9. cvičení z PST

16. listopadu 2022

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

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

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

platí, že

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

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 $\mathrm{norm}(\overline{X}_n) = \mathrm{norm}(Z_n)$.

Poznámka: V rámci Centrální limitní věty (níže) se vyskytuje posloupnost nezávislých náhodných veličin, která nejčastěji vzniká následujícím způsobem:

Mějme náhodnou veličinu $X : \Omega \to \mathbb{R}$ na pravděpodobnostním prostoru Ω (např. pro házení mincí je $\Omega = \{\textit{rub, líc}\}$ a veličina třeba $X(\textit{líc}) = 1$ a $X(\textit{rub}) = 0$ s rozdělením $\mathrm{Alt}(p)$). Jestliže nyní budeme opakovat (nekonečně) nezávislých pokusů, pak jejich výsledky tvoří posloupnost $\widetilde{\omega} = (\omega_1, \omega_2, \dots)$, kde $\omega_i \in \Omega$ pro $i \in \mathbb{N}$. Množina všech takovýchto možných posloupností je tedy $\Omega^{\mathbb{N}}$ (tj. spočetná kartézská mocnina množiny Ω).

Na této množině $\Omega^{\mathbb{N}}$ lze opět vybudovat pravděpodobnostní prostor tj. σ-algebru $\widetilde{\mathcal{A}}$ na $\Omega^{\mathbb{N}}$ (která se bude skládat ze spočetných sjednocení množin typu $\bigtimes_{i=1}^{\infty} A_i = A_1 \times A_2 \times \cdots$, kde $A_i \subseteq \Omega$ je jev pro každé i) a pravděpodobnost bude dána jako $\widetilde{P}\left(\bigtimes_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} P(A_i)$.

Výsledek při i-tém pokusu nyní bude veličina $X_i : \Omega^{\mathbb{N}} \to \mathbb{R}$, definovaná prostě jako $X_i(\widetilde{\omega}) = \omega_i$ pro $\widetilde{\omega} = (\omega_1, \omega_2, \dots)$. Takovéto veličiny pak budou nezávislé a budou mít rozdělení stejné jako veličina X.

- -

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 chyby (pro všechna $t \in \mathbb{R}$ a $n \in \mathbb{N}$):

$$\Big|F_{\mathrm{norm}(Z_n)}(t) - \Phi(t)\Big| < 0.4748 \cdot \frac{\varrho}{\sigma^3\sqrt{n}}$$

- -

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

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

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.)

9.1 *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

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

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

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

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

$$\Big|F_{\text{norm}(Z_n)}(t) - \Phi(t)\Big| \ < \ \frac{0.62}{\sqrt{n}} \quad \text{pro všechna } t \in \mathbb{R}.$$

9.2 *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

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

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

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

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

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

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

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

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

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

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

- -

Odhad chyby v CLV pro alternativní rozdělení: Pokud mají veličiny X_i alternativní rozdělení s parametrem p, tj. $P(X_i = 1) = p$, pak pro binomické rozdělení $Z_n = \sum_{i=1}^{n} X_i \sim \text{Bi}(n,p)$ dostáváme odhad

$$\left|F_{\text{norm}(Z_n)}(t) - \Phi(t)\right| \quad < \quad 0.4748 \cdot \frac{p^2 + (1-p)^2}{\sqrt{np(1-p)}} = 0.4748 \cdot \frac{p^2 + (1-p)^2}{\sqrt{D(Z_n)}}\ .$$

Aproximace CLV se obvykle používá pro $D(Z_n) \geq 9$. Pak je odhad chyby nejvýše: $\left|F_{\text{norm}(Z_n)}(t) - \Phi(t)\right| < \frac{0.4748}{\sqrt{9}} = 0.159$.

9.3 *Letecká společnost prodává letenky a chce co nejvíce utržit. Letadlo má* 216 *míst, ale ví se, že zhruba* 5% *lidí se k odletu nedostaví.*

(a) Jaká je pravděpodobnost, že pokud společnost prodá 220 letenek, nepřesáhne počet cestujících kapacitu letadla?

(b) Kolik nejvíce může společnost prodat letenek na jeden let, aby pravděpodobnost, že nepřesáhne kapacitu, byla alespoň 90 %?

