10. cvičení z STP

22. dubna 2021

Rychlost konvergence v CLV: Pokud pro veličiny X_i v CLV navíc ještě je $\varrho := E(|X_i - \mu|^3) < \infty$, pak platí Berry–Esseenův odhad pro $Z_n = \sum_{i=1}^n X_i$ (pro všechna $t \in \mathbb{R}$ a $n \in \mathbb{N}$):

$$\left|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\right| \le C_1 \cdot \frac{\varrho}{\sigma^3\sqrt{n}}$$

kde C_1 je nějaké konstanta. Nejlepší současný odhad pro C_1 zatím je, že $C_1 < 0.4748$.

Odhad chyby v CLV pro alternativní rozdělení: Pokud mají veličiny X_i alternativní rozdělení s parametrem p, tj. $P(X_i = 1) = p$, pak

$$\mu = E(X_i) = p, \qquad \sigma = \sqrt{D(X_i)} = \sqrt{p(1-p)}$$
$$\varrho = E(|X_i - p|^3) = p(1-p)\big(p^2 + (1-p)^2\big) = \sigma^2\big(p^2 + (1-p)^2\big)$$

čímž dostáváme odhad

$$\left|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\right| \quad \le \quad C_1 \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}} \quad < \quad 0.4748 \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}}$$

Nyní můžeme dále použít ještě, že pro $p \in (0,1)$ je $p^2 + (1-p)^2 \le 0.5$. Pro použití CLV pro aproximaci binomického rozdělení $Z_n \sim \mathrm{Bi}(n,p)$ se obvykle používá kritérium, že

$$\mathrm{var}(Z_n) = np(1-p) \ge 9$$

což pak dává odhad chyby:

$$\left|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\right| \quad \le \quad \ldots \quad < \quad 0.4748 \cdot \frac{0.5}{\sqrt{9}} < 0.08$$

- -

Obvyklý způsob použití CLV: Veličina $\mathrm{norm}(Z_n) = \mathrm{norm}(\overline{X}_n)$ má přibližně rozdělení $N(0,1)$. Pro výpočty se tedy užívá, že

- veličina Z_n se střední hodnotou $E(Z_n) = n\mu$ a rozptylem $D(Z_n) = n\sigma^2$ má přibližně rozdělení $N(n\mu, n\sigma^2)$,
- veličina $\overline{X}_n = \frac{Z_n}{n}$ se střední hodnotou $E(\overline{X}_n) = \mu$ a rozptylem $D(\overline{X}_n) = \frac{\sigma^2}{n}$ má přibližně rozdělení $N(\mu, \frac{\sigma^2}{n})$.

Příklad 10.1 *Počet bodů ze zkouškové písemky je náhodná veličina pohybující se v rozmezí* $0 - 100$*, s průměrem* 53 *a rozptylem* 839*. Celkem* 200 *studentů psalo zkouškový test. Určete pravděpodobnost, že průměrný počet bodů u těchto studentů byl menší než* 50*, a uveďte za jakých předpokladů.*

Řešení:
Označme si X_i počet bodů i-tého studenta, $i = 1, 2, \ldots, n$, pro $n = 200$. Ze zadání je $E(X_i) = 53$ a $\mathrm{var}(X_i) = 839$. Veličina $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ (tzv. výběrový průměr) představuje průměrný počet bodů skupiny n studentů. Zajímá nás tedy pravděpodobnost

$$P\left(\overline{X}_n < 50\right).$$

Předpokládáme, že veličiny X_i jsou nezávislé (tj. např. že studenti opisují), a odsud dostaneneme

$$E\left(\overline{X}_n\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{n}{n}E(X_1) = 53$$

$$\mathrm{var}\left(\overline{X}_n\right) = \frac{1}{n^2}\mathrm{var}(\sum_{i=1}^n X_i) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{var}(X_i) = \frac{\mathrm{var}(X_1)}{n} = \frac{839}{200}$$

Podle CLV tedy máme

$$P\left(\overline{X}_n < 50\right) = P\left(\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\mathrm{var}(\overline{X}_n)}} < \frac{50-53}{\sqrt{\frac{839}{200}}}\right) = P\left(\mathrm{norm}(X) < \underbrace{-\frac{3\sqrt{200}}{\sqrt{839}}}_{\doteq -1.46}\right) \doteq$$

$$\doteq \Phi(-1.46) = 1 - \Phi(1.46) \doteq 1 - 0.93 = 0.07.$$

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

Příklad 10.2 *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 $\overline{\boldsymbol{x}} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{\sum_i i\cdot n_i}{\sum_i n_i} = \frac{0\cdot 17+1\cdot 4+2\cdot 1+3\cdot 2+5\cdot 1}{17+4+1+2+1} = \frac{17}{25}$

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

Příklad 10.3 *Počet neúspěšných zásahů terče předtím, než se střelec trefí, má geometrické rozdělení* 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 pravděpo-

dobnost úspěšného zásahu.

Řešení:

Veličina

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

má geometrické rozdělení 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 $(x_1, \ldots, x_n)$ 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

$$L(p) = P_p(X_1 = x_1, \ldots, X_n = x_n) = \prod_{j=1}^{n} P_p(X_j = x_j) =$$

$$= \Big((1-p)^0 p\Big)^{20} \Big((1-p)^1 p\Big)^{10} \Big((1-p)^2 p\Big)^{7} \Big((1-p)^3 p\Big)^{3} =$$

$$= \ (1-p)^{0\cdot 20+1\cdot 10+2\cdot 7+3\cdot 3} \cdot p^{20+10+7+3} \ = \ (1-p)^{33} \cdot p^{40}$$

kde X_j jsou jednotlivé nezávislé veličiny (odpovídající jednotlivým pokusům) a x_j naměřené hodnoty.

