LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

BUTSTRAPA METODE SADALĪJUMA PĀRBAUDES TESTIEM

DIPLOMDARBS

Autors: **Anna Trautmane**

Stud. apl. at06013

Darba vadītājs: asoc.prof. Dr.math. Jānis Valeinis

RĪGA 2011

Anotācija

Darba tēma ir butstrapa metode sadalījuma pārbaudes testiem. Tiek apskatīta šīs metodes ideja un pierādījumi. Ar piemēru palīdzību gan neparametriskā, gan parametriskā butstrapa metode pielietota statistikas sadalījumu meklēšanai. Darba nobeigumā ir apskatīts butstrapa metodes pielietojums sadalījuma pārbaudes testiem kā Kolmogorova - Smirnova un Neimaņa. Apskatīts arī bloku butstrapa metode atkarīgiem datiem.

Atslēgas vārdi: datu pārkārtošna, neparametriskais butstraps, parametriskais butstraps, bloku butstraps

Abstract

The subject of this thesis is bootstrap methods for goodness-of-fit tests. The idea and the proof of bootstrap method are examined. With examples are shown non-parametric and parametric methods for define the distribution of statistic. Finally, bootstrap methods are introduced for goodness of fit tests like Kolmogorov - Smirnov and Neymann. The block bootstrap methods are also examined.

Keywords: resampling, non-parametric bootstrap, parametric bootstrap, block bootstrap

Saturs

Apzīmējumi

F_n	empīrsikā sadalījuma funkcija
F	teorētiskā sadalījuma funkcija
$\xrightarrow{g.d.}$	gandrīz droša konverģence
$\xrightarrow{d}$	konverģence pēc sadalījuma
CRT	Centrālā robežteorēma
$B(t)$	Brauna tilts
$W(t)$	Brauna kustība
$N(\mu, \sigma^2)$	Normāli sadalīts gadījuma lielums ar vidējo vērtību μ un dispersiju σ^2
$LogN(\mu, \sigma^2)$	LogNormāli sadalīts gadījuma lielums ar parametriem μ un σ^2
$Exp(\lambda)$	Eksponenciāli sadalīts gadījuma lielums ar parametru λ
χ^2_k	χ^2 sadalīts gadījuma lielums ar k brīvības pakāpēm
$U(a, b)$	Vienmērīgi sadalīts gadījuma lielums intervālā (a, b)
$NegB(r, q)$	Negatīvi binomiāli sadalīts gadījuma lielums ar parametriem r un q
$Po(\lambda)$	Puasona sadalīts gadījuma lielums ar parametru λ
ARMA	Autoregresīvs slīdošā vidējā process
AR	Autoregresīvs process

Ievads

Darbā aplūkota butstrapa metode, kuru piedāvāja un noformulēja Efrons (1979) ([1]). Datoru attīstība tieši ietekmē statistikas nozari, kurā plaši tiek izmantota datoru iespējas. Ar datoru palīdzību ir iespējams pētīt lielas un sarežģītas datu kopas. Viens no Efrona galvenajiem ieguldījumiem bija parādīt, ka butstrapa metodi var kombinēt ar mūsdienu datoru skaitļošanas jaudu un iespējām. Tāpēc butstrapa metodes pievilcīgums ir saistīts mūsdienu datoru ietekmi uz statistikas attīstību. Efrons arī pirmoreiz pielietoja vārdu "bootstrap", kas ir saistīts ar amerikāņu versiju vienam no stāstiem par baronu fon Minhauzenu, kurā paziņoja, ka pats sevi izvilcis no purva, velkot sevi aiz kurpju (*angļu val. - boot*) auklām (*angļu val. - starp*). Eiropas versijā viņš izvilka sevi aiz matiem. Tas arī atspoguļo šīs metodes būtību, kur asimptotika tiek aizstāta ar simulācijām vai butstrapa metodēm.

Viens no šī darba mērķiem ir iepazīties ar butstrapa metodi, izpētīt tā būtību un darbības procesu. Šajā darbā iepazīsimies gan ar neparametrisko butstrapu, gan ar parametrisko butstrapu. Pierādīsim neparametrisko butstrapa metodi statistikai $\sqrt{n}(\bar{X} - \mu)$ izmantojot Kolmogorova un Vaseršteina metriku. Pielietosim abus butstrapa veidus centrētai vidējai vērtībai un Kolmogorova - Smirnova testam nepārtrauktiem sadalījumiem, šo pielietojumu aprakats ir atrodams Dekking, Kraaikamp un Lopuhaa grāmatā ([2]).

Otrs mērķis ir butstrapa metodes pielietojums Kolmogorova - Smirnovas testam pie diskrētiem sadalījumiem. Diskrētu sadalījumu gadījumā tiek pielietots paramertriskais butstraps un pārbaudīta saliktā hipotēze (skatīt [3]). Ar butstrapa metodi tiek meklēta statistikas kvantile, pēc kuras nosakām, vai hipotēzi pieņemt vai noraidīt. Praktiski apskatīsim negatīvo binomiālo sadalījumu un Puasons sadalījumu.

Kā pēdējais darba mērķis ir butstrapa metods pielietošana Neimaņa testam neatkarīgiem un atkarīgiem datiem. Jāpiebilst, ka Neimaņa testam ir sarežģīts robežsadalījums, kas satur novērtētus parametrus. Neatkarīgu datu gadījumā pielietosim neparametrisko butstrapa metodi, kura atkarīgiem datiem dod aplamus rezultātus. Bet tā kā ar Neimaņa testu iespējams pārbaudīt atkarīgus datus, tad pielietosim bloku butstrapa metodes tādas kā slīdošais bloku butstraps un nešķeļosais bloku butstraps. Butstrapa pielietošana ļauj izvairīties no datu kovariāciju novērtēšanas.

Darba 1., 2, un 2. nodaļā ir butstrapa idejas apraksts, teorētiski pierādījumi un

vienkārši piemēri, lai labāk saprastu butstrapa metodi. 4. nodaļā ir aprakstīts butstraps Kolmogorova - Smirnova testam, kā arī parādīts tā praktiskais pielietojums gan nepārtrauktu, gan diskrētu sadalījumu gadījumā. Butstrapa metode Neimaņa testam ir aprakstīta darba 5. nodaļā.

Praktiskā daļa tiks veikta ar programmas R palīdzību.

1. Butstrapa ideja

Butstrapa metodes būtība ir konstruēt attiecību starp populāciju un izlasi. Tā pieņem doto izlasi $X_1, X_2, \ldots, X_n$ kā labi reprezentējošu priekš populācijas, un veido butstrapa izlases $X_1^*, X_2^*, \ldots, X_n^*$ kā realizācijas no populācijas. Visbiežāk butstrapa izlases elementi tiek iegūti kā gadījuma izlase no $X_1, \ldots, X_n$ ar vienādām varbūtībām (neparametriskais butstraps). Taču ir iespējams konstruēt parametrisku butstrapa metodi, kad butstrapa izlase tiek veidota pēc kāda fiksēta sadalījuma likuma $F_{\hat{\theta}}$, kur $\hat{\theta}$ piemēram ir $\{\bar{x}, s^2\}$, ja $F = N(\mu, \sigma^2)$.

Vispirms aprakstīsim *neparametriskā jeb empīriskā butsrapa* ideju sīkāk. Pieņemsim, ka $X_1, X_2, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti (turpmāk iid) ar $X_1 \sim F$ un $T(X_1, X_2, \ldots, X_n, F)$ ir kāds funkcionālis, piemēram, $T = T(X_1, X_2, \ldots, X_n, F) = \sqrt{n}(\bar{X}_n - \mu))/\sigma$, kur $\mu = \mathbb{E}_F(X_1)$ un $\sigma^2 = D_F(X_1)$. Bieži sastopama statistikas problēma ir T sadalījuma atrašana, tas ir atrast $P_F(T(X_1, X_2, \cdots, X_n, F) \leq t)$. Butstrapa metodes ideja ir sekojoša: ģenerēt daudz izlašu no dotās izlases un aproksimēt statistikas T sadalījumu. Tas ir, aproksimēt statistikas $T(X_1, X_2, \cdots, X_n, F)$ varbūtību sadalījumu ar statistikas $T(X_1^*, X_2^*, \ldots, X_n^*, F_n)$ varbūtību sadalījumu, kad gadījuma izlasi $X_1, X_2, \ldots, X_n$ no F tiek aizvietota ar butstrapa gadījuma izlasi $X_1^*, X_2^*, \ldots, X_n^*$ no empīriskās sadalījuma funkcijas

$$F_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

kur

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

Apzīmēsim iegūtās butstrapa izlases $\{X_{11}^*, \ldots, X_{1n}^*\}$, $\{X_{21}^*, \ldots, X_{2n}^*\}, \ldots,$ $\{X_{B1}^*, X_{B2}^*, \ldots, X_{Bn}^*\}$, kur B apzīmē butstrapoto izlašu skaitu. Statistikā funkcionāļa T sadalījumu punktā t, tas ir, $P_F(T \leq t)$ aproksimē ar $\{j \text{ skaits} : T_j^* \leq t\}/B$, kur $T_1^*, T_2^*, \ldots, T_B^*$ apzīmē statistikas T vērtības B dažādām butstrapa izlasēm. Iepriekš apskatītajam piemēram attiecīgā butstrapotā statistika ir $T(X_1, \ldots, X_n, F) = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)/S$, kur S apzīmē izlases (empīrisko) dispersiju.

Aprakstīsim arī *parametriskā butstrapa* ideju. Pieņemsim, ka mēs uzskatām, ka iegūtai izlasei ir kāds parametrisks sadalījums. Šādā gadījumā sadalījuma funkcija ir pilnībā noteikta ar parametru vai vektoru ar parametriem $\theta : F = F_\theta$. Tad sadalījuma funkcija F nav jānovērtē kā pilnībā nezināma funkcija, bet tā vietā pietiek novērtēt parametru (vai

vektoru) θ ar $\hat{\theta}$ un tad novērtēt F ar $\hat{F} = F_{\hat{\theta}}$.

No pirmā acu uzmetiena, butstrapa ideja liekas par vieglu, lai tā strādātu. Bet tiklīdz ir jādefinē butstrapa definīcija kādai konkrētai situācijai, tad tas sagādā grūtības. Tas viss ir atkarīgs no tā, ko mēs sagaidām no butstrapa metodes. Vai mēs gribam, lai novērtē statistikas sadalījumu, vai citu interesējošu parametru. Vēlāk parādīsim gan neparametriskā butstrapa, gan parametriskā butstrapa metodi centrētās vidējās vērtības gadījumā, lai labāk saprastu kā šīs metodes strādā, bet vispirms nākošā nodaļā apskatīsim neparametriskā butstrapa metodes pierādījumus balstoties uz konkrētām metrikām.

2. Butstrapa metode statistikai $\sqrt{n}(\bar{X} - \mu)$

Lai labāk saprastu butstrapa metodi, tad šajā nodaļā apskatīsim pierādījumus, kuri parādīs, ka butstrapa metode strādā statistikai $\sqrt{n}(\bar{X} - \mu)$. Pierādījumos izmantosim Kolmogorova metriku un Vaseršteina metriku. Šiem pierādījumiem izmantosim sekojošas definīcijas un pieņēmumus.

Definīcija 1. ([4], 462. lpp) Pieņemsim, ka $(X_1, \cdots, X_n)$ ir neatkarīgi, vienādi sadalīti, $X_1 \sim F$ un $T(X_1, \ldots, X_n, F)$ ir kāds funkcionālis. Funkcionāļa T neparametrisko butstrapa sadalījumu definē

$$H_{Boot}(x) = P_{\widehat{F}_n}(T(X_1^*, \ldots, X_n^*, \widehat{F}_n) \leq x),$$

kur $(X_1^*, \ldots, X_n^*)$ ir iid izlase apjomā n no $\widehat{F}_n$.

Ar P_* turpmāk apzīmēsim varbūtības attiecībā pret butstrapa sadalījumu.

Definīcija 2. ([4], 463. lpp) Pieņemsim, ka F, G - sadalījuma funkcijas un $\rho(F, G)$-metrika starp tām. Dotai izlasei $X_1, \ldots, X_n \sim F$ un funkcionālim $T(X_1, \ldots, X_n, F)$ definēsim

$$H_n(x) = P_F(T(X_1, \ldots, X_n, F) \leq x)$$

$$H_{Boot}(x) = P_*(T(X_1^*, \ldots, X_n^*, \widehat{F}_n) \leq x).$$

Saka, ka butstraps ir vāji konsistents attiecībā pret ρ funkcionālim T, ja $\rho(H_n, H_{Boot}) \to 0$ pēc varbūtības. Butstraps ir stingri konsistents attiecībā pret ρ ,ja $\rho(H_n, H_{Boot}) \to 0$ gandrīz droši.

Ja $X_1, \ldots, X_n \sim$ iid un $X_1 \sim F$, $\mu = \mathbb{E}(X_1)$, tad butstrapa pierādīšanas tehniskai izpratnei izmatosim statistiku $T_n = \sqrt{n}(\bar{X} - \mu)$. Pieņemsim, ka $H_n(x) = P_F(T_n \leq x)$ un $\widehat{H}_n(x) = \mathbb{P}_*(T_n^* \leq x)$, kur $T_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ ir butstapa novērtējums un $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ ir butstrapa izlase. Mūsu uzdevums ir parādīt, ka $\sup_x |H_n(x) - \widehat{H}_n(x)| \xrightarrow{g.d.} 0$, lai pierādītu ka butstraps strādā.

2.1. Pierādījums izmatojot Kolmogorova metriku

Pirmo metodi pamatoja Singh(1981) izmatojot Kolmogorova metriku. Šis pierādījums ir atrodams DasGupta(2008) grāmatā ([4]).

