VÝPOČETNÉ ASPEKTY METÓDY BOOTSTRAP

Matúš Maciak, Jaroslav Ševčík

mmatthew@matfyz.cz, jarik@matfyz.cz

Matematicko-fyzikální fakulta, UNIVERZITA KARLOVA V PRAZE
Katedra Pravděpodobnosti a matematické statistiky

Abstrakt: Účelom tejto práce bolo poskytnúť stručný prehľad modernej a v praxi stále čoraz častejšie používanej štatistickej metódy bootstrap, a tiež konkrétne ilustrovať jej použitie v štatistickom software. K tomuto účelu sme zvolili v akademickej oblasti často používaný, voľne dostupný program R a jeho komerčnú verziu S-PLUS. Okrem niekoľkých vzorových príkladov, ktoré slúžia k lepšej názornosti tejto metódy, sme pomocou simulácií aj zisťovali, ako je potrebné vhodne voliť počet boostrapových výberov. Keďže táto populárna metóda je na aplikáciu pomerne jednoduchá, snažili sme sa upozorniť aj na to, že v pozadí tohoto prístupu stojí náročná teória a mala by byť preto používaná s rozvahou.

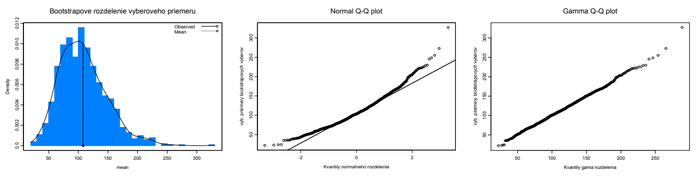
Čo je to BOOTSTRAP?

Medzi klasické výpočetné úlohy v matematickej štatistike patrí nepochybne zisťovanie bodového odhadu pre nejaký parameter, prípadne štatistiku. Často nás však zaujímajú aj **variácia** alebo **konfidenčné intervaly** tohoto odhadu. Skúmame preto jeho rozdelenie. Tradične sa v štatistike k tomuto účelu používa CLV, teda za určitých predpokladov hľadané rozdelenie pri dostatočne veľkom rozsahu dát aproximujeme normálnym rozdelením. Uplatnením moderných výpočtových prostriedkov sa však ponúka aj ďalšia možnosť, a to použiť k odhadu rozdelenia skúmaného parametru, či štatistiky tzv. resample metódy, označované aj ako computer intensive methods. Jednou z takýchto metód je aj metóda **bootstrap**, uvedená Efronom v roku 1979. Pre úplnosť spomeňme zvyšné základné resamplové techniky. Sú to **subsampling**, ktorého špeciálnym prípadom je jackknife, ďalej **permutačné testy** a **krížové overovanie**.

1. Parametrický bootstrap:

Majme náhodný výber $X_1, X_2, \ldots, X_N$ z rozdelenia $F(X, \xi)$. Skúmame štatistiku $\theta = \theta(F)$, ktorú odhadneme z dát. Parametrický bootstrap vychádza z predpokladu apriornej informácie o rozdelení $F(X, \xi)$. Predpokladáme teda konkrétny tvar tohoto rozdelenia, ktoré je určené vektorom parametrov $\xi = (\xi_1, \ldots, \xi_K)$, ktorý odhadneme z náhodného výberu $X_1, X_2, \ldots, X_N$. Z takto získaného odhadnutého rozdelenia $\widehat{F}_N(X, \xi)$ simulujeme B náhodných výberov (bootstrapové výbery, dodávame náhodu), kde $N < B << N^N$ a pre každý náhodný výber spočítame $\widehat{\theta}^*_{N,i}, i = 1, \ldots, B$. Z týchto odhadov z každého bootstrapového výberu získavame aproximáciu rozdelenia odhadu $\widehat{\theta}_N$.