Ekvivalentně budeme hledat maximum funkce

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

Z její derivace

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

dostáváme řešení

$$\ell'(\widehat{p}) = 0 \qquad \Longrightarrow \qquad \widehat{p} = \frac{40}{73} \doteq 0.54795$$

které vyhovuje zadání, tj. $\widehat{p} \in (0,1)$. Ze znamének derivace je snadno vidět, že v $\widehat{p} = \frac{40}{73}$ 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 $k = 1, 2, \ldots$.

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

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

a její odhad z realizace je

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

Porovnáním dostaneme

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

což dává opět řešení

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

jako v předchozí metodě.

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

Příklad 10.4 *Náhodná veličina X nabývá hodnot s pravděpodobnostmi dle tabulky, kde c, q jsou reálné parametry rozdělení. Z četností hodnot v náhodném výběru, uvedených v tabulce, odhadněte parametry c a q.*

hodnota i	1	2	3
pravděpodobnost $P_{(c,q)}(X = i)$	$c - q$	c	$c + q$
četnost n_i	8	10	5

Řešení:
Protože součet pravděpodobností všech hodnot je 1, musí být

$$1 = (c - q) + c + (c + q) = 3c$$

tedy $c = \frac{1}{3}$. Současně musí být pravděpodobnosti nezáporné, tj. $0 \leq c - q = \frac{1}{3} - q$ a $0 \leq c + q = \frac{1}{3} + q$, takže $|q| \leq \frac{1}{3}$. Zbývá tedy odhadnout parametr q.

Metoda maximální věrohodnosti:

Hledáme hodnotu q, která maximalizuje funkci věrohodnosti

$$L(q) = P_{(\frac{1}{3},q)}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{j=1}^{n} \underbrace{P_{(\frac{1}{3},q)}(X_j = x_j)}_{P_{(\frac{1}{3},q)}(X = x_j)} =$$

$$= \left(\frac{1}{3} - q\right)^{n_1} \cdot \left(\frac{1}{3}\right)^{n_2} \cdot \left(\frac{1}{3} + q\right)^{n_3} = \left(\frac{1}{3} - q\right)^{8} \cdot \left(\frac{1}{3}\right)^{10} \cdot \left(\frac{1}{3} + q\right)^{5}$$

kde X_j jsou jednotlivé nezávislé veličiny (v pokusech) a x_j naměřené hodnoty. Funkce L je nezáporná a spojitá na uzavřené množině $\langle -\frac{1}{3}, \frac{1}{3} \rangle$, takže zde nabývá maxima. To nemůže být v krajních bodech (tam je nulová) a proto je nabyto uvnitř dané množiny. To odpovídá hledání maxima funkce

$$\ell(q) = \ln\big(L(q)\big) = 8 \cdot \ln\left(\frac{1}{3} - q\right) + 5 \cdot \ln\left(\frac{1}{3} + q\right) + \textit{konst.}$$

na intervalu $(-\frac{1}{3}, \frac{1}{3})$. Protože maximum existuje, musí pro něj platit

$$0 = \ell'(q) = \frac{-8}{\frac{1}{3} - q} + \frac{5}{\frac{1}{3} + q}$$

Odhad parametru q je

$$q = -\frac{1}{13} \doteq -0.077 \ .$$

Odhady pravděpodobností hodnot $1, 2, 3$ jsou tedy

$$p_X(1) = \frac{16}{39} \doteq 0.41 \qquad p_X(2) = \frac{1}{3} \doteq 0.333 \qquad p_X(3) = \frac{10}{39} \doteq 0.256$$

což vyhovuje zadání.

Metoda momentů:

Střední hodnota je

$$E(X) = \left(\frac{1}{3} - q\right) + 2 \cdot \frac{1}{3} + 3 \cdot \left(\frac{1}{3} + q\right) = 2 + 2\,q$$

její odhad z realizace je

$$\overline{x} = \frac{1}{n} \sum_i i \cdot n_i = \frac{1}{8 + 10 + 5} \cdot (1 \cdot 8 + 2 \cdot 10 + 3 \cdot 5) = \frac{43}{23} \ .$$

Porovnáním dostaneme

$$2 + 2q = E(X) = \overline{x} = \frac{43}{23}$$

což odpovídá hodnotě

$$q = -\frac{3}{46} \doteq -0.065 \ .$$

Odhady pravděpodobností hodnot $1, 2, 3$ jsou tedy

$$p_X(1) = \frac{55}{138} \doteq 0.399 \qquad p_X(2) = \frac{1}{3} \doteq 0.333 \qquad p_X(3) = \frac{37}{138} \doteq 0.268$$

což opět vyhovuje zadání.

Jak je vidět, metoda max. věrohodnosti by vyšla stejně, ať by veličina X měla jakékoliv hodnoty (dokonce i nečíselné), zatímco metoda momentů velmi podstatně závisí právě na tom, jaké hodnoty veličina X má (např. kdybychom místo hodnot $\{1, 2, 3\}$ zvolili třeba $\{-3, 10, 4\}$, byl by výsledek úplně jiný a dokonce bychom ani žádný přijatelný odhad nemuseli takto získat.)

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).

Příklad 10.5 *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

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

(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 $\overline{\boldsymbol{x}}$.