11. cvičení z PST

25. - 29. dubna 2022

Odhad chyby v CLV pro Poissonovo rozdělení: Pro veličinu $Z \sim \text{Poiss}(\lambda)$ platí

$$\left|F_{\text{norm}(Z)}(t) - \Phi(t)\right| \quad \leq \quad \frac{0.4748}{\sqrt{\lambda}} \text{ pro všechna } t \in \mathbb{R}.$$

V praxi se obvykle CLV používá už pokud $\lambda \geq 10$ jako dobrá aproximace (v tomto případě je odhad chyby $\leq \frac{0.4748}{\sqrt{10}} = 0.1502$, ale ve skutečnosti je tento odhad příliš nadsazený a skutečná chyba je menší.)

Důkaz: Veličinu Z můžeme rozepsat jako $Z = \sum_{i=1}^{n} X_i$, kde $X_i \sim \text{Poiss}\left(\frac{\lambda}{n}\right)$ jsou nezávislé veličiny. Položme $\omega = \frac{\lambda}{n}$. Pro použití odhadu Berry-Esseen máme $\sigma = \sqrt{\text{var}(X_i)} = \sqrt{\omega} = \sqrt{\frac{\lambda}{n}}$ a potřebujeme ještě odhadnout tuto hodnotu:

$$\varrho := E\left(|X_i - E(X_i)|^3\right) = E\left(|X_i - \omega|^3\right) = \sum_{k=0}^{\infty} |k-\omega|^3 \frac{\omega^k}{k!} e^{-\omega} =$$

$$= \sum_{k=0}^{[\omega]} (\omega - k)^3 \frac{\omega^k}{k!} e^{-\omega} + \sum_{k=[\omega+1]}^{\infty} (k-\omega)^3 \frac{\omega^k}{k!} e^{-\omega} = 2\sum_{k=0}^{[\omega]} (\omega - k)^3 \frac{\omega^k}{k!} e^{-\omega} + \underbrace{\sum_{k=0}^{\infty} (k-\omega)^3 \frac{\omega^k}{k!} e^{-\omega}}_{=E\left(\left(X_i - E(X_i)\right)^3\right) = \omega} =$$

$$= \omega + 2\sum_{k=0}^{[\omega]} \underbrace{(\omega - k)^3}_{\leq \omega^3} \frac{\omega^k}{k!} e^{-\omega} \leq \omega + 2\omega^3 \underbrace{\sum_{k=0}^{[\omega]} \frac{\omega^k}{k!} e^{-\omega}}_{\leq 1} \leq \omega + 2\omega^3 = \frac{\lambda}{n} + 2\left(\frac{\lambda}{n}\right)^3$$

V Berry-Esseen odhadu tedy pro všechna $n \in \mathbb{N}$ a všechna $t \in \mathbb{R}$ máme

$$\left|F_{\text{norm}(Z)}(t) - \Phi(t)\right| \quad < \quad 0.4748 \cdot \frac{\varrho}{\sigma^3 \sqrt{n}} = 0.4748 \cdot \frac{\frac{\lambda}{n} + 2\left(\frac{\lambda}{n}\right)^3}{\left(\sqrt{\frac{\lambda}{n}}\right)^3 \sqrt{n}} = 0.4748 \cdot \left(\frac{1}{\sqrt{\lambda}} + 2\left(\frac{\lambda^{3/2}}{n^2}\right)\right)$$

a protože λ zůstává pevné, zatímco s n můžeme jít libovolně vysoko, dostaneme v limitě odhad $\left|F_{\text{norm}(Z)}(t) - \Phi(t)\right| \quad \leq \quad \frac{0.4748}{\sqrt{\lambda}}$.

11.1 *V lese se narodí průměrně 4 zajíci denně. Předpokládejme, že počet narozených zajíců se řídí Poissonovým rozdělením. Jaká je pravděpodobnost, že v následujících 7 týdnech se v lese narodí alespoň 175 zajíců?*

Řešení:
Pro veličinu

$$Z = \text{“počet narozených zajíců za 49 dnů”}$$

nás zajímá $P(Z \geq 175)$. U této veličiny sice snadno zjistíme její rozdělení (bude to $Z \sim \text{Poiss}(4 \cdot 49)$), ale k přesnějšímu vyčíslení by bylo při tomto přístupu potřeba sečíst kolem 175 velmi malých čísel, což by bylo jednak náročné a také by vznikalo hodně chyb.