Príklad 1: (výberový priemer)
Skúmajme dáta, ktoré udávajú počet hodín medzi poruchami klimatizácie v Boeingu 720 jet aircraft $(3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487)$. Zaujíma nás rozdelenie odhadu strednej hodnoty - výberového priemeru. Zdá sa, že EDF sa dá dobre nafitovať exponenciálnym, alebo gamma rozdelením. Simulujme teda náh. výbery z odhadnutého gamma rozdelenia a pomocou nich aproximujeme rozdelenie výberového priemeru. Z obrázku (obr. 1) je vidieť, že výberový priemer má výrazne zošikmené rozdelenie, čo potvrdzuje aj normal Q-Q plot, podľa ktorého nejde o normálne rozdelenie. Naproti tomu Gamma Q-Q plot naznačuje tomu, že ide o gamma rozdelenie.

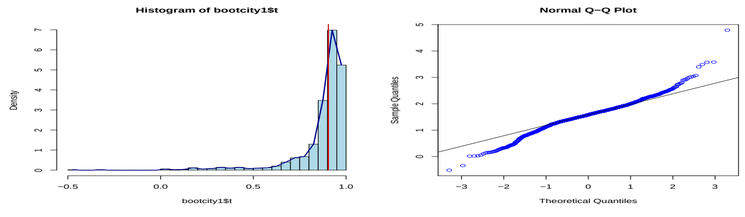


Obr.1 Aroximácia výberového priemeru dát o klimatizácii parametrickým bootstrapom, normal a gamma Q-Q plot štatistík z bootstrapových výberov.

2. Neparametrický bootstrap:

Podobne ako v prechádzajúcom prípade majme náhodný výber $X_1, X_2, \ldots, X_N$ z rozdelenia $F(X)$ a skúmame štatistiku $\theta = \theta(F)$, ktorú odhadneme z dát. V prípade neparametrického bootstrapu však nepredpokladáme apriórnu informáciu o tvare rozdelenia F. Neznámu F aproximujeme pomocou EDF, ako $\widehat{F_N}$ a z tejto simulujeme N pozorovaní, čo znamená, že z náhodného výberu $X_1, X_2, \ldots, X_N$ vyberáme N krát s opakovaním. Dostávame prvý bootstrapový výber $X^*_{1,1}, \ldots, X^*_{1,N}$. Postup opakujeme B-krát. Pre každý takto získaný bootstrapový výber spočítame odhad skúmanej štatistiky $\widehat{\theta}^*_{N,i}$, $i = 1, \ldots, B$. Z týchto odhadov napokon aproxinujeme rozdelenie odhadu $\widehat{\theta}_N$.

Príklad 2: (korelačný koeficient)
Skújmame korelačný koeficient medzi dvoma výbermi (data city z knižnice `boot` v `R` o raste populácie v rôznych amerických mestách). Použijeme neparametrický bootstrap na aproxiáciu rozdelenia korelačného koeficientu. (Obr. 2)



Obr.2 Odhad rozdelenia pre korelačný koeficient a normal Q-Q plot pre Fisherové transformácie korelačných koeficientov z boostrapových výberov.

Kedy možno bootstrap použiť?

Bootstrap možno aplikovať len na štatistiky,o ktorých vieme, že bootstrapová aproximácia spĺňa určité teoretické vlastnosti, a síce konzistenciu a asymptotickú presnosť (rýchlosť konvergencie ku skutočnému rozdeleniu, Edgeworthov rozvoj). Tieto požiadavky sú splnené u väčšiny bežných štatistík, ako sú napr. **hladké funkcie výberových momentov** (priemer, rozptyl, atď), **odhady v lineárnych a zobecnených lineárnych modeloch**, väčšina **max. vierohodných odhadov** a veľa štatistík odvodených z **časových rád**. V lineárnych modeloch existujú dva prístupy pri aplikácii bootstrapu na odhad regresných koeficientov a to **resampling residuals** a **resampling cases**. V poslednej dobe sa často využíva nový spôsob generovania náhodných výberov, označovaný ako **wild bootstrap**. Tento vykazuje lepšie vlastnosti a je vhodný v prípade heteroskedastických dát. U časových rád sa pre konštrukciu bootstrapových výberov používa resamplovanie blokov. Základné prístupy sú **based resampling, block resampling** a metóda **blocks of blocks**.

Bootstrap verzus konvenčný prístup(CLV)

Výhody Bootstrapu: Je jednoduchý na použitie. Oslobodzuje nás od dvoch obmedzení klasického prístupu, a síce predpokladu zvonového rozdelenia dát a nutnosti zamerať sa na štatistiky, ktorých teoretické vlastnosti môžu byť analyzované matematicky. Často pozorujeme rýchlejšiu konvergenciu ku skutočnému rozdeleniu ak v prípade CLV, je teda vhodný aj pre menšie výbery.

