# 10. cvičení z PST

21. - 25. listopadu 2022

**Připomenutí:** Mějme náhodný výběr $(X_1, \dots, X_n)$ závislý na parametru $\vartheta$ (tj. máme vektor z nezávislých *stejně rozdělených* náhodných veličin $X_i$ s distribuční funkcí $F_\vartheta$ závislou na parametru $\vartheta$). Můžeme uvažovat i závislost na více parametrech, ale většinou budeme pracovat jen s jedním.

V praxi máme hodnotu parametru danou (označme si ji $\vartheta_0$), ale bohužel ji neznáme. Snažíme se ji proto určit (jako hodnotu $\widehat{\vartheta}$) z naměřených hodnot $(x_1, \dots, x_n) \in \mathbb{R}^n$ a to co "nejlépe" (tím, že si stanovíme nějaké vhodné podmínky, které chceme splnit). Hodnotě $\widehat{\vartheta}$ pak říkáme *bodový odhad* (té skutečné hodnoty parametru $\vartheta_0$).

Možných metod odhadu je více. Obvykle se používají

- metoda maximální věrohodnosti
  - \+ *výhody:* dává (v podstatě) vždy výsledek; je možné ji použít i pro veličiny, co nemají číselné hodnoty (což znamená, že nezáleží na hodnotách, ale na jejich pravděpodobnostech)
  - − *nevýhody:* není vytvořena pro veličiny se smíšeným rozdělením (tj. jiným než buď diskrétním nebo spojitým)
- metoda momentů
  - \+ *výhody:* dá se použít na jakýkoliv typ veličiny $X$ (která má konečné hodnoty $E(X^k)$ pro prvních několik $k = 1, 2, 3, \dots$)
  - − *nevýhody:* obecně nemáme zaručeno, že dostaneme nějaký výsledek; při pouhém přejmenování hodnot veličiny (tj. zachováme pravděpodobnosti, jen změníme číselné hodnoty, které jsou jim přiřazeny) se výsledek metody značně mění

**10.1** *Počet kazů $X$ na tabulkách skla se řídí Poissonovým rozdělením. Bylo pozorováno*

| $i$ = *počet kazů na dané tabulce* | 0 | 1 | 2 | 3 | 5 |
|---|---|---|---|---|---|
| $n_i$ = *pozorovaná četnost* | 17 | 4 | 1 | 2 | 1 |

*Metodou maximální věrohodnosti a metodou momentů určete parametr $\lambda$ tohoto Poissonova rozdělení.*

**Řešení:**

Celkový počet měření je $n = \sum_i n_i = 17 + 4 + 1 + 2 + 1 = 25$. Naměřené hodnoty $(x_1, \dots, x_n)$ se skládají z hodnot $i \in \{0, 1, 2, 3, 5\}$, kde každá z nich se vyskytuje se svojí četností $n_i$. Protože nebude záležet na pořadí, v jakém jsme hodnoty $x_i$ naměřili, můžeme si pro jednoduchost představit, že je

$$(x_1, \dots, x_n) = \Big(\underbrace{0, \dots, 0}_{17-krát}, \underbrace{1, \dots, 1}_{4-krát},\ 2,\ 3,\ 3,\ 5\Big)\ .$$

Pro náhodnou veličinu $X$ s rozdělením Poiss($\lambda$) je $P_\lambda(X = k) = \dfrac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}$.

**Metoda maximální věrohodnosti:**

Hledáme takové $\hat{\lambda} > 0$, které maximalizuje funkci věrohodnosti $L(\lambda)$, která je definována jako

$$L(\lambda) = P_\lambda(X_1 = x_1, \dots, X_n = x_n) \overset{(nezav.)}{=} \prod_{j=1}^{n} P_\lambda(X_j = x_j) = \prod_{j=1}^{n} \frac{\lambda^{x_j}}{x_j!}\,\mathrm{e}^{-\lambda} =$$

