12. cvičení z PSt — 7. 5. 2026

Definice 1. Nechť $\theta \in \Theta$ je nějaký (nám neznámý) parametr, který udává nějaké rozdělení F_θ. Definujeme

- *náhodný výběr* jako posloupnost nezávislých n. v. $X_1, \dots, X_n \sim F_\theta$ reprezentující naměřená data,
- *estimátor* jako funkci $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n) \in \Theta$ reprezentující odhad hodnoty θ z naměřených dat.

Definice 2. Pro estimátor $\hat{\theta}$ definujeme

- *bias* (vychýlení) jako $\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$,
- *varianci* (rozptyl) jako $\text{var}(\hat{\theta}) = \mathbb{E}\Big((\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\Big)$,
- *střední kvadratickou chybu* (MSE) jako $\text{MSE}(\hat{\theta}) = \mathbb{E}\Big((\hat{\theta} - \theta)^2\Big)$.

Věta 3. *Pro každý estimátor $\hat{\theta}$ platí*

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

Definice 4. Nechť $X_1, \dots, X_n \sim F_\theta$ je náhodný výběr. Definujeme *věrohodnost* (likelihood) estimátoru $\hat{\theta}$ pro $x = (x_1, \dots, x_n)$ následujícím způsobem:

(1) pokud F_θ je *diskrétní rozdělení* s pravděpodobnostní funkcí p_θ, pak

$$L(x; \hat{\theta}) = p_{\hat{\theta}}(x_1) \cdot p_{\hat{\theta}}(x_2) \cdot \ldots \cdot p_{\hat{\theta}}(x_n).$$

(2) pokud F_θ je *spojité rozdělení* s pravděpodobnostní hustotou f_θ, pak

$$L(x; \hat{\theta}) = f_{\hat{\theta}}(x_1) \cdot f_{\hat{\theta}}(x_2) \cdot \ldots \cdot f_{\hat{\theta}}(x_n).$$

V obou případech vyjadřuje to, jak pravděpodobná jsou námi naměřená data $x_1, \dots, x_n$, pokud je estimátor $\hat{\theta}$ roven skutečné hodnotě θ.

Definice 5 (MLE)**.** *Maximum likelihood estimator* pro naměřená data $x = (x_1, \dots x_n)$ je estimátor $\hat{\theta}_{\text{MLE}}$ maximalizující likelihood, tedy

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\hat{\theta} \in \Theta} L(x; \hat{\theta}).$$

MLE, bias, rozptyl

1. Připomeňte si příklad z minulého cvičení: Chtěli jsme zjistit, kolik tanků Němci vyrobili, měli jsme přitom jen čtyři náhodná sériová čísla. Ujasněte si, co zde odpovídá pojmům z definice 1.

2. V přednášce jsme odvodili, že lineární regrese metodou nejmenších čtverců (least squares) je MLE, pokud předpokládáme, že šum ε_i v modelu $Y_i = ax_i + b + \varepsilon_i$ je gaussovský. Předpokládejme teď místo toho, že šumy ε_i mají Laplaceovu hustotu $f(\varepsilon) = \frac{1}{2}e^{-|\varepsilon|}$. Ukažte, že MLE odhad $(\hat{a}, \hat{b})$ minimalizuje

$$\sum_{i=1}^{n} |y_i - ax_i - b|.$$

Jinými slovy: z čtverců reziduí se stanou absolutní hodnoty reziduí. Výslednému vzorečku se říká *L1 regrese*. *(Hint dole)*[1]

[1] Dosaďte do funkce f a zlogaritmujte. Postup je téměř identický tomu co bylo na přednášce pro metodu nejmenších čtverců.

3. Máme náhodný výběr $X_1, \ldots, X_n \sim N(\mu, 1)$, kde μ je neznámé. Klasický MLE estimátor je $\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Představme si ale, že z nějakého důvodu věříme, že skutečné μ je malé, a tak chceme odhad trochu „stáhnout“ k nule. Uvažujme proto estimátor $\hat{\mu}' = \frac{1}{2}\hat{\mu}_{\text{MLE}}$.

(a) Pro oba estimátory spočtěte jejich bias, varianci, a MSE.

(b) Zkuste je porovnat.

4. Máme náhodný výběr $X_1, \ldots, X_n \sim \text{Ber}(\theta)$, kde $\theta \in [0, 1]$ je neznámý parametr. Označme $S = \sum_{i=1}^{n} X_i$. Uvažujte čtyři estimátory parametru θ:

$$\hat{\theta}_1 = \frac{S}{n}, \qquad \hat{\theta}_2 = \frac{S+1}{n+2}, \qquad \hat{\theta}_3 = 0.42, \qquad \hat{\theta}_4 = X_1.$$

(a) Které z těchto estimátorů jsou nestranné jako estimátory parametru θ?

(b) Které z těchto estimátorů mají nulovou varianci pro libovolné θ?

(c) Pokud $\theta = 1/2$ a $n > 1000$, tipněte si, který estimátor má nejmenší MSE.

(d) Vyberte si aspoň jeden estimátor $\hat{\theta}_j$ a spočtěte pro něj $\mathbb{E}(\hat{\theta}_j)$, $\text{bias}(\hat{\theta}_j)$, $\text{var}(\hat{\theta}_j)$ a $\text{MSE}(\hat{\theta}_j)$.

Testování hypotéz

5. Udělali jsme statistický test nulové hypotézy H_0. Test vyšel signifikantně, H_0 jsme zamítli a p-hodnota byla $p = 0.01$.

U každého tvrzení označte, zda z výsledku logicky plyne.

(a) Vyvrátili jsme nulovou hypotézu H_0, tj. určitě nemůže platit. ☐ ano ☐ ne

(b) Zjistili jsme, že pravděpodobnost nulové hypotézy je 1 %. ☐ ano ☐ ne

(c) Dokázali jsme alternativní hypotézu. ☐ ano ☐ ne

(d) Zjistili jsme, že pravděpodobnost alternativní hypotézy je 99 %. ☐ ano ☐ ne

(e) Když jsme zamítli H_0, pravděpodobnost, že jsme se rozhodli špatně, je 1 %. ☐ ano ☐ ne

(f) Výsledek je spolehlivý v tom smyslu, že kdybychom experiment mnohokrát opakovali, dostali bychom signifikantní výsledek v 99 % případů. ☐ ano ☐ ne

Tento dotazník dostali učitelé metodologie budoucích vědeckých pracovníků v medicíně; tipněte si, kolik procent učitelů a kolik procent studentů odpovědělo správně na všechny otázky.

6. Děláme 20 nezávislých testů. Předpokládejme, že všech 20 nulových hypotéz je ve skutečnosti pravdivých. Každý test děláme na hladině $\alpha = 0.05$.

(a) Jaká je pravděpodobnost, že jeden konkrétní test falešně zamítne svoji nulovou hypotézu?

(b) Jaká je pravděpodobnost, že alespoň jeden z 20 testů falešně zamítne?

(c) Proč je to problém, když používáme hranici 0.05 u každého testu zvlášť?

(d) (Bonferroniho korekce) Nyní si představme, že každý z 20 testů děláme na hladině $\alpha' = 0.05/20$. Dokažte, že pravděpodobnost, že aspoň jeden z 20 testů falešně zamítne nulovou hypotézu, je nejvýše 0.05, ať už jsou jednotlivé testy nezávislé nebo ne.