Gludinošais butstraps

Juris Cielēns

Latvijas Universitāte

2011

Aplūkotie modeļi

Pieņēmumi

Pieņemsim, ka dota izlase $X_1, X_2, ..., X_n$ ar sadalījuma funkciju $F_1(x)$ un izlase $Y_1, Y_2, ..., Y_m$ ar sadalījuma funkciju $F_2(y)$, izlašu apjomi ir attiecīgi n un m.
Definīcija. Par divu izlašu varbūtību - varbūtību grafiku sauc funkciju

$$PP(t) := F_1(F_2^{-1}(t)),$$

kur $t \in [0, 1]$ un $F_2^{-1}(t)$ ir otrās izlases kvantiļu funkcija.

Aplūkotie modeļi

Lokācijas-skalēšanas modelis

Definīcija. Starp divām izlasēm pastāv lokācijas-skalēšanas modelis, ja

$$\mathrm{F}_1(\mathrm{x}) = \mathrm{F}_2\left(\frac{\mathrm{x}-\mu}{\sigma}\right), \mathrm{jeb}\ \mathrm{F}_1^{-1}(\mathrm{t}) = \mu + \sigma \mathrm{F}_2^{-1}(\mathrm{t}).$$

Biežāk sastopamie sadalījumi, kas pieder lokācijas-skalēšanas modeļu klasei:

- Normālais sadalījums: $\mathrm{f}(\mathrm{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(\mathrm{x}-\mu)^2}{2\sigma^2}}$
- Vienmērīgais sadalījums: $\mathrm{f}(\mathrm{x}; \mathrm{a}; \mathrm{b}) = \frac{1}{\mathrm{b}-\mathrm{a}}$, ja $\mathrm{a} \leq \mathrm{x} \leq \mathrm{b}$ un 0 - citur
- Koši sadalījums: $\mathrm{f}(\mathrm{x}; \mathrm{x}_0; \gamma) = \frac{1}{\pi} \frac{\gamma}{(\mathrm{x}-\mathrm{x}_0)^2 + \gamma^2}$

Aplūkotie modeļi

Lēmaņa alternatīvu modelis

Definīcija. Starp divām izlasēm pastāv Lēmaņa alternatīvu modelis, ja

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

Sekas. Ja funkcijas F_1 un F_2 pieder Lēmaņa alternatīvu klasei, tad ir spēkā

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

ko sauc par arī par proporcionālā riska modeli.
Biežāk pieminētais sadalījums ir Veibulla sadalījums:

$$f(x;\lambda;k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \text{kad } x \geq 0.$$

Empīriskie procesi

Par empīrisko procesu sauc funkciju

$$\mathrm{EP}(t) := \sqrt{n}(\mathrm{PP}_n(t) - \mathrm{PP}(t)).$$

- Lokācijas-skalēšanas modelim $\mathrm{PP}(t) = F_1(\sigma F_2^{-1}(t) + \mu)$
- Lēmaņa alternatīvu modelim $\mathrm{PP}(t) = F_1(F_2^{-1}(1 - (1 - t)^h))$
- Īpašgadījums: $\mathrm{PP}(t) = F_1(F_2^{-1}(t))$

$\mathrm{PP}_n(t)$ - atbilstošās funkcijas empīriskā versija ar novērtētiem parametriem.

Ticamības joslas

Kritisko vērtību noteikšanas metodes

- Kritiskās vērtības aprēķināšana izmantojot teorētisko statistikas sadalījumu - šo metodi pielieto, ja asimptotiskais sadalījums ir vienkāršs, piemēram t-sadalījums vai χ^2 sadalījums;
- Kritiskās vērtības iegūšana ar simulāciju palīdzību - pielieto, ja statistikas sadalījums nav atkarīgs no izlases (datiem), tomēr tā aprēķināšana ir samērā sarežģīta (statistikas $\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$ sadalījums ir $\sup_{0<t<1} |B(t)|$, kur $B(t)$-Brauna tilts);
- Kritiskās vērtības iegūšana ar butstrapa palīdzību - pielieto, ja statistikas sadalījums ir sarežģīts un atkarīgs no datiem

Simulācijas, butstraps un gludinošais butstraps

Simulācijas

1. No zināma sadalījuma ģenerē gadījuma lielumu izlasi;
2. Aprēķina testa statistiku un to saglabā;
3. Atkārto 1. un 2. soli 10 000 reizes, tādējādi iegūstot statistikas sadalījumu.

Butstraps

1. Izmantojot dotos datus, ģenerē gadījuma izlasi ar atkārtošanos;
2. Aprēķina testa statistiku un to saglabā;
3. Atkārto 1. un 2. soli 10 000 reizes, tādējādi iegūstot statistikas sadalījumu.

