LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

VIENLAICĪGĀS TICAMĪBAS JOSLAS VARBŪTĪBU-VARBŪTĪBU UN KVANTIĻU-KVANTIĻU GRAFIKIEM

Kursa darbs

Autors: **Juris Cielēns**

Stud. apl. jc05001

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2009

Saturs

1. Ievads

Šī darba mērķis ir vienlaicīgo ticamības joslu konstruēšana varbūtību - varbūtību un kvantiļu - kvantiļu grafikiem. Pieņemsim, ka ir dotas divas neatkarīgas izlases: $X_1, ..., X_n$ ar teorētisko sadalījuma funkciju $F_1(x) = P(X_1 \leq x)$ un $Y_1, ..., Y_m$ ar teorētisko sadalījuma funkciju $F_2(y) = P(Y_1 \leq y)$, pieņemsim, ka izlašu elementi ir neatkarīgi un vienādi sadalīti. Darbā tiks aplūkoti varbūtību-varbūtību (PP) un kvantiļu-kvantiļu (QQ) grafiki, kurus var definēt attiecīgi $PP(t) = F_1(F_2^{-1}(t)), 0 < t < 1$ un $QQ(x) = F_1^{-1}(F_2(x)), -\infty < x < \infty$ [1]. Pasaulē plaši tiek pētīts līdzīgs grafiks, ko sauc par ROC līkni un kas tiek definēta $R(t) = 1 - F_1(F_2^{-1}(1 - t))$.

1990. gadā Beirlants un Deheuvels [1] pierādīja PP un QQ empīriskō procesu asimptotisko sadalījumu gadījumā, kad abu izlašu sadalījuma funkcijas sakrīt. Šis rezultāts ir svarīgs, jo dod iespēju grafiski attēlot PP un QQ grafiku vienlaicīgās ticamības joslas pie fiksēta nozīmības līmeņa. Šāds grafiks ir ekvivalents Kolmogorova-Smirnova testam divu izlašu gadījumā.

Lajos Horvats (2008) [2] sniedza metodi, ar kuru tiek konstruētas vienlaicīgās ticamības joslas ROC līknēm. Izmantojot šo metodi šajā darbā tika konstruētas ticamības joslas PP un QQ grafikiem. Nezināmam P-P grafikam tiks apskatīta ticamības joslu konstruēšana,izmantojot gludināto butstrapa metodi. Publikācijā horwath sniegtā gludinātā butstrapa pierādījuma ideja tika izmantota pierādot PP un QQ procesu asimptotiku. Veicot plašu literatūras avotu izpēti šāds pierādījums atrast netika, tika sniegti tikai rezultāti.

Darbs satur trīs nodaļas un pielikumu. Pirmajā nodaļā tiek aplūkoti P-P un Q-Q grafiki un dažas to īpašības. Otrā nodaļa veltīta empīrisko P-P un Q-Q procesu asimptotikai. Trešā nodaļa satur praktisku metodes pielietojumu. Pielikumā ir kompaktdisks, kurā ievietota metodes veikšanai izmantoto programmu bibliotēka, kas konfigurēta tā, ka var tikt instalēta statistikas programmā R.

2. P-P un Q-Q grafiki un empīriskie procesi

2.1. PP grafiks

Praksē parasti nav zināmas doto izlašu sadalījuma funkcijas, tāpēc tiek izmantoti empīriskie novērtējumi.

Definīcija 1. Par empīrisko sadalījuma funkciju sauc funkciju

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

kur X_i ir izlases elements, n - izlases apjoms, I - indikātorfunkcija.

Definīcija 2. Par empīrisko kvantiļu funkciju sauc funkciju

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

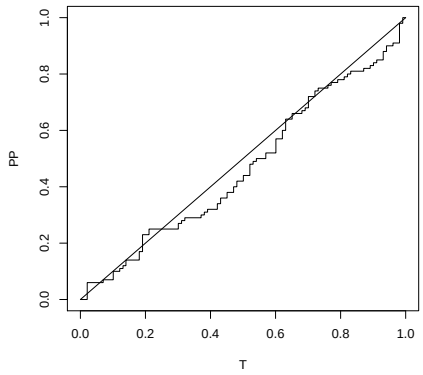
Definīcija 3. [1, 242. lpp.] Par empīrisko P-P grafiku sauc funkciju

$$PP_{nm}(t) = F_{1n}(F_{2m}^{-1}(t)),$$

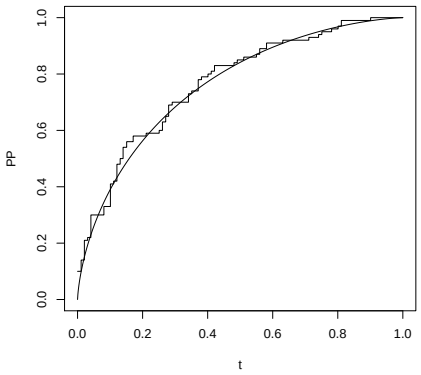
kur $0 < t < 1$, F_{1n} - pirmās izlases empīriskā sadalījuma funkcija, F_{2m}^{-1} - otrās izlases empīriskā kvantiļu funkcija.

