

10. cvičení z PST

18. - 22. dubna 2022

Příklad 10.1 Tramvaj má intervaly mezi příjezdy 10 minut. Jaká je pravděpodobnost, že během 24 pracovních dnů stráví člověk při cestách do práce a zpět čekáním na tramvaj nejvýše 3 hodiny? V jakém symetrickém intervalu se při těchto cestách bude nacházet celková čekací doba s pravděpodobností 90%?

Řešení:

Pro veličinu

$$Z = \text{“celková doba čekání během 24 dnů při cestách tam a zpět” [v hodinách]}$$

nás zajímá $P(Z \leq 3)$.

K řešení opět použijeme centrální limitní větu. Označme si tedy pro $i = 1, 2, \dots, n$, kde $n = 24 \cdot 2 = 48$, veličiny

$$X_i = \text{“doba strávená čekáním při } i\text{-tém příchodu na zastávku” [v hodinách]}$$

které pokládáme za nezávislé. Tramvaj jezdí přesně po 10 minutách, zatímco naše příchody na zastávku budeme pokládat za náhodné s rovnoměrným rozdělením v rámci 10 minutového intervalu. Proto i doba čekání X_i bude mít rovnoměrné rozdělení (v jednotkách hodin) tvaru $\text{Ro}(a, b) = \text{Ro}\left(0, \frac{1}{6}\right)$.

Protože opět platí $Z = \sum_{i=1}^n X_i$, dostaneme

$$E(X_i) = \frac{a+b}{2} = \frac{0 + \frac{1}{6}}{2} = \frac{1}{12} \quad \Rightarrow \quad E(Z) = n \cdot E(X_1) = 48 \cdot \frac{1}{12} = 4$$

$$\text{var}(X_i) = \frac{(b-a)^2}{12} = \frac{\left(\frac{1}{6} - 0\right)^2}{12} = \frac{1}{12 \cdot 36} \quad \Rightarrow \quad \text{var}(Z) = n \cdot \text{var}(X_1) = 48 \cdot \frac{1}{12 \cdot 36} = \frac{1}{9}$$

$$\Rightarrow \sqrt{\text{var}(Z)} = \sqrt{\frac{1}{9}} = \frac{1}{3}.$$

(a) Podle CLV bude mít veličina $\text{norm}(Z) = \frac{Z - E(Z)}{\sqrt{\text{var}(Z)}} = 3(Z - 4)$ přibližně rozdělení $N(0, 1)$. Můžeme proto psát

$$\begin{aligned} P(Z \leq 3) &= P\left(\underbrace{3 \cdot (Z - 4)}_{\text{norm}(Z)} \leq 3 \cdot (3 - 4)\right) = P\left(\text{norm}(Z) \leq -3\right) \stackrel{\text{(CLV)}}{=} \\ &\stackrel{\text{(CLV)}}{=} \Phi(-3) = 1 - \Phi(3) \doteq 1 - 0.9987 = \mathbf{0.0013}. \end{aligned}$$

(b) Hledáme $\varepsilon > 0$ tak, aby $0.9 = P(X \in \langle E(X) - \varepsilon, E(X) + \varepsilon \rangle)$ tedy

$$\begin{aligned} 0.9 &= P(X \in \langle E(X) - \varepsilon, E(X) + \varepsilon \rangle) = P(-\varepsilon \leq X - E(X) \leq \varepsilon) = \\ &= P\left(-\frac{\varepsilon}{\sqrt{\text{var}(X)}} \leq \frac{X - E(X)}{\sqrt{\text{var}(X)}} \leq \frac{\varepsilon}{\sqrt{\text{var}(X)}}\right) = P(-3\varepsilon \leq \text{norm}(X) \leq 3\varepsilon) \stackrel{\text{(CLV)}}{=} \Phi(3\varepsilon) - \Phi(-3\varepsilon) = \\ &= 2\Phi(3\varepsilon) - 1 \end{aligned}$$

Řešíme proto přibližnou rovnost

$$0.9 \doteq 2\Phi(3\varepsilon) - 1$$

$$\Phi(3\varepsilon) \doteq \frac{1.9}{2} = 0.95$$

$$\varepsilon \doteq \frac{1}{3} \cdot \Phi^{-1}(0.95) = \frac{1.645}{3} = 0.548 \text{ h}$$

Celková čekací doba tedy bude s pravděpodobností 90% v intervalu $\langle 4.548 \text{ h}, 3.452 \text{ h} \rangle$.