Definīcija 3. Kolmogorova metrika

$$K(F,G) = \sup_{-\infty<x<\infty} |F(x) - G(x)|.$$

Lai varētu pierādīt butstrapa konsistenci attiecībā pret Kolmogorova metriku mums būs nepieciešami sekojoši fakti.

Teorēma 1. (Polya) *([4]) Ja* $G_n \xrightarrow{d} G$, G *ir nepārtraukta sadalījuma funkcija, tad* $\sup_{-\infty<x<\infty} |G_n(x) - G(x)| \to 0$, *kad* $n \to \infty$.

No Polya [1] teorēmas seko, ka

$$\sup_{-\infty<x<\infty} \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le x \right) - \Phi(x) \right| \to 0, n \to \infty,$$

kur $\Phi(\cdot)$ apzīmē $N(0,1)$ sadalījuma funkciju.

Teorēma 2. (Berry-Esseen) *([4])* $X_1, \cdots, X_n$, *iid,* $\mathbb{E}(X_1) = \mu, D(X_1) = \sigma^2, \beta_3 = \mathbb{E}|X_1 - \mu|^3$. *Tad eksistē universāla konstante* C $(C \approx \frac{4}{5})$, *kas nav atkarīga no* n *vai no* X_i *sadalījuma, ka*

$$\sup_x \left| P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le x \right) - \Phi(x) \right| \le \frac{C\beta_3}{\sigma^3\sqrt{n}}.$$

Teorēma 3. (Zygmund - Macienkiewitz SLSL) *([4])* $Y_1, \ldots, Y_n$, *iid,* $Y_i \sim F, 0 < \delta < 1, \mathbb{E}|Y_1|^\delta < \infty$. *Tad* $n^{-\frac{1}{\delta}} \sum_{i=1}^n Y_i \to 0$ *gandrīz droši.*

Teorēma 4. *([4]) Pieņemsim, ka* $X_1, \ldots, X_n \sim F$ *iid,* $\mathbb{E}_F(X_1^2) < \infty, T(X_1, \ldots, X_n, F) = \sqrt{n}(\bar{X} - \mu)$ *un* $T(X_1^*, \ldots, X_n^*, \widehat{F}_n) = \sqrt{n}(\bar{X}_n^* - \bar{X})$. *Tad*

$$K(H_n, H_{Boot}) \to 0$$

gandrīz droši, kad $n \to \infty$.

Pierādījums. No sākuma pēc Kolmogorova metrikas definīcijas pielietosim to sadalījumiem H_n un H_{Boot}. Ar vienkāršu paņēmienu: pieskaitot un atņemot $\Phi\left(\frac{x}{\sigma}\right)$ un $\Phi\left(\frac{x}{s}\right)$, sadalīsim šo metriku trīs daļās un tad apskatīsim katru daļu atsevišķi.

$$K(H_n, H_{Boot}) = \sup_x |P_F(T_n \le x) - P_*(T_n^* \le x)|$$

$$= \sup_x \left| P_F\left(\frac{T_n}{\sigma} \le \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) + \Phi\left(\frac{x}{s}\right) - P_*\left(\frac{T_n^*}{s} \le \frac{x}{s}\right) \right|$$

$$\le \sup_x \left| P_F\left(\frac{T_n}{\sigma} \le \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) \right| + \sup_x \left| \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) \right| + \sup_x \left| \Phi\left(\frac{x}{s}\right) - P_*\left(\frac{T_n^*}{s} \le \frac{x}{s}\right) \right|$$

$$= A_n + B_n + C_n.$$

$A_n \to 0$ gandrīz droši pēc Polya teorēmas 1., $B_n \to 0$ gandrīz droši, jo $s^2 \to \sigma^2$ gandrīz droši un $\Phi(\cdot)$ ir vienmērīgi nepārtraukta. Tālāk pielietosim Berry-Esseen teorēmu, lai parādītu, ka $C_n \to 0$.

Tā kā butstrapa izlase ir diskrētu gadījumu lielumi izlase ar vienādām varbūtībām $1/n$ katram gadījuma lielumam, tad matemātiskā cerība ir vienāda ar $\mathbb{E}(X^*) = \bar{X} = \frac{\sum_{i=1}^n y_i}{n}$, kas arī ir pielietots šajā gadījumā. No fakta, ka butstrapa izlase ir diskrēta gadījuma lieluma izlase ar vienādām varbūtībām seko, ka $D_{\hat{F}_n}(X_1^*) = s^2 = \frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2$. Modulī esošai izteiksmei pieskaitīsim un atņemsim μ, pēc tam sadalīsim to 2 daļās pēc moduļa īpašības. Ar formulām šie pārveidojumi attēloti sekojoši:

$$\begin{aligned} C_n &\leq \frac{4}{5\sqrt{n}} \frac{\mathbb{E}_{\hat{F}_n}|X_1^* - \bar{X}_n|^3}{(D_{\hat{F}_n}(X_1^*))^{3/2}} = \frac{4}{5\sqrt{n}} \frac{\sum_{i=1}^n |X_i - \bar{X}_n|^3}{ns^3} \\ &\leq \frac{4}{5n^{3/2}s^3} \cdot 2^3 \left(\sum_{i=1}^n |X_i - \mu|^3 + n|\mu - \bar{X}_n|^3 \right) \\ &= \frac{M}{s^3} \left(\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i - \mu|^3 + \frac{|\bar{X}_n - \mu|^3}{\sqrt{n}} \right), \end{aligned}$$

kur $M = \frac{32}{5}$. Tā kā $s \to \sigma > 0, \bar{X}_n \to \mu$, tad $|\bar{X}_n - \mu|^3/(\sqrt{n}s^3) \to 0$ gandrīz droši. No Zygmund - Macienkiewitz teorēmas seko, ka

$$\frac{1}{n^{3/2}} \sum_{i=1}^n |X_i - \mu|^3 = n^{-1/\delta} \sum_{i=1}^n Y_i \to 0$$

gandrīz droši, kad $n \to \infty, \delta = 2/3, Y_i$ ir iid, $\mathbb{E}(Y_i)^\delta = \mathbb{E}_F|X_i - \mu|^{3 \cdot 2/3} = D_F(X_1) < \infty$. Tādejādi $A_n + B_n + C_n \to 0$ un $K(H_n, H_{Boot}) \to 0$ gandrīz droši.

Esam pierādījuši, ka butstrapa metode strādā statistikai $T_n = \sqrt{n}(\bar{X} - \mu)$ attiecībā pret Kolmogorova metriku.

2.2. Pierādījums izmantojot Vaseršteina metriku

Otro metodi pamatoja Bickel un Freedman(1981)([5]) izmatojot Vaseršteina metriku. Šis pierādījums ir atrodams Shao un Tu(1995) grāmatā ([6]).

Definīcija 4. Ja X un Y ir gadījuma lielumi ar sadalījumiem F un G, tad Vaseršteina metrika definēta sekojoši

$$d_r(F, G) = d_r(X, Y) = \inf (\mathbb{E}|X - Y|^r)^{1/r},$$

kur infīmums tiek meklēts visiem kopīgajiem sadalījumiem ar galīgiem F un G.

Tālāk formulēsim Vaseršteina metrikas īpašības kā lemmas, kuras izmantosim butstrapa metodes pierādīšanā.

Lemma 5. *Pieņemsim, ka $X_n \sim F_n$ un $X \sim F$, tad $d_r(F_n, F) \to 0$ tad un tikai tad, ja $X_n \rightsquigarrow X$ un $\int |x|^r dF_n(x) \to \int |x|^r dF(x)$.*

Lemma 6. *Ja $\mathbb{E}(|X_1|^r) < \infty$, tad $d_r(\hat{F}_n, F) \xrightarrow{g.d.} 0$.*

Lemma 7. *Jebkurai konstantei a izpildās $d_r(aX, aY) = |a| d_r(X, Y)$.*

Lemma 8. *Ja $\mathbb{E}(X_j) = \mathbb{E}(Y_j)$ un $\mathbb{E}(|X_j|^2) < \infty, \mathbb{E}(|Y_j|^2) < \infty$, tad*

$$\left(d_2 \left(\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} Y_j \right) \right)^2 \leq \sum_{j=1}^{n} d_2(X_j, Y_j)^2.$$

Pierādījums. Ņemsim vēra, ka infīms eksistē. Pieņemsim, ka (U_j, V_j) ir neatkarīgi un $\mathbb{E}(|X_j - Y_j|^p)^{1/p} = d_p(X_j, Y_j)$.
Izmatosim Minkovski nevienādību

$$d_2(\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} Y_j)^2 \leq \mathbb{E}\left(\left\langle \sum_{j=1}^{n} (X_j - Y_j), \sum_{j=1}^{n} (X_j - Y_j) \right\rangle \right) = \sum_{j=1}^{n} d_2(X_j, Y_j)^2.$$

Lemma 9. *Ja $\mathbb{E}(X^2) < \infty$ un $\mathbb{E}(Y^2) < \infty$, tad*

$$d_2(X, Y)^2 = (d_2(X - \mathbb{E}(X), Y - \mathbb{E}(Y)))^2 + |\mathbb{E}(X - Y)|^2.$$

Pierādījums. Apzīmēsim $a = \mathbb{E}(X)$ un $b = \mathbb{E}(Y)$. Izvēlēsimies tādus X un Y, ka $\mathbb{E}(|X - Y|^2) = d_2(X, Y)^2$. Ir zināma sekojoša izteiksme

$$\mathbb{E}(|(X - a) - (Y - b)|^2) = \mathbb{E}(|(X - Y|^2) - |a - b|^2$$

un no iepriekšējās lemmas seko

$$d_2(X - a, Y - b)^2 \leq d_2(X, Y)^2 - |a - b|^2.$$

Tā kā $\mathbb{E}(|(X - a) - (Y - b)|^2) = d_2(X - a, Y - b)^2$, esam pierādījuši šo īpašību.

Izmantojot šīs īpašības, iegūsim:

$$d_2(\hat{H}_n, H_n) = d_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu))$$

Lai iegūtu nākošo izteiksmi, vidējo vērtību izteiksim caur summu (līdzīgi, kā šajā gadījumā $\bar{X} + \frac{n\bar{Y}}{n} = 1/n \sum_{i=1}^{n}(x_i - \bar{Y})$) un pielietosim 3.lemmu

$$= \frac{1}{\sqrt{n}} d_2 \left(\sum_{i=1}^{n} (X_i^* - \bar{X}_n), \sum_{i=1}^{n} (X_i - \mu) \right)$$

Pielietojot 4. lemmu iegūsim

$$\leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} d_2(X_i^* - \bar{X}_n, X_i - \mu)^2}$$

$$= d_2(X_1^* - X_1, X_1 - \mu)$$

Pielietojot 5. lemmu iegūsim

$$= \sqrt{d_2(X_1^*, X_1)^2 - (\mu - \mathbb{E}_* X_1^*)^2}$$

$$= \sqrt{d_2(\widehat{F}_n, F)^2 - (\mu - \bar{X}_n)^2}$$

$$\xrightarrow{g.d.} 0$$

jo $d_2(\widehat{F}_n, F) \xrightarrow{g.d.} 0$ un $\bar{X}_n \xrightarrow{g.d.} \mu$. Esam pierādījuši, ka butstrapa metode strādā statistikai $T_n = \sqrt{n}(\bar{X} - \mu)$ attiecībā pret Vaseršteina metriku.

3. Butstrapu metodes piemēri

Šajā nodaļā apskatīsim piemērus, kas palīdzēs mums labāk saprast butstrapa metodi. Apskatīsim gan parametrisko, gan neparametrisko butstrapu vidējai vērtībai un centrētai vidējai vērtībai. Kā arī nodaļas beigās apskatīsim piemērus, kad butstrapa metode nestrādā.

3.1. Neparametriskai butstraps

Sāksim ar neparametriskā butstrapa metodes atbilstību vidējai vērtībai (skatīt [2]). Pieņemsim, ka dota gadījuma lieluma izlase $X_1, X_2, \ldots, X_n$. Aprēķināsim vidējo vērtību, kas ir viens no iespējamiem gadījuma lielumiem, kas raksturo doto izlasi.

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Mūsu mērķis ir atrast $\bar{X}_n$ sadalījuma funkciju F. Šo sadalījumu meklēsim ar butstrapa metodi, tātad izveidosim gadījuma izlases $X_1^*, X_2^*, \ldots, X_n^*$, kurām aprēķināsim vidējo vērtību un apzīmēsim kā $\bar{X}_n^*$. Ideja ir izmantot $\bar{X}_n^*$ sadalījumu, lai aproksimētu $\bar{X}_n$ sadalījumu. Rodas jautājums, cik laba ir šī aproksimācija.

