārināt uz divu izlašu gadījumu. Tomēr ir vērts apskatīt arī nedaudz savādāku pieeju. Pieņemsim, ka dotas divas neatkarīgu un vienādi

sadalītu gadījuma lielumu izlases $X_1, ..., X_n$ un $Y_1, ..., Y_m$ attiecīgi ar sadalījuma funkcijām F_1 un F_2. Veicam hipotēžu pārbaudi:

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

1961. gadā Vatsons [19][109. lpp] aprakstījis statistiku, kuru mēdz saukt par U^2 testu. Šo testu var izmantot kā alternatīvu Krāmera-fon Mises statistikas divu izlašu gadījumam:

$$U_n^2 = n \int_{-\infty}^{\infty} \left\{ F_{1n}(x) - F(x) - \int_{-\infty}^{\infty} (F_{2m}(y) - F(y)) dF(y) \right\}^2 dF(x).$$

Pielietojot empīrisko procesu teoriju, iegūst

$$U_n^2 \to_d \int_{-\infty}^{\infty} \left\{ B(F_1(x)) - \int_{-\infty}^{\infty} (B(F_2(y)) dF(y) \right\}^2 dF(x).$$

Iepriekš tika aplūkoti testi izlases sadalījuma pērbaudei, tomēr uz empīriskajiem procesiem iespējams balstīt arī citus testus. Piemēram Krāmera-fon Mises testu var pārveidot par sadalījuma simetrijas testu [20]

$$H_0 : \ F(x) = 1 - F(-x) \text{ pret } H_1 : F(x) \neq 1 - F(-x).$$

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - (1 - F_n(-x)))^2 dF(x).$$

Izteiksmi iespējams pērveidot sekojoši

$$\omega^2 = \int_{-\infty}^{\infty} (\sqrt{n}(F_n(x) - F(x) + F(x)) + \sqrt{n}(F_n(-x) - F(-x) + F(-x)) - \sqrt{n})^2 dF(x)$$

līdz ar to

$$\omega^2 \to_d \int_{-\infty}^{\infty} (B(F(x)) + B(F(-x)) + \sqrt{n}(F(x) + F(-x) - 1))^2 dF(x).$$

Ja izpildās hipotēze, tad izteiksme vienkāršojas uz integrāli pa divu neatkarīgu Brauna tiltu summu.

$$\omega^2 \to_d \int_{-\infty}^{\infty} (B(F(x)) + B(F(-x)))^2 dF(x).$$

Pretējā gadījumā trešais saskaitāmais tiecas uz bezgalību un integrālis diverģē. Minētajā literatūras avotā [20] pieminēti arī testi par vienādu sadalījuma likumu izlasē, kā arī neatkarības testi. Tie abi tiek definēti nedaudz pārveidojot Krāmera-fon Mises ω^2 kritēriju.

Statistikas teorijā tiek aplūkotas ne tikai vienkāršas hipotēzes, bet arī saliktas. Šādas ir hipotēzes par modeļiem, kuri iekļauj sevī parametru novērtēšanu. Piemēram,

Kolmogorova-Smirnova testu ar parametru novērtēšanu analizējis Lilliefors [21] un šāds tests normalitātes pārbaudei tiek saukts par Lilliefora testu. Parametru novērtēšanas gadījumā izmainās statistika t.i. iegūst

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

kur $\hat{\theta}$ - novērtētie parametri. Līdz ar statistikas maiņu mainās arī asimptotiskais sadalījums. 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 konverģence uz Brauna tiltu vairs nav spēkā. To viegli pamatot ar Teilora rindas palīdzību. 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

$$\sqrt{n}(F_n(x)-F(x,\hat{\theta})) = \sqrt{n}(F_n(x)-F(x,\theta_0)) - \sqrt{n}(\hat{\theta}-\theta_0)\frac{\partial}{\partial\theta}F(x,\theta)|_{\theta=\theta_0} + o_p(\sqrt{n}(\hat{\theta}-\theta_0)).$$

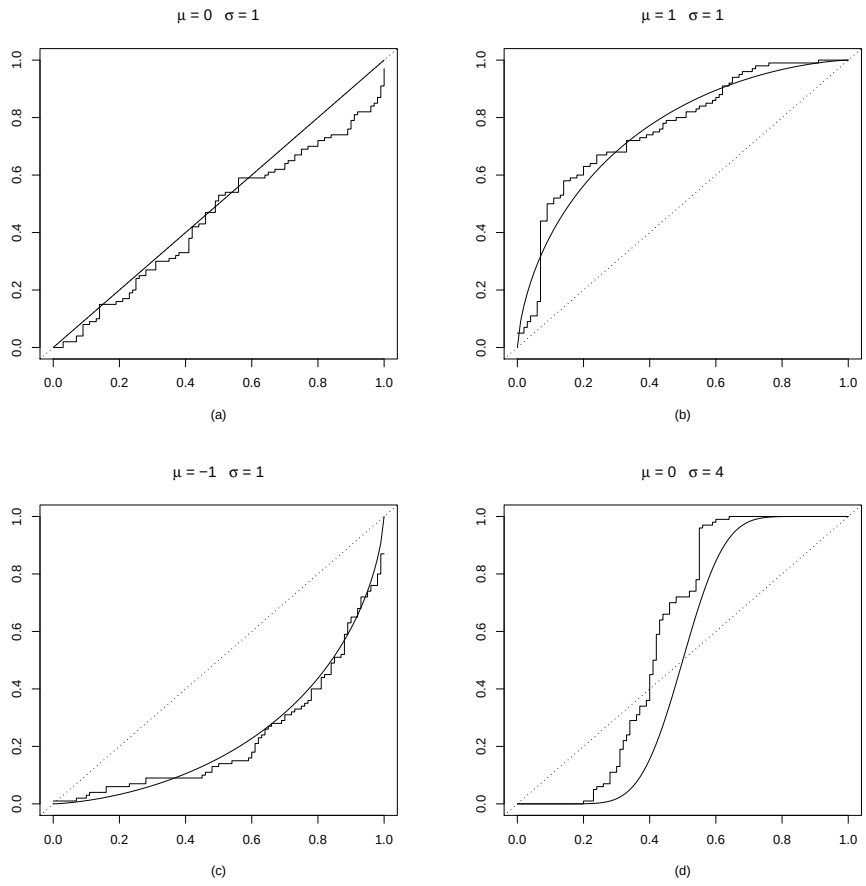
Rezultātā var secināt, ka saliktās statistikas asimptotiskais sadalījums ir atkarīgs no parametru novērtējumiem un pārbaudāmās izlases sadalījuma veida.