Lai salīdzinātu teorētiskos un empīriskos grafikus, tie tiks apkopoti vienā grafikā. Empīriskais P-P grafiks tiks iegūts ģenerējot gadījuma izlases ar apjomu 100.

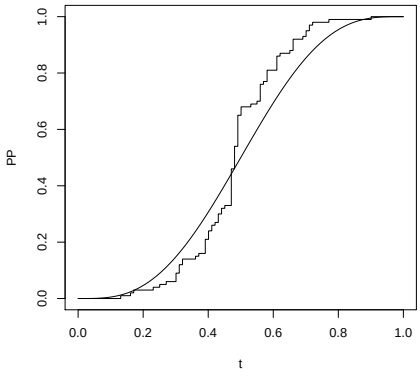
Grafikos gludā funkcija ir teorētiskais P-P grafiks, kāpņveida funkcija - empīriskais P-P grafiks. Ja izlašu sadalījuma likumi ir vienādi, tad empīriskais P-P grafiks tuvs taisnei $y = x$. Ja izlašu sadalījuma likumi ir no vienas klases, bet atšķiras matemātiskā cerība, tad grafiks noliecas virs vai zem diagonāles, atkarībā no tā, vai otrās izlases matemātiskā



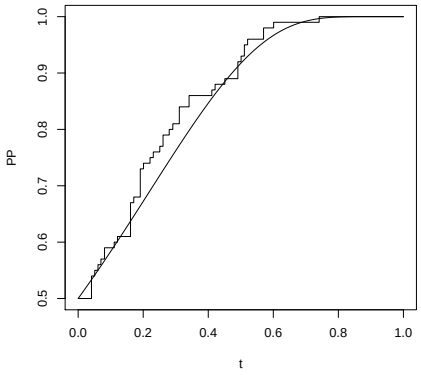
2.1. att. N(0,1) pret N(0,1)



2.2. att. N(0,1) pret N(1,1)



2.3. att. N(0,1)pret N(0,4)



2.4. att. N(0,1) pret χ^2_4

cerība ir lielāka vai mazāka par pirmās izlases matemātisko cerību (attēls 2.2).Ja izlašu sadalījuma likumi ir no vienas klases, bet atšķiras dispersija, grafiks tiek saspiest uz vidu (attēls 2.3). Attēlā 2.4 redzams patvaļīgi izvēlētu izlašu PP grafiks. Grafikos redzams, ka starp teorētisko un empīrisko versiju ir atšķirība.

Definīcija 4. [3, 28. lpp.] Par empīrisko P-P procesu sauc

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

2.2. Q-Q grafiks

Definīcija 5. [1, 242. lpp.] Par empīrisko Q-Q grafiku sauc funkciju

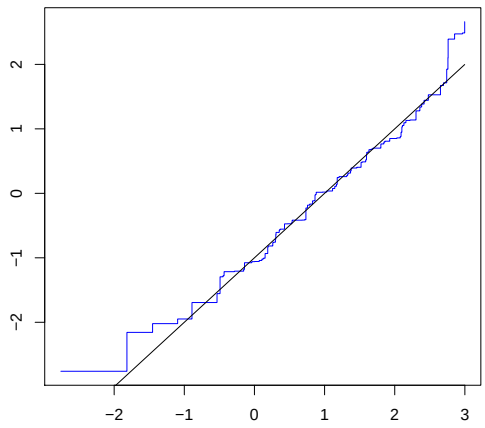
$$QQ_{nm}(t) = F_{1n}^{-1}(F_{2m}(t)),$$

kur $0 < t < 1$, F_{1n}^{-1} - pirmās izlases empīriskā kvantiļu funkcija, F_{2m} - otrās izlases empīriskā sadalījuma funkcija.