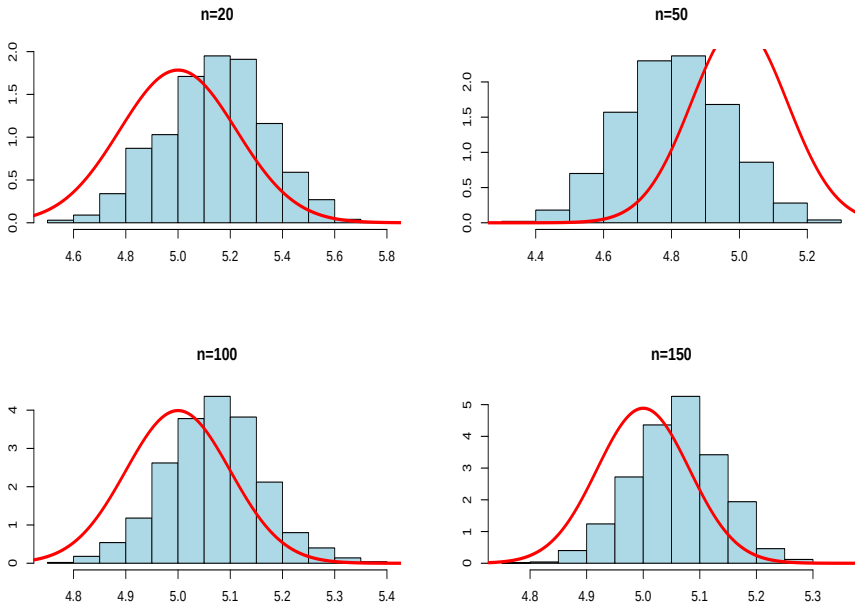
Parādīsim, ka butstrapa aproksimācija $\bar{X}_n$ nestrādā. Apskatīsim izlasi $x_1, x_2, \ldots, x_n$, kas ir realizācija gadījumu izlasei $X_1, X_2, \cdots, X_n$ no $N(\mu, 1)$ sadalījuma. Tādā gadījumā izlases vidējai vērtībai $\bar{X}_n$ ir $N(\mu, 1/n)$ sadalījums. Novērtēsim μ ar $\bar{x}_n$ un aizvietosim gadījuma izlasi no $N(\mu, 1)$ sadalījuma ar butstrapa izlasi $X_1^*, X_2^*, \ldots, X_n^*$ no $N(\bar{x}_n, 1)$ sadalījuma. Attiecīgi butstrapotai izlases vidējai vērtībai $\bar{X}_n^*$ ir $N(\bar{x}_n, 1/n)$ sadalījums. Tāpēc gadījumu lielumu $\bar{X}_n$ un $\bar{X}_n^*$ sadalījuma funkcijas G_b un G_n^*, var tikt noteiktas sekojoši

$$G_n(a) = \Phi(\sqrt{n}(a - \mu)) \text{ un } G_n^*(a) = \Phi(\sqrt{n}(a - \bar{x}_n)).$$

Šajā gadījumā izrādās, ka maksimālais attālums starp abām sadalījumu funkcijām ir vienāds ar

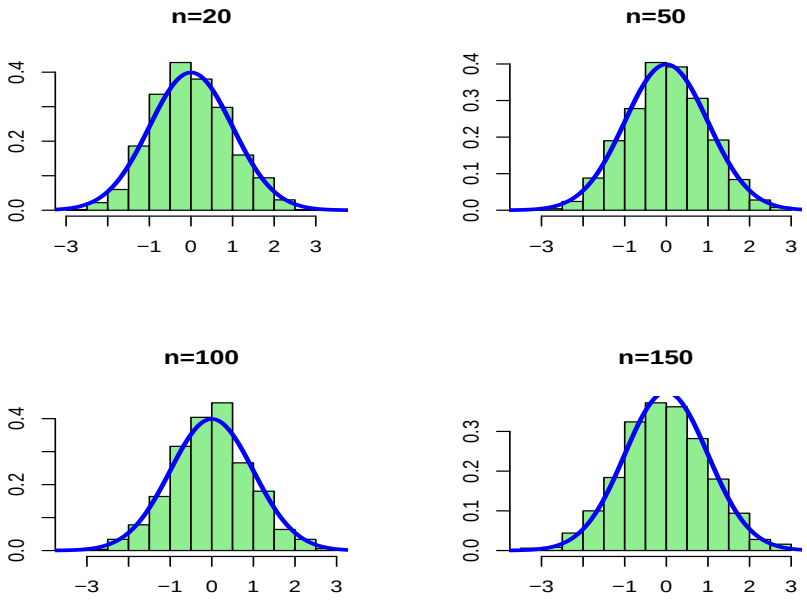
$$2\Phi\left(\frac{1}{2}\sqrt{n}|\bar{x}_n - \mu|\right) - 1.$$

Tā kā $\bar{X}_n$ ir $N(\mu, 1/n)$ sadalījums, tad šī vērtība ir aptuveni vienāda ar $2\Phi(|z|/2) - 1$, kur z ir realizācija gadījuma lielumam $Z \sim N(0, 1)$. Šī sakarība ir vienāda ar 0 tikai tad, ja $z = 0$. Tas nozīmē, kas attālums starp gadījumu lielumiem $\bar{X}_n$ un $\bar{X}_n^*$ vienmēr būs stingri lielāka par 0, pat lieliem n.



1. att. Butstrapa metode $\bar{X}$

Ar datorprogrammas R palīdzību tiek ģenerētas gadījumu lielumu izlases ar sadalījumu $N(5,1)$ un nobutstrapota šīm izlasēm vidējā vērtība $\bar{X}$. Izvēlēsimies 4 iespējamos izlases apjomus $n = 20$, 50, 100 un 150. Pie šiem apjomiem salīdzināsim iegūtos rezultāts, kuri ir redzams 1. attēlā. No histogrammām, kas aproksimētas ar normālā sadalījuma $N(5, 1/n)$ blīvuma funkcijām, redzams, ka butstrapa metode nav bijusi veiksmīga nevienam izlases apjomam. Ar grafiku palīdzību esam parādījuši, ka palielinot izlases apjomu butstrapa novērtējums vidējai vērtībai nemainās.



2. att. Neparametriskā butstrapa metode $\sqrt{n}(\bar{X} - \mu)$

Butstrapa aproksimācija var uzlabot, ja aproksimēsim sadalījumu centrētai izlases vidējai vērtībai $\bar{X}_n - \mu$, kur μ ir sagaidāmā vērtība attiecībā pret F. Attiecīgā butstrapotā versija ir $\bar{X}_n^* - \bar{X}$, kur $\bar{X}$ ir sagaidāmā vērtība attiecībā uz $\hat{F}$. Tā kā gadījuma izlase $X_1, X_2, \ldots, X_n$ ir no $N(\mu, 1)$ sadalījuma, tad $(\bar{X}_n - \mu) \sim N(0, 1/n)$ un $(\bar{X}_n^* - \bar{X}) \sim N(0, 1/n)$. Izdalot šīs izteiksmes ar $\frac{1}{\sqrt{n}}$ iegūsim, ka $\sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$ un $\sqrt{n}(\bar{X}_n^* - \bar{X}) \sim N(0, 1)$. Tātad mums ir jānobutstrapo $T_n = \sqrt{n}(\bar{X} - \mu)$. To, ka butstrapa metode šīs statistikas sadalījuma noteikšanai strādā, pierādījām iepriekšējā nodaļā. Jau iepriekš izveidotajām gadījuma lieluma izlasēm ar sadalījumu $N(5, 1)$, apjomā $n = 20,\ 50,\ 100$ un 150, nobutstrapoim statistiku T_n. Rezultāti ir apskatāmi 2. attēlā, kurā attēlota statistikas T_n aproksimācija ar histogrammu un $N(0, 1)$ blīvuma funkciju. No attēla varam secināt, ka statistikas T_n butstrapa aproksimācijai izpildās CRT, tātad tā uzskatāma par labu.

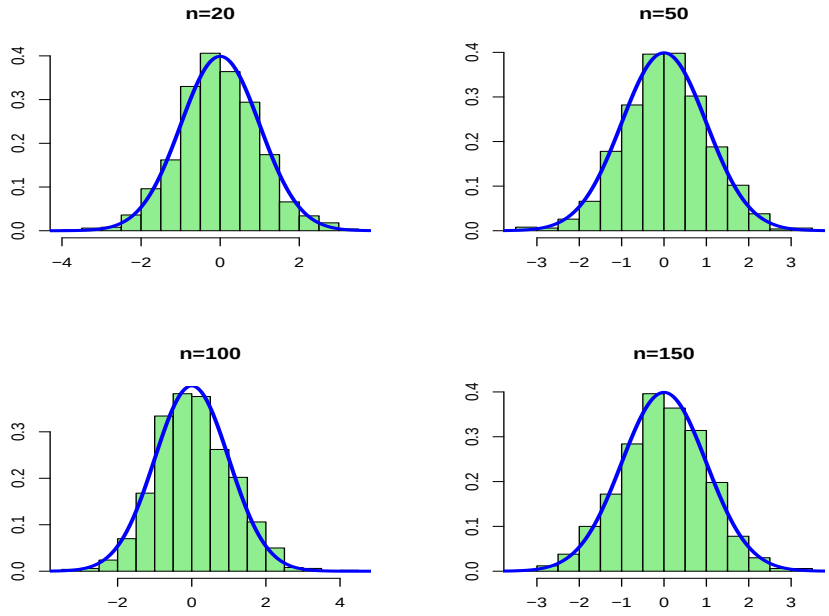
3.2. Parametriskais butstraps

Apskatīsim, ko parametriskais butstraps nozīmē centrētās vidējās vērtības gadījumā. Pieņemsim, ka izlase $x_1, x_2, \cdots, x_n$ ir gadījuma izlases $X_1, X_2, \cdots, X_n$ realizācija no $N(\mu, 1)$ sadalījuma. Novērtēsim μ ar $\bar{x}_n$ un apskatīsim butstrapa izlasi $X_1^*, X_2^*, \cdots, X_n^*$ no $N(\bar{x}_n, 1)$. Uzdevums ir varbūtību sadalījumu statistikai $\bar{X}_n - \mu$ aproksimēt ar satistikas $\bar{X}_n^* - \mu^*$ sadalījumu, kur $\mu^* = \bar{x}_n$. Šo statistiku sadalījumi ir vienādi : $N(0, 1/n)$ sadalījumi. Pārveidosim šīs statistikas izdalot ar $\frac{1}{\sqrt{n}}$, iegūsim $\sqrt{n}(\bar{X}_n - \mu) \sim N(0, 1)$ un $\sqrt{n}(\bar{X}_n^* - \bar{x}_n) \sim N(0, 1)$. Gadījuma izlasēm no sadalījuma $N(5, 1)$ ar dažādiem apjomiem veicam parametriskā butstrapa simulācijas statistikai $T_n = \sqrt{n}(\bar{X}_n - \mu)$, rezultāts ir redzams 3. attēlā, kurā attēlota butstrapotās statistikas sadalījums ar histogrammu, tai pievienota teorētiskā $N(0, 1)$ blīvuma funkcija. Kā redzams, tad parametriskais butstraps labi aproksimē statistikas T_n sadalījuma funkciju.

3.3. Piemēri, kad neparametriskais butstraps nestrādā

Ir piemēri, kuriem butstrapa metode, balstīta uz izlasēm no F_n, nestrādā. Šiem piemēriem ir raksturīgs, ka funkcionālim T_n neizpildās centrālā robežteorēma. Atzīmēsim, dažas konkrētas situācijas, kurās butstrapa metode neatrod funkcionāļa T_n sadalījuma funkciju.

(a) $T_n = \sqrt{n}(\bar{X} - \mu)$, kad $\mathrm{Var}_F(X_1) = \infty$.



3. att. Parametriskā butstrapa metode $\sqrt{n}(\bar{X} - \mu)$

(b) $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$ un $\nabla g(\mu) = 0$.

(c) $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$ un g nav diferencējam punktā μ.

(d) $T_n = \sqrt{n}(F_n^{-1}(p) - F^{-1}(p))$ un $f(F^{-1}(p)) = 0$ vai F ir atšķirīgs labais un kreisais atvasinājums.

Piemērs. Pieņemsim, ka $X_1, X_2, \cdots, X_n$ ir iid un $X_1 \sim F$, $\sigma^2 = \mathrm{Var}_F(X) = 1$. Pieņemsim, ka $g(x) = |x|$ un $T_n = \sqrt{n}(g(\bar{X}) - g(\mu))$. Ja μ patiesā vērtība ir 0, tad $T_n \xrightarrow{d} |Z|$, kur $Z \sim N(0, \sigma^2)$. Lai parādītu, ka butstrapa metode šajā gadījumā nestrādā, nepieciešams pieminēt dažus papildus faktus.

(a) Gandrīz visām virknēm $\{X_1, X_2, \cdots\}$ statistikas $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ nosacītais sadalījums konverģē uz $N(0, \sigma^2)$ pateicoties CRT.

(b) Statistikas $(\sqrt{n}(\bar{X}_n - \mu), \sqrt{n}(\bar{X}_n^* - \bar{X}_n))$ kopējais asimptotiskais sadalījums tiecas uz (Z_1, Z_2), kur Z_1, Z_2 ir iid un $N(0, \sigma^2)$.

Pieņemsim, ka (X_n, Y_n) ir virkne no gadījuma vektoriem, tādiem, ka $X_n \xrightarrow{d} Z \sim H($ jebkurš $Z)$ un $Y_n | X_n \xrightarrow{d} Z($ tas pats $Z)$ gandrīz droši. Tātad $(X_n, Y_n) \xrightarrow{d} (Z_1, Z_2)$, kur Z_1, Z_2 ir iid un $\sim H$.