Nevýhody Bootstrapu: Niekedy môže poskytnúť klamlivý obraz skutočnosti (napr. u dát s odľahlými pozorovaniami)

Aplikácia bootstrapu v R 2.2.0

Existujú knižnice **boot**, **bootstrap** a **simpleboot**, ktoré umožňujú aplikáciu bootstrapových štatistických metód v `R` (podľa odporúčaní sa vyžaduje verzia `R 2.2.0` a viac).

K dispozícii sú oba prístupy, parametrický aj neparametrický bootstrap. Funkcie pre jednoduchý bootstrap (boot(), bootstrap() a pairs.boot()) s možnými metódami **ordinary**, **balanced**, **anthitetic bootstrap** a **permutation**. Na bootstrap v lineárnom modele slúži funkcia lm.boot() s voliteľnými metódami **model-based resampling** a **resampling cases**. K použitiu bootstrapu v časových radách sú nutné ešte dodatočné knižnice (tseries, zoo, quadprog).Funkcie tsboot() a tsbootstrap() umožňujú využiť metódy **Model-based resampling**, **Block resampling** a tiež metódu **Blocks-of-blocks**. V prípade niektorej z metód blokov je k dispozícia voľba pevnej dĺžky bloku, alebo dĺžky bloku s geometrickým rozdelením s voliteľnou strednou hodnotou l.

Konfidenčné intervaly

Klasické konfidenčné intervaly sú založené na asymptotickej aproximácii, ktorá je v praxi často nepresná (interval $\widehat{\theta} \pm z^{(\alpha)}\widehat{\sigma}$, kde $z^{(\alpha)}$ sú percentily normálneho rozdelenia). Pri bootstrape sa okrem empirických percentilov využívajú aj vylepšené konfidenčné intervaly, a to **ABC**, $\mathbf{BC}_\alpha$ a **t-bootstrap**.

1. BC_α Intervaly pre parameter θ:
$\rightarrow$ bootstrapové výbery $X^*_{i,N}, X^*_{i,N}, \ldots, X^*_{i,N}$ pre $i = 1, \ldots, B \rightarrow$ odhady $\widehat{\theta}^*_{N,1}, \ldots, \widehat{\theta}^*_{N,B} \rightarrow$ odhad distribučnej funkcie θ , $\widehat{G}(\theta)$ ako $\widehat{G}(c) = \sum_{i=1}^{B} \mathbb{I}_{\{\widehat{\theta}^*_{N,i} \leq c\}}$

$\Rightarrow BC_\alpha$ interval je $\left[\widehat{\theta}^*_{BC_\alpha}[\alpha],\ \widehat{\theta}^*_{BC_\alpha}[1-\alpha]\right]$, kde

$$\widehat{\theta}^*_{BC_\alpha}[\alpha] = \widehat{G}^{-1}\phi\left(z_0 + \frac{z_0 + z^\alpha}{1 - a(z_0 + z^\alpha)}\right),$$

kde $z^\alpha = \phi^{-1}(\alpha)$, z_0 je biasová korekcia a a je akcelerácia.

2. ABC Intervaly pre parameter θ:
$\rightarrow$ analytická verzia BC_α konfidenčných intervalov.
$\Rightarrow ABC$ konfidenčný interval je $\left[\widehat{\theta}^*_{ABC}[\alpha],\ \widehat{\theta}^*_{ABC}[1-\alpha]\right]$, kde

$$\widehat{\theta}^*_{ABC}[\alpha] = \widehat{\theta} + \frac{\omega}{(1 - \widehat{a}\omega)^2}\sqrt{\widehat{\theta}}, \quad \text{kde } \omega = z_0 + z^\alpha.$$

3. t-bootstrap intervaly pre parameter θ:
$\rightarrow$ náhodný výber $X^*_1, X^*_2, \ldots, X^*_N \rightarrow$ odhadneme $\widehat{\theta}(X)$ a smerodatnú odchýlku $\widehat{\sigma}_\theta \rightarrow$ statistika $T = \frac{\widehat{\theta} - \theta}{\widehat{\sigma}_\theta} \rightarrow$ z bootstrapových výberov odhadneme kvantily $T^{(\alpha)}$.