Īepriekš tika minēts, ka gan Krāmera-fon Mises, gan Kolmogorova-Smirnova testus var izmantot divu izlašu gadījumā. Tomēr ir arī cits veids kā pārbaudīt divu izlašu sadalījuma funkciju vienādību. Aplūkojam hipotēzi

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

Definīcija 2.1. [22, 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.

Attēlā 2.1. 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. Ja izlašu sadalījuma likumi ir no vienas



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.

klases, bet atšķiras dispersija, grafiks tiek saspiests uz vidu. 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 2.2. [23, 28. lpp.] Par empīrisko P-P procesu sauc

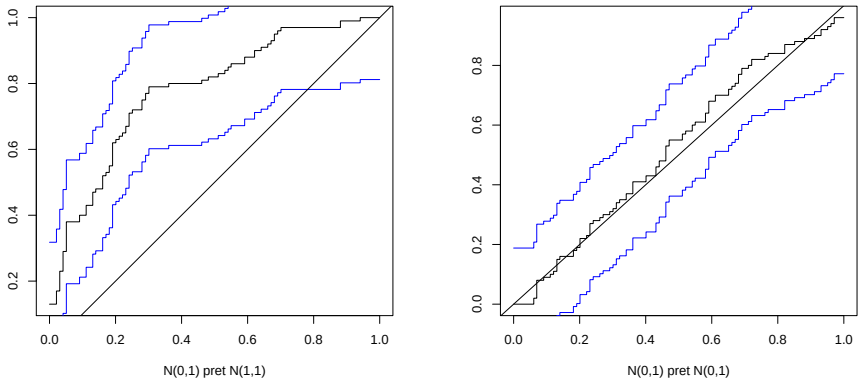
$$PP(t) = \sqrt{n}(F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t))).$$

Šo procesu var lietot gan divu izlašu sadalījuma funkciju vienādības pārbaudei, gan arī lai noteiktu doto izlašu empīriskā P-P grafika ticamības joslu. Šajā nodaļā apstāsimies pie pārbaudāmās hipotēzes. Veicot ekvivalentus pārveidojumus var iegūt, ka

$$\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_2^{-1}(t))| \to \sup_{0<t<1} |B^{(n)}(F_1F_2^{-1}(t))) + \sqrt{\frac{n}{m}}\frac{f_1(F_2^{-1}(t))}{f_2(F_2^{-1}(t))}B^{(m)}(t))|,$$

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. Pie nosacījuma, ka izpildās hipotēze H_0, iegūst

$$\sup_{0<t<1} |\sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_2^{-1}(t))| \to \sup_{0<t<1} |(B^{(n)}(F_1F_2^{-1}(t))) + \sqrt{\frac{n}{m}}B^{(m)}(t))|.$$



2.2. att.: Nenoraidītas un noraidītas hipotēzes piemēri, konstruējot vienlaicīgās ticamības joslas P-P grafikiem. Izlašu apjomi $n = m = 100$.

Šādam sadalījumam var aprēķināt kritiskās vērtības un, konstruēt vienlaicīgās ticamības joslas empīriskajam P-P grafikam. Hipotēze H_0 tiek noraidīta, ja konstruētās joslas neiekļauj taisni $y = t$, kur $t \in [0, 1]$ kaut vienā punktā. Nenoraidītas un noraidītas hipotēzes piemērus var aplūkot attēlā 2.2..

2.2. Empīriskā procesa butstraps lokācijas-skalēšanas modelim

Vienkāršākais veids, kā aprēķināt kādas statistikas kritiskās vērtības, ir Monte Carlo simulācijas. Simulāciju laikā daudz reižu tiek ģenerētas gadījuma izlases no zināma sadalījuma un aprēķināta interesējošā statistika vai arī uzreiz tiek simulēts asimptotiskais sadalījums. Tomēr šī metode nav izmantojama ja asimptotiskais sadalījums atkarīgs no datiem. Protams, veicot simulācijas robežsadalījums tiks aproksimēts, tomēr to nevarēs izmantot reālu datu gadījumā, jo tad nebūs zināms datu sadalījums.

Lai apietu šo problēmu statistikā bieži izmanto procedūru, ko sauc par bustrapošanu. Šo metodi 1979. gadā publicējis Efrons [24]. Metodes galvenā būtība ir tā, ka paši dotie dati tiek uzskatīti par teorētisko sadalījumu un jaunas gadījuma izlases tiek ģenerētas no dotajiem datiem. Pēc tam ģenerētajām izlasēm aprēķina testa statistiku un nosaka kritisko vērtību dažādiem ticamības līmeņiem.

Kā piemēru aplūkosim lokācijas-skalēšanas modeli. Mūsu mērķis būs veikt hipotēžu pārbaudi par šāda modeļa eksistenci starp divām dotām izlasēm. Sākumā tiks sniegtas dažas definīcijas.

Definīcija 2.3. [25] Starp divu izlašu $X_1, ..., X_n$ un $Y_1, ..., Y_m$ sadalījuma funkcijām pastāv lokācijas-skalēšanas modelis, ja

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

Šo attiecību var izteikt arī ar kvantiļu funkcijām

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

kur F_1 un F_2 attiecgās izlašu teorētiskās sadalījuma funkcijas. Ievērosim, ka veicot vienkāršo hipotēžu pārbaudi, t.i. fiksējot konkrētus parametrus μ un σ, tiek iegūts ļoti līdzīgs tests iepriekš aplūkotajai divu izlašu sadalījuma funkciju pārbaudei. Turklāt jāpiemin, ka statistikas asimptotiskais sadalījums nav atkarīgs no tā, kādus μ un σ izvēlas. Tomēr no parametru izvēles ir atkarīgs šī modeļa atbilstošais varbūtību-varbūtību grafiks, kurš šajā gadījumā tiek definēts kā $F_1(\sigma F_2^{-1}(t) + \mu)$.