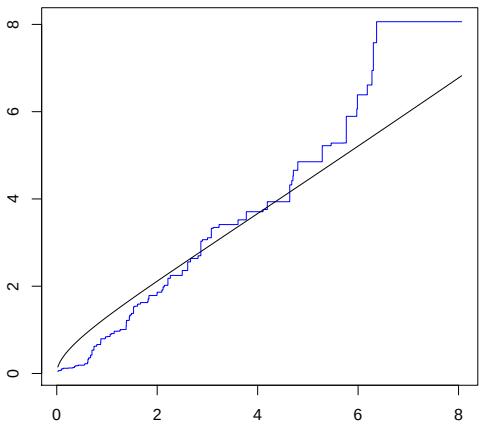
Definīcija 6. [3, 28 lpp.]. Par empīrisko kvantiļu - kvantiļu procesu sauc

$$\delta_{nm}(x) = \sqrt{n} \sup_{-\infty<x<\infty} |f(F_1^{-1}(F_2(x)))(F_{1n}^{-1}(F_{2m}(x)) - F_1^{-1}(F_2(x)))|$$

Atšķirībā no P-P grafika, Q-Q grafiks nav ierobežots. Sekojošajos attēlos redzami QQ grafiku piemēri.



2.5. att. N(0,1) pret N(1,1)



2.6. att. exp(0.5) pret χ_2

2.3. Pamatteorēmas

Teorēma 1 (Glivenko-Kantelli). *Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju $F(x)$, empīrisko sadalījuma funkciju $F_n(x)$, teorētisko kvantiļu funkciju $F^{-1}(t)$,empīrisko kvantiļu funkciju $F_n^{-1}(t)$, teorētisko blīvuma funkciju $f(x)$ un empīrisko blīvuma funkciju $f_n(x)$ tad ir spēkā*

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

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

$$\sup_{0<t<1}(|f_n(x)-f(x)|) \xrightarrow[n\to\infty]{} 0 \ g.d. \ ,$$

kur g.d. nozīmē gandrīz droši.

Teorēma 2. *Pieņemsim, ka $X_1, ..., X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju $F(x)$ un empīrisko sadalījuma funkciju $F_n(x)$, tad*

$$\sup_{-\infty<x<\infty}|\sqrt{n}(F_n(x)-F(x))| \to \sup_{-\infty<x<\infty}|B(F(x))| \ g.d., jeb$$

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

Teorēma 3 (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 4. *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*

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

3. Empīrisko P-P un Q-Q procesu asimptotiskie sadalījumi

Šajā nodaļā tiks aprakstīts, kā iegūt empīrisko P-P un Q-Q procesu asimptotisko sadalījumu. Pārmeklējot vairākus literatūras avotus, norādes, kā iegūt asimptotisko sadalījumu P-P un Q-Q procesiem netika sniegtas. Bieži vien tika parādīti tikai gala rezultāti. Lajosa Horvata publikācijā [2] ir pierādīts,ka asimptotiskais sadalījums ROC līknes empīriskajam procesam sakrīt ar šīs pašas līknes asimptotisko sadalījumu gludinātajam empīrisko procesu. Izmantojot tur sniegtā pierādījuma ideju, tika iegūts pierādījums P-P un Q-Q empīriskajiem procesiem. Gludinātais empīriskais process tiks aplūkots vēlāk.

3.1. Kritiskās vērtības novērtēšana

Definīcija 7. [4, 11. lpp] Par standarta Brauna kustību $\{W(t) : t \geq 0\}$ sauc procesu, kuram :

1) neatkarīgi pieaugumi, t.i. $\forall\ t_i, i = \overline{0,n}\ W(t_i) - W(t_{i-1})$ - neatkarīgi; 2) pieaugumi ir stacionāri, t.i. pieaugums $W(t+h)-W(t)$, $h > 0$, nav atkarīgs no t; 3) procasam $W(t, t \geq 0)$ ir gandrīz droši nepārtrauktas trajektorijas; 4) $W(t+h)-W(t) \sim N(0,h)\ \forall h \geq 0, t \geq 0$

Definīcija 8. par Brauna tiltu sauc procesu $\{B(t) : t \geq 0\}$, kura sadalījums sakrīt ar $B(t) = W(t) - tW(1)$ sadalījumu.