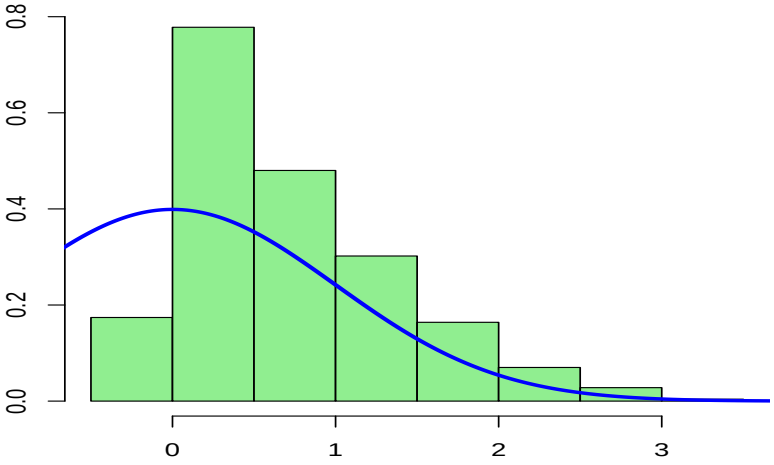
Tāpēc, atgriežoties pie mūsu piemēra, kad μ patiesā vērtība ir 0, iegūsim

$$T_n^* = \sqrt{n}(|\bar{X}_n^*| - |\bar{X}_n|)$$

$$= |\sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n| - |\sqrt{n}\bar{X}_n|$$

$$\xrightarrow{d} |Z_2 + Z_1| - |Z_1|,$$

kur Z_1, Z_2 ir iid un $\sim N(0, \sigma^2)$. Bet tas nav absolūtās vērtības no $N(0, \sigma^2)$ sadalījuma. Butstrapa virknei CRT nestrādā, kad $\mu = 0$. un līdz ar to arī butstrapa metode šim piemēram nestrādā.



4. att. Butstrapa metode $\sqrt{n}(g(\bar{X}) - g(\mu))$

Šo piemēru arī pārbaudīsim ar datorprogrammas R palīdzību, attēlojot statistikas T_n aproksimāciju ar histogrammu un attiecīgi blīvumu funkciju $N(0, 1)$. Kā redzams 4. attēlā, tad $N(0, 1)$ sadalījuma blīvuma funkcija neatbilst statistikas T_n histogrammai.

4. Butstrapa metode Kolmogorova - Smirnova testam

Kolmogorova - Smirnova tests pieder sadalījuma pārbaudes testiem (*angļu.val - Goodness of fit test*), kuri pārbauda hipotēzi par datu sadalījumu funkciju,

$$H_0 : F = F_0(x) \text{ pret } H_1 : F \neq F_0(x).$$

Šajos testos bieži izmanto empīrisko sadalījuma funkciju $F_n(x)$. Pēc Glivenko - Kantelli teorēmas seko, ka lieliem n, F_n ir tuva patiesam sadalījumam F. Lai pārbaudītu hipotēzi H_0 ir jāanalizē novirze starp F_n un F, šī novirze arī būs statistikas vērtība.

Teorēma 10. (Glivenko - Kantelli) *Pieņemsim, ka $X_1, \cdots, X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar teorētisko sadalījuma funkciju $F(x)$ un empīrisko sadalījuma funkciju $F_n(x)$, tad ir spēkā*

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

kur g.d nozīmē gandrīz droši.

Plaši izplatīti sadalījuma pārbaudes testi ir

(a) Cramer - von - Mises tests

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(t) - F_0(t))^2 dF_0(t)$$

Pieņemsim, ka $X_1 < X_2 < \cdots < X_n$ un $U_i = F_0(X_i)$, tad šī testa vērtība ir vienāda ar

$$\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(U_i - \frac{2i-1}{n} \right)^2.$$

(b) Anderson - Darling tests

$$A = \int_{-\infty}^{\infty} \frac{(F_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$$

Pieņemsim, ka $X_1 < X_2 < \cdots < X_n$ un $U_i = F_0(X_i)$, tad šī testa vērtība ir vienāda ar

$$A = -n - \frac{1}{n} \left[\sum_{i=1}^{n} (2i-1)(\log U_i + \log(1 - U_{n-i+1})) \right].$$

(c) Kolmogorova - Smirnova tests

$$D = \sqrt{n} \sup_x |F_n(x) - F_0(x)|.$$

Pirms apskatām Kolomogorova - Smirnova testa butstrapošanu, aprakstīsim testa būtību un uz ko tas konverģē pie $n \to \infty$.

Teorēma 11. (Kolmogorova - Smirnova tests) *Pieņemsim, ka $X_1, X_2, \cdots, X_n$ iid gadījuma lielumi ar sadalījuma funkciju F, tad*

$$D = \sqrt{n} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow{d} \sup_x |B(x)|, \tag{4.1}$$

kur $B(x)$ ir Brauna tilts un $F_n(x)$ ir empīriskā sadalījuma funkcija.

Definīcija 5. Par Brauna tiltu sauc Gausa procesu $\{B(t) : t \in [0,1]\}$, kura kovariāciju struktūra ir $cov(B(s), B(t)) = s(1-t)$, kur $s < t$. Šī procesa sadalījums ir tāds pats kā $W(t) - tW(1)$ sadalījums, kur W - standarta Brauna kustība.

Ja ar Kolmogorova - Smirnova testu pārbauda vienkāršo hipotēzi, piemēram $N(0,1)$, tad Kolmogorova - Smirnova tests tiecas uz Brauna tiltu. Saliktā hipotēze ietver sevī parametru novērtēšanu, kas izmaina statistiku. Tā ir sekojoša:

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

kur $\hat{\theta}$ ir novērtētie parametri. Piemēram, $N(\mu, \sigma^2)$ sadalījumam $\hat{\theta} = \{\bar{x}, s^2\}$, kur $\bar{x}$ ir vidējās vērtības novērtējums un s^2 ir dispersijas novērtējums. Saliktās hipotēzes gadījumā asimptotiskais sadalījums (4.1) vairs nav spēkā.

4.1. Neparametriskais butstraps Kolmogorova - Smirnova testam

Vispirms ar neparametrisko butstrapa metodi nobutstraposim Kolmogorova - Smirnova testu. Iegūsim, ka abām statistikām

$$\sqrt{n} \sup_x |F_n(x) - F(x)| \text{ un } \sqrt{n} \sup_x |F_n^*(x) - F_n(x)|$$

ir vienāds sadalījums gandrīz visām virknēm $X_1, \cdots, X_n$. Par to pārliecināsimies ar datorprogrammas R palīdzību, ģenerējot normāla sadalījuma gadījuma izlase apjomā 1000 ar dažādiem parametriem. Bet pirms rezultātu analizēšanas, apskatīsim teoriju par kodola

blīvuma funkciju, jo histogrammas tiks aproksimētas kodolu blīvuma funkcijas novērtējumu.

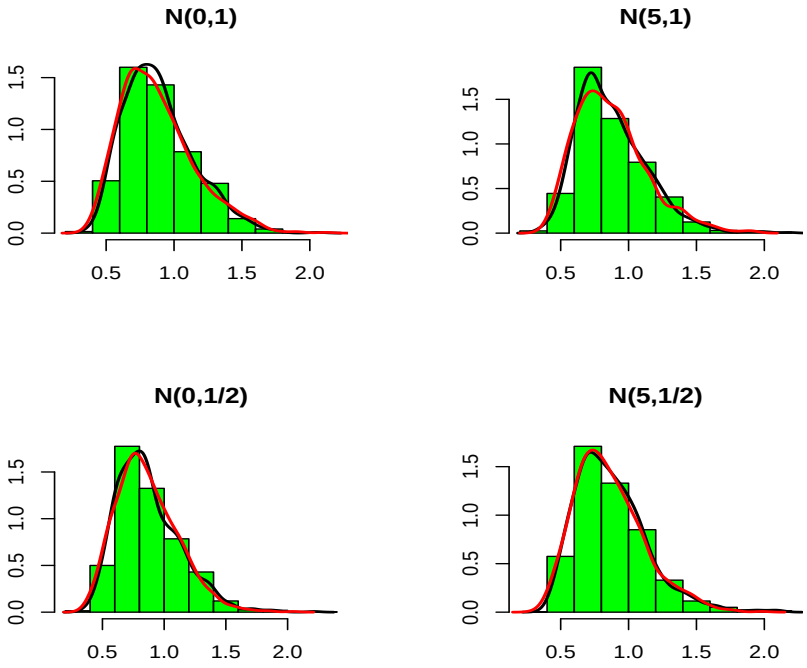
Ar jēdzienu **kodols** tiek saprasta jebkura gluda funkcija K, kurai izpildās sekojošas īpašības:

1. $K(x) \geq 0$,
2. $\int_{-\infty}^{\infty} K(x)dx = 1$,
3. K ir simetriska, t.i., $K(x) = K(-x)$.

Definīcija 6. Uzdotai izlasei $X_1, X_2, \cdots, X_n$, kodolam K un pozitīvam skaitlim h, kuru sauc par gludinošo parametru, kodola blīvuma funkcijas novērtējums ir

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

Arī kodola blīvuma funkcija ir gluda funkcija un tā konverģē uz patieso blīvuma funkciju ātrāk nekā histogramma. Tāpēc analizējot butstrapa metodi, histogrammai pievienosim kodola blīvuma funkciju, lai pārliecinātos, ka butstrapa metode strādā.



5. att. Neparametriskais butstraps KS testam

Melnā līnija ir kodola blīvuma novērtējums, kas aprēķināts butstrapotam Kolmogorova - Smirnova testam, bet sarkanā līnija ir kodola novērtējums, kas aprēķināts simulētam

Kolmogorova - Smirnova testam. Par kodola funkciju abiem novērtējumiem ir izvēlēts noklusētais Gausa kodols. Kā redzams, tad šie novērtējumi ir ļoti tuvi. Šeit mēs apskatījām butstrapoto Kolomogorova - Smirnova testu normālajam sadalījumam ar dažādiem parametriem, bet 1.1. pielikumā varam apskatīt arī attēlus ar bustrapoto Kolomogorova - Smirnova testu citiem sadalījumiem.

4.2. Parametriskais butstraps Kolmogorova - Smirnova testam nepārtrauktiem sadalījumiem

Statistiskās procedūras balstītas uz empīriskajiem procesiem ir plaši pielietotas, lai pārbaudītu parametriskos sadalījumus. Šīs metodes parasti nav sadalījumā neatkarīgas, tāpēc asimptotiskās kritiskās vērtības balstās uz nezināmiem parametriem. Šo problēmu varam atrisināt izmantojot parametrisko butstrapu. Kolmogorova - Smirnova tests izmanto empīrisko sadalījuma funkcijas aproksimāciju īstajai sadalījuma funkcijai, tāpēc pielietosim parametrisko butstrapu Kolmogorova - Smirnova testam.

Vispirms apskatīsim parametrisko butstrapu eksponenciālajam sadalījumam. Atgādināsim, ka $F_{\hat{\lambda}}(a) = 0$, ja $a < 0$ un $F_{\hat{\lambda}}(a) = 1 - e^{-\hat{\lambda}a}$ visiem $a \geq 0$, kur $\hat{\lambda} = 1/\bar{x}_n$ tiek novērtēts no izlases datiem. Kolmogorova - Smirnova testa statistika ir formā:

$$T_{ks} = \sqrt{n} \sup_{a \in \mathbb{R}} |\hat{F}_n(a) - F_{\hat{\lambda}}(a)|,$$

Statistikas T_{ks} varbūtību sadalījumu nav iespējams noteikts, jo parametrs λ eksponenciālajam sadalījumam nav zināms. Tomēr mēs varam aproksimēt T_{ks} ar parametrisko butstrapu. Tas ir, lietosim datus, lai noteiktu λ novērtējumu $\hat{\lambda} = 1/\bar{x}_n$ un aizvietot gadījuma izlasi $X_1, X_2, \cdots, X_n$ no F_λ ar butstrapa izlasi $X_1^*, X_2^*, \cdots, X_n^*$ no $F_{\hat{\lambda}}$. Pēc tam aproksimēsim T_{ks} sadalījumu ar tās butstrapa versiju, tas ir,

$$T_{ks}^* = \sqrt{n} \sup_{a \in \mathbb{R}} |\hat{F}_n^*(a) - F_{\hat{\lambda}^*}(a)|,$$

kur $\hat{F}_n^*$ ir empīriskā sadalījuma funkcija butstrapa izlasei un $F_{\hat{\lambda}^*}$ apzīmē novērtēto eksponenciālo sadalījuma funkciju, kur $\hat{\lambda}^* = 1/\bar{X}_n^*$ tiek aprēķināts no butstrapa izlases. Butstarpotā statistitka ir pārāk sarežģīta, lai tiešā veidā aprēķinātu tās varbūtību sadalījumu, tāpēc mēs lietosim parametriskā butstrapa simulācijas:

1. Ģenerēsim butstrapa izlasi $x_1^*, x_2^*, \cdots, x_n^*$ no eksponenciālā sadalījuma ar novērtēto parametru $\hat{\lambda}$.

2. Izrēķināsim butstrapoto statistiku T_{ks}^*.

Atkārtojot soļus 1 un 2 ļoti daudz reižu, piemēram 1000, iegūsim butstrapotās statistikas vērtības, nu kurām noteiksim izvēlēto kvantili, piemēram 0.95 kvantili.