Aplūkosim saliktu hipotēzi lokācijas skalēšanas modelim

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

šajā gadījumā tā būs hipotēžu pārbaude par to, vai starp divām dotām izlasēm eksistē lokācijas-skalēšanas modelis. Pirmais solis, kas jāveic ir zināmo parametru novērtēšana. Lokācijas-skalēšanas modeļa gadījumā parametrus var novērtēt sekojoši:

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

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

Ja izlašu apjomi sakrīt, novērtējumus var uzrakstīt vienkāršākā formā

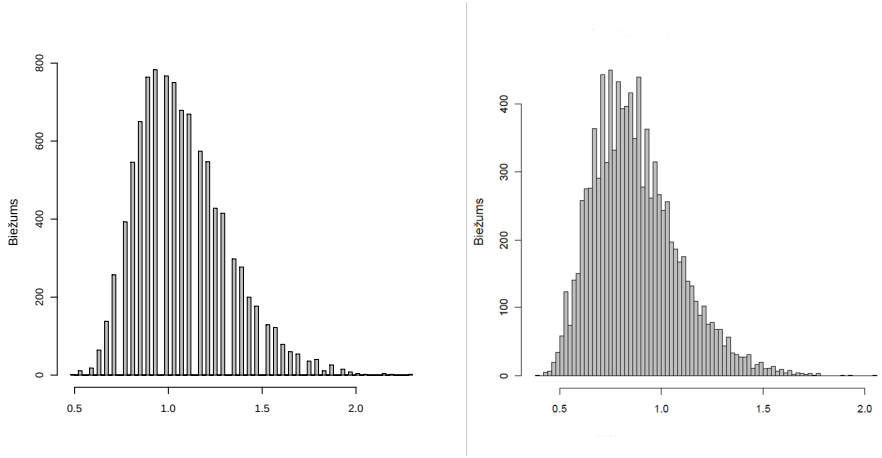
$$\hat{\sigma} = \frac{\frac{1}{n}\sum_{i=1}^n X_{(i)}Y_{(i)} - \frac{1}{n}\sum_{i=1}^n X_i \frac{1}{n}\sum_{i=1}^n Y_i}{\frac{1}{n}\sum_{i=1}^n Y_i^2 - (\frac{1}{n}\sum_{i=1}^n Y_i)^2},$$

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

kur $X_{(i)}$ un $Y_{(i)}$ - augošā secībā sakārtoti doto izlašu elementi. Šie novērtējumi iegūti minimizējot tā saukto Mallova attālumu, kurš aprakstīts publikācijā [25].

Definīcija 2.4. [25] Lokācijas-skalēšanas modelim Mallova attālums tiek definēts kā

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



2.3. att.: Lokācijas-skalēšanas empīriskā procesa asimptotiskais sadalījums. kreisajā pusē - rezultāts, kas iegūts ar butstrapa palīdzību, labajā pusē - ar simulāciju palīdzību. Ģenerētas izlases ar sadalījumiem $N(0,1)$ un $N(1,4)$ apjomā 500, simulāciju un bustrapa skaits - 10000.

Pēc parametru novērtēšanas, transformējam otro izlasi

$$\widetilde{Y}_i = \hat{\sigma} Y_i + \hat{\mu}$$

un iegūstam vienkāršāk pierakstāmu hipotēžu pārbaudi:

$$H_0 : \ F_1(\widetilde{F}_2^{-1}(t)) = t \text{ pret } H_1, F_1(\widetilde{F}_2^{-1}(t)) \neq t, \ t \in [0,1].$$

Lai konstruētu vienlaicīgās ticamības joslas lokācijas-skalēšanas modelim, var izmantot statistiku

$$\sup_{0 \leq t \leq 1} |\sqrt{n}(F_{1n}(\widetilde{F}_{2m}^{-1}(t)) - F_1(\widetilde{F}_2^{-1}(t)))|,$$

kuras asimptotiskais sadalījums ir

$$\sup_{0 \leq t \leq 1} \left| B^{(n)}(t) + \sqrt{\frac{n}{m}} B^{(m)}(t) + \sqrt{n} f_2(F_2^{-1}(t))((\hat{\mu} - \mu) - F_2^{-1}(t)(\hat{\sigma} - \sigma)) \right|.$$

Ar detalizētu lokācijas-skalēšanas modeļa empīriskā procesa asimptotiskā sadalījuma izvedumu var iepazīties manā diplomdarbā [26]. Attēlā 2.3. redzamas divas lokācijas skalēšanas modeļa asimptotiskā sadalījuma histogrammas. Kreisajā pusē redzama histogramma, kura iegūta sākotnēji simulējot vienu izlašu pāri un pēc tam ar bustrapa palīdzību aprēķinātas statistikas vērtības, labajā pusē redzams simulāciju rezultāts. Tika ģenerēti dati ar sadalījumiem $N(0,1)$ un $N(1,4)$ ar izlašu apjomu 500. Bustraps un simulācijas atkārtotas 10000 reižu. Kā redzams attēlā, bustraps rada caurumus histogrammā, galvenais iemesls ir tāds, ka butstrapojot tiek izmantota dota izlase, kura satur ierobežotu skaitu datu un rezultātā nevar iegūt nepārtrauktu sadalījumu kā tas ir simulāciju gadījumā, kur dati tiek ģenerēti no nepārtraukta sadalījuma.

Šī ir liela problēma, ja darbs ir ar reāliem datiem, jo tad nebūs iespējams veikt simulācijas. Jāpiemin, ka bustrapa gadījumā būtiski pieaug statistikas kritiskā vērtība pie jebkura ticamības līmeņa, tātad iegūstam plašākas ticamības joslas. Rezultātā mēs nevaram noteikt gadījumus, kad lokācijas-skalēšanas modelis tiešām nav spēkā. Lai izvairītos no šādas situācijas, tiek izmantots gludinošais bustraps. No jaunākajām publikācijām, kur izmantots gludinošais bustraps, var minēt Horvata publikāciju par vienlaic;igo ticamības joslu konstruēšanu ROC līknēm [7]. Ideja par gludinošā butstrapa izmantošanu lokācijas-skalēšanas modelim ir ņemta tieši no minētās publikācijas.

Gludinātā bustrapa idejas pamatā ir empīriskas funkcijas gludināšana ar kodolu funkcijām. Tālāk nedaudz tiks aprakstīta pati gludināšanas ideja. Pieņemsim, ka ir dota izlase $X_1, ..., X_n$, kur n - izlases apjoms.

Definīcija 2.5. 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: $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. Mūsu apskatītajā piemērā nepieciešams gludināt gan izlašu sadalījuma funkcijas, gan kvantiļu funkcijas.

Definīcija 2.6. [27, 684. lpp] Pieņemsim, ka k ir kodols un h - joslas platums, tad par gludināto blīvuma funkciju sauc funkciju

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

Definīcija 2.7. 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 2.8. 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 $t \in (0; 1)$.