Teorēma 5 (varbūtību - varbūtību process). *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 $F_1(x)$ un $F_2(x)$, tad ir spēkā*

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

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. $\sqrt{n}(F_{1n}F_{2m}^{-1} - F_1F_2^{-1}(t)) = \sqrt{n}(F_{1n}F_{2m}^{-1}(t) - F_1F_{2m}^{-1}(t)) + \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(F_1F_{2m}^{-1}(t) - F_1F_2^{-1}(t))$. Pielietojot Lemmu 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 un Brauna tilts ir vienmērīgi nepārtraukts[2, 1901. lpp], tad

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

no kurienes seko, ka

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

Līdz arto primais saskaitāmais ir novērtēts. Pielietojot Teorēmu 3 otrajam saskaitāmajam, iegūst

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

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 Lemmas 6

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

Līdz ar to, otro saskaitāmo var novērtēt šādi:

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

□

Līdzīgs rezultāts ir spēkā empīriskajam kvantiļu - kvantiļu procesam. Corgo [5, 6. lpp.] aplūko kvantiļu procesus vienai izlasei. Šeit kvantiļu funkcijas vietā ir kvantiļu - kvantiļu grafika funkcija.

Teorēma 6. *(kvantiļu - kvantiļu process)*
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* $F_1(x)$ *un* $F_2(x)$*, 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.*

Pierādījums. Līdzigi kā iepriekš statistika tiks sadalīta divos saskaitāmajos.

$$\sqrt{n} f_1(F_1^{-1}F_2(x))(F_{1n}^{-1}F_{2m}(x) - F_1^{-1}F_2(x)) = \sqrt{n} f_1(F_1^{-1}F_2(x))(F_{1n}^{-1}F_{2m}(x) - F_1^{-1}F_{2m}(x)) + \frac{\sqrt{m}}{\sqrt{n}} \sqrt{m} f_1(F_1^{-1}F_2(x))(F_1^{-1}F_{2m}(x) - F_1^{-1}F_2(x))$$

Pēc Lemmas 6 pirmajam saskaitāmajam ir spēkā

$$\forall \epsilon > 0\ P(\sup_{-\infty<x<\infty} |\sqrt{n} f_1(F_1^{-1}F_2(x))(F_{1n}^{-1}F_{2m}(x) - F_1^{-1}F_{2m}(x)) - \frac{f_1(F_1^{-1}F_2(x))}{f_1(F_1F_{2m}(x))} B^{(n)}(F_{2m}(x))| > \epsilon) = 0 g.d.$$

Izmantojot Teorēmu 1, ir zināms, ka $F_{2m}(x)$ tiecas uz $F_2(x)$ gandrīz droši, kas dod rezultātu

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

Pielietojam otrajam saskaitāmajam Teorēmu 3. Tā kā $(F^{-1}(t))' = \frac{1}{f(F^{-1}(t))}$, tad iegūstam

$$\frac{\sqrt{n}}{\sqrt{m}} \sqrt{m} f_1(F_1^{-1}F_2(x))(F_1^{-1}F_{2m}(x) - F_1^{-1}F_2(x)) = \frac{\sqrt{n}}{\sqrt{m}} \sqrt{m} f_1(F_1^{-1}F_2(x))(F_{2m}(x) - F_2(x)) \frac{1}{f_1(F_1^{-1}(\xi)},$$

kur $\xi = \xi_m(x)$ ir starp $F_{2m}(x)$ un $F_2(x)$. Pēc Teorēmas 1 $\xi_m(x) \to F_2(x)$, kad $m \to \infty$. Līdz ar to ir iegūts sekojošais:

$$\frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}f_1(F_1^{-1}F_2(x))(F_1^{-1}F_{2m}(x) - F_1^{-1}F_2(x)) = \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(F_{2m}(x) - F_2(x))$$

Pēc Teorēmas 4

$$\forall \epsilon > 0 \; P(\sup_{-\infty<x<\infty} |\frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(F_{2m}(x) - F_2(x)) - \frac{\sqrt{n}}{\sqrt{m}}B^{(m)}(F_2(x))| > \epsilon) = 0 \; g.d.$$

Līdz ar to teorēma ir pierādīta.

□

Lai gan skaidri redzams, ka kvantiļu - kvantiļu procesa gadījumā asimptotikā Brauna tilta arguments ir funkcija, tomēr katrai sadalījuma funkcijai ir spēkā $0 < F(x) < 1$ un var uzskatīt, ka funkcijas $F_2(x)$ vietā ir t, kas dod ērtāku skatījumu uz rezultātu. Šādu Brauna tiltu kombināciju var viegli novērtēt ar programmu R.

