10. cvičení z STP

25. duben 2019

Připomeňme si, co říká **Centrální limitní věta (CLV):**

Nechť X_i, pro $i = 1, 2, \ldots$ je posloupnost nezávislých náhodných veličin, které mají stejné rozdělení se střední hodnotou μ a (konečným) rozptylem σ^2. Pak pro veličiny

$$Z_n = \sum_{i=1}^{n} X_i$$

platí, že

$$\lim_{n \to \infty} P\Big(\mathrm{norm}(Z_n) \leq t\Big) = \Phi(t) \quad \text{pro každé} \quad t \in \mathbb{R} \; .$$

Neboli: pro velká n má veličina $\mathrm{norm}(Z_n)$ přibližně normální rozdělení $N(0, 1)$.

Centrální limitní větu můžeme formulovat (namísto pro Z_n) také pro tzv. výběrový průměr, tj. veličiny

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \cdot Z_n \; .$$

protože pro ně platí $\mathrm{norm}(\overline{X}_n) = \mathrm{norm}(Z_n)$.

Příklad 10.1 *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

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

nás zajímá $P(Z \geq 175)$. U této veličiny sice snadno zjistíme její rozdělení (bude to $Z \sim \mathrm{Poiss}(4 \cdot 49)$), 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 co nejvíce stejně rozdělených nezávislých veličin. Označme si tedy pro $i = 1, 2, ..., n$, kde $n = 7 \cdot 7 = 49$, veličiny

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

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

$$\begin{aligned} E(Z) &= n \cdot E(X_1) = 49 \cdot 4 = 196 \\ \mathrm{var}(Z) &= n \cdot \mathrm{var}(X_1) = 49 \cdot 4 = 196 \quad \Rightarrow \quad \sqrt{\mathrm{var}(Z)} = \sqrt{196} = 14 \end{aligned}$$

což v případě rozptylu platí díky nezávislosti veličin.

Podle CLV bude mít veličina $norm(Z) = \frac{Z-E(Z)}{\sqrt{var(Z)}} = \frac{Z-196}{14}$ přibližně rozdělení $N(0,1)$. Můžeme proto psát

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

Příklad 10.2 *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

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

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

$$X_i = \text{“}\textit{doba strávená čekáním při i-té cestě}\text{”}\ [\text{v hodinách}]$$

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 $\text{Ro}\,(a,b) = \text{Ro}\,\big(0, \frac{1}{6}\big)$.

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

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

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

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

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

$$P(Z \leq 3) = P\Big(\underbrace{3 \cdot (Z-4)}_{norm(Z)} \leq 3 \cdot (3-4)\Big) = P\Big(norm(Z) \leq -3\Big) \doteq$$
$$\doteq \Phi(-3) = 1 - \Phi(3) \doteq 1 - 0.9987 = 0.0013\ .$$

Příklad 10.3 *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.

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

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

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

$$= \frac{\lambda^{0\cdot17+1\cdot4+2\cdot1+3\cdot2+5\cdot1}}{konst.}\,\mathrm{e}^{-\lambda(17+4+1+2+1)} = \frac{\lambda^{17}}{konst.}\,\mathrm{e}^{-25\lambda},$$

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

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

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 $\overline{\boldsymbol{x}} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{\sum_i i\cdot n_i}{\sum_i n_i} = \frac{0\cdot17+1\cdot4+2\cdot1+3\cdot2+5\cdot1}{17+4+1+2+1} = \frac{17}{25}$

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 pouze za předpokladu, že je spojitá v oboru hodnot dané veličiny. Např. pro exponenciální rozdělení (které modeluje dobu čekání) je obor hodnot $(0, +\infty)$ a tam už hustotu spojitou máme.

Příklad 10.4 *Doba do poruchy starého výtahu má exponenciální rozdělení. Bylo zjištěno, že se výtah 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

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

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

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

Metoda maximální věrohodnosti:

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

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

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

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

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

(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 $k = 1, 2, \ldots$.

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 $\hat{\lambda} = \frac{1}{10}$. 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 $\overline{\boldsymbol{x}}$.