Řešení:
Nechť $n \in \mathbb{N}$ je počet cestujících, kteří si koupili letenku. Pro $i = 1, \ldots, n$ si zavedeme veličiny

$$X_i = \begin{cases} 1 & , \ i\text{-tý cestující se dostaví k odletu,} \\ 0 & , \ i\text{-tý cestující se nedostaví k odletu.} \end{cases}$$

Veličiny X_i považujeme za nezávislé (i když to ve skutečnosti nemusí být až úplně splněno), s alternativním rozdělením, kde $P(X_i = 1) = 0.95$. Počet cestujících, kteří se dostaví k odletu bude veličina

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

která má binomické rozdělení $\mathrm{Bi}\Big(n, 0.95\Big)$, tedy $E(Z_n) = 0.95 \cdot n$ a $\mathrm{var}(Z_n) = 0.05 \cdot 0.95 \cdot n$. Použijeme centrální limitní větu na normovanou veličinu

$$\mathrm{norm}(Z_n) = \frac{Z_n - E(Z_n)}{\sqrt{\mathrm{var}(Z_n)}} = \frac{Z_n - 0.95 \cdot n}{\sqrt{0.95 \cdot 0.05 \cdot n}} \ .$$

(a) Máme $n = 220$. Chceme určit $P(Z_n \leq 216)$. Tedy

$$\mathrm{norm}(Z_n) = \frac{Z_n - 0.95 \cdot 220}{\sqrt{0.95 \cdot 0.05 \cdot 220}} = \frac{Z_n - 209}{\sqrt{10.45}}$$

a

$$P\Big(Z_n \leq 216\Big) = P\left(\mathrm{norm}(Z_n) \leq \frac{216 - 209}{\sqrt{10.45}}\right) = P\left(\mathrm{norm}(Z_n) \leq 2.165\right) \doteq \Phi(2.165) = 0.9848 \ .$$

V tomto příkladu můžeme využít i přímo binomické rozdělení, protože hodnota 220 není příliš vzdálená od 216). Máme totiž

$$P\Big(Z_n \leq 216\Big) = 1 - P\Big(Z_n \geq 217\Big) = 1 - \sum_{i=217}^{220} \binom{220}{220-i} 0.95^i \cdot 0.05^{220-i} = 0.9958.$$

(b) Zajímá nás teď největší n takové, že

$$0.9 \leq P\Big(Z_n \leq 216\Big) \ .$$

Pomocí úprav nerovností opět můžeme psát:

$$0.9 \leq P\Big(Z_n \leq 216\Big) = P\left(\mathrm{norm}(Z_n) \leq \frac{216 - 0.95 \cdot n}{\sqrt{0.95 \cdot 0.05 \cdot n}}\right) \doteq \Phi\left(\frac{216 - 0.95 \cdot n}{\sqrt{0.95 \cdot 0.05 \cdot n}}\right)$$

tedy přibližně má platit nerovnost

$$\Phi\left(\frac{216 - 0.95 \cdot n}{\sqrt{0.95 \cdot 0.05 \cdot n}}\right) \geq 0.9$$

$$\frac{216 - 0.95 \cdot n}{\sqrt{0.95 \cdot 0.05 \cdot n}} \geq \Phi^{-1}(0.9) = 1.282$$

neboli

$$\left(\sqrt{0.95 \cdot n}\right)^2 + \left(\sqrt{0.05} \cdot 1.282\right) \cdot \sqrt{0.95 \cdot n} - 216 \leq 0.$$

Dostaneme tak kvadratickou nerovnici

$$x^2 + \left(\sqrt{0.05} \cdot 1.282\right)x - 216 \leq 0$$

kde $x = \sqrt{0.95 \cdot n}$. Určíme kořeny a dostaneme, že nerovnost platí pro $x \in \langle -14.841,\ 14.5543 \rangle$. Největší $n \in \mathbb{N}$ tedy splňuje

$$\sqrt{0.95 \cdot n} \leq 14.5543$$

neboli

$$n \leq \frac{14.5543^2}{0.95} = 222.98$$

Letenek se tak může prodat $n = 222$.

Pokud by nám na přesném výsledku více záleželo, můžeme si ho v tomto případě zkontrolovat opět pomocí přesného vzorce jako výše. Pro $n = 222$ je $P(Z_n \leq 216) = 0.9679 \geq 0.9$, pro $n = 223$ je $P(Z_n \leq 216) = 0.9323 \geq 0.9$ a teprve pro $n = 224$ je $P(Z_n \leq 216) = 0.8757 < 0.9$.

Skutečný počet je tedy $n = 223$ (což trochu naznačoval i výsledek 222.98). Ale i CLV dává dostatečně dobrý odhad.