Odhad chyby v CLV pro rovnoměrné rozdělení: Pokud mají veličiny X_i rovnoměrné rozdělení na intervalu $\langle a, b \rangle$, pak

$$\mu = E(X_i) = \frac{a+b}{2}, \quad \sigma = \sqrt{D(X_i)} = \sqrt{\frac{(b-a)^2}{12}} = \frac{b-a}{2\sqrt{3}}, \quad \varrho = E(|X_i - \mu|^3) = \frac{(b-a)^3}{32}$$

čímž pro $Z_n = \sum_{i=1}^n X_i$ dostáváme odhad

$$\left| F_{\text{norm}(Z_n)}(t) - \Phi(t) \right| < 0.4748 \cdot \frac{3\sqrt{3}}{4\sqrt{n}} < \frac{0.62}{\sqrt{n}}.$$

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

V praxi máme hodnotu parametru danou (označme si ji ϑ_0), ale bohužel ji neznáme. Snažíme se ji proto určit (jako hodnotu $\hat{\vartheta}$) 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 $\hat{\vartheta}$ pak říkáme *bodový odhad* (té skutečné hodnoty parametru ϑ_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

Příklad 10.2 Počet kazů X na tabulkách skla se řídí Poissonovým rozdělením. Bylo pozorováno

$i = \text{počet kazů na dané tabulce}$	0	1	2	3	5
$n_i = \text{pozorovaná četnost}$	17	4	1	2	1

Metodou maximální věrohodnosti (příp. metodou momentů) určete parametr λ 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

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

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

Metoda maximální věrohodnosti:

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

$$\begin{aligned} L(\lambda) &= P_\lambda(X_1 = x_1, \dots, X_n = x_n) \stackrel{(\text{nezav.})}{=} \prod_{j=1}^n P_\lambda(X_j = x_j) = \prod_{j=1}^n \frac{\lambda^{x_j}}{x_j!} e^{-\lambda} = \\ &= \left(\frac{\lambda^0}{0!} e^{-\lambda} \right)^{17} \left(\frac{\lambda^1}{1!} e^{-\lambda} \right)^4 \left(\frac{\lambda^2}{2!} e^{-\lambda} \right)^1 \left(\frac{\lambda^3}{3!} e^{-\lambda} \right)^2 \left(\frac{\lambda^5}{5!} e^{-\lambda} \right)^1 = \\ &= \frac{\lambda^{0 \cdot 17 + 1 \cdot 4 + 2 \cdot 1 + 3 \cdot 2 + 5 \cdot 1}}{\text{konst.}} e^{-\lambda(17 + 4 + 1 + 2 + 1)} = \frac{\lambda^{17}}{\text{konst.}} e^{-25\lambda}, \end{aligned}$$

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.

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

Z její derivace

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

získáme řešení

$$\frac{17}{\hat{\lambda}} - 25 = 0 \quad \implies \quad \hat{\lambda} = \frac{17}{25} = 0.68.$$

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

Metoda momentů:

Chceme, aby platily rovnosti teoretických momentů $E(X^k)$, závislých na parametru λ , 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 $k = 1, 2, \dots$.

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

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

- střední hodnotu $E(X) = \lambda$

- výběrový průměr $\bar{x} = \frac{\sum_{j=1}^n x_j}{n} = \frac{\sum_i i \cdot n_i}{\sum_i n_i} = \frac{0 \cdot 17 + 1 \cdot 4 + 2 \cdot 1 + 3 \cdot 2 + 5 \cdot 1}{17 + 4 + 1 + 2 + 1} = \frac{17}{25}$

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

Příklad 10.3 Počet neúspěšných zásahů terče předtím, než se střelec trefí, má geometrické rozdělení $\text{Geom}(p)$. Zaznamenali jsme, že terč byl zasažen

20 krát napoprvé
 10 krát až napodruhé
 7 krát až napotřetí
 3 krát až napočtvrté.

Metodou maximální věrohodnosti (příp. metodou momentů) odhadněte parametr p , představující pravděpodobnost úspěšného zásahu.

Řešení:

Veličina

$X = \text{“počet neúspěšných zásahů, než se trefíme”}$

má geometrické rozdělení $\text{Geom}(p)$ pro $p \in (0, 1)$ a

$$P_p(X = i) = (1 - p)^i p, \quad i \in \mathbb{N}_0.$$

Z tabulky

hodnota i veličiny X	0	1	2	3
pozorovaná četnost n_i	20	10	7	3

vidíme, že počet měření je $n = \sum_i n_i = 20 + 10 + 7 + 3 = 40$. Naměřené hodnoty (x_1, \dots, x_n) se skládají z hodnot $i \in \{0, 1, 2, 3\}$, kde každá se vyskytuje se svojí četností n_i .