$$= \left(\frac{\lambda^0}{0!}\,\mathrm{e}^{-\lambda}\right)^{17} \left(\frac{\lambda^1}{1!}\,\mathrm{e}^{-\lambda}\right)^{4} \left(\frac{\lambda^2}{2!}\,\mathrm{e}^{-\lambda}\right)^{1} \left(\frac{\lambda^3}{3!}\,\mathrm{e}^{-\lambda}\right)^{2} \left(\frac{\lambda^5}{5!}\,\mathrm{e}^{-\lambda}\right)^{1} =$$

$$= \frac{\lambda^{0\cdot 17+1\cdot 4+2\cdot 1+3\cdot 2+5\cdot 1}}{konst.}\,\mathrm{e}^{-\lambda(17+4+1+2+1)} = \frac{\lambda^{17}}{konst.}\,\mathrm{e}^{-25\lambda}\,,$$

kde $X_j$ jsou jednotlivé nezávislé veličiny (v pokusech) a $x_j$ naměřené hodnoty.

Pro vyšetření maxima je vhodnější přejít k logaritmu této funkce, tj.

$$\ell(\lambda) = \ln L(\lambda) = 17\ln\lambda - 25\lambda - \ln(konst.)$$

Z její derivace

$$\ell'(\lambda) = \frac{17}{\lambda} - 25\,.$$

získáme řešení

$$\frac{17}{\hat\lambda} - 25 = 0 \qquad \Longrightarrow \qquad \widehat\lambda = \frac{17}{25} = 0.68\,.$$

a ze znamének derivace je snadno vidět, že v $\widehat\lambda = \frac{17}{25}$ je skutečně maximum.

**Metoda momentů:**

Chceme, aby platily rovnosti teoretických momentů $E(X^k)$, závislých na parametru $\lambda$, a výběrových momentů $m_k := \frac{1}{n}\sum_{i=1}^n x_i^k$, tedy $E(X^k) = m_k$ pro co nejvíce počátečních hodnot $k = 1, 2, \ldots$.

Počet rovnic volíme tak, abychom dostali co nejmenší (nenulový) počet řešení (ideálně jen jedno) pro parametr $\lambda$ . Existenci řešení ale obecně zaručenou nemáme.

V našem případě budeme tedy požadovat rovnost $E(X) = m_1$ $(= \overline{\boldsymbol{x}})$ . Přitom máme

- střední hodnotu $E(X) = \lambda$
- výběrový průměr $\overline{\boldsymbol{x}} = \frac{\sum\limits_{j=1}^{n} x_j}{n} = \frac{\sum\limits_{i} i\cdot n_i}{\sum\limits_{i} n_i} = \frac{0\cdot 17+1\cdot 4+2\cdot 1+3\cdot 2+5\cdot 1}{17+4+1+2+1} = \frac{17}{25}$

Takže dostáváme opět odhad $\widehat\lambda = \frac{17}{25}$, což není příliš překvapivé, protože parametr $\lambda$ má význam střední hodnoty $X$ a ta se nejlépe odhaduje pomocí výběrového průměru $\overline{\boldsymbol{x}}$.

**Poznámka k věrohodnostní funkci pro spojitá rozdělení:** Pro metodu max. věrohodnosti se u diskrétního rozdělení využívá pravděpodobnosti, že daná hodnota $x_0$ bude *přesně* nabyta, tj. $P(X = x_0)$. Tyto pravděpodobnosti by ale byly v případě spojitého rozdělení vždy nulové. Musíme tedy použít nějakou jinou charakteristiku v daném bodě a zde se nabízí hustota $f_X$. Jak ale víme, hustota není určena svými hodnotami, ale jen svými integrály. My ovšem nebudeme ani tak chtít zkoumat hustotu v bodě $x_0$, nýbrž spíše chování výrazu $P\Big(X \in (x_0-\varepsilon, x_0+\varepsilon)\Big)$ pro $\varepsilon \to 0+$. Dá se ukázat, že pokud je hustota $f_X$ spojitá v $x_0$, pak platí

$$\lim_{\varepsilon\to 0+} \frac{1}{2\varepsilon}\cdot P\Big(X \in (x_0-\varepsilon, x_0+\varepsilon)\Big) = f_X(x_0)\ .$$

Tedy v tomto případě je chování daného výrazu skutečně přibližně úměrné hodnotě $f_X(x_0)$.