Atgriezīsimies pie piemēra ar lokācijas-skalēšanas modeli un atcerēsimies šī modeļa atbilstošo empīrisko procesu:

$$\sup_{0\leq t\leq 1} |\sqrt{n}(F_{1n}(\widetilde{F}_{2m}^{-1}(t)) - F_1(\widetilde{F}_2^{-1}(t)))|.$$

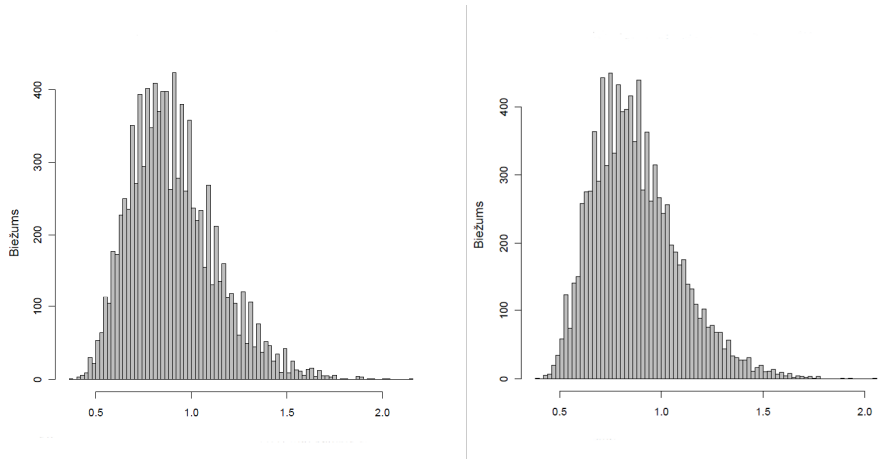
Kā redzams, tiek pētīta starpība starp empīrisko un teorētisko varbūtību-varbūtību grafiku. Ja mēs pielietojam parasto neparametrisko bustrapu, tad patiesībā šis process izmainās uz šādu:

$$\sup_{0\leq t\leq 1} |\sqrt{n}(F_{1n}^{*}(\widetilde{F}_{2m}^{*-1}(t)) - F_{1n}(\widetilde{F}_{2m}^{-1}(t)))|,$$

kur F_{1n}^{*} un F_{2m}^{*} ir no izlasēm ar empīriskajām sadalījuma funkcijām attiecīgi F_{1n} un F_{2m} bustrapoto izlašu sadalījuma funkcijas. Tas nozīmē, ka mēs vairs nesalīdzinām teorētisko sadalījuma funkciju un kādas izlases empīrisko sadalījuma funkciju, bet gan divas empīriskās sadalījuma funkcijas un tādā situācijā empīrisko procesu rezultāti vairs nav spēkā un patiesībā mēs pat nevaram pateikt, kāds šādai statistikai būs asimptotiskais sadalījums.

Tā kā praksē mēs nekad nezinam dotās izlases teorētisko sadalījuma funkciju, varam pielietot gludināšanas metodes šīs pašas izlases epmīriskajai sadalījuma funkcijai un iegūto gludo funkciju pasludināt par īsto teorētisko sadalījuma funkciju. Precizitāte, protams, būs atkarīga no izlases apjoma. Nākošais solis ir palīgizlašu bustrapošana no gludinātajām funkcijām un tālak jau seko statistikas aprēķināšana un ticamības joslu konstruēšana. Pieņemsim, ka dotas divas izlases $X_1, ., X_n$ un $Y_1, ..., Y_n$. Gludinātā bustrapa metodes realizācijas algoritms līdz ar to ir sekojošs:

1. no dotajiem datiem iegūstam nogludinātas sadalījuma funkcijas $\hat{F}_{1n}$ un $\hat{F}_{2m}$;
2. ģenerējam neatkarīgu izlašu pāri $\xi_1, ..., \xi_n$, $\eta_1, ..., \eta_n$ no sadalījuma $U(0,1)$;
3. ar inversās transformācijas palīdzību iegūstam īstās bustrapotās izlases $X_i^{*} = \hat{F}_{1n}^{-1}(\xi_i)$, $Y_i^{*} = \hat{F}_{2n}^{-1}(\eta_i)$;
4. veicam parametru novērtēšanu un transformāciju butstrapotajām izlasēm;
5. aprēķinam testa statistiku, kura šajā gadījumā ir
$$\sup_{0\leq t\leq 1} \left|\sqrt{n}(\hat{F}_{1n}^{*}(\widetilde{\hat{F}}_{2m}^{*-1}(t)) - \hat{F}_{1n}(\widetilde{\hat{F}}_{2n}^{-1}(t)))\right|$$
6. atkārtojam darbības $2-5$ un aprēķinam statistikas kritisko vērtību pie fiksēta ticamības līmeņa.



2.4. att.: Lokācijas-skalēšanas empīriskā procesa asimptotiskais sadalījums. kreisajā pusē - rezultāts, kas iegūts ar gludinošā butstrapa palīdzību, labajā pusē - ar simulāciju palīdzību. Ģenerētas izlases ar sadalījumiem $N(0,1)$ un $N(1,4)$ apjomā 500, simulāciju un bustrapa skaits - 10000.

Horvats publikācijā par ROC līknēm [7] ir pierādījis, ka empīriskam procesam, kurā izmantots gludinošais butstraps, asimptotiskais sadalījums sakrīt sadalījumu, ko iegūstam veicot Monte Carlo simulācijas. Līdzīgi kā gadījumā ar parasto neparametrisko butstrapu, varam konstruēt histogrammu jaunajam empīriskajam procesam un salīdzināt to ar asimptotisko sadalījumu. Salīdzinājumam apskatīsim kritiskās vērtības lokācijas-skalēšanas modeļa vienlaicīgo ticamības joslu konstrukcijai. Vienkāršas hipotēzes gadījumā aplūkosim, kas tiek iegūts butstrapojot asimptotisko sadalījums, veicot simulācijas, veicot parasto un gludinošo butstrapu, saliktas hipotēzes gadījumā aplūkosim simulācijas, parasto un gludinošo butstrapu. Kā redzams tabulā 2.1. Monte Carlo simulācijas dod vistuvāko rezultātu asimptotiskajam sadalījumam un labu novērtējumu iegūstam arī ar gludinātā butstrapa palīdzību. Kā jau iepriekš bija redzams histogrammās, parastais neparametriskais butstraps šajā situācijā dod sliktus rezultātus. Tabulā 2.2. redzama šī pati analīze saliktas hipotēzes gadījumā. Šoreiz mēs nevaram precīzi pateikt, kādas vērtības dod asimptotiskais sadalījums, tomēr varam izmantot faktu, ka Monte Carlo simulācijas dod tuvu rezultātu teorētiskajam un salīdzināt butstrapa metodes attiecībā pret simulācijām. Arī šajā gadījumā gludinātais butstraps ir daudz labāks par parasto neparametrisko butstrapu. Šajā nodaļā tika apskatīts tikai lokācijas-skalēšanas modelis, tomēr līdzīgu analīzi var veikt jebkuru empīrisko procesu gadījumā.