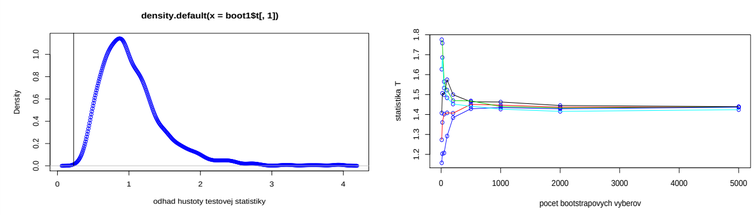
$\Rightarrow t$-bootstrap interval je $\left[\widehat{\theta} - \widehat{\sigma}T^{(\alpha)},\ \widehat{\theta} - \widehat{\sigma}T^{(1-\alpha)}\right]$

Príklad 3: (výberový korelačný koeficient)
Porovnanie konfidenčných intervalov, počítaných rôznymi metódami pre odhad korelačného koeficientu ρ, kde $\widehat{\rho} = \mathbf{0.723}$:

Parametrický prístup				
α	ABC	BC_α	t-bootstrap	Standard
0.05	0.47	0.47	0.45	0.55
0.95	0.86	0.86	0.87	0.90
Neparametrický prístup				
α	ABC	BC_α	t-bootstrap	Standard
0.05	0.56	0.55	0.51	0.59
0.95	0.83	0.85	0.86	0.85

K výpočtu $\mathbf{BC}_\alpha$, **ABC** a **t-bootstrap** konfidenčných limít sú určené funkcie abc.ci(), boot.ci(), abc.par(), abc.non(), bcanon() a boott().

Príklad 4: (testová štatistika - časová rada)
Máme časovú radu $\{X_t,\ t = \{1, 2, \ldots, 309\}\}$, kde X_t je logaritmus podielu ceny ovčej vlny na austrálskom trhu a minimálnej ceny stanovenej na daný týždeň. Testujeme hypotézu H_0 že ide o náhodnú prechádzku proti alternatíve H_1, že ide o autoregresnú postupnosť prvého radu. Volíme testovú štatistiku $\mathbf{T} = \frac{\mathbf{1} - \widehat{\alpha}}{\mathbf{s}_{\widehat{\alpha}}}$ kde $\widehat{\alpha}$ je odhad autoregresného koeficientu a $s_{\widehat{\alpha}}$ je smerodatná odchýlka pre $\widehat{\alpha}$. Metódou **Block resampling** s geometrickým rozdelením dĺžky blokov so strednou hodnotou $l = 20$ a počtom opakováni $R = 5000$ odhadneme rozdelenie pre testovú štatistiku T.



Obr.3 Odhad hustoty rozdelenia testovej štatistiky T, spočítaný na základe 1000 bootstrapových výberov a odhad štatistiky spočítaný z pôvodných dát. Na druhom obrázku je bootstrapový odhad testovej štatistiky T, počítaný pre rôzne hodnoty počtu opakovaní bootstrapových výberov v piatich prípadoch.

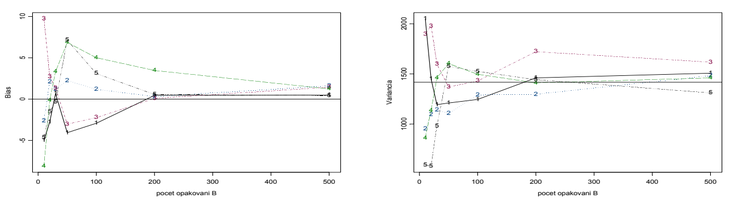
Hodnoty štatistiky T blízko nule by svedčili v prospech nulovej hypotézy. Na základe zkonštruovaných konfidenčných intervalov pre štatistiku T so spoľahlivosťou 95% však zamietneme nulovú hypotézu H_0.

