

Metoda BOOTSTRAP

Jana Veverková

Statistický seminář III.

22. 11. 2011

Základní myšlenka

- ▶ Mějme náhodný výběr z rozdělení s distribuční funkcí $F(\theta)$.

$$\mathbb{X} = (X_1, \dots, X_n) \quad X_i \stackrel{iid}{\sim} F(\theta)$$

- ▶ Běžně odhadem parametru θ rozumíme libovolnou statistiku T_n , která je funkcí náhodného výběru $\mathbb{X}$.

$$\hat{\theta}_n = T_n(X_1, \dots, X_n)$$

- ▶ Jak vypadá odhad metodou BOOTSTRAP?

- ▶ Zkonstruujeme empirickou distribuční funkci

$$F_n = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{[X_i < x]}$$

- ▶ Provedeme náhodný výběr o rozsahu n z rozdělení s distr. funkcí F_n .

$\mathbb{X}^* = (X_1^*, \dots, X_n^*)$... bootstrapový výběr

- ▶ Bootstrapovým odhadem parametru $\theta^* = \theta(F_n)$ rozumíme statistiku T_n^* , která je funkcí bootstrapového výběru

$$T_n^* = T_n(X_1^*, \dots, X_n^*)$$

- ▶ Podobně definujeme empirickou distribuční funkci bootstrapového výběru

$$F_n^* = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[X_i^* < x]$$

Rozdělení bootstrapového výběru

- ▶ Pro marginální pravděpodobnosti platí

$$P[X_i^* = X_j \mid X_1 = x_1, \dots, X_n = x_n] = \frac{1}{n} \quad \forall i, j = 1, \dots, n$$

- ▶ Vzhledem k nezávislosti bootstrapového výběru pro všechny výběry platí

$$P[X_1^* = X_1, \dots, X_n^* = X_n \mid X_1 = x_1, \dots, X_n = x_n] = \frac{1}{n^n}$$

Charakteristiky bootstrapového odhadu

- ▶ Definujeme následující charakteristiky

$$E^* T_n^* = \int T_n(x_1, \dots, x_n) d(F_n(x_1), \dots, F_n(x_n))$$

$$\text{var}^* T_n^* = \int [T_n(x_1, \dots, x_n) - E^* T_n^*]^2 d(F_n(x_1), \dots, F_n(x_n))$$

- ▶ a distribuční funkci

$$H_n^*(x) = P^*(T_n(X_1^*, \dots, X_n^*) \leq x)$$

- ▶ Jde o tzv. teoretické charakteristiky a teoretickou distribuční funkci získané metodou bootstrap.

Příklad I. (1)

- ▶ Mějme náhodný výběr z neznámého rozdělení $F_X(x)$

$$X_1 = 3, X_2 = 1, X_3 = 4$$

- ▶ Zajímá nás odhad střední hodnoty μ . Běžně užívaný výběrový průměr má hodnotu

$$\bar{X}_3 = \frac{8}{3}$$

Odhad metodou bootstrap

- ▶ Existuje 27 různých bootstrapových výběrů, k nim spočteme příslušné bootstrapové průměry $\bar{X}_3^*$.

Příklad I. (2)

Bootstrapový výběr	Bootstrapový průměr
1, 1, 1	$\overline{X}_{3,1}^* = 1$
1, 1, 3	$\overline{X}_{3,2}^* = \frac{5}{3}$
1, 1, 4	$\overline{X}_{3,3}^* = 2$
1, 3, 1	$\overline{X}_{3,4}^* = \frac{5}{3}$
$\vdots$	$\vdots$
4, 4, 4	$\overline{X}_{3,27}^* = 4$

- Vypočteme střední hodnotu bootstrapového průměru

$$E^* \overline{X}_n^* = \frac{\sum_{i=1}^{27} \overline{X}_{n,i}^*}{27} = \frac{8}{3}$$

Aproximace bootstrapového rozdělení

- ▶ Počet bootstrapových výběrů je n^n a nemusí být tedy vždy možné vypočítat všechny hodnoty bootstrapového odhadu.
- ▶ Řešení nabízí metoda MONTE CARLO:
 - ▶ Z náhodného výběru $X_1, \dots, X_n$ provedeme B bootstrapových výběrů.
 - ▶ Získáme B odhadů parametru θ^* .
 - ▶ Z těchto hodnot odhadneme rozdělení parametru θ^* .

Proč bootstrap?

- ▶ Zajímá nás rozdělení H_n statistiky $T_n(X_1, \dots, X_n)$.
- ▶ F známá
 - Explicitně vypočítat H_n - nelze vždy.
 - Aproximovat H_n metodou Monte Carlo.
- ▶ F neznámá
 - Aproximovat H_n asymptotickým rozdělením na základě limitních vět.
 - Aproximovat H_n metodou BOOTSTRAP (kombinace tzv. substitučního principu a metody Monte Carlo).