2.1. tabula: Kritiskās vērtības lokācijas-skalēšanas empīriskajam procesam vienkāršas hipotēzes gadījumā izlasēm ar apjomu n=500

Metode	90%	95%	99%
Asimptotiskais sadalījums	1.70	1.89	2.27
Monte Carlo simulācijas	1.69	1.87	2.27
Parastais butstraps	2.05	2.23	2.68
Gludinātais butstraps	1.73	1.93	2.32

2.2. tabula: Kritiskās vērtības lokācijas-skalēšanas empīriskajam procesam saliktas hipotēzes gadījumā izlasēm ar apjomu n=500

Metode	90%	95%	99%
Monte Carlo simulācijas	1.15	1.24	1.45
Parastais butstraps	1.43	1.56	1.79
Gludinātais butstraps	1.18	1.24	1.47

2.3. Empīriskie procesi dažādu modeļu novērtēšanas problēmu risināšanā

Kā iepriekš tika minēts, empīriskos procesus var izmatot hipotēžu pārbaudei un nepieciešamības gadījumā iespējams tos butstrapot. Tomēr tas nav vienīgais empīrisko procesu pielietošanas veids. Reizēm lai nokļūtu līdz kādam rezultātam empīrisko procesu teoriju var izmantot kā rīku starprezultātu pierādīšanai. Šajā nodaļā tiks aplūkoti gadījumi, kad empīriskais process izmantots kāda modeļa novērtēšanā.

Piemērs 2.1 (Empīriskais process GARCH(1,1) modelim)**.**

Definīcija 2.9. [28] Modelis GARCH(1,1) (angliski - *Generalized AutoRegresive Conditional Heteroscedasticity*) tiek uzdots sekohojoši:

$$y_k = \sigma_k \epsilon_k,\ \sigma_k^2 = \alpha + \beta y_{k-1}^2 + \gamma \sigma_{k-1}^2,\ k \in Z,$$

kur ϵ_k ie neatkarīgi, vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju $G(x)$, $E\epsilon_1^2 = 1$ un $a = (\alpha, \beta, \gamma)$ ir vektors no nezināmiem parametriem, $\alpha, \beta, \gamma > 0$ un $\beta + \gamma < 1$. Pie šādiem nosacījumiem eksistē viens unikāls, stacionārs un ergodisks atrisinājums y_k.

Sekojošais rezultāts minēts Vjazilova [28] publikācijā, kura veltīta GARCH(1,1) procesa parametru novērtēšanai.

Teorēma 2.1. [25][990. lpp] Pieņemsim, ka $y_0, y_1, ..., y_n$ ir novērojumi stacionāram atrisinājumam un

$$\sigma_k^2 = t_1 + t_2 y_{k-1}^2 + t_3 \sigma_{k-1}^2,\ \epsilon_k := \begin{cases} \frac{y_k}{\sigma_k(t)} \text{ ja } \sigma_k^2(t) > 0 \\ 1 \text{ ja } \sigma_k^2(t) \leq 0 \end{cases},\ k = 1, ..., n,$$

kur $t = (t_1, t_2, t_3) \in \mathbb{R}^3$ un $\sigma_0 \equiv 0$. Ievieš atlikumu empīrisko procesu

$$U_n(x,t) := \sqrt{n} \sum_{k=1}^{k} f(\hat{y}_{k-1}; t) 1_{\epsilon_k(t) \leq x},\ U_n(x) := \sqrt{n} \sum_{k=1}^{k} f(\hat{y}_{k-1}; a) 1_{\epsilon_k(t) \leq x},$$

kur $y_k = (..., y_{k-1}, y_k)$, $\hat{y}_k = (..., 0, y_0, ..., y_k)$. Ja izpildās sekojoši nosacījumi:

- $G(x)$ ir stingri monotona kopā $\{x \in \mathbb{R} : 0 < G(x) < 1\}$;
- eksistē diferencējama blīvuma funkcija $g(x) = G'(x)$ un $sup_{x \in \mathbb{R}} x^2 |g(x)| < \infty$;
- $E|f(y_1; a)|^4 < \infty$ un $\sqrt{n} \sum_{k=1}^{n} |f(y_k; a) - f(\hat{y}_k; a)| = o_p(1)$;
- eksistē $\dot{f} = \partial f / \partial t_i$, eksistē $E||\dot{f}(y_1; a)||^2 < \infty$, $n^{-1} \sum_{k=1}^{n} ||\dot{f}(y_k; a) - \dot{f}(\hat{y}_k; a)|| = o_p(1)$, $||\dot{f}(y; t) - \dot{f}(y; a)|| \leq H(y, a)||t - a||$, kur t ir kādā punkta a apkārtnē, $EH(y_1, a) < \infty$ un $sup_{k \in \mathbb{N}} EH(\hat{y}_1, a) < \infty$, $||.||$ ir Eiklīda norma.

Tad ir spēkā

$$\sup_{x \in \mathbb{R}, ||t|| \leq T} |U_n(x, a + \sqrt{n}t - U_n(x) - \frac{1}{2} x g(x) \langle E|f(y_1; a) e(y_1; a)|, t \rangle - G(x) \left\langle E\dot{f}(y_1; a), t \right\rangle| = o_p(1),$$

visiem $0 \leq T < \infty$, kad $n \to \infty$, kur

$$e(y; a) := \left(\frac{1}{(1-\gamma)\sigma^2(y; a)}, \frac{Y(y, \gamma)}{\sigma^2(y; a)}, \frac{Z(y, a)}{\sigma^2(y; a)} \right),$$

$Y(y_k, \gamma) := \sum_{m=0}^{\infty} \gamma^m y_{k-m}^2, \sigma^2(y_k; a) := \alpha/(1-\gamma) + \beta Y(y_k; \gamma), Z(y_k; a) := \sum_{m=0}^{\infty} \gamma^m \sigma^2(y_{k-m-1}; a)$.

Piemērs 2.2 (Ilglaicīgās un īslaicīgās atmiņas laiktindu atlikumu empīriskais process)**.** Aplūkojam laikrindu $\{y_t\}$, kuru ģenerē modelis

$$y_t = \beta' X_t + \epsilon_t \text{ un } \epsilon_t = \sum_{i=0}^{\infty} a_i e_{i-1},$$

kur X_t ir p-dimensionālu laikrindu virkne, kura mērojama attiecībā pret $\mathcal{F}_{t-1} = \sigma(\epsilon_{t-1}, \epsilon_{t-2}, ..)$ vai ir neatkarīga no $\{\epsilon_t\}$. Koeficienti a_i apmierina nosacījumus $\sum_{i=0}^{\infty} a_i < \infty$, $a_0 = 1$ un $a_k = k^{H-3/2}L_0(k)$ kādai lēni variējošai funkcijai L_0 ar $H < 1$. Kā arī ϵ_t ir neatkarīgu un vienādi sadalītu gadījuma lielumu virkne ar vidējo vērtību 0 un $\sigma_e^2 = Ee_t^2 < \infty$.