K řešení proto použijeme centrální limitní větu a tudíž budeme chtít veličinu Z “rozsekat” na více stejně rozdělených nezávislých veličin. Označme si tedy pro $i = 1, 2, ..., n$, kde $n = 7 \cdot 7 = 49$, veličiny

$$X_i = \text{“počet narozených zajíců v } i\text{-tý den”}\ .$$

Veličiny pokládáme za nezávislé s rozdělením $X_i \sim \text{Poiss}(4)$, tedy $E(X_i) = 4 = \text{var}(X_i)$. Protože platí $Z = \sum_{i=1}^{n} X_i$, dostaneme

$$E(Z) = n \cdot E(X_1) = 49 \cdot 4 = 196$$
$$\text{var}(Z) = n \cdot \text{var}(X_1) = 49 \cdot 4 = 196 \quad \Rightarrow \quad \sqrt{\text{var}(Z)} = \sqrt{196} = 14$$

což v případě rozptylu platí díky nezávislosti veličin.

Podle CLV (a kritéria použitelnosti CLV pro Poissonovo rozdělení, tj. $196 = E(Z) \geq 10$) bude mít veličina $norm(Z) = \frac{Z-E(Z)}{\sqrt{var(Z)}} = \frac{Z-196}{14}$ přibližně rozdělení $N(0,1)$. Můžeme proto psát

$$P(Z \geq 175) = P\left(\frac{Z-196}{14} \geq \frac{175-196}{14}\right) = P\Big(norm(Z) \geq -1.5\Big) =$$
$$= 1 - P\Big(norm(Z) < -1.5\Big) \overset{(CLV)}{\doteq} 1 - \Phi(-1.5) = 1 - \Big(1 - \Phi(1.5)\Big) =$$
$$= \Phi(1.5) \doteq \mathbf{0.9332}\ .$$

(Pro srovnání: skutečná hodnota pro Poissonovo rozdělení je **0.9398**.)

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

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

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

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

11.2 *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?*

Řešení:
Pro veličinu

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

nás zajímá $P(Z \leq 3)$.

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

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

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

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

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

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

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

$$P(Z \le 3) = P\Big(\underbrace{3 \cdot (Z-4)}_{norm(Z)} \le 3 \cdot (3-4)\Big) = P\Big(norm(Z) \le -3\Big) \overset{(CLV)}{\doteq}$$

$$\overset{(CLV)}{\doteq} \Phi(-3) = 1 - \Phi(3) \doteq 1 - 0.9987 = \mathbf{0.0013} \ .$$

Odhad chyby je maximálně $\left|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\right| < \frac{0.62}{\sqrt{n}} = \frac{0.62}{\sqrt{48}} \doteq 0.0895$. Ale pro $t = -3$, kde nás hodnota pravděpodobnosti zajímá, je tento odhad zbytečně hrubý (protože pravděpodobnost už bude blízká k 0).

Připomenutí: Mějme náhodný výběr $(X_1, \ldots, X_n)$ 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 $\widehat{\vartheta}$) z naměřených hodnot $(x_1, \ldots, x_n) \in \mathbb{R}^n$ 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 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 $k = 1, 2, 3, \ldots$)
  - – *nevýhody:* obecně nemáme zaručeno, že dostaneme nějaký výsledek

11.3 *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 (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 $(x_1, \ldots, x_n)$ 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

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

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

Metoda maximální věrohodnosti:

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

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

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

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

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 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 $k = 1, 2, \ldots$.

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 $E(X) = m_1$ $(= \overline{\boldsymbol{x}})$. Přitom máme

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

Takže dostáváme opět odhad $\widehat\lambda = \frac{17}{25}$, což není příliš překvapivé, protože parametr λ 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).

11.4 *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

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

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 $\lambda > 0$, 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

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

(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 $\frac{1}{\widehat{\lambda}} = E(X) = \overline{\boldsymbol{x}} = 10$ 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 $\frac{1-\hat{p}}{\hat{p}} = E(X) = \overline{x} \;\Rightarrow\; \hat{p} = \frac{1}{\overline{x}+1}$
- λ pro Poissonovo $Poiss(\lambda)$, odhad je $\hat{\lambda} = E(X) = \overline{x}$
- τ pro exponenciální $Exp(\tau)$, odhad je $\hat{\tau} = E(X) = \overline{x}$
- μ 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 pouze kontrola, že nám to vyšlo správně.