3.2. Ticamības intervāli P-P un Q-Q procesiem

Zinot procesa asimptotisko sadalījumu, var iegūt kritisko vērtību c pie fiksēta nozīmības līmeņa α. Ticamības joslas konstruē šādi: 1) P-P grafikam:

$$P(\Delta_{nm}(t) \leq c) = \alpha,$$

$$P(\sqrt{n}\sup_{0<t<1}|(F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t)))| \leq c) = \alpha,$$

$$\forall\ 0 < t < 1 \; P(\sqrt{n}|(F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t)))| \leq c) = \alpha,$$

$$\forall\ 0 < t < 1 \; P(-\frac{c}{\sqrt{n}} \leq F_{1n}(F_{2m}^{-1}(t)) - F_1(F_2^{-1}(t)) \leq \frac{c}{\sqrt{n}}) = \alpha,$$

$$\forall\ 0 < t < 1 \; P(F_{1n}(F_{2m}^{-1}(t)) - \frac{c}{\sqrt{n}} \leq F_1(F_2^{-1}(t)) \leq F_{1n}(F_{2m}^{-1}(t)) + \frac{c}{\sqrt{n}}) = \alpha.$$

2) Līdzīgi iegūst ticamības joslu Q-Q grafikam, rezultāts ir šāds:

$$\forall\ x \; P(F_{1n}^{-1}(F_{2m}(x)) - \frac{c}{\sqrt{n}f(F_1^{-1}(F_2(x)))} \leq F_1^{-1}(F_2(x)) \leq F_{1n}^{-1}(F_{2m}(x)) + \frac{c}{\sqrt{n}f(F_1^{-1}(F_2(x)))}) = \alpha.$$

Galvinais iegūto ticamības joslu pielietojums ir hipotēžu pārbaudē, lai noteiktu vai divas izlases ir sadalītas pēc viena sadalījuma likuma.

$H_0 : F_1(x) = F_2(x)$

$H_1 : F_1(x) \neq F_2(x)$

Pie nulles hipotēzes empīriskajam varbūtību - varbūtību procesam ir spēkā:

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

Kvantiļu - Kvantiļu procesam ir spēkā:

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

Tātad procesa apimptotiskais sadalījums ir divu Brauna tiltu lineāra kombinācija. šī kritiskā vērtība tika izrēķināta veicot brauna tiltu simulācijas un tika iegūts, ka pie $\alpha = 0.95$ $c = 1.892441$. Lai noraidītu hipotēzi ie pietiekami, lai kaut vienā punktā teprētiskais PP vai QQ grafiks izietu ārpus ticamības joslas

3.3. Empīrisko funkciju gludināšana

Iepriekš tika iegūti asimptotiskie sadalījumi (Teorēmas 5 un 6) empīriskajiem varbūtību - varbūtību un kvantiļu - kvantiļu grafikiem, tomēr parādās jauna problēma. Asimptotiskais rezultāts sevī ietver izlašu teorētiskās sadalījuma, kvantiļu un blīvuma funkcijas dažādās kombinācijās. Praksē šīs funkcijas nav zināmas, tāpēc tās jānovērtē no datiem. Metode, kura tiks aplūkota ir gludināšana ar kodoliem. Pieņemsim, ka mums ir dota izlase $X_1, ..., X_n$, kur n - izlases apjoms un $x \in (-\infty; \infty)$.

Definīcija 9. Funkciju k $:R \to R^+$ sauc par kodolu, ja:

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

Biežāk izmantotās kodolfunkcijas:

1. Vienmērīgais(Uniform): $k(u) = \frac{1}{2}I_{|u|\leq 1}$, kur I ir indikatorfunkcija;
2. Epanečnikova (Epanechnikov): $k(u) = \frac{3}{4}(1 - u^2)I_{|u|\leq 1}$;
3. Gausa (Gaussian): $k(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$;

Definīcija 10. [6, 684. lpp] Pieņemsim, ka k ir kodols un h - zināma konstante, kas atkarīga no izlases, tad par gludināto blīvuma funkciju cauc funkciju

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

Definīcija 11. 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 12. 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 turpmāk tiks izmantots Gausa kodols.

3.4. Butstraps

Pašlaik ir zināma aplūkoto procesu asimptotika, zināms, ka tā satur attiecīgo izlasi raksturojošās funkcijas, kā arī zināms, ka varam tās novērtēt ar kodolu palīdzību. Nākas saskarties ar nākamo problēmu - novērtējumi ir pārāk neprecīzi. Gludinot ar kodoliem vienu pašu funkciju, tiek iegūts labs rezultāts, bet, gludinot funkciju kompozīciju, pieļautās kļūdas summējas. Publikācijā [2], kur līdzīgi tiek novērtēta asimptotika ROC līknēm, kodolu gludināšanas metode tiek kombinēta ar butstrapa metodi, lai novērtētu statistikas kritisko vērtību.