Definīcija 2.10. [29] Process ϵ_t ir ar ilglaicīgu atmiņu ja $H \in (1/2, 1)$ un ar īslaicīgu atmiņu ja $H < 1/2$. Čans un Lings 2008. gada publikācijā [29] aplūko empīriskos procesus abos gadījumos, tālāk tiks aprakstīti rezultāti no minētās publikācijas.

Aplūkojam gadījumu $H \in (1/2, 1)$ un definējam

$$K_n(x) = \frac{1}{\sigma_n} \sum_{i=1}^{n} [1_{\epsilon_t \leq x} - F(x)|,$$

kur $\sigma_n^2 = var(\sum_{i=1}^{n} \epsilon_t)$ un $F(x)$ ir ϵ_t teorētiskā sadalījuma funkcija.

Teorēma 2.2. *Ir spēkā sakarība*

$$\sup_{x \in \mathbb{R}} \left| K_n(x) + \frac{1}{\sigma_n} F'(x) \sum_{i=1}^{n} \epsilon) t \right| = o(1) \ g.d.,$$

$$\sigma_n^2 / k(H) n^{2H} L_0^2(n) \to 1, kad\ n \to \infty\ un\ \sigma_n^{-1} \sum_{i=1}^{n} \to_d N(0,1),$$

kur $k(H) = \int_0^{\infty} (x + x^2)^{H-3/2} dx$.

No šī rezultāta izriet

$$[\sup_{x \in \mathbb{R}} F'(x)]^{-1} \sup_{x \in \mathbb{R}} |K_n(x)| \to_d \ \ N(0,1), ja\ \sup_{x \in \mathbb{R}} |F'(x)| < \infty.$$

Uzliekot modelim daudz komplicētākus nosacījumus, iespējams iegūt rezultātu arī pie $H < 1/2$ un $H = 1/2$. Vēl minētajā publikācijā aplūkoti atlikumu empīriskie procesi nestabiliem AR(p) modeļiem. Tomēr laikrindas nav vienīgie modeļi, kuriem tiek pētīts atlikumu lglaicīgās atmiņas empīriskais process, to var veikt arī regresijas modeļiem un šajā virzienā publikāciju žurnālam Annals of Statistics ienieguši Kuliks un Loreks [30].

Iepriekšējos piemēros tika aplūkoti modeļi, kuros empīriskais process nekonverģē uz Brauna tiltu. Tomēr atcerēsimies kā konverģence uz Brauna tiltu ir raksturīga tikai tiem empīriskajiem procesiem, kuriem ir spēkā Donskera Teorēma 1.2. Turpmākajos piemēros aplūkosim gadījumus, kad tiešām statistikas asymptotiskais sadalījums ir saistīts ar Brauna tiltu. Tiks aplūkotas dažas hipotēžu pārbaudes un lai gan šāds pielietojums jau tika aprakstīts vienā no iepriekšējām nodaļām, šie piemēri īpaši interesanti ir tādēļ, ka tie apkopo empīriskos procesus un empīriskās ticamības modeļiem. Rezultātā tiek iegūts empīriskās ticamības modeļa novērtējums ar empīrisko procesu metodi.

Piemērs 2.3. [31][268. lpp] Aplūkojam neatkarīgu un vienādi sadalītu gadījuma lielumu izlasi $X_1, ..., X_n$ un pārbaudam hipotēzi $H_0 : F(x) = F_0(x)$, kur $F_0(x)$ ir zināma, nepārtraukta sadalījuma funkcija. Definējam lokalizēto empīriskās ticamības attiecību

$$R(x) = \frac{\sup_x \left\{ L(\widetilde{F}) : \widetilde{F} = F_0(x) \right\}}{\sup_x \left\{ L(\widetilde{F}) \right\}},$$

kur $L(\widetilde{F}) = \prod_{i=1}^{n} (F(X_i) - \widetilde{F}(X_i-))$. Empīriskā sadalījuma funkcija F_n atbilst suprēmam saucējā un suprēms skitītājā atbilst situācijai, kad tiek uzlikti svari $F_0(x)/(nF_n(x))$ visiem novērojumiem līdz X un to ieskaitot un svari $(1 - F_0(x)/(n(1 - F_n(x)))$ visiem novērojumiem aiz x. No tā seko

$$logR(x) = nF_n(x) log \frac{F_0(x)}{F_n(x)} + n(1 - F_n(x)) log \frac{1 - F_0(x)}{1 - F_n(x)},$$

un izmantojot to, ka $0 < F_0(x) < 1$, iegūst

$$-2logR(x) = \frac{n(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))} + o_p(1) \to_d \chi_1^2,$$

ja ir spēkā H_0.

Integrējot šo izteiksmi attiecībā pret F_0, iegūst statistiku

$$T_n = -2 \int_{-\infty}^{\infty} logR(x) dF_0(x) \to_d \int_0^1 \frac{B^2(t)}{t(1-t)} dt,$$

kur B - Brauna tilts.

Piemērs 2.4. [31][270. lpp] Aplūkosim testu par simetriju. $H_0 : F(-x) = 1 - F(x-)$, katram $x > 0$. Līdzīgi kā iepriekšējā piemērā, tiek definēta lokālā ticamības attiecība

$$R(x) = \frac{\sup_x \left\{ L(\widetilde{F}) : \widetilde{F(-x)} = 1 - \widetilde{F}(x-) \right\}}{\sup_x \left\{ L(\widetilde{F}) \right\}}, x > 0.$$

Var pierādīt, ka

$$logR(x) = nF_n(x) log \frac{F_n(-x) + 1 - F_n(x-)}{2F_n(-x)} + n(1 - F_n(x-)) log \frac{F_n(-1) + 1 - F_n(x-)}{2(1 - F_n(x-))}.$$

Sekojošs rezultāts ir spēkā

Teorēma 2.3. [31][270. lpp] Ja F ir nepārtraukta sadalījuma funkcija un ir spēkā hipotēze par simetriju, tad

$$T_n = -2 \int_0^{\infty} logR(x) d \left\{ F_n(x) - F_n(-x) \right\} \to_d \int_0^1 \frac{W^2(t)}{t} dt,$$

kur W-standarta Brauna kustība jeb standarta Vīnera process. Minētajā publikācijā ar līdzīgu pieeju risinātas problēmas testam par neatkarību kā arī par eksponencialitāti.

