

11. cvičení

Úloha 1 (MLE společně)

Máme náhodný výběr $X_1, \dots, X_n \sim \text{Geom}(p)$, jako parametr nás zajímá $\vartheta = p$. Navrhněte bodový odhad ϑ metodou maximální věrohodnosti.

Řešení

$L_X(x; \vartheta) = \prod_{i=1}^n (1 - \vartheta)^{x_i - 1} \vartheta = \vartheta^n \cdot \prod_{i=1}^n (1 - \vartheta)^{x_i - 1}$, log-likelihood je $\log(L_X(x; \vartheta)) = n \log(\vartheta) + \log(1 - \vartheta) \sum_{i=1}^n (x_i - 1)$. Když toto zderivujeme podle ϑ , dostaneme $\frac{\partial}{\partial \vartheta} \log(L_X(x; \vartheta)) = \frac{n}{\vartheta} + \frac{(\sum_{i=1}^n x_i) - n}{1 - \vartheta}$, což je rovno nule pro $\vartheta = \frac{n}{\sum_{i=1}^n x_i}$ (a navíc se dá ověřit, že to je maximum).

Úloha 2 (Opět mince)

Hodíme 100krát spravedlivou mincí. Kolikrát nám v průměru padne orel? Pomocí CLV dále odhadněte pravděpodobnost, že padne více než 60 orlů.

Řešení

$X \sim \text{Bin}(100, 1/2)$, tedy $\mathbb{E}[X] = 50$, a rozptyl je 25. Navíc $X = \sum_{i=1}^{100} X_i$, kde $X_i \sim \text{Bern}(1/2)$, a X_i mají rozptyl $1/4$. Potom víme, že $\frac{X-50}{\sqrt{25}}$ je přibližně rozdělena podle $N(0, 1)$, a tedy $P[X \geq 60] = P[\frac{X-50}{\sqrt{25}} \geq 2]$, což můžeme z CLV aproximovat pomocí $1 - \Phi(2) \approx 0.02$.

Úloha 3

Němci vyrábějí tanky s pořadovými čísly $1, \dots, N$ pro neznámé N . Ukořistíme k z nich a vidíme pořadová čísla X_1, \dots, X_k , tj. rovnoměrně náhodnou k -prvkovou podmnožinu $\{1, \dots, N\}$. Necht' $M = \max(X_1, \dots, X_k)$.

- Ukažte, že $P(M = m) = \frac{\binom{m-1}{k-1}}{\binom{N}{k}}$ pro $m \in \{k, \dots, N\}$.
- Připomeňte si, že M je MLE pro N (ukazovali jsme na přednášce).
- Spočtete $\mathbb{E}(M) = \frac{k(N+1)}{k+1}$. Může se hodit hockey-stick identity: $\sum_{m=k}^N \binom{m}{k} = \binom{N+1}{k+1}$. Pak si připomeňte, jak z toho plyne nestranný odhad $\hat{N}_{unbiased} = \frac{k+1}{k} M - 1$ (**Pozor: na přednášce bylo místo -1 napsáno $-\frac{k+1}{k}$ — to byla chyba.**)

Řešení a) Celkem máme $\binom{N}{k}$ možností výběru, a z nich právě $\binom{m-1}{k-1}$ splňuje, že m je maximum.

b) Dle přednášky

c) $\mathbb{E}[M] = \sum_{m=k}^N m \cdot \frac{\binom{m-1}{k-1}}{\binom{N}{k}} = \frac{\sum_{m=k}^N k \cdot \binom{m}{k}}{\binom{N}{k}} = \frac{k \sum_{m=k}^N \binom{m}{k}}{\binom{N}{k}} = \frac{k \cdot \binom{N+1}{k+1}}{\binom{N}{k}} = \frac{k(N+1)}{k+1}$. Nestranný odhad z toho plyne tak, že vidíme, že střední hodnota se od N liší, tak algebraicky upravujeme M , aby ve střední hodnotě vyšlo N .

Úloha 4 (MLE once more)

Máme náhodný výběr $X_1, \dots, X_n \sim N(\mu, 1)$ – data na vstupu x_1, \dots, x_n tedy pochází z normální distribuce s neznámým středem μ , ale známou směrodatnou odchylkou 1.