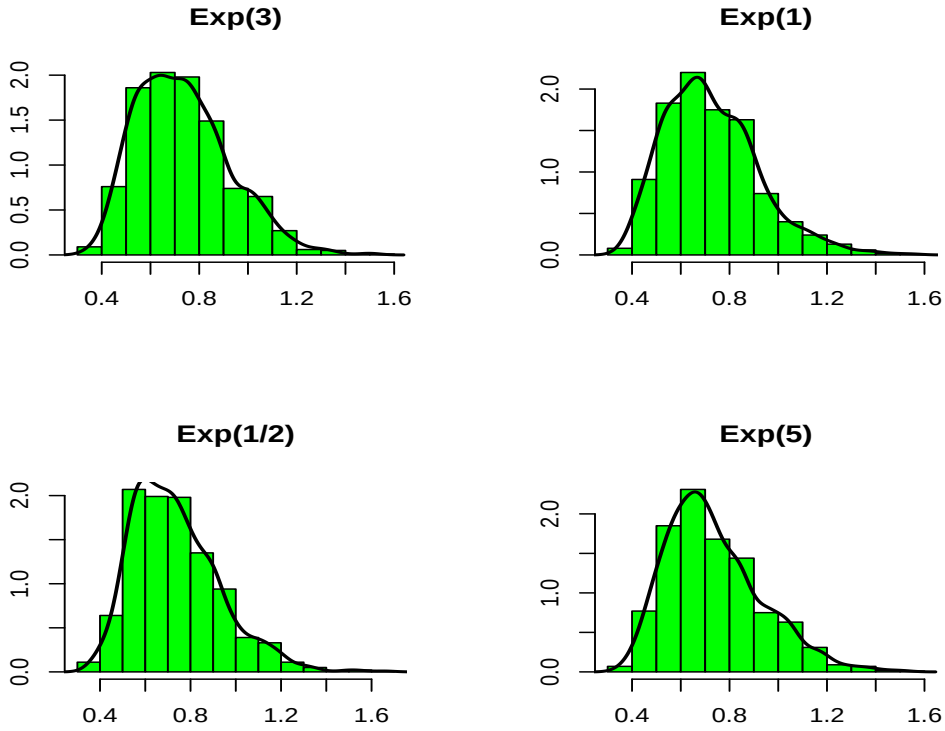
1. tabula Butstrapotās kvantiles

N(0,1)	0.907388	LogN(0,1)	0.893208
N(5,1)	0.897476	LogN(5,1)	0.884958
N(0,1/2)	0.899302	LogN(0,1/2)	0.900324
N(5,1/2)	0.908969	LogN(5,1/2)	0.876943
Exp(3)	1.071561	χ_1^2	1.332308
Exp(1)	1.085137	χ_2^2	1.167698
Exp(0.5)	1.089907	χ_3^2	1.075808
Exp(5)	1.093469	χ_4^2	1.073652

Ar datorprogrammu R ģenerēsim eksponenciālos sadalījumus ar parametriem λ= 3, 1, 0.5, 5, un šiem sadalījumiem butstraposim Kolmogorova - Smirnova testu. 6. attēlā ir redzamas bustrapotā testa histogrammas ar kodolu blīvumu funkciju novērtējumiem. Kā redzams, tad statistikas vērtības nav atkarīgas no parametra λ. Kolmogorova - Smirnova testu pielietojām arī cietiem sadalījumiem kā normālajam, lognormāljam un χ^2, šo dalījumu histogrammas varat apskatīt 1.2.pielikumā. Izveidosim tabulu ar kvantilēm, kuras iegūsim no butstrapotā testa. 1. tabulā redzamas šīs vērtības dažādiem sadalījumiem, kā redzams, tad tās aproksimē teorētiskās kritiskā vērtības pie attiecīgā nozīmības līmeņa. Tā kā ar parametrisko butstrapu tiek pārbaudīta saliktā hipotēze, tad 0.95 kvantile normāli sadalītām gadījuma izlasēm ir 0.9, eksponenciāli sadalītām 1.07 un χ^2 sadalītām 1.06.

4.3. Parametriskais butstraps Kolmogorova - Smirnova testam diskrētiem sadalījumiem

Šajā nodaļā apskatīsim Gabor Szucs publikāciju (skatīt [3]), kurā ir aprakstīts parametriskais butstraps diskrētiem sadalījumiem.



6. att. Parametriskais bustraps KS testam

Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju $F_0(x) = P(X_1 \leq x)$. Dota parametrisku sadalījumu funkciju saime $\mathcal{F} = \left\{F(x,\theta) : x \in \mathbb{R}, \theta \in \Theta \subseteq \mathbb{R}^d\right\}$, ar $d \in \mathbb{N}$ parametriem. Mērķis ir pārbaudīt hipotēzi $H_0 : F_0 \in \mathcal{F}$, tas nozīmē, ka jāpārbauda vai $F_0(\cdot) = F(\cdot, \theta_0)$. Pārbaudei izmantosim Kolmogorova - Smirnova testu. Attiecīgajam θ novērtējumam $\hat{\theta}$ apzīmēsim

$$V_n = \sqrt{n} \sup_x |F_n(x) - F(x, \hat{\theta})|, x \in \mathbb{R},$$

kur V_n ir testa vērtība, kurai vēlamies nobutstrapot. Pieņemsim, ka $X_1^*, \ldots, X_n^*$ neatkarīgi lielumi ģenerēti no sadalījuma $F(\cdot, \hat{\theta})$ un $F_n^*(\cdot)$ ir attiecīgā empīriskā sadalījuma funkcija. No ģenerētās izlases aprēķināsim parametra $\hat{\theta}$ novērtējumu $\hat{\theta}^*$, rezultātā butstrapotā statistika ir formā

$$V_n^* = \sqrt{n} \sup_x |F_n(x)^* - F(x, \hat{\theta}^*)|, x \in \mathbb{R}.$$

Kad esam definējuši kā izskatās statistika, kuru butstraposim, aprakstīsim butstrapa precedūras soļus ar kuru varēsim pārbaudīt hipotēzi $H_0 : F_0 \in \mathcal{F}$. Pēc šiem šoļiem arī veidosim praktiski apskatīsim butstrapa metodi diskrētiem sadalījumiem.

1. Aprēķinām novērtējumu $\hat{\theta}$ no izlases $X_1, \cdots, X_n$.

2. Aprēķinām V_n.

3. Ģenerējam gadījuma lielumus $X_1^*, \ldots, X_n^*$ no sadalījuma $F(\cdot, \hat{\theta})$.

4. Aprēķinām parametra $\hat{\theta}$ novērtējumu $\hat{\theta}^*$ no butstrapa izlases.

5. Aprēķinām V_n^*.

6. Atkārtojam soļus 3-5 B reizes, pieņemsim, ka $V_{n,1}^* \leq \cdots \leq V_{n,B}^*$ ir sakārtota virkne no B statistikām V_n^*. Pieņemsim, ka $x_{1-\alpha}$ ir $(1-\alpha)$ empīriskā kvantile no V_n^*, tas ir, tā ir $\lceil B(1-\alpha) \rceil$ lielākā statistikas vērtība, kur $\lceil y \rceil = \min\{j \in \mathbb{Z} : y \leq j\}$, ja $y \in \mathbb{R}$.

7. Noraidām H_0, ja V_n vērtība ir lielāka par $x_{1-\alpha}$.

Praktiski pārbaudīsim hipotēzi

$$H_0 : F_0 \in \mathcal{NB}(r) \text{ pret } H_1 : F_0 \in Po(\lambda),$$

kur $\mathcal{NB}(r)$ ir r kārtas negatīvais binomiālais sadalījums un $Po(\lambda)$ ir Puasona sadalījums.

2. tabula Pārklājuma precizitāte

q	r=1	3	10	30
0.10	0.113	0.108	0.105	0.123
0.25	0.119	0.109	0.115	0.102
0.50	0.118	0.106	0.085	0.111
0.75	0.112	0.096	0.097	0.094
0.90	0.098	0.107	0.104	0.101

Vispirms ģenerēsim $\mathcal{NB}_r(q)$ gadījuma lielumu izlases ar dažādiem parametriem apjomā $n = 50$, un pārbaudīsim hipotēzi H_0 ar butstrapa metodi, kad $B = 500$ un nozīmības līmenis ir $\alpha = 0.1$. Pārklājuma precizitāte pēc 1000 Monte Carlo simulācijām ir apskatāma 2..tabulā. Redzams, ka visas vērtības ir tuvas nozīmības līmenim $\alpha = 0.1$.

Pēc tam apskatīsim parametriskā butstrapa jaudu. Ģenerēsim $Po(\lambda)$ izlases apjomā $n = 50$, un pārbaudīsim hipotēzi H_0 par datu atbilstību negatīvajam binomiālajam sadalījumam, ja $r = 1, 3, 10$ un 30. Arī šeit nozīmības līmenis $\alpha = 0.1$ un pielietosim 1000 Monte Carlo simulācijas ar $B = 500$. Noraidīto hipotēžu daudzums ir parādīts

3. tabula Testa jauda

λ	r=1	3	10	30
1	0.880	0.309	0.114	0.109
3	1.000	0.812	0.211	0.115
5	1.000	0.980	0.405	0.137
10	1.000	1.000	0.750	0.201
30	1.000	1.000	1.000	0.733
50	1.000	1.000	1.000	0.950

3..tabulā. Secinām, ka testa jauda samazinā, ja parametrs r palielinās. Tas izskaidrojams, ka r palielinoties negatīvais binomiālasi sadalījums paliek arvien līdzīgāks Puasona sadalījumam.

5. Neimaņa tests

Šajā nodaļā apskatīsim Neimaņu, kurš tika formulēts jau 1937. gadā, bet salīdzinoši nesen formulētie rezultāti, padarīja šo testu plašāk pielietojumu. Arī Neimaņa tests pieder sadalījuma pārbaudes testiem. Apskatīsim šo testu gan neatkarīgiem, gan atkarīgiem datiem. Neimaņa testa ir aprakstīts piemēram ([7]) publikācijā.

Vispirms definēsim vispārēju atkarības koeficientu. Pieņemsim, ka $(X_t)_{t\in\mathbb{Z}}$ ir stacionārs process varbūtību telpā $(\Omega, \mathcal{F}, P)$. Jebkurām divām σ- algebrām $\mathcal{A}$ un $\mathcal{B} \subset \mathcal{F}$ atkarības koeficientu ir

$$\alpha(\mathcal{A}, \mathcal{B}) := \sup|P(A \cap B) - P(A)P(B)|,$$

kur $A \in \mathcal{A}$ un $B \in \mathcal{B}$. Jebkuriem J, L $(-\infty \leq J \leq L \leq \infty)$ definēsim σ- algebru $F_J^L = \sigma(X_k, J \leq k \leq L)$. Jaukto procesu atkarības koeficienti $\forall n$ procesam $(X_t)_{t\in\mathbb{Z}}$ tiek definēts sekojoši:

$$\alpha(n) := \sup_{J\in\mathbb{Z}} \alpha(F_{-\infty}^J, F_{J+n}^\infty)$$

Definīcija 7. ([8]) Saka, ka process $(X_t)_{t\in\mathbb{Z}}$ ir α jauktais process, ja $\alpha(n) \to 0$, kad $n \to \infty$.

Pieņemsim, ka stacionāram α- jauktajam procesam $(X_t)_{t\in\mathbb{Z}}$ ir galīga sadalījuma funkcija F. Ar Neimaņa testu mēs vēlamies pārbaudīt sekojošu vienkāršu hipotēzi

$$H_0 : F = U[0,1] \text{ pret } H_1 : F \neq U[0,1],$$

kur $U[0,1]$ ir vienmērīgais sadalījums intervālā [0,1]. Atzīmēsim, ka pārbaudot hipotēzi $H_0 : F = F_0$ vispārīgam nepārtrauktam sadalījumam F_0, tas var tikt reducēt uz vienmērīgo sadalījumu, pārveidojot datus $F_0(X_t), t \in \mathbb{Z}$.

Neimaņa tests šai hipotēzei ir sekojošā formā (neatkarīgiem un vienādi sadalītiem datiem)

$$R_k = \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2 , k = 1, 2, \cdots ,$$

kur $\phi_0, \phi_1, \cdots$, ir ortonormēta sistēma telpā $L_2[0,1]$ ar $\phi_0 = 1$, pārējie ir Lagranža polinomi. Šos polinomus var definēt rekurenti, pie $j = 0, 1, 2, 3, 4$,

$\phi_0 = 1$,

$\phi_1 = \sqrt{12}(x - 1/2)$,

$\phi_2 = \sqrt{5}(6(x - 1/2)^2 - 1/2)$,

$\phi_3 = \sqrt{7}(20(x-1/2)^3 - 3(x-1/2))$,

$\phi_4 = 210(x-1/2)^4 - 45(x-1/2)^2 + 9/8$.

Neimaņa tests atkarīgiem datiem (α - jauktajiem procesiem) ir formā

$$N_k = (12\sigma^2)^{-1} R_k = (12\sigma^2)^{-1} \sum_{j=1}^{k} \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_j(X_i) \right\}^2$$

ar

$$\sigma^2 = \sum_{t=-\infty}^{+\infty} \mathrm{Cov}(X_0, X_t).$$

Vispārīgi sakot, faktors $(12\sigma^2)^{-1}$ pielāgo R_k atkarību tā, ka pieņemot H_0 statistika N_k ir ar tādu pašu galīgu sadalījumu kā R_k neatkarīgu datu gadījumā.

Lai pielietotu Neimaņa testu atkarīgiem datiem, ir jāaprēķina testa statistika neatkarīgu datu gadījumā un jānovērtē σ^2, tad statistikas vērtība ar šo novērtējumu jāizdala.

Visbeidzot, lai konstruētu visaptverošu testa statistiku mums ir jāizvēlas konkrēta k vērtība. Ir ļoti daudz literatūras, par to kā izvēlēties atbilstošo k neatkarīgu datu gadījumā. Ledwina (1994) pamatoja metodi, kas balstīti uz maksimālās ticamības funkciju, lai noteiktu piemērotu k vērtību.

$$S_{mod} = \min\left\{k : 1 \le k \le d(n), R_k - k\log n \ge R_j - j\log n, j = 1, \cdots, d(n)\right\},$$

kur ar $d(n)$ apzīmē k augšējo robežu, kas var tiekties arī uz bezgalību, kad $n \to \infty$. Kad ir izvēlēts k, tad tests $N_{S_{mod}}$ ir pielietojams.