Līdz šim tika aplūkoti piemēri empīriskajiem procesiem, kad tie tiecas uz citu sadalījumu - ne Brauna tiltu, kā arī gadījumus, kad rezultātā tiek iegūta izteiksme, kura satur Brauna tiltu vai Brauna kustību. Nobeigumā jāmin, ka liela daļa rezultātiem, kas sastopami van de Gīras grāmatā [14], iegūti izmantojot tieši Vapnika-Červonenkis teoriju un entropijas rezultātus. Minētajā grāmatā galvenā uzmanība vērsta uz novērtējumu konverģences ātrumu.

Lai gan lielākā daļa šāda tipa testu iegūta 20. gadsimta otrajā pusē, tomēr jaunu testu izveide joprojām turpinās. Nesen 2011. gadā izveidots uz empīrisko procesu pieejas balstīts tests procesa stacionaritātes pārbaudei [32]. Tiek definēts mērs, lai novērtētu attālumu no stacionaritātes:

$$D := sup_{(v,\omega)\in[0,1]^2}|D(v,\omega)|,$$

kur

$$D(v,\omega) := \frac{1}{2\pi}(\int_0^v \int_0^{\pi\omega} f(u,\lambda)d\lambda du - v\int_0^{\pi\omega}\int_0^1 f(u,\lambda)dud\lambda),$$

kur $f(u,\lambda)$ apzīmē laikā mainīgo spektrālo blīvumu. D ir 0, ja process ir stacionārs.

Nobeigums

Izstrādājot maģistra darbu tika iepazīta metriskās un Vapnika-Červonenkis entropijas teorija kā arī to pielietojumi. Vapnika-Červonenkis teorija mūsdienās ir ļoti svarīga dažādu modeļu sarežģītības novērtēšanā un to lieto arī ārpus empīrisko procesu teorijas, piemēram, ļoti plašs VC teorijas pielietojums ir mašīnu mācīšanas teorijā. Varētu pat teikt, ka šīs teorijas pielietojums saistībā ar empīriskajiem procesiem ir sava veida speciālgadījums.

Diemžēl VC teorijai pat mūsdienās ir vairāk teorētisks nekā praktisks pielietojums. šo situāciju raksturo, piemēram, centieni empīriski novērtēt VC dimensiju. Šajā darbā gan tika sniegts algoritms VC dimensijas novērtēšanai, tomēr joprojām notiek aktīvi pētījumi, kā to uzlabot.

Jāpiemin, ka viena no svarīgākajām šā darba sastāvdaļām ir empīrisko procesu butstrapa pētījums kā arī iegūtais rezultāts, ka parastais neparametriskais butstraps empīrisko procesu gadījumā var nestrādāt un nepieciešams pielietot gludinošo butstrapu. Kā piemēru var minēt aplūkoto lokācijas-skalēšanas modeli, kuram neparametriskais bustraps deva lielākas kritiskās vērtības nekā gludinātais butstraps. Šis fakts ir ļoti svarīgs, jo empīrisko procesu asimptotiskie sadalījumi mēdz būt ļoti sarežģīti un vienīgais veids, kā tos novērtēt no datiem var izrādīties tieši bustraps.

Izmantotā literatūra un avoti

[1] J. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Instituto Italiano degli Attuari*, 4:421–424, 1933.

[2] J. L. Doob. Heuristic approach to the kolmogorov-smirnov theorems. *Annals of Math. Stat.i*, 20(3):393–403, 1949.

[3] M.D Donsker. Justification and extension of doob's heuristic approach to the kolmogorov-smirnov theorems. *Annals of Math. Stat.i*, 23:277–281, 1952.

[4] M. Akahira and K Takeuchi. Bootstrap method and empirical process. *Ann. Inst. Statist. Math.*, 43(2):297–310, 1991.

[5] 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.

[6] J.A. Wellner. Empirical processes in action: Review. *International Statistical Review*, 60(3):247–269, 1992.

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

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

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

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

[11] P. Morters and Y. Peres. Brownian motion, 2008.

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

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

[14] S. van de Geer. *Empirical processes in M-Estimation*. Cambridge University Press, 2000.

[15] V. N. Vapnik. *The nature of Statistical Learning Theory*. Springer, 2000.

[16] M. McDonald, C.R. Shalizi and D.J. Shervish. Time series forecasting: model evaluation and selection using nonparametric risk bounds, 2012.

[17] V. Vapnik, E. Levin and Y.L. Cun. Measuring the vc dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.

[18] X. Shao, V. Cherkassky and W. Li. Measuring the vc dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986, 2000.

[19] G.S. Watson. Goodness-of-fit tests on a circle. *Biometrika*, 48(1):109–114, 1961.

[20] G.V. Martynov. statistical tests based on empirical processes and related questions. *Journal of Soviet Mathematics*, 61(4):2195–2271, 1992.

[21] 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.

[22] 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.

[23] 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.

[24] B. Efron. Boostrap methods: another look at the jacknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[25] 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*, pages 123–137, 2008.

[26] J. Cielēns. Emprisko procesu pielietojums strukturālo attiecību modeļos. Diplomdarbs, Rīga, 2010.

[27] 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.

[28] A. E.. Vyazilov. Empirical processes in the garch(1,1) model and robust estimation of parameters. *Russ. Math. Surv.*, 56:900–991, 2001.

[29] N.H. Chan and S. Ling. Residual empirical process form long and short memory time series. *The Annals of Statistics.*, 38(5):2453–2470, 2008.

[30] R. Kulik and P Lorek. Empirical process of residuals for regression models with long memory errors. iesniegts urn"lam Annals of statistics.

[31] J. Einmahl and S. McKeague. Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2):267–290, 2003.

[32] P. Preuss, M. Vetter and H. Dette. A test on stationarity based on empirical processes, 2011.

A Pielikums

A1. R kods Brauna tilta un Brauna kustības realizāciju iegūšanai

```
#Lai realizētu brauna tilta un Brauna kustības simulācijas, nepieciešams
#programmā R ielādēt bibliotēku e1071

##Attēla 1.1. iegūšanā izmantotas komandas

# Brauna tilta realizācijas

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

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")
```