Butstrapa metode ir daudzu citu izlašu ģenerēšana no dotās izlases un statistikas aprēķināšana iegūtajām papildizlasēm. To veic šādi:

1. Pieņemsim, ka mums ir dota izlase ar apjomu n, kā arī zināmas varbūtības, ar kādām var nejauši izvēlēties katru no izlases elementiem;

2. Tiek ģenerēti n gadījuma lielumi no dotās izlases;

3. Izmantojot iegūto jauno izlasi, tiek aprēķināta pētāmā statistika;

4. Procedūru atkārtojot tiek iegūts vektors, kas sastāv no aprēķinātajām statistikām. Sakārtojot vektoru augošā secībā, nosakam statistikas kritisko vērtību pie izvēlēta nozīmības līmeņa.

Realizējot šo izdeju programmā R nācās saskarties ar problēmu, kas saistīta ar to, kā tieši iegūt šīs papildizlases. Pirmais variants ir ņemt elementus no dotās izlases ar vienādu varbūtību. Rezultātā tek iegūta jauna izlase, kas sastāv tikai no dotajā izlasē esošajiem datiem, turklāt empīriskā sadalījuma funkcija var nebūt tuva dotās izlases empīriskajai sadalījuma funkcijai. Otrs variants ir novērtēt precīzi katra elementa svaru albilstošajā izlasē un ģenerējot palildizlases izmantot šo svaru vektoru. Rezultātā tiek iegūta izlase, kuras empīriskā sadalījuma funkcija ir tuvāka dotās izlases empīriskajai sadalījuma funkcijai. Tomēr pastāv liela iespēja izlasēm atkārtoties (tas pats ir arī pirmajā variantā), tātad sakārtotas statistikas vērtības nedod vektoru, kas būtu stringri monotons, bet gan tādu, kas gabaliem konstants. Metode, kuru izmantojis Lajos Horvats [2, 1901. lpp] ir labāks variants, kā risināt problēmu. Metode tiks aprakstīta pa soļiem:

1. Pieņemsim, ka ir dotas izlases $X_1, ..., X_n$ un $Y_1, ..., Y_m$ ar empīriskajām sadalījuma funkcijām attiecīgi $F_{1n}(x)$ un $F_{2m}(x)$;

2. Tiek ģenerētas divas neatkarīgas un vienmērīgi sadalītas izlases $U_1^1, ..., U_n^1$ un $U_1^2, .., U_m^2$;

3. Izmantojot gludināto kvantiļu funkcijas (atkarīgas no dotajām izlasēm), iegūstam divas jaunas izlases. Sauksim tās par butstrapotajām izlasēm ar elementiem attiecīgi
$X_i^* := \hat{F}_{1n}^{-1}(U_i^1), i = 1, ..., n$ un $Y_i^* := \hat{F}_{2m}^{-1}(U_i^2), i = 1, ..., m$;

4. Tiek aprēķināta statistika $\sup_{0<t<1} |\sqrt{n}(F_{1n}^* F_{2n}^{*-1}(t) - \hat{F}_{1n}\hat{F}_{2m}^{-1}(t))|$, kur ar $*$ apzīmētas butstrapa izlases raksturojošās empīriskās funkcijas, ar $\wedge$ apzīmētas dotās izlases raksturojošās gludinātās funkcijas. Jau minētajā publikācijā [2] ir pierādījums, ka ROC līknēm šīs statistikas asimptotika sakrīt ar attiecīgā empīriskā procesa statistiku ;

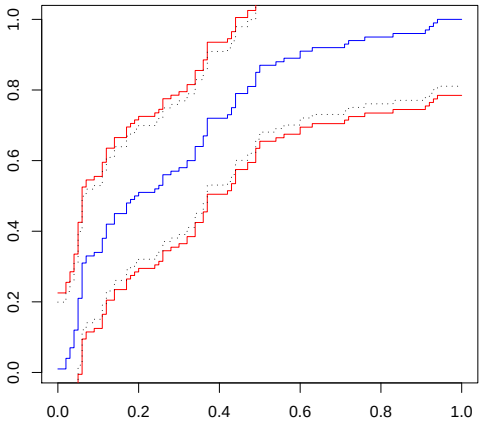
5. Procedūru atkārtojam 1000 reizes, katru reizi saglabājot statistiku. Iegūtās statistikas sakārtojam augošā secībā un par kritisko vērtību ņemam to elementu, aiz kura pa labi atrodas $1 - \alpha\%$ kritisko vērtību (α - sākotnēji notaiktais nozīmības līmenis).