Apskatīsim gadījumu, kad Neimaņa statistika R_k ir aizvietota ar N_k,tad iegūsim sekojošu formulu

$$S_{mod2} = \min\left\{k : 1 \le k \le d(n), N_k - k\log n \ge N_j - j\log n, j = 1, \cdots, d(n)\right\},$$

Acīmredzami, ka pozitīvas atkarības gadījumā ($\sigma^2 > 1/12$), S_{mod2} vairāk koncentrēsies uz vērtību 1 saskaņā ar H_0. Tāpēc, izdevīgi lietot S_{mod2}, ja iepriekš ir pieejama informācija par pozitīvu korelāciju. Atzīmēsim, ka neatkarīgu datu gadījumā $12\sigma^2 = 1$, ja hipotēze H_0 ir patiesa.

Novērtēsim σ^2, kas ietver autkovariācijas struktūras novērtēšanu. ARMA procesu gadījumā autokovariāciju funkcijas ir $\gamma(h) := \mathrm{Cov}(X_{t+h}, X_t)$ visiem $t, h \in \mathbb{Z}$. Stacionāriem procesiem un ARMA procesiem $\gamma(h)$ novērtējums ir formā:

$$\hat{\gamma}(h) = (n-h)^{-1} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X}), \text{ ja } 0 \le h \le n-1,$$

kuru izmantosim, lai novērtētu σ^2. Novērtējums ir sekojošs:

$$\hat{\sigma}^2 = \hat{\gamma}(0) + 2\sum_{j=1}^{q} \hat{\gamma}(j),$$

kur q norāda pēdējo lagu, kuram tiek novērtēta kovariācija $\gamma(q)$. Atkarīgiem datiem autokovariācija eksponenciāli dilst. ([7]) publikācijā ir ieteikums apskatīt un lietot tikai tās autkovariācijas, kuras ir lielākas par 0.001. Šo novērtējumu vēlāk apskatīsim arī praktiski.

5.1. Pielietojums neatkarīgiem datiem

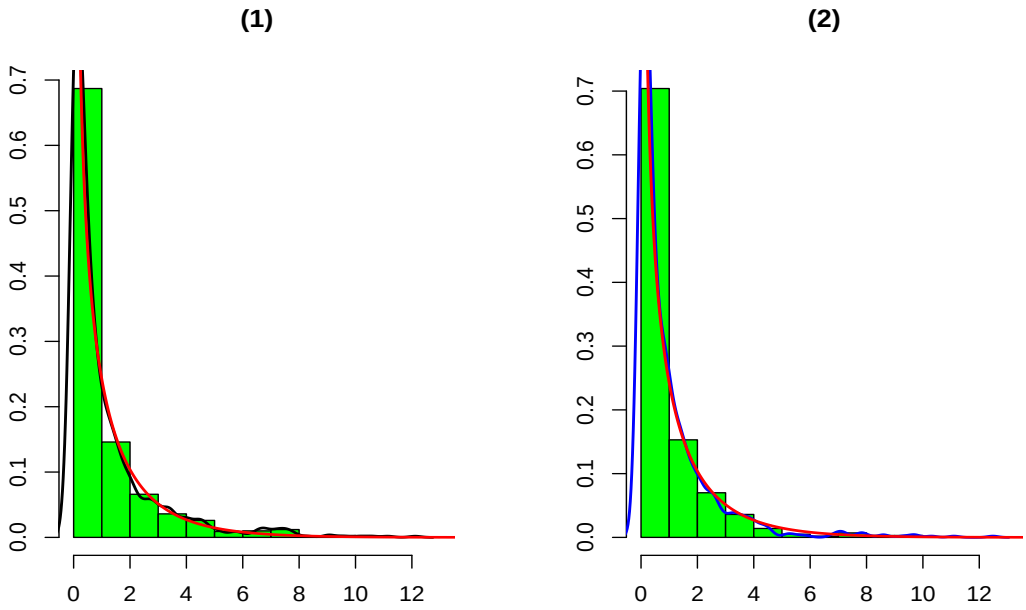
Vispirms apskatīsim neparametrisko butstrapa metodi Neimaņa testam, kad dati ir neatkarīgi. Lai pārbaudītu vai butstrapa metode strādā, butstrapotās statistikas vērtības salīdzināsim ar simulētajām statistika vērtībām, pie izvēlēta $k = 1$. Ar datorprogrammu R ģenerēsim neatkarīgu vienmērīgi sadalītu gadījuma lieluma izlasi apjomā $n = 1000$. No tās izveidosim $B = 1000$ butstrapa izlases $\{X_{11}^*, \cdots, X_{1n}^*\}$, $\{X_{21}^*, \cdots, X_{2n}^*\}$, $\cdots$, $\{X_{B1}^*, \cdots, X_{Bn}^*\}$, kurām aprēķināsim

$$R_1^* = n^{-\frac{1}{2}} \sum_{i=1}^{n} \phi_1(X_i^*)$$

7. attēlā grafikā (1) ir redzamas butstrapotais Neimaņa tests attēlots ar histogrammu, kurai pievienots klāt kodola blīvuma funkcijas novērtējums, melnā krāsā, un χ_1^2 blīvuma funkcija, sarkanā krāsā, uz ko teorētiski tiecas Neimaņa tests. Bet grafikā (2) ir redzamas simulētā Neimaņa testa attēlojums ar histogrammu, kurai arī pievienots klāt kodola blīvuma funkcijas novērtējums, zilā krāsā, un χ_1^2 blīvuma funkcija, sarkanā krāsā. Kā redzams, tad abas histogrammas ir līdzīgas un χ_1^2 aproksimācija ir laba. No tā secinām, ka butstrapa metodi var pielietot Neimaņa testam neatkarīgu datu gadījumā.

5.2. Pielietojums atkarīgiem datiem

Singh ([9]) parādīja piemēru, kad butstrapa metode dod aplamus rezultātus atkarīgiem novērojumiem. Taču ir butstrapa metodes, kas dod labus rezultātus arī datiem ar atkarības struktūru. Šajā gadījumā butstrapa metodes mērķis ir nepazaudēt datu atkarības struktūru, pārkārtojot datus. Viena no efektīvākajām un vienkāršākajām ir bloku butstrapa metode. Pretēji iepriekš aprakstītajai butstrapa procedūrai, kad tiek pārkāroti izlases elementi, bloku butstrapa metode pārkārto izlases elementus, sakārtotus blokos. Rezultātā,



7. att. (1) Butstrapotais Neimaņa tests, (2) Simulētais Neimaņa tests

bloka iekšienē datu atkarība ir saglabāta. Bloku butstrapa aprakstam izmantosim Mārča Bratka maģistra darbu ([10]).

Slīdošo bloku butstraps

Viens no bloku butstrapa veidiem ir *slīdošā bloka butstraps* (no angļu valodas "moving block bootstrap", turpmāk MBB), kuru piedāvāja Kunsch ([11]) un Liu un Singh ([12]). Pieņemsim, ka $X_1, X_2, \cdots$ ir stacionāra gadījuma lieluma virkne un $\mathcal{X}_n = (X_1, \cdots, X_n)$ ir dotā izlase. Bloka garums $l \in [1, n]$ ir vesels skaitlis. $\mathcal{B}_i = (X_i, \cdots, X_{i+l-1})$ ir bloks garumā l sākot ar elementu X_i, $1 \leq i \leq N$, kur $N = n - l + 1$. Lai iegūtu MBB izlasi, izvēlamies vajadzīgo bloku skaitu no $\mathcal{B}_1, \cdots, \mathcal{B}_N$ un pārkārtojam gadījuma izlasē $\mathcal{B}_1^*, \cdots, \mathcal{B}_k^*$, kur katrs no izvēlētajiem blokiem satur l elementus. Apzīmēsim $\mathcal{B}_i^*$ elementus ar $X^*_{(i-1)l+1}, \cdots, X^*_{il}, i = 1, \cdots, k$. Tādejādi $X_1^*, \cdots, X_m^*$ sastāda MBB izlasi ar apjomu $m \equiv kl$.

Līdzīgi, kā butstrapa metodei ar neatkarīgiem un vienādi sadalītiem gadījuma lielumiem, MBB izlases apjoms parasti ir izvēlēts ar tādu pašu kārtu, kā sākotnējās izlases apjoms. Ja b ir mazākais veselais skaitlis, kurš apmierina $bl \geq n$, tad iespējams izvēlēties b kā vajadzīgo bloku skaitu, un lietot tikai pirmos n elementus. Kad ir iegūta MBB izlase ar apjomu n, tad rīkojās kā tāpat kā neatkarīgu datu gadījumā, aprēķinām izvēlētās statistikas vērtību un visu procedūru atkārtojam ļoti daudz reižu.

Nešķeļošo bloku butstraps

Vēl apskatīsim *nešķeļošo bloku butstrapa* metodi (no angļu valodas ņonoverlapping block bootstrap", turpmāk NBB), kuru piedāvāja Carlstein ([13]). Galvenā NBB ideja ir lietot blokus, kuri nešķeļas. Pieņemsim, ka $l \in [1, n]$ ir vesels skaitlis un $b \geq 1$ ir lielākais veselais skaitlis, kurš apmierina $bl \leq n$. Definēsim bloku $\mathcal{B}_i = (X_{(i-1)l+1}, \ldots, X_{il}), i = 1, \ldots, b$. Atzīmēsim, ka MBB izvēlētie bloki šķeļas, bet NBB bloki nešķeļas. Rezultātā, iespējamo atšķirīgo bloku skaits ir mazāks nekā MBB metodei.

Tālākā procedūra ir identiska MBB metodei. Veidojam gadījuma izlasi ar atkārtojumiem $\mathcal{B}_1^*, \cdots, \mathcal{B}_k^*$ no $\mathcal{B}_1, \ldots, \mathcal{B}_b$, kur $k \geq 1$. Apzīmēsm butstrapa izlasi ar $X_1^*, \ldots, X_m^*$, kura iegūta no blokiem $\mathcal{B}_1^*, \ldots, \mathcal{B}_k^*$, kur $m = kl$. Ja n dalās ar l bez atlikumiem, tad butstrapa izlases apjoms ir vienāds ar sākotnējās izlases apjomu n. Tāpat kā MBB gadījumā, kad esam ieguvuši NBB izlasi, aprēķinām izvēlētās statistikas vērtību un visu procedūru atkārojam ļoti daudz reižu.

Abas aprakstītās metodes pielietosim Neimaņa testa butstrapošanai, atkarīgu datu gadījumā. Vispirms aprakstīsim soļus, pēc kuriem iespējam ģenerē atkarīgus datus ar $U(0, 1)$ sadalījumu, izmantojot AR(1) procesu, kas ir α - jauktais process (skatīt Definīciju 7).

1. Simulēsim $X_1, \cdots, X_n$ no stacionāra AR(1) procesa, kur $\{X_t\}_{t \in Z}$ definēts kā

$$X_t - \theta X_{t-1} = Z_t, \tag{5.1}$$

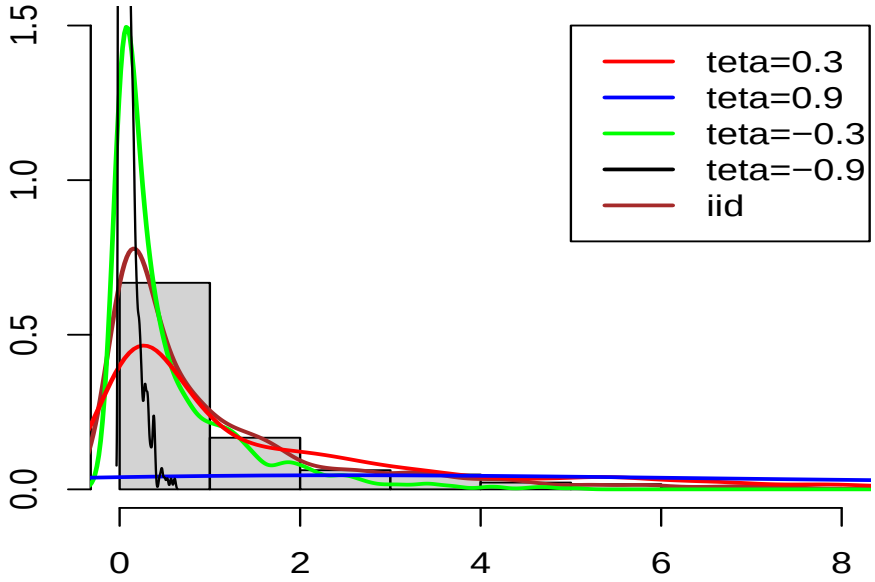
   kur $\{Z_t\}_{t \in Z}$ ir vāji stacionārs process ar vidējo vērtību 0 un autokovariāciju $\mathbb{E}(Z_t Z_{t+h}) = \sigma_Z^2 < \infty$, ja h=0 un 0 pretējā gadījumā, un procesa koeficients $|\theta| \leq 1$.

2. Ģenerē datus no 5.1 ar $Z_t \sim N(0, 1 - \theta^2)$. Ģenerētajam procesam $\{X_t\}_{t \in Z}$ būs $N(0, 1)$ sadalījums.

3. Visbeidzot transformēsim datus ar Φ, kur Φ apzīmē $N(0, 1)$ sadalījuma funkciju. Iegūsim izlasi $X_1^{'}, \ldots, X_n^{'} = \Phi(X_1), \cdots, \Phi(X_n)$, kurai būs $U(0, 1)$ sadalījums.