Příklad II. (1)

- ▶ Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení $N(\mu, \sigma^2)$.
- ▶ Přirozeným odhadem parametru

$$\theta = e^{\mu}$$

je statistika

$$T_n = e^{\bar{X}_n},$$

kde $\bar{X}_n$ je výběrový průměr.

- ▶ Statistika T_n má logaritmicko normální rozdělení s parametry μ a $\frac{\sigma^2}{n}$.

Příklad II. (2)

- ▶ Zabývejme se rozdělením statistiky

$$R_n = \sqrt{n}(T_n - \theta) = \sqrt{n}(e^{\bar{X}_n} - e^{\mu})$$

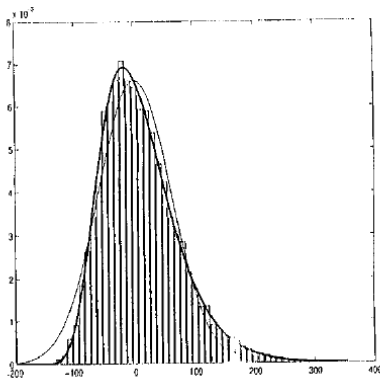
- ▶ Za předpokladu normálního rozdělení má R_n hustotu

$$h_n(x) = \frac{\sqrt{n}}{\sigma(x + e^{\mu}\sqrt{n})} \varphi\left(\left(\ln\left(\frac{x}{\sqrt{n}} + e^{\mu}\right) - \mu\right) \frac{\sqrt{n}}{\sigma}\right)$$

- ▶ Z Taylorova rozvoje lze odvodit, že R_n má asymptoticky $N(0, e^{2\mu}\sigma^2)$
- ▶ Metodou bootstrap: Nechť $X_1^*, \dots, X_n^*$ je bootstrapový výběr, pak bootstrapová statistika je

$$R_n^* = \sqrt{n}(e^{\bar{X}_n^*} - e^{\mu^*}) = \sqrt{n}(e^{\bar{X}_n^*} - e^{\bar{X}_n})$$

Příklad II. (3)



Hustota skutečného rozdělení statistiky pro $n = 100$ z $N(3,9)$, hustota asymptotického normálního rozdělení, histogram odpovídající bootstrapové statistiky pro $B=10\,000$.

Bootstrapová empirická distribuční funkce

- ▶ Pro empirickou distribuční funkci $F_n(x)$ platí:

$$F_n(x) \xrightarrow{s.j.} F(x) \quad n \rightarrow \infty$$

$$EF_n(x) = F(x)$$

- ▶ Analogická tvrzení platí i pro bootstrapovou empirickou distribuční funkci $F_n^*(x)$:

$$E^* F_n^*(x) = \frac{1}{n} \sum_{i=1}^n E^* \mathbb{I}_{[X_i^* < x]} = F_n(x)$$

- ▶ Pro bootstrapový rozptyl rozdílu funkcí platí

$$\text{var}^* [F_n^*(x) - F_n(x)] = \frac{1}{n} F_n(x)(1 - F_n(x))$$

- ▶ Jelikož $F_n(x)(1 - F_n(x)) \leq \frac{1}{4}$, platí dále

$$\text{var}^* [F_n^*(x) - F_n(x)] \leq \frac{1}{4n} \xrightarrow{n \rightarrow \infty} 0$$

- ▶ Odtud plyne

$$F_n^*(x) \xrightarrow{P^*} F_n(x) \quad n \rightarrow \infty$$

Bootstrap pro průměr

- ▶ Mějme náhodný výběr z rozdělení F s konečnou střední hodnotou μ a konečným rozptylem σ^2

$$X_1, \dots, X_n \stackrel{iid}{\sim} F$$

- ▶ Provedme bootstrapový výběr

$$X_1^*, \dots, X_n^* \stackrel{iid}{\sim} F_n$$

- ▶ Bootstrapovým odhadem parametru μ je v tomto případě

$$\hat{\mu}^* = \bar{X}_n^* = \frac{1}{n} \sum_{i=1}^n X_i^*$$

Charakteristiky bootstrapového průměru

- Podmíněná střední hodnota

$$E^* \left[\bar{X}_n^* \right] = \frac{1}{n} \sum_{i=1}^n E^* X_i^* = E^* X_1^* = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

- Podmíněný rozptyl

$$\text{var}^* \left[\bar{X}_n^* \right] = \frac{1}{n} \text{var}^* X_1^* = \frac{s_n^2}{n}$$

kde $s_n^2 = \frac{1}{n} \sum_{i=1}^n \left(X_i - \bar{X}_n \right)^2$.