Aplikácia v S-plus

Pre účely bootstrapu je nutné stiahnúť knižnicu **S+Resample**. K dispozícii je na stránke http://www.insightful.com/downloads/libraries. Knižnica podporuje **parametrický** aj **neparametrický** bootstrap, **premutačné testy**, metódu **jackknife**, odhady v **lineárnych modeloch**, atď. Implementácia funkcií a význam parametrov je podobný ako v **R**-ku. K dispozícii sú funkcie sample(), bootstrap(), pbootstrap(), sbootstrap(), pre permutačné testy je určená funkcia permutationTest() a pre empirické percentily a iné konfidenčné intervaly funkcie limits.emp(), limits.bca().

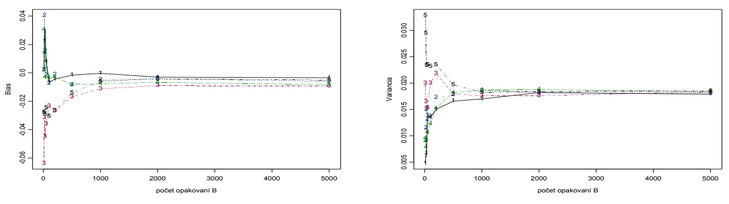
Voľba počtu opakovaní

V prípade jednoduchých štatistík, ako je napr. **výberový priemer**, dobré výsledky dostávame už pri $B = 500$ opakovaniach (Obr. 4). K tomuto účelu sme simulovali hodnotu vychýlenia a rozptylu bootstrapového odhadu tejto štatistiky (data o klimatizácii) pri 10, 20, 50, 100, 200 a 500 opakovaniach. V tomto prípade ľahko teoreticky zistíme, k čomu by malo konvergovať toto vychýlenie a rozptyl, $\mathbf{E}^*(\overline{X^*_N}) = \overline{X_N}$, $\mathbf{Var}^*(\overline{X^*_N}) = \frac{S^2_N}{N}$, kde $S^2_N = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X}_N)^2$.



Obr.4 Empirický bias a rozptyl bootstrapového odhadu strednej hodnoty v piatich prípadoch pre rôzny počet opakovaní B (data o klimatizácii).

V prípade komplikovanejších štatistík občas potrebujeme väčší počet bootstrapových opakovaní, aby sme dosiahli menšie **vychýlenie** a **rozptyl** bootstrapového odhadu. Pre konfidenčné intervaly sa doporučuje hodnota B aspoň 1000, v prípade niektorých iných štatistík to môže byť aj 10000 a viac (Nezabudnime, že počet bootstrapových výberov B musí byť väčší ako je rozsah dát N). Treba preto pri aplikácii bootstrapu myslieť aj na túto skutočnosť (v `S-PLUS` je defaultne nastavené $B = 1000$, v `R` nie je defaultne nastavenie). Počet opakovaní bootstrapového odhadu skúmali aj v prípade testovej štatistiky z príkladu 4, (Obr. 3). Na záver ešte pridávame simuláciu počtu opakovaní bootstrapového odhadu pre korelačný koeficient (Obr. 5).



Obr.5 Empirický bias a rozptyl bootstrapového odhadu korelačného koeficientu v piatich prípadoch pre rôzny počet opakovaní B (data o štandardizovaných testoch z helpu ku knižnici `S+Resample`).

Vplyv odľahlých pozorovaní

Napokon sa ešte pozrime, ako odhaliť, či bootstrapový odhad môže byť poškodený odľahlým pozorovaním, teda či existuje v datovom súbore odľahlé pozorovanie. Na tento účel sa používa metóda **jackknife**, ktorá spočíta bootsrapový odhad štatistiky N-krát pre každý datový súbor, ktorý vznikne z pôvodného vynechaním jedného pozorovania. Ak sa po vynechaní nejakého pozorovania bootstrapový odhad výrazne odlišuje od ostatných klasifikuje sa ako odľahlé. V `R` k dispozícii funkcia **jackknife()** a v `S-PLUS` funkcia **jack.after.bootstrap()**

Literatúra

[1] A.C. Davison a D.V. Hinkley. (1997). *Bootstrap Methods and their application*, Cambridge University Press

Poďakovanie:
Na záver by sme sa radi poďakovali ČSOB banke za grant, vďaka ktorému sme sa mohli zúčastniť na konferencii ROBUST 2006.