Visus soļus īstenosim ar statistikas programmu R, izmantojot procedūru arima.sim pie apjoma $n = 1000$. Izvēlēsimies arī dažādas $\theta = 0.3, 0.9, -0.3, -0.9$ vērtības. Sākumā simulēsim Neimaņa testu pie jau minētajām θ vērtībā. Izmantosim nevis N_1 statistiku, bet iid gadījuma statistiku $R_1 \xrightarrow{d} 12\sigma^2\chi_1^2$. Simulācijas ir redzamas 8. attēlā,

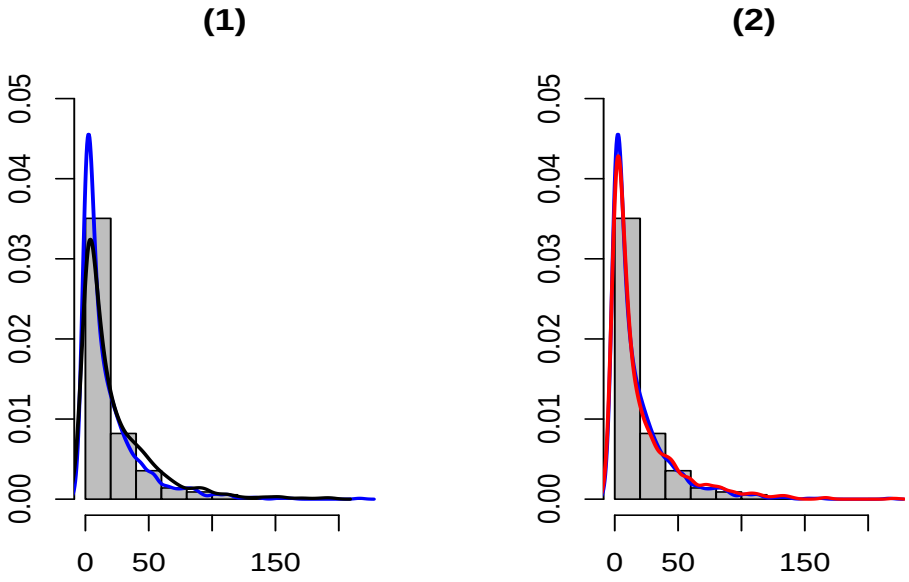
kur ar histogrammu attēlotas simulētās Neimaņa testa vērtības iid gadījumam. Tā tiek aproksimēta ar kodolu bīlvumu funkciju novērtējumiem gan iid gadījumā, gan gadījumā, kad $\theta = 0.9, 0.3, -0.3, -0.9$ vērtībām. Kā redzams, tad robežsadalījums mainās.



8. att. Simulētā Neimaņa testa asimptotika

Tālāk izvēlēsimies $\theta = 0.9$ vērtību, un ar iepriekš aprakstītajiem soļiem ģenerēsim atkarīgus datus ar sadalījumu $U(0,1)$ pie apjoma $n = 1000$. Ģenerētajiem datiem pielietosim bloka butstrapa metodes MBB un NBB. Abām metodēm izvēlēsimies bloka garumu vienādu ar $l = 50$. Lai butstrapotās izlases lielums būtu vienāds ar sākotnējās izlases apjomu, tad bloku skaits, kuri tiks iekļauti butstrapa izlasē, ir vienāds ar $k = 20$. Ar butstrapa metodi iegūtās statistikas vērtības salīdzināsima ar simulētajām statistika vērtībām, kas parādīs vai butstrapa metodes aproksimācija ir laba. Rezultāts ir redzams 9. attēlā, kur grafikā (1) ir attēlotas simulētās Neimaņa testa vērtības, kuras aproksimētas ar kodola blīvuma funkcijas novērtējumu (zilā krāsā) un kodola blīvuma funkcijas novērtējumu butstarpotajam Neimaņa testam ar NBB metodi (melnā krāsā). Grafikā (2) arī ir attēlots simulētās Neimaņa testa vērtības ar kodola blīvuma funkcijas novērtējumu, kā arī kodola blīvuma funkcijas novērtējums butstrapotajam Neimaņa testam ar MBB metodi (sarkanā krāsā). Redzams, ka bustrapa metode labi aproksimē Neimaņa testa teorētisko sadalījumu.

Grafiski esam parādījuši, ka butstrapa metodi var pielietot Neimaņa testam un ka tā



9. att. (1) NBB metode, (2) MBB metode

labi aproksimē testa statsitikas sadalījumu.

6. Secinājumi

Šajā darbā parādīts, ka butstrapa metode ir plaši pielietojama praksē neatkarīgiem un vienādi sadalītiem datiem. Darbā aprakstīts neparametriskais butstraps, kas balstās uz atkārtotām izlasēm, kuas iegūst no sākotnējiem datiem, un parametriskais butstraps, kad butstrapa izlase tiek veidota pēc kāda fiksēta sadalījuma likuma $F_{\hat{\theta}}$. Ar butstrapa metodes palīdzību var atbildēt uz jautājumiem, kas ir par sarežģitu tradicionālai statistikas analīzei.

Ar parametrisko butstrapu ir iespējams pielietot Kolmogorova - Smirnova testam diskrētiem sadalījumiem. Neparametriskais butstaps tika pielietots Neimaņa testam neatkarīgu datu gadījumā, bet atkarīgu datu gadījumā neparametriskais butstraps dod aplamus rezultātus. Tāpēc darbā ir aprakstītas un pielietotas bloku butstrapa metodes kā MBB un CBB. Šīs metodes labi aproksimē testa statistikas sadalījumu. Jāatzīmē, ka visas butstrap metodes aproksimācijas var padarīt precīzākas, palielinot atkārtojumu skaitu butstrapa procedūrai.

Jāuzsver, ka visas butstrapa metodes ir ļoti darbietilpīgas no datoru resursu viedokļa. Tāpēc butstrapa metodes pievilcīgums ir saistīts ar datoru jaudu un iespējām. Butstrapa metodi var pielietot netikai statistikā, bet arī dabas zinātnēs, medicīnā, ekonometrijā un citās nozarēs. Varam secināt, ka butstrapa metode ir plaši pielietojama.

Izmantotā literatūra un avoti

[1] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[2] M. Dekking, C. Kraaikamp, and HP Lopuhaa. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Verlag, 2005.

[3] G. Szucs. Parametric bootstrap tests for continuous and discrete distributions. *Metrika*, 67(1):63–81, 2008.

[4] A. DasGupta. *Asymptotic theory of statistics and probability*. Springer Verlag, 2008.

[5] P.J. Bickel and D.A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

[6] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer, 1995.

[7] A. Munk, J.P. Stockis, J. Valeinis, and G. Giese. Neyman smooth goodness-of-fit tests for the marginal distribution of dependent data. *Annals of the Institute of Statistical Mathematics*, pages 1–21, 2010.

[8] R.C. Bradley. *Introduction to strong mixing conditions*. Kendrick Press, 2007.

[9] K. Singh. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, 9(6):1187–1195, 1981.

[10] M. Bratka. Butstrapa metodes analīze atkarīgiem datiem. *LU, Fizikas un matemātikas fakultāte*, 2009.

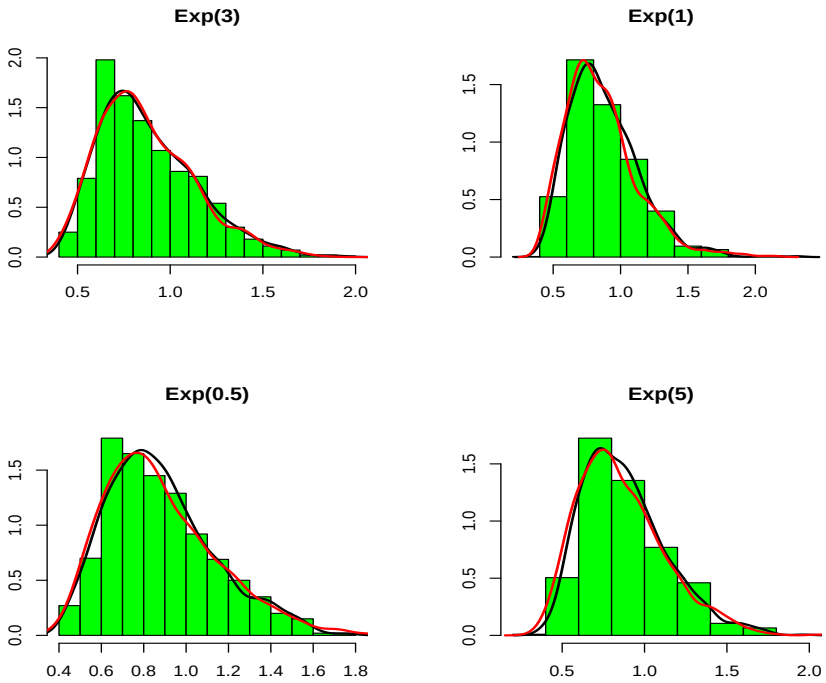
[11] H.R. Kunsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

[12] R. Liu and K. Singh. Moving blocks bootstrap and jackknife capture weak dependence. *Exploring the Limits of Bootstrap*, pages 225–248, 1992.

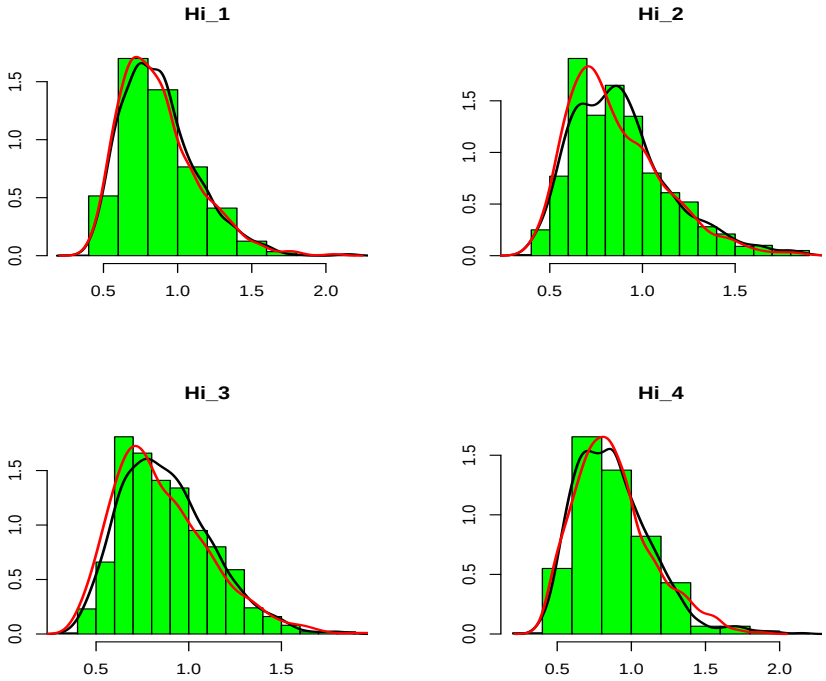
[13] E. Carlstein. The use of subseries methods for estimating the variance of a general statistic from a stationary time series. *Annals of Statistics*, 14:1171–1179, 1986.

1. Pielikums

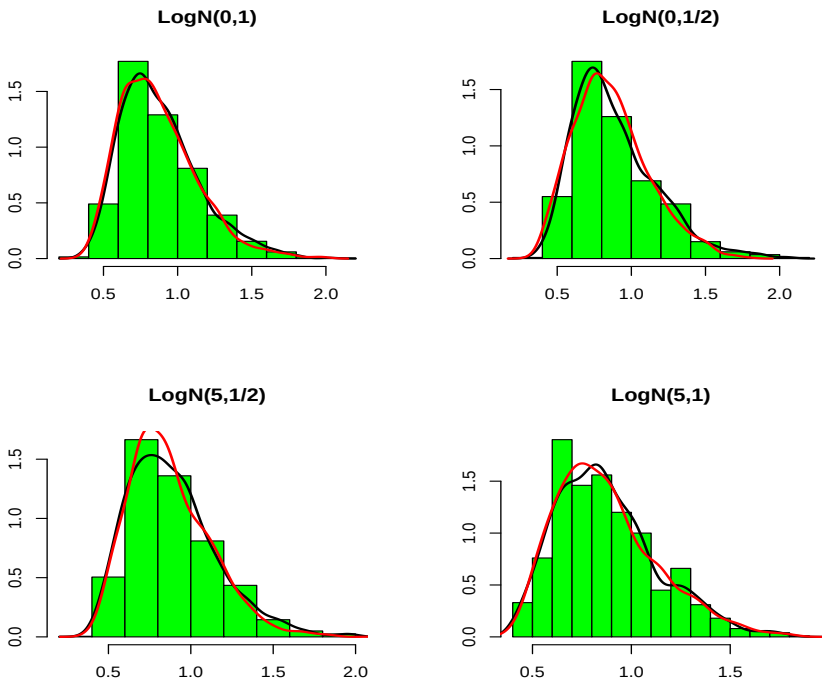
1.1. Kolmogorova - Smirnova tests ar neparametrisko butstrapu