Asymptotické vlastnosti bootstrapového průměru

Věta 1

Platí-li $E|X|^3 < \infty$ a není-li funkce F řádkovitá, pak pro všechna x skoro jistě platí

$$P^* \left[\frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{s_n} < x \right] = \Phi(x) + \frac{\mu_3(1 - x^2)}{6\sigma^3\sqrt{n}}\phi(x) + o\left(\frac{1}{\sqrt{n}}\right),$$

kde $\Phi(x)$ je distribuční funkce $N(0, 1)$ rozdělení a $\phi(x)$ jeho hustota.

Pozn.: Rozdíl mezi standardní aproximací $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ a bootstrapovou $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/s_n$ se zvyšuje s rostoucí šikmostí původního rozdělení.

Redukce vychýlení odhadu bootstrapem

- ▶ Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení s parametry μ a σ^2 , g je spojitá funkce taková, že $E|g(\bar{X}_n)| < \infty$, uvažujme parametr $\theta = g(\mu)$.
- ▶ Výběrový průměr $\bar{X}_n$ je nestranný a konzistentní odhad μ , potom $g(\bar{X}_n)$ je konzistentní odhad $g(\mu)$, ale obecně není nestranný (pokud g není lineární).
- ▶ Vychýlení označme

$$b_n = Eg(\bar{X}_n) - g(\mu)$$

- ▶ Z Taylorova rozvoje¹ plyne

$$b_n = E(g(\bar{X}_n) - g(\mu)) = \frac{1}{2} \frac{\sigma^2}{n} g''(\mu) + O(n^{-2})$$

¹ $g(\bar{X}_n) - g(\mu) = (\bar{X}_n - \mu)g'(\mu) + \frac{1}{2}(\bar{X}_n - \mu)^2 g''(\mu) + R_n$

- ▶ Je-li $X_1^*, \dots, X_n^*$ bootstrapový výběr, potom bootstrapová verze vychýlení je

$$b_n^* = E^* g(\bar{X}_n^*) - g(\bar{X}_n)$$

- ▶ Z Taylorova rozvoje² opět plyne

$$E^*(g(\bar{X}_n^*) - g(\bar{X}_n)) = \frac{1}{2} \frac{\sigma^{*2}}{n} g''(\bar{X}_n) + O(n^{-2})$$

- ▶ Z dalšího Taylorova rozvoje³ získáme

$$Eb_n^* = \frac{1}{2} \frac{\sigma^2}{n} g''(\mu) + O(n^{-2})$$

² $g(\bar{X}_n^*) - g(\bar{X}_n) = (\bar{X}_n - \bar{X}_n)g'(\bar{X}_n) + \frac{1}{2}(\bar{X}_n^* - \bar{X}_n)^2 g''(\bar{X}_n) + R_n$

³ $g''(\bar{X}_n) = g''(\mu) + \frac{1}{2}(\bar{X}_n - \mu)g'''(\mu) + R_n$

- Uvažujme nyní místo odhadu $g(\bar{X}_n)$ jeho opravu $g(\bar{X}_n) - b_n^*$. Vychýlení takto opraveného odhadu je

$$E[g(\bar{X}_n) - b_n^*] - g(\mu) = b_n - Eb_n^* = O(n^{-2}),$$

což je zřejmě lepší výsledek.

Příklad III.

- ▶ Uvažujme parametr $\theta = e^\mu$ a jeho výběrový odhad $\hat{\theta}_n = e^{\bar{X}_n}$.
- ▶ Srovnáme vychýlení odhadu $b_n = E\hat{\theta}_n - \theta$ a vychýlení opraveného odhadu $\hat{\theta}_n^c = \hat{\theta}_n - b_n^*$, kde $b_n^* = E^*\hat{\theta}_n^* - \hat{\theta}_n$.
- ▶ Náhodné výběry byly generovány z $N(3,9)$, skutečná hodnota $\theta = 20,086$. Pro každý výběr o rozsahu n bylo generováno 500 bootstrapových výběrů.

$\hat{\theta}_{20} - \theta$	5,186
$\hat{\theta}_{20}^c - \theta$	-1,011
$\hat{\theta}_{50} - \theta$	1,863
$\hat{\theta}_{50}^c - \theta$	-0,145
$\hat{\theta}_{100} - \theta$	0,940
$\hat{\theta}_{100}^c - \theta$	-0,041

Bootstrapové konfidenční intervaly

- ▶ Pro nalezení konfidenčních intervalů pomocí bootstrapu existuje několik postupů.