Līdzīgi jārīkojas arī Q-Q procesa gadījumā.

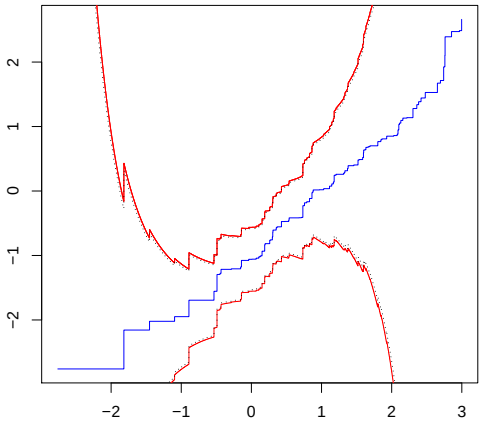
4. Rezultātu pielietojums

4.1. Grafiskie rezultāti

Iepriekšējās nodaļā tika aprakstīts, kā iegūt P-P un Q-Q grafiku, ja dotas divas izlases, kā arī, kā iegūt vienlaicīgās ticamības joslas šiem grafikiem. Šajā nodaļā aplūkosim dažus rezultātu piemērus. Vispārīgā gadījumā grafikā redzama joslu, kurā atrodas teorētiskais P-P vai Q-Q grafiks ar sākumā uzdoto varbutību α. Attēlos 4.1 un 4.2 redzami attiecīgi P-P un Q-Q grafiki izlasēm, kuras satur neatkarīgi ģenerētus gadījuma lielumus. Dotas izlases $X \sim N(0,1)$ un $Y \sim N(1,1)$, izlašu apjomi $n = 100$, $\alpha = 0.95$, butstrapojumu skaits $N = 1000$. Ar nepārtrauktu līniju uzzīmēta vienlaicīgā ticamības josla, novērtēta no izlasēm, ar pārtraukto - pie hipotēzes H_0

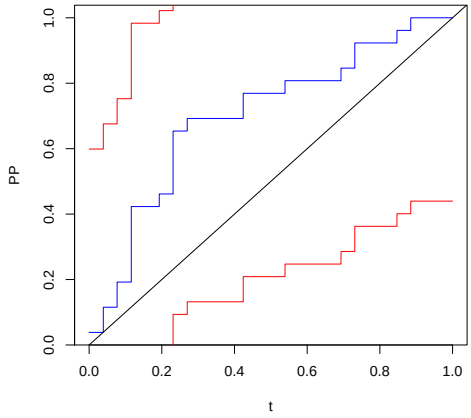


4.1. att. empīriskais P-P grafiks

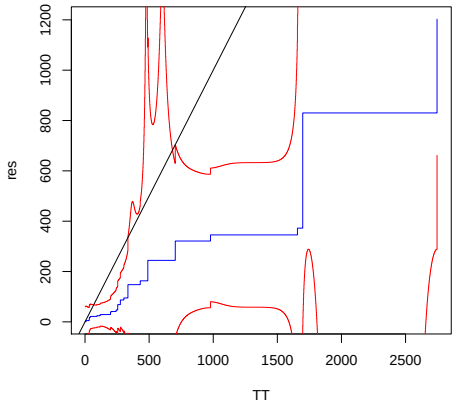


4.2. att. empīriskais Q-Q grafiks

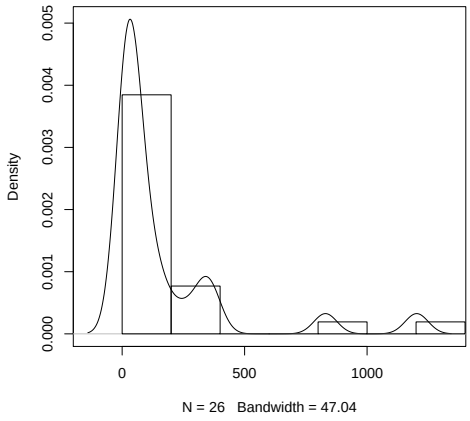
Veiksim hipotēžu pārbaudi reāliem datiem. Izmantosim datus no [7, 420. lpp.]. Eksperimetā mākoņi tika apsmidzināti ar sudraba jodītu un pārbaudīts, vai novērojams nokrišņu pieaugums. Pirmā izlase satur nokrišņu daudzumu pirms apstrādes, otrā - pēc apstrādes. Izmantojot šos datus, veicam divus testus. Tiek konstruētas ticamības joslas P-P un Q-Q grafikiem pie nozīmības līmeņa $\alpha = 0.95$. Rezultāti redzami attēlos. Hipotēze tiek noraidīta, ja attēlos redzamā taisne iziet ārpus joslas kaut vienā punktā. Ļoti interesanta is QQ grafika ticamības josla, to ietekmē pirmās izlases blīvuma funkcija.