10. att. Neparametriskais butstraps KS testam

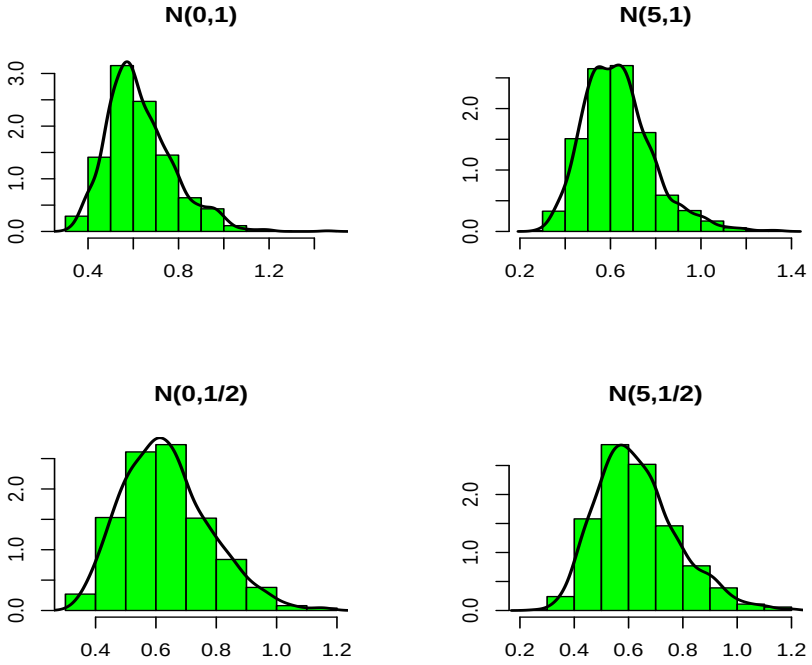


11. att. Neparametriskais butstraps KS testam

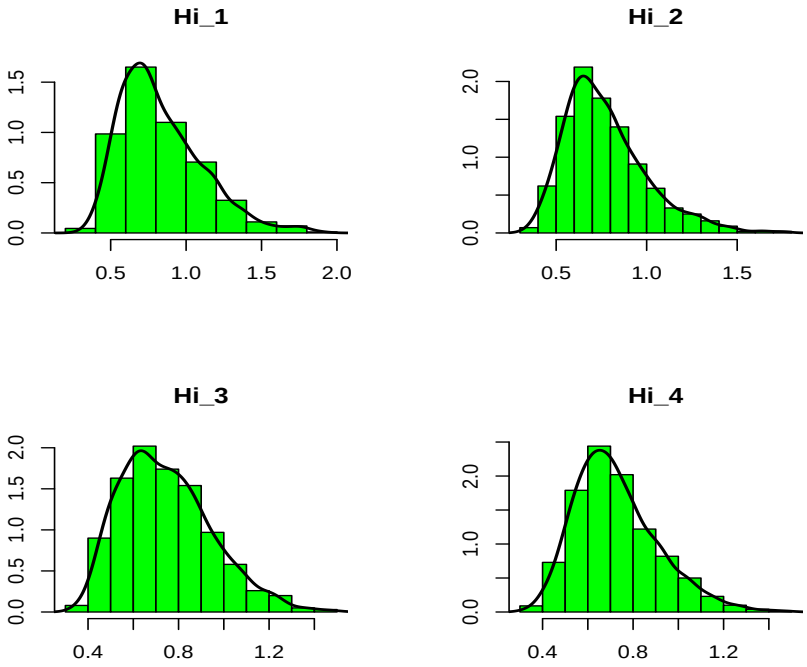


12. att. Neparametriskais butstraps KS testam

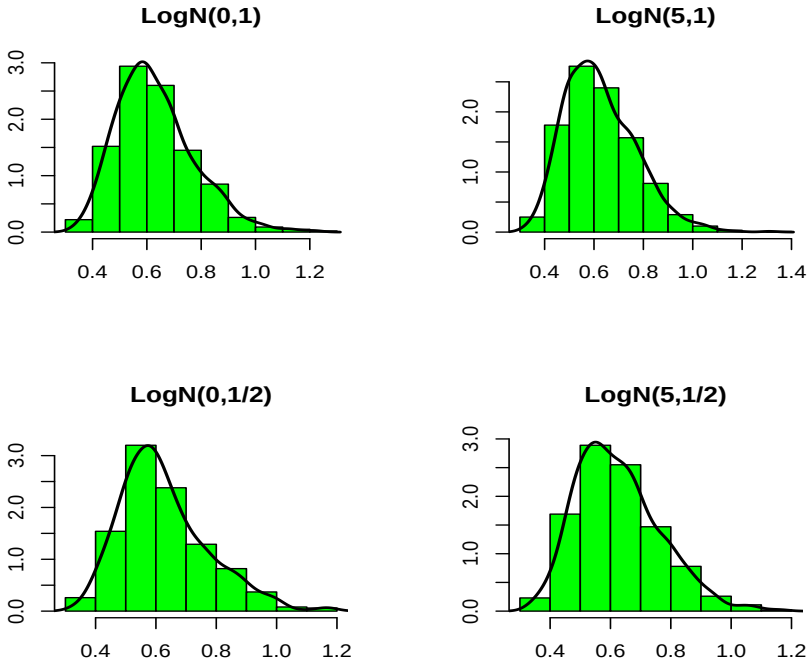
1.2. Kolmogorova - Smirnova tests ar parametrisko butstrapu



13. att. Parametriskais butstraps KS testam



14. att. Parametriskais butstraps KS testam



15. att. Parametriskais butstraps KS testam

1.3. Izveidoto programmu kodi

```
### NEPARAMETRISKAIS BUTSTRAPS ###

# Vidējai vertibai #

# izlases apjoms
```

```
n<-20 #ievieto pēc tam 50, 100 un 150
izl1<-rnorm(n,5,1)
# cik reizes butstrapot
B<-1000
# butstraps vid.vert
mean.boot1<-replicate(B,mean(sample(izl1,replace=TRUE)))
par(mfrow=c(2,2))
hist(mean.boot1,prob=TRUE,col='lightblue',main="n=20",xlab='',ylab='')
xx<-seq(4,6,len=1000)
points(xx,dnorm(xx,5,sqrt(1/n)),type="l",lwd=3,col='red')

# Centrētai vidējai vertibai #
n<-20 #ievieto pēc tam 50, 100 un 150
boot.sta1t<-replicate(B,sqrt(n)*(mean(sample(izl1,replace=TRUE))-mean(izl1)))
hist(boot.stat1,prob=TRUE,col='lightgreen',main="n=20",xlab='',ylab='')
xx<-seq(-6,5,len=1000)
points(xx,dnorm(xx,0,1),type='l',lwd=3,col='blue')

### PARAMETRISKAIS BUTSTRAPS ###
# Centretai videjai vertibai #
par(mfrow=c(2,2))
n<-20 #ievieto pēc tam 50, 100 un 150
boot.stat2<-replicate(B,sqrt(n)*(mean(rnorm(n,mean(izl1),1))-mean(izl1)))
hist(boot.stat2,prob=TRUE,col='lightgreen',main="n=20",xlab='',ylab='')
xx<-seq(-6,5,len=1000)
points(xx,dnorm(xx,0,1),type='l',lwd=3,col='blue')

##### Piemērs, kam butstrapa metode nestrādā####
# izlases apjoms
n<-100
# ģenerē izlasi
izl<-rnorm(n,0,1)
```

```
# cik reizes butstrapot
B<-1000
boot.stat<-replicate(B,sqrt(n)*(abs(mean(sample(izl,replace=TRUE)))-abs(mean(izl))))
hist(boot.stat,prob=TRUE,col='lightgreen',main="",xlab='',ylab='')
xx<-seq(-6,5,len=1000)
points(xx,dnorm(xx,0,1),type='l',lwd=3,col='blue')

####### NEPARAMETRISKAIS BUTSTRAPS #########
## Normalais sadalijums ##
n<-1000
izl<-rnorm(n,5,1/2)
B<-1000
alpha<-0.05
par(mfrow=c(2,2))
ff<-function(n)
{
dati<-sample(izl,replace=TRUE)
fn<-ecdf(dati)
a1<-max(abs(fn(sort(dati))-fn(sort(izl))))
a2<-max(abs(c(0,fn(sort(dati))[-n])-fn(sort(izl))))
max(a1,a2)
}
rez<-replicate(1000,sqrt(n)*ff(n))
hist(rez,prob=T,main="N(5,1/2)",xlab='',ylab='', col="green")
lines(density(rez),type="l",lwd=2)
# simuletais sadalijums
rez2<-replicate(1000,ks.test(rnorm(n,5,1/2),"pnorm",5,1/2)$statistic[[1]]*sqrt(n))
points(density(rez2),type="l",col="red",lwd=2)
###### Līdzīgi pārējiem sadalījumiem#####

####### PARAMETRISKAIS BUTSTRAPS #########
n<-1000
```

```
B<-1000
alpha<-0.05
par(mfrow=c(2,2))
## Normalais sadalijums  ##
izl<-rnorm(n,5,1/2)
vv<-mean(izl)
disp<-var(izl)
ff<-function(n)
{
dati<-rnorm(n,vv,sqrt(disp))
fn<-ecdf(dati)
a1<-max(abs(fn(sort(dati))-pnorm(sort(dati),mean(dati),sd(dati))))
a2<-max(abs(c(0,fn(sort(dati))[-n])-pnorm(sort(dati),mean(dati),sd(dati))))
max(a1,a2)
}
rez3<-replicate(1000,sqrt(n)*ff(n))
hist(rez3,prob=T,main="N(5,1/2)",xlab='',ylab='', col="green")
lines(density(rez3),type="l",lwd=2)
stat3<-ks.test(izl,"pnorm",vv,sqrt(disp))$statistic[[1]]*sqrt(n)
kvant3<-sort(rez3)[(1-alpha)*B]
###### Līdzīgi pārējiem sadalījumiem#####

######### DISKRETIE SADALIJUMI #########
n<-50
B<-500
alpha<-0.1
N<-1000
## Neg. Bin. sadalijums ###
r<-1
q<-0.5
v.nbin<-matrix(rep(0,N*2),N,2)
b<-0
```

```
for (i in 1:N)
{
dati.nbin<-rnbinom(n,r,q)
vv<-mean(dati.nbin)
q.novert<-r/(r+vv)
ff.nbin<-function(n)
{
dati<-rnbinom(n,r,q.novert)
fn<-ecdf(dati)
vv1<-mean(dati)
q.novert1<-r/(r+vv1)
a1<-max(abs(fn(sort(dati))-pnbinom(sort(dati),r,q.novert1)))
a2<-max(abs(c(0,fn(sort(dati))[-n])-pnbinom(sort(dati),r,q.novert1)))
max(a1,a2)
}
rez<-replicate(B,sqrt(n)*ff.nbin(n))
V<-ks.test(dati.nbin,"pnbinom",r,q.novert)$statistic[[1]]*sqrt(n)
### kvantiles atrasana ###
kvant<-sort(rez)[B*(1-alpha)]
v.nbin[i,1]<-V
v.nbin[i,2]<-kvant
if (v.nbin[i,1]>v.nbin[i,2])
{b<-b+1}
}


############## NEIMAŅA TESTS#############
### 1.daļa ####
## Sakuma parametri
nn<-1000
N<-1000
## Defineejam Polinomu
```

```
P1<-function(x) (x-0.5)*sqrt(12);
#### Neatkarigiem datiem
sim.neat<-c()
for (i in 1:N)
{
# Datu generesana
dati.neat<-runif(nn)
R_1<-(sqrt(nn)^(-1)*sum(P1(dati.neat)))^2
sim.neat[i]<-R_1
}
hist(sim.neat,prob=T,main=" ",xlab=" ",ylab=" ",xlim=c(0,8),ylim=c(0,2),col="light
lines(density(sim.neat),col="brown",lwd=2)
#
sim<-c()
theta<- 0.9 # Maina uz 0.3, -0.3 un -0.9
for (i in 1:N)
{
# Datu generesana
ar_1<-arima.sim(list(ar=c(theta)),n=nn, rand.gen = function(n, ...)\\
 sqrt(1-theta^2)*rnorm(nn,0,1))
dati.atk<-pnorm(ar_1)
R_1<-(sqrt(nn)^(-1)*sum(P1(dati.atk)))^2
sim[i]<-R_1
lines(density(sim),col="blue",lwd=2) #Maina krasu
legend(5,1.5,c("teta=0.3","teta=0.9","teta=-0.3","teta=-0.9","iid"),lwd=2,\\
col=c("red","blue","green","black","brown"))

### 2.daļa ###
par(mfrow=c(1,2))
# Atkārto 2 rezies, katrā grafikā pa histrogrammai
sim<-c()
theta<- 0.9
```

```
for (i in 1:N)
{
# Datu generesana
ar_1<-arima.sim(list(ar=c(theta)),n=nn, rand.gen = function(n, ...) sqrt(1-theta^2)
dati.atk<-pnorm(ar_1)
R_1<-(sqrt(nn)^(-1)*sum(P1(dati.atk)))^2
sim[i]<-R_1
}
hist(sim,prob=T, main="(2)",xlab=" ",ylab=" ",col="grey",ylim=c(0,0.05))
lines(density(sim),col="blue",lwd=2)

#################
# Atkarīgais bloku nešķeļošu butstraps
#################
nbb.fun<-function(fdata,fn,fl,fk0,fbl){
bootsamp<-c()
samp<-sample(fbl,fk0,replace=TRUE)
for (fi in 1:fk0){
bootsamp<-c(bootsamp,fdata[(fl*(samp[fi]-1)+1):(fl*(samp[fi]-1)+fl)])}
(sqrt(fn)^(-1)*sum(P1(bootsamp)))^2
}
l<-50
rez<-replicate(1000,nbb.fun(dati.atk,nn,l,ceiling(nn/l),nn/l))
lines(density(rez),type="l",lwd=2)


#################
# Atkarīgais bloku slīdošais butstraps
#################
mbb.fun<-function(fdata,fn,fl,mbb.k,mbb.b){
bootsamp<-c()
samp<-sample(mbb.b,mbb.k,replace=TRUE)
```

```
for (fi in 1:mbb.k){
bootsamp<-c(bootsamp,fdata[samp[fi]:(samp[fi]+fl-1)])}
(sqrt(fn)^(-1)*sum(P1(bootsamp)))^2
}
l<-50
b<-nn-l+1
rez1<-replicate(1000,mbb.fun(dati.atk,nn,l,nn/l,b))
lines(density(rez1),type="l",lwd=2, col="red")
```

Diplomdarbs “Butstrapa metode sadalījuma pārbaudes testiem” 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: Anna Trautmane

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.

Vadītājs: asoc.prof. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: lektors Jānis Smotrovs

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Diplomdarbs aizstāvēts valsts gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretāre: asoc. prof. Dr.math. Inese Bula _____

(paraksts)