A2. R kods empīrisko procesu piemēru sagatavošanai

```
# Attēlā 1.2. redzmie piemēri tika iegūti ar zemāk redzamo komandu palīdzību

n<-1000  #izlases apjoms, kuru var mainīt
x<-runif(n)  #ģenerē izlasi
X<-seq(min(x),max(x),by=0.0001)
XX<-ecdf(x)# novērtē empīrisko sadalījuma funkciju
#empīriskā procesa grafiks


#Attēlā 1.3. redzamie empīriskā kvantiļu procesa grafiki iegūti, izmatojot
#zemāk redzamās komandas. Katrai izlasei kostruēts standarta empīriskais
#proess un koriģētais.

# U(0,1) sadalīta izlase

n<-500  #izlases apjoms
x<-runif(n)  #ģenerētā izlase
T<-seq(0.001,0.999,by=0.0001)

plot(T,sqrt(n)*(qunif(T)-quantile(x,probs=T,type=1)),cex=0.2,
xlab=bquote(paste(n==.(n),"; sadalījums U(0,1)")),ylab="",cex.lab=2,type="l")

plot(T,dunif(qunif(T))*sqrt(n)*(qunif(T)-quantile(x,probs=T,type=1)),cex=0.2,


# N(0,1) sadalīta izlase

n<-500  #izlases apjoms
x<-rnorm(n)  #ģenerētā izlase
T<-seq(0.001,0.999,by=0.0001)

plot(T,sqrt(n)*(qnorm(T)-quantile(x,probs=T,type=1)),cex=0.2,
xlab=bquote(paste(n==.(n),"; sadalījums N(0,1)")),ylab="",cex.lab=2,type="l")

plot(T,dnorm(qnorm(T))*sqrt(n)*(qnorm(T)-quantile(x,probs=T,type=1)),cex=0.2,


# Exp(1) sadalīta izlase

n<-500  #izlases apjoms
x<-rexp(n)  #ģenerētā izlase
T<-seq(0.001,0.999,by=0.0001)

plot(T,sqrt(n)*(qexp(T)-quantile(x,probs=T,type=1)),cex=0.2,
xlab=bquote(paste(n==.(n),"; sadalījums Exp(1)")),ylab="",cex.lab=2,type="l")

plot(T,dexp(qexp(T))*sqrt(n)*(qexp(T)-quantile(x,probs=T,type=1)),cex=0.2,
```

A3. R kods bustrapa metožu salīdzināšanai lokācijas-skalēšanas empīriskajam procesam

```
#Simulācijas lokācijas-skalēšanas modeļa empīriskajam procesam

PP_simul<-function(n){
x<-rnorm(n,0,1)
y<-rnorm(n,1,2)

#atkarībā no tā, vai tiek veikta vienkāršā vei saliktā hipotēžu pārbaude,
#parametrus attiecīgi uzdod vai novērtē

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
#s<-0.5
#m<- -0.5
y<-y*s+m #veic otrās izlases transformāciju

XX<-ecdf(x)
T<-seq(0,1,length=1000)
rez<-XX(quantile(y,T,type=1))

#plot(T,rez,type="s")
#points(t,p,type="l",lwd=2,col="blue")

return(rez)
}
T<-seq(0,1,length=1000)
p<-T
#Simulācijas

N<-10000
n<-500
R<-replicate(N,sqrt(n)*max(abs(PP_simul(n)-p)))

sort(R)[c(0.9,0.95,0.99)*N]

hist(R,breaks=sqrt(N),main="Teorētiskā statistikas histogramma",xlab="",
ylab="Biežums",col="grey",cex.lab=1.2)
```

```
# Neparametriskā bustrapa metode lokācijas skalēšanas modelim

PP_boot<-function(x,y){

XX<-ecdf(x)

#atkarībā no tā, vai tiek veikta vienkāršā vei saliktā hipotēžu pārbaude,
#parametrus attiecīgi uzdod vai novērtē

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
#s<-0.5
#m<- -0.5
y<-y*s+m
T<-seq(0,1,length=1000)
rez<-XX(quantile(y,T,type=1))
#plot(T,rez,type="s")
#points(t,p,type="l",lwd=2,col="blue")
return(rez)
}

#Empīriskais PP-grafiks
N<-500
x<-rnorm(N,0,1)
y<-rnorm(N,1,2)
t<-seq(0,1,length=1000)
XX<-ecdf(x)

#atkarībā no tā, vai tiek veikta vienkāršā vei saliktā hipotēžu pārbaude,
#parametrus attiecīgi uzdod vai novērtē

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
#s<-0.5
#m<- -0.5
Y<-y*s+m
p<-XX(quantile(Y,t,type=1))
plot(p,t)

#Butstraps

N<-10000 #butstrapa reižu skaits
n<-500   # butstrapoto izlašu apjoms

R<-replicate(N,sqrt(n)*max(abs(PP_boot(sample(x,n,replace=TRUE),sample(y,n,
replace=TRUE))-p)))
```

```
# Gludinārā butstrapa metode lokācijas skalēšanas empīriskajam procesam

K<-function(x,X,h,t){
#x-punkts
#X-izlase
mean(pnorm((x-X)/h,0,1))-t
}
K<-Vectorize(K,vectorize.args="x")

Uroot<-function(X,h,t){
if (t<1/10^9){t<-1/10^9}
if (t>1-1/10^9){t<-1-1/10^9}
uniroot(K,c(-10^6,10^6),X,h,t)$root
}
Uroot<-Vectorize(Uroot,vectorize.args="t")

inv<-function(u,X){
k<-length(X[X[,2]<u,2])
if(k==0){x<-X[1,1]}
else
{x<-X[k,1]}
}
inv<-Vectorize(inv,vectorize.args="u")

PP_sm_boot<-function(n,X,Y){
u<-runif(n,0,1)
v<-runif(n,0,1)
x<-inv(u,X)
y<-inv(v,Y)

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
#s<-0.5
#m<--0.5
y<-y*s+m
XX<-ecdf(x)
T<-seq(0,1,length=1000)
rez<-XX(quantile(y,T,type=1))
#plot(T,rez,type="s")
#points(t,p,type="l",lwd=2,col="blue")

return(rez)
}

PP_sm_boot(100,X,Y)
```

```
#Gludais PP-grafiks

N<-500
x<-rnorm(N,0,1)
y<-rnorm(N,1,2)
h1<-bw.SJ(x)
h2<-bw.SJ(y)
t<-seq(0,1,length=1000)
T<-seq(min(x,y),max(x,y),length=1000)
X<-matrix(c(T,K(T,x,h1,0)),ncol=2,nrow=1000)
Y<-matrix(c(T,K(T,y,h2,0)),ncol=2,nrow=1000)

s<-(mean(sort(x)*sort(y))-mean(x)*mean(y))/(mean(y^2)-mean(y)^2)
m<-mean(x)-s*mean(y)
#s<-0.5
#m<--0.5
y<-y*s+m

p<-K(Uroot(y,h2,t),x,h2,0)
plot(t,p,type="l")

#Butstraps

N<-10000
n<-500
R<-replicate(N,sqrt(n)*max(abs(PP_sm_boot(n,X,Y)-p)))
sort(R)[c(0.9,0.95,0.99)*N]
```

Maģistra darbs "Empīriskie procesi ar pielietojumiem statistikā" 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: docents Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: Mārcis Bratka

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretāre: Margarita Buiķe _____

(paraksts)