4.3. att. PP tests



4.4. att. QQ tests



4.5. att.: pirmās izlases histogramma un novērtētā blīvuma funkcija

4.2. Metodes pārbaude

Lai pārbaudītu metodi tika veikta pārklājumu precizitātes pārbaude. Metodes ideja ir veikt simulācijas, konstruēt attiecīgos grafikus un to ticamības joslas un veikt hipotēžu pārbaudi. Nenoraidīto gadījumu proporcija parāda, cik precīzi strādā modelis. Pie dažādiem izlašu apjomiem tika ģenerētas gadījuma lielumu izlases ar vienādiem sadalījumiem N(0,1). QQ grafika gadījumā ticamības intervāla konstruēšanā tika izmantota teorētiskā blīvuma funkcija. Metode tika pārbaudīta pie hipotēzes par vienādām sadalījuma funkcijām. Rezultāts redzams tabulā.

4.1. tabula Pārklājumu precizitātes pārbaude

	PP grafiks			QQ grafiks		
	90%	95%	99%	90%	95%	99%
$n = 10$	0,955	0,972	0,997	0,843	0,895	0,952
$n = 20$	0,927	0,967	0,992	0,854	0,904	0,963
$n = 50$	0,932	0,967	0,995	0,858	0,92	0,971
$n = 100$	0,921	0,957	0,985	0,893	0,943	0,981
$n = 200$	0,916	0,957	0,989	0,898	0,945	0,991
$n = 500$	0,901	0,951	0,993	0,897	0,955	0,989
$n = 1000$	0,904	0,952	0,992	0,898	0,952	0,988

5. Nobeigums

Metodes precizitāte tika pādbaudīta tikai pie hipotēzes H_0, tā strādā labo un ja izlašu apjomi tiecas uz bezgalību, tiek iegūts precīzs rezultāts. Tika veikti mēģinājumi pārbaudīt, cik ātri konverģē process vispārīgā gadījumā, tomēr nācās saskarties ar dažām problēmām. Ja tiek veikta pārklājumu precizitātes vārbaude vispārīgā gadījumā, ir jāizvēlas vai nu fiksēt joslas platumu h, kad tiek gludinātas funkcijas, vai arī to katru reizi aprēķināt. Veicot pārbaudi ir jāzin arī teorētiskais grafiks. Protams, to var definēt atsevišķi un pārbaudīt, vai tas atrodas ticamības joslā. Tomēr nav skaidrs vai konstruētajai ticamības joslai jāiekļauj šis teorētiskais grafiks, jo veicot 'pārbaudi netika iegūts labs rezultāts. 95% ticamības joslas teorētisko grafiku iekļāva tikai 90% gadījumu.

Kā bija redzams grafiskajā piemērā ar reāliem datiem, izlases ne vienmēr ir pateicīgas analīzei un lēmumu pieņemšanai. PP grafika gadījumā nevar noraidīt hipotēzi, ka sadalījuma funkcijas sakrīt, toties QQ grafika ticamības josla parāda, kakvantiļu funkcijas izlasēm atšķiras.

pētot aprakstīto metodi tika izveidots programmu kopums, ar kura palīdzību ērtāk apstrādāt dotos datus. Darba gaitā tika iepazīts veids, kā labāk saglabāt izveidotās programmas vēlākai lietošanai. Apgūstot programmā R instalējamo bibliotēku izveidi, metodi izpildošās programmas tika apkopotas vienā arhīvā, kuru iespējams instalēt programmā R un lietot kā pārējās standarta bibliotēkas.

Izmantotā literatūra un avoti

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

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

[3] 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.

[4] Peres Y. Morters, P. Brownian motion, 2009.

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

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

[7] Kraaikamp C. Lopuhaa H.P. Meester L.E. Dekking, F.M. *A Modern Introduction to Probability and Statistics*. Springer, 2005.

Kursa darbs “Vienlaicīgās ticamības joslas varbūtību-varbūtību un kvantiļu-kvantiļu grafikiem” 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:

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts kursa darbs gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)