- Napište věrohodnostní (likelihood) funkci $p_\theta(x)$; později se může hodit pracovat s funkcí $\log p_\theta(x)$, které se říká log-likelihood.
- Derivací spočtete $\hat{\mu}_{MLE}$, mělo by vyjít $\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$.
- Ověřte, že $\hat{\mu}_{MLE}$ je nestranný.
- Přesvědčte se, že kdyby směrodatná odchylka nebyla 1, ale byl to jakýkoliv (nám známý) parametr σ^2 , $\hat{\mu}_{MLE}$ by vyšlo úplně stejně; volba $\sigma^2 = 1$ jen zjednodušuje výpočet.

Řešení a) $p_\theta(x) = \Pr[X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_n = x_n] = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu)^2 / 2}$, $\log p_\theta(x) = \sum (\log(\frac{1}{\sqrt{2\pi}}) - (x_i - \mu)^2 / 2)$

- b) Derivujeme sčítance jednotlivě, tedy dostaneme $\sum -x_i + 2\mu$, a při hledání extrému chceme derivaci rovnou nule, což je právě tehdy, když $\mu = \frac{1}{n} \sum_{i=1}^n x_i$, což je i náš odhad.
- c) Zjevně střední hodnota je z linearity μ .
- d) σ se objevuje ve jmenovateli, takže po zlogaritmování z toho budou jenom aditivní konstanty $-\log(\sigma^2)$, což extrémy nijak nemění.

Úloha 5 (Nestranný \neq dobrý)

Najděte příklad nějakého nestranného estimátoru (třeba pro problémy z předchozích úloh), který je zjevně hodně špatný.

Řešení

Třeba estimátor, který odhaduje průměr jako hodnotu prvního čísla, nebo estimátor co přičte C pokud $X_1 > X_2$ a jinak odečte C .

Tahák

- **Centrální limitní věta:** Označme $Y_n = ((X_1 + \dots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$. Pak $Y_n \xrightarrow{d} N(0, 1)$. Neboli

$$\lim_{n \rightarrow \infty} F_{Y_n}(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$

- Na počítání $\Phi(x)$: <https://t.ly/JRQ2>
- Zkoumáme posloupnost n.n.v. se stejným rozdělením, např. $Geom(\theta)$, $U(0, \theta)$, kde θ je parametr.
- Zapisujeme $X_1, \dots, X_n \sim F_\theta$, tzv. náhodný výběr z F_θ (model s parametrem).
- Naměříme $X_1 = x_1, \dots$, chceme odhadnout θ .
- $\hat{\theta}$... nějaká metoda jak odhadnout θ pomocí naměřených dat (hodnot X_1, \dots, X_n). Angl. *estimator* – jeden získaný odhad je *estimate*, ten značíme $\hat{\theta}$.
- $L(\theta; x_1, \dots, x_n) = P[X_1 = x_1 \wedge \dots \wedge X_n = x_n]$... pravd. pozorovaných dat závislá na parametru θ .
- nebo $L(\dots) = f_{X_1, \dots, X_n}(x_1, \dots, x_n)$... hustota pravděpodobnosti ...
- $\ell(\theta; x_1, \dots, x_n) = \log L(\dots)$... pro snazší výpočty.
- *Odhad metodou maximální věrohodnosti (Maximal Likelihood)* hledáme θ , pro které je maximální $L(\theta; x_1, \dots, x_n)$, resp. $\ell(\dots)$. Obvykle pomocí derivací funkce L , resp. ℓ .
- bias (vychýlení): $\mathbb{E}(\hat{\theta} - \theta)$... θ skutečný parametr, $\hat{\theta}$ náš odhad (náhodná veličina, protože závisí na naměřených datech)
- odhad je nevychýlený/nestranný/unbiased: bias = 0
- odhad je asymptoticky nevychýlený: bias konverguje k 0, neboli $\mathbb{E}(\hat{\theta}) \rightarrow \theta$
- odhad je konzistentní: $\hat{\theta}$ konverguje k 0 v pravděpodobnosti: pro všechna $\varepsilon > 0$: $P(|\hat{\theta} - \theta| > \varepsilon) \rightarrow 0$
- MSE (mean square error, střední kvadratická odchylka): $\mathbb{E}((\hat{\theta} - \theta)^2)$
- Věta: $MSE = \text{bias}^2 + \text{var}(\hat{\theta})$.