Proto se ve věrohodnostní funkci nakonec opravdu hustota používá, ale za předpokladu, že je $f_X$ je *spojitá* buď všude nebo v oboru hodnot, který je otevřenou množinou (důvodem je to, že limitu děláme z obou stran). Např. pro exponenciální rozdělení (které modeluje dobu čekání) je obor hodnot $(0, +\infty)$ a tam už hustotu spojitou máme (přestože na celém $\mathbb{R}$ spojitá není).

**10.2** *Doba do poruchy přístroje má exponenciální rozdělení. Bylo zjištěno, že se přístroj porouchal postupně za* 4 *dny,* 7 *dní,* 12 *dní,* 2.5 *dne a* 24.5 *dne. Metodou maximální věrohodnosti (příp. metodou momentů) určete parametr* $\lambda$ *tohoto exponenciálního rozdělení.*

**Řešení:**
Máme tedy veličinu

$$X = \text{“doba do poruchy přístroje”} \text{ [ve dnech]}$$

s exponenciálním rozdělením $Exp(\lambda)$ a hustotou $f_\lambda(x) = \begin{cases} \lambda\,\mathrm{e}^{-\lambda x} & \text{pro } x > 0 \\ 0 & \text{pro } x \le 0. \end{cases}$

Počet měření je $n = 5$ a jejich hodnoty jsou $x_1 = 4$ dny, $\ldots, x_5 = 24.5$ dne.

**Metoda maximální věrohodnosti:**
Obor hodnot $X$ je $(0, +\infty)$, což je otevřený interval a hustota je zde spojitá. (Hodnotu 0 neuvažujeme, protože jako čekací dobu má smysl brát jen kladné hodnoty.)

Hledáme takové $\hat\lambda > 0$, které maximalizuje věrohodnostní funkci

$$L(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} \lambda\,\mathrm{e}^{-\lambda x_i} = \lambda\,\mathrm{e}^{-\lambda\cdot 4} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 7} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 12} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 2.5} \cdot \lambda\,\mathrm{e}^{-\lambda\cdot 24.5} =$$

$$= \lambda^5\,\mathrm{e}^{-\lambda\cdot(4+7+12+2.5+24.5)} = \lambda^5\,\mathrm{e}^{-\lambda\cdot 50}\ .$$

Logaritmicko-věrohodnostní funkce je

$$\ell(\lambda) = \ln L(\lambda) = 5\ln\lambda - 50\lambda.$$

Z její derivace

$$\ell'(\lambda) = \frac{5}{\lambda} - 50.$$

získáme řešení

$$\frac{5}{\hat\lambda} - 50 = 0 \qquad \Longrightarrow \qquad \widehat\lambda = \frac{1}{10}\ [\mathrm{den}^{-1}] \qquad \Longrightarrow \qquad \widehat\tau = \frac{1}{\hat\lambda} = 10\ [\text{dnů}]$$

(ve kterém skutečně nastává maximum, jak je vidět ze znamének derivace.)

**Metoda momentů:**
Chceme, aby platily rovnosti $E(X^k) = m_k$ teoretických a výběrových momentů pro co nejvíce počátečních hodnot $k = 1, 2, \ldots$.

Máme

- střední hodnotu $E(X) = \frac{1}{\lambda}$
- výběrový průměr $\overline{\boldsymbol{x}} = m_1 = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{4+7+12+2.5+24.5}{5} = \frac{50}{5} = 10$

Z požadované rovnosti $\frac{1}{\hat\lambda} = E(X) = \overline{\boldsymbol{x}} = 10$ dostáváme opět odhad $\widehat\lambda = \frac{1}{10}$. Tato shoda je opět způsobena tím, že parametr $\tau = \frac{1}{\lambda}$ má význam střední hodnoty $X$ a ta se nejlépe odhaduje pomocí výběrového průměru $\overline{\boldsymbol{x}}$.