Simulācijas, butstraps un gludinošais butstraps

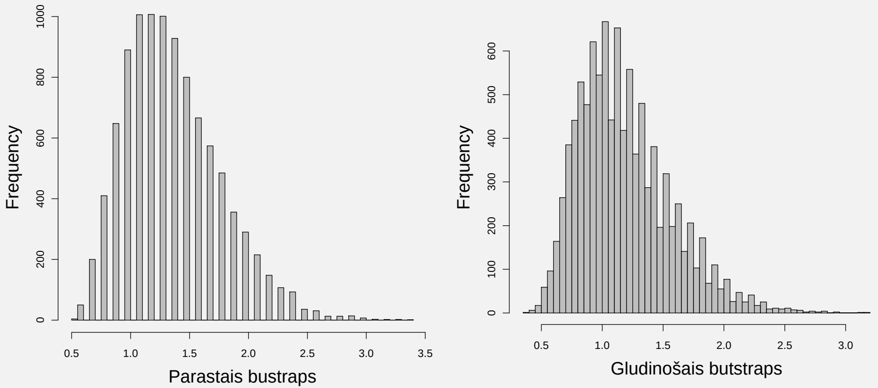
Blīvuma funkcijas un sadalījuma funkcijas kodolu novērtējums

- $\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})$
- $\hat{F}_n(x) = \int_{-\infty}^{x} \hat{f}_n(y)dy$

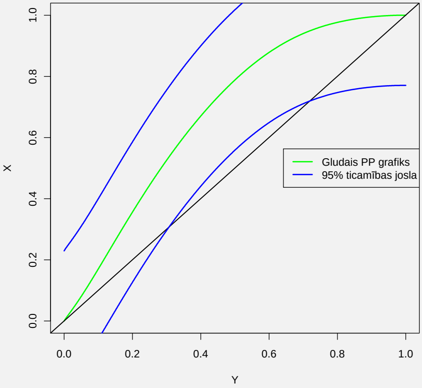
Gludinošā butstrapa algoritms

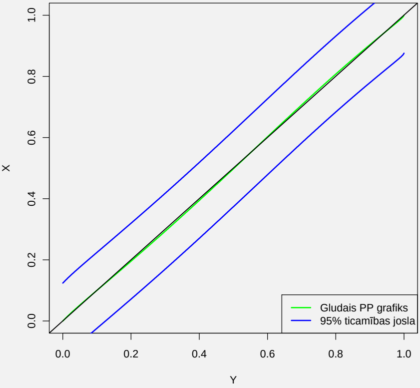
1. Izmantojot dotos datus, novērtē sadalījuma funkciju $\hat{F}_n(x)$;
2. Ģenerē intervālā $[0,1]$ vienmērīgi sadalītu gadījuma lielumu izlasi;
3. Ar inversās transformācijas palīdzību iegūst gadījuma izlasi no sadalījuma $\hat{F}_n(x)$;
4. Aprēķina interesējošo statistikas vērtību;
5. Atkārto 2.-4. soli 10 000 reizes, lai iegūtu statistikas sadalījumu un noteiktu kritisko vērtību.

Butstrapa un gludinošā butstrapa salīdzinājums



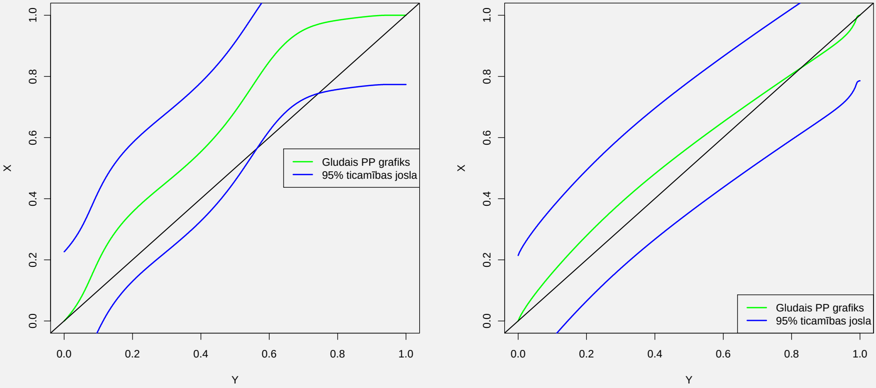
Gludās 95% ticamības joslas Lokācijas skalēšanas modelim





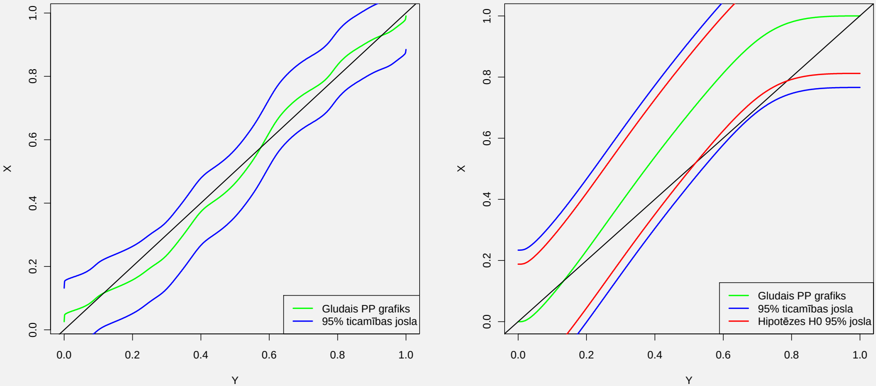
att.: N(0,1) pret N(1,1.5)

Gludās 95% ticamības joslas Lēmaņa alternatīvu modelim



att.: Weib(2,6) pret Weib(2,8)

Gludās 95% ticamības joslas



att.: N(0,1) pret U(0,1) un N(0,1) pret N(0.55,1.7)

Izmantotā literatūra

1. J. Valeinis. Confidence bands for structural relationship models. Dissertation, Goet- tingen, 2007.
2. Z. Horwath, L. Horwath and W. Zhou. Confidence bands for roc curves. Journal of Statistical Planning and Inference, 138:1894-1904, 2008.
3. G.R. Shorak and J.A. Wellner. Empirical Processes with Applications to Statistics. John Wiley Sons, New York, 1986.