Metoda maximální věrohodnosti:

Hledáme hodnotu $p \in (0, 1)$, která maximalizuje funkci věrohodnosti

$$\begin{aligned} L(p) &= P_p(X_1 = x_1, \dots, X_n = x_n) = \prod_{j=1}^n P_p(X_j = x_j) = \\ &= \left((1-p)^0 p \right)^{20} \left((1-p)^1 p \right)^{10} \left((1-p)^2 p \right)^7 \left((1-p)^3 p \right)^3 = \\ &= (1-p)^{0 \cdot 20 + 1 \cdot 10 + 2 \cdot 7 + 3 \cdot 3} \cdot p^{20+10+7+3} = (1-p)^{33} \cdot p^{40} \end{aligned}$$

kde X_j jsou jednotlivé nezávislé veličiny (odpovídající jednotlivým pokusům) a x_j naměřené hodnoty. Ekvivalentně budeme hledat maximum funkce

$$\ell(p) = \ln(L(p)) = 33 \cdot \ln(1-p) + 40 \cdot \ln p.$$

Z její derivace

$$\ell'(p) = -\frac{33}{1-p} + \frac{40}{p} = \frac{-73p + 40}{(1-p)p}$$

dostáváme řešení

$$\ell'(\hat{p}) = 0 \quad \implies \quad \hat{p} = \frac{40}{73} \doteq 0.54795$$

které vyhovuje zadání, tj. $\hat{p} \in (0, 1)$. Ze znamének derivace je snadno vidět, že v $\hat{p} = \frac{40}{73}$ je skutečně maximum.

Metoda momentů:

Porovnáváme teoretické k -té momenty $E(X^k)$ s jejich odhady $m_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ pro prvních několik $k = 1, 2, \dots$

Střední hodnota geometrického rozdělení $X \sim \text{Geom}(p)$ je

$$E(X) = \frac{1-p}{p}$$

a její odhad z realizace je

$$m_1 = \bar{x} = \frac{\sum_{j=1}^n x_j}{n} = \frac{\sum_i i \cdot n_i}{\sum_i n_i} = \frac{0 \cdot 20 + 1 \cdot 10 + 2 \cdot 7 + 3 \cdot 3}{20 + 10 + 7 + 3} = \frac{33}{40}.$$

Porovnáním dostaneme

$$\frac{1-\hat{p}}{\hat{p}} = E(X) = \bar{x} = \frac{33}{40}$$

což dává opět řešení

$$\hat{p} = \frac{40}{73} \doteq 0.54795$$

jako v předchozí metodě.

Jak je vidět, v případě geometrického rozdělení dostáváme pro jeho parametr p stejné výsledky pro obě metody.

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 $P(X \in (x_0 - \varepsilon, x_0 + \varepsilon))$ pro $\varepsilon \rightarrow 0+$. Dá se ukázat, že pokud je hustota f_X spojitá v x_0 , pak platí

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

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).

Příklad 10.4 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 λ tohoto exponenciálního rozdělení.

Řešení:

Máme tedy veličinu

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

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

Počet měření je $n = 5$ a jejich hodnoty jsou $x_1 = 4$ dny, \dots , $x_5 = 24.5$ 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 $\lambda > 0$, které maximalizuje věrohodnostní funkci

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^n f_\lambda(x_i) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda e^{-\lambda \cdot 4} \cdot \lambda e^{-\lambda \cdot 7} \cdot \lambda e^{-\lambda \cdot 12} \cdot \lambda e^{-\lambda \cdot 2.5} \cdot \lambda e^{-\lambda \cdot 24.5} = \\ &= \lambda^5 e^{-\lambda \cdot (4+7+12+2.5+24.5)} = \lambda^5 e^{-\lambda \cdot 50}. \end{aligned}$$

Logaritmicke-věrohodnostní funkce je

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

Z její derivace

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

získáme řešení

$$\frac{5}{\hat{\lambda}} - 50 = 0 \quad \implies \quad \hat{\lambda} = \frac{1}{10} \text{ [den}^{-1}\text{]} \quad \implies \quad \hat{\tau} = \frac{1}{\hat{\lambda}} = 10 \text{ [dnů]}$$

(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 $k = 1, 2, \dots$

Máme

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

Z požadované rovnosti $\frac{1}{\hat{\lambda}} = E(X) = \bar{x} = 10$ dostáváme opět odhad $\hat{\lambda} = \frac{1}{10}$. 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 \bar{x} .

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

- p pro alternativní $Alt(p)$, odhad je $\hat{p} = E(X) = \bar{x}$
- p pro binomické $Bi(n, p)$, odhad je $n\hat{p} = E(X) = \bar{x} \Rightarrow \hat{p} = \frac{\bar{x}}{n}$
- p pro geometrické $Geom(p)$, odhad je $\frac{1-\hat{p}}{\hat{p}} = E(X) = \bar{x} \Rightarrow \hat{p} = \frac{1}{\bar{x}+1}$
- λ pro Poissonovo $Poiss(\lambda)$, odhad je $\hat{\lambda} = E(X) = \bar{x}$
- τ pro exponenciální $Exp(\tau)$, odhad je $\hat{\tau} = E(X) = \bar{x}$

- μ pro normální $N(\mu, \sigma^2)$, odhad je $\hat{\mu} = E(X) = \bar{x}$

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 pouze kontrola, že nám to vyšlo správně.