Percentilová metoda

- ▶ Nechť $X_1, \dots, X_n \stackrel{iid}{\sim} F$.
- ▶ Chceme určit $(1-\alpha)$ konfidenční interval pro $\theta = \theta(F)$ na základě odhadu $\hat{\theta} = \theta(F_n)$ a bootstrapového odhadu $\hat{\theta}^* = \theta(F_n^*)$.
- ▶ Označme rozdělení $\hat{\theta}^*$ jako

$$H_n^*(x) = P^* \left[\hat{\theta}^* < x \right]$$

- ▶ Pokud toto rozdělení nelze určit, odhadneme ho metodou Monte Carlo

$$H_n^*(x) \doteq \frac{1}{B} \sum_{b=1}^B \mathbb{I}_{[\hat{\theta}_b^* < x]}$$

- ▶ Pro $\alpha \in (0, 1)$ definujeme kvantily rozdělení $H_n^*(x)$ jako

$$\hat{\theta}_D^* = \hat{\theta}^*\left(\frac{\alpha}{2}\right) = H_n^{*-1}\left(\frac{\alpha}{2}\right)$$

$$\hat{\theta}_H^* = \hat{\theta}^*\left(1 - \frac{\alpha}{2}\right) = H_n^{*-1}\left(1 - \frac{\alpha}{2}\right)$$

- ▶ Pak platí

$$P^* \left[\hat{\theta}_D^* \leq \theta \leq \hat{\theta}_H^* \right] \doteq 1 - \alpha$$

- Interval

$$\left\langle H_n^{*-1}\left(\frac{\alpha}{2}\right), H_n^{*-1}\left(1 - \frac{\alpha}{2}\right) \right\rangle$$

je přibližným centrálním konfidenčním intervalem parametru θ na hladině spolehlivosti $1 - \alpha$.

BC-percentilová metoda

- V případě, že

$$P^*[\hat{\theta}^* < \hat{\theta}] \neq 0.5,$$

tj. je-li bootstrapové rozdělení vychýlené vzhledem ke skutečné hodnotě parametru.

- ▶ Tato metoda opravuje toto vychýlení (BC - "*bias corrected*")
- ▶ Definujme

$$z_0 = \Phi^{-1}(H_n^*(\hat{\theta})),$$

kde $\Phi(x)$ je distribuční funkce $N(0,1)$.

- ▶ Potom interval

$$\left\langle H_n^{*-1} \left(\Phi(2z_0 - \mu_{(\frac{\alpha}{2})}) \right), H_n^{*-1} \left(\Phi(2z_0 + \mu_{(\frac{\alpha}{2})}) \right) \right\rangle,$$

kde $\mu_{(\frac{\alpha}{2})}$ je kvantil $N(0,1)$, je přibližný $(1 - \alpha)$ konfidenční interval parametru θ .

Studentizované intervaly spolehlivosti

- Uvažujme studentizovanou statistiku odhadu $\hat{\theta}_n$ parametru θ , tj.

$$R_n = \frac{\hat{\theta}_n - \theta}{S_n}$$

a její bootstrapový protějšek

$$R_n^* = \frac{\hat{\theta}_n^* - \theta_n}{S_n^*}$$

- Běžný interval spolehlivosti pro θ je

$$\left(\hat{\theta}_n - \gamma_{1-\frac{\alpha}{2}} S_n, \hat{\theta}_n - \gamma_{\frac{\alpha}{2}} S_n \right),$$

kde γ_p je p-kvantil distribuční funkce statistiky R_n .

- ▶ Bootstrapový interval spolehlivosti je

$$\left(\hat{\theta}_n^* - \gamma_{1-\frac{\alpha}{2}}^* S_n^*, \hat{\theta}_n^* - \gamma_{\frac{\alpha}{2}}^* S_n^* \right)$$

- ▶ Lze ukázat, že takto sestrojené intervaly pokrývají neznámý parametr s lepší přesností, než klasické intervaly využívající asymptotickou normalitu statistiky R_n .
- ▶ Nevýhodou je velký počet výpočetních operací. Je třeba B_1 bootstrapových výběrů k odhadu S_n^* a dalších B_2 výběrů pro odhad kvantilů, tj. celkem $B_1 B_2$ výběrů.

Použitelnost metody Bootstrap

Přednosti

- ▶ Schopnost redukovat vychýlení odhadů.
- ▶ Schopnost zpřesňovat rozdělení statistik a tedy i přesnost pokrytí parametrů konfidenčními intervaly pro studentizované statistiky.

Omezení použitelnosti

- ▶ Pro rozdělení s nekonečným rozptylem neodhaduje bootstrap konzistentně rozdělení hladkých funkcí výběrového průměru.
- ▶ Neodhaduje konzistentně rozdělení odhadů parametrů, které leží na hranici parametrické množiny.
- ▶ Neodhaduje konzistentně rozdělení extrémálních odhadů.
- ▶ Jestliže X_n nejsou nezávislé, bootstrapové statistiky nedávají konzistentní odhady parametrů a rozdělení.