**Poznámka:** Pro následující rozdělení veličiny $X$ dávají obě výše probírané metody stejné výsledky pro daný parametr:

- $p$ pro alternativní $Alt(p)$, odhad je $\hat p = E(X) = \overline{x}$

- $p$ pro binomické $Bi(n,p)$, odhad je $n\hat{p} = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{\overline{x}}{n}$
- $p$ pro geometrické $Geom(p)$, odhad je $\frac{1-\hat{p}}{\hat{p}} = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{1}{\overline{x}+1}$
- $\lambda$ pro Poissonovo $Poiss(\lambda)$, odhad je $\hat{\lambda} = E(X) = \overline{x}$
- $\tau$ pro exponenciální $Exp(\frac{1}{\tau})$, odhad je $\hat{\tau} = E(X) = \overline{x}$
- $\mu$ pro normální $N(\mu, \sigma^2)$, odhad je $\hat{\mu} = E(X) = \overline{x}$

Metoda maximální věrohodnosti je v těchto případech výpočetně složitější než metoda momentů, která je zde velmi snadná. V písemkách je ovšem smyslem zadání prověřit znalost použití zvolené metody (obvykle právě metody maximální věrohodnosti), takže znalost výsledku získaného jiným způsobem je za těchto okolností pouze kontrola, že nám to vyšlo správně.

**10.3** *V krabici je mnoho hracích kostek, z nichž některé jsou v pořádku, některé falešné. Na falešných padá šestka s pravděpodobností* $1/2$*, zbývající čísla mají stejnou pravděpodobnost. Opakovaně jsme vytáhli kostku, hodili jí a vrátili ji zpět. Četnost výsledků byla:*

| *hodnota* $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *četnost* $n_i$ | 18 | 20 | 12 | 15 | 10 | 25 |

*(1) Odhadněte metodou momentů a metodou maximální věrohodnosti, kolik procent kostek je falešných.*

*(2) Okomentujte stručně, co bychom dostali metodou momentů, kdybychom měli napozorované hodnoty:*

| *hodnota* $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| *četnost* $n_i$ | 18 | 20 | 12 | 15 | 15 | 20 |

**Řešení:**
Podíl falešných kostek označme $c \in \langle 0, 1\rangle$. Naše náhodná veličina je

$X =$"*hodnota, která padne na dané kostce*"

a můžeme ji vyjádřit jako směs $X = \mathrm{Mix}_c(X_1, X_2)$ složenou z náhodných veličin

$X_1 =$"*hodnota, která padne na falešné kostce*"

$X_1$ : "množina falešných kostek" $\to \mathbb{R}$

s rozdělením pravděpodobnosti $p_{X_1}(i) = \begin{cases} \frac{1}{2}, & i = 6 \\ \frac{1}{10}, & i = 1, \ldots, 5. \end{cases}$

$X_2 =$"*hodnota, která padne na správné kostce*"

$X_2$ : "množina správných kostek" $\to \mathbb{R}$ .

s rozdělením pravděpodobnosti $p_{X_2}(i) = \frac{1}{6}$ pro $i = 1, \ldots, 6$.

(1) **Metoda momentů:** Máme

$$E(X_1) = \tfrac{1}{10}(1 + \cdots + 5) + \tfrac{1}{2} \cdot 6 = \tfrac{9}{2}$$

a

$$E(X_2) = \tfrac{1}{6}(1 + \cdots + 6) = \tfrac{7}{2}$$

a z definice směsi tak dostaneme

$$E(X) = c \cdot E(X_1) + (1-c) \cdot E(X_2) = c \cdot \tfrac{9}{2} + (1-c) \cdot \tfrac{7}{2} = c + \tfrac{7}{2} \ .$$

Realizace výběrového průměru je

$$\overline{\boldsymbol{x}} = \frac{\sum_i n_i \cdot i}{\sum_i n_i} = \frac{18 \cdot 1 + 20 \cdot 2 + 12 \cdot 3 + 15 \cdot 4 + 10 \cdot 5 + 25 \cdot 6}{18 + 20 + 12 + 15 + 10 + 25} = \frac{354}{100} = 3.54 \ .$$

Srovnáním dostaneme

$$\widehat{p} + 3.5 = E(X) = \overline{\boldsymbol{x}} = 3.54 \ ,$$

což dává $\widehat{p} = 0.04 \in \langle 0, 1 \rangle$, a to vyhovuje zadání.

**Metoda maximální věrohodnosti:**
Z definice směsi máme pro její pravděpodobnostní funkci, že

$$p_X(i) = c \cdot p_{X_1}(i) + (1-c) \cdot p_{X_2}(i) = \begin{cases} c\,\frac{1}{10} + (1-c)\,\frac{1}{6} = \frac{5-2\,c}{30} & , \ i = 1, \ldots, 5 \ , \\ \\ c\,\frac{1}{2} + (1-c)\,\frac{1}{6} = \frac{1+2\,c}{6} & , \ i = 6 \ . \end{cases}$$

Víme, že šestka padla 25×, ostatní čísla padla dohromady 75× (není třeba mezi nimi rozlišovat, protože mají stejnou pravděpodobnost). Tedy věrohodnostní funkce je

$$L(c) = \left(\frac{5-2\,c}{30}\right)^{75} \cdot \left(\frac{1+2\,c}{6}\right)^{25} \ ,$$
$$\ell(c) = \ln(L(c)) = 75\,\ln(5-2\,c) + 25\,\ln(1+2\,c) + \textit{konst.}$$

Maximum nastává pro $\widehat{c}$ takové, že

$$0 = \ell'(\widehat{c}) = \frac{-150}{5-2\,\widehat{c}} + \frac{50}{1+2\,\widehat{c}} = 50 \cdot \frac{2-8\widehat{c}}{(5-2\,\widehat{c})(1+2\,\widehat{c})} \ ,$$
$$\widehat{c} = \tfrac{1}{4} \in \langle 0, 1 \rangle \ .$$

protože na intervalu $\langle 0, \frac{1}{4}\rangle$ je $\ell' > 0$ a na $(\frac{1}{4}, 1\rangle$ je $\ell' < 0$. Tato hodnota je i v souladu s počátečními omezujícími podmínkami.

Vidíme také, že obě metody dávají odlišné výsledky. Kdybychom si čísla na kostkách jen přejmenovali na jiné hodnoty (např. $\sqrt{2}$, $-89$, atd.) dostali bychom metodou momentů opět jiný výsledek (pokud by vůbec nějaké řešení existovalo), zatímco metodou maximální věrohonosti bychom obdrželi opět stejné řešení, které by respektovalo fyzický stav kostek a ne to, co je na nich napsáno.

(2) Pro hodnoty

| hodnota $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| četnost $n_i$ | 18 | 20 | 12 | 15 | 15 | 20 |

bychom v metodě momentů dostali

$$\overline{\boldsymbol{x}} = \frac{18 \cdot 1 + 20 \cdot 2 + 12 \cdot 3 + 15 \cdot 4 + 15 \cdot 5 + 20 \cdot 6}{18 + 20 + 12 + 15 + 15 + 20} = \frac{349}{100} = 3.49 \ .$$

Protože ale $E(X) = c \cdot 4.5 + (1-c) \cdot 3.5 \in \langle 3.5,\ 4.5 \rangle$, tak rovnice $c + 3.5 = E(X) = \overline{\boldsymbol{x}} = 3.49$ nemá řešení pro $c \in \langle 0, 1 \rangle$.

V tomto případě metoda momentů prostě nedává žádnou odpověď. Důvodem je obecně to, že zatímco pro střední hodnotu máme omezení $E(X) \in \langle 3.5,\ 4.5 \rangle$, tak pro výběrový průměr je zde pouze omezení $\overline{\boldsymbol{x}} \in \langle 1,\ 6 \rangle$.