9. cvičení z STP

15. dubna 2021

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.

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

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

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

Pro lepší představu o tom, jakou roli pro veličinu s normálním rozdělením $X \sim N(\mu, \sigma^2)$ hraje směrodatná odchylka σ se používá tzv.

pravidlo tří-sigma (`https://cs.wikipedia.org/wiki/Pravidlo_t\%C5\%99\%C3\%AD\_sigma`)

které je ovšem čistě jen technickou pomůckou:

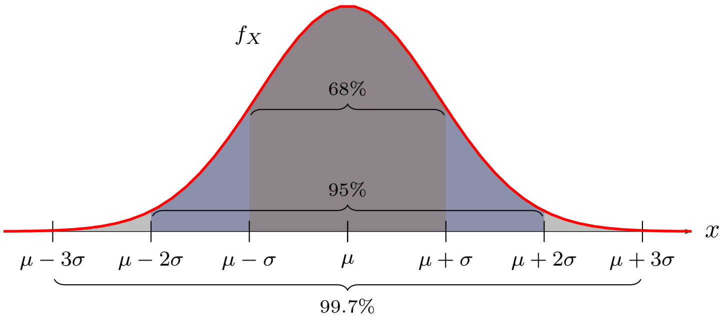
Jestliže si budeme počítat pravděpodobnosti

$$P(|X-\mu| \le k\cdot\sigma) = P\Big(\Big|\underbrace{\tfrac{X-\mu}{\sigma}}_{\sim N(0,1)}\Big| \le k\Big) = \Phi(k) - \Phi(-k) = 2\cdot\Phi(k) - 1 \quad \text{pro} \quad k = 1, 2, 3, \dots$$

dostaneme postupně

$$P(|X-\mu| \leq \sigma) = 2 \cdot \Phi(1) - 1 \doteq 2 \cdot 0.8413 - 1 = 0.6826 \doteq 68\%$$
$$P(|X-\mu| \leq 2 \cdot \sigma) = 2 \cdot \Phi(2) - 1 \doteq 2 \cdot 0.9772 - 1 = 0.9544 \doteq 95\%$$
$$P(|X-\mu| \leq 3 \cdot \sigma) = 2 \cdot \Phi(3) - 1 \doteq 2 \cdot 0.99865 - 1 = 0.9973 \doteq 99.7\%$$

Pro vyšší hodnoty, tj. $k \geq 4$ už jsou pravděpodobnosti vpodstatě rovny 1, takže se v praxi příliš nepoužívají (záleží samozřejmě na zvolené přesnosti).



V dalším příkladu se podíváme, jak můžeme lepší představu o pravděpodobnostech získat.

Příklad 9.1 *Pravděpodobnost toho, že se za dobu T porouchá přístroj je $p = 0.2$. S jakou pravděpodobností se za dobu T ze 100 (nezávisle pracujících) přístrojů porouchá*

(a) alespoň 20,

(b) méně než 28,

(c) 14 *až* 26 *přístrojů?*

Řešení:
Pro $i = 1, \ldots, n$ (kde $n = 100$) si zavedeme veličiny

$$X_i = \begin{cases} 1 & , \ i\text{-tý přístroj se porouchá,} \\ \\ 0 & , \ i\text{-tý přístroj bude v pořádku.} \end{cases}$$

Veličiny X_i budou nezávislé s alternativním rozdělením $\mathrm{Alt}(p) = \mathrm{Alt}(0.2)$, protože $P(X_i = 1) = 0.2$. Počet porouchaných přístrojů je tedy veličina

$$Z = \sum_{i=1}^{100} X_i$$

která má tudíž binomické rozdělení $\mathrm{Bi}(n,p) = \mathrm{Bi}(100, 0.2)$. To sice umíme přesně popsat, ale vyčíslování součtu mnoha velmi malých členu by vedlo ke značným numerickým chybám (a bez softwaru by ani nebylo možné). Proto použijeme CLV, která velmi dobře aproximuje hledané pravděpodobnosti.

Pro $Z = \sum_{i=1}^{n} X_i$ tedy máme

$$E(Z) = n \cdot E(X_1) = n \cdot p = 100 \cdot 0.2 = 20$$
$$\mathrm{var}(Z) = n \cdot \mathrm{var}(X_1) = n \cdot p \cdot (1-p) = 100 \cdot 0.2 \cdot 0.8 = 16$$
$$\Rightarrow \quad \sqrt{\mathrm{var}(Z)} = \sqrt{16} = 4$$

Podle CLV můžeme předpokládat, že veličina $norm(Z) = \frac{Z-E(Z)}{\sqrt{\mathrm{var}(Z)}} = \frac{Z-20}{\sqrt{16}}$ má přibližně normované normální rozdělení $N(0,1)$.

To také můžeme chápat tak, že veličina Z má *přibližně* normální rozdělení

$$N\Big(E(Z), \mathrm{var}(Z)\Big) = N(20, 4^2)$$

tedy že

$$F_Z(t) \doteq \Phi\left(\frac{t-20}{\sqrt{16}}\right) \quad \text{pro} \quad t \in \mathbb{R} \ .$$

.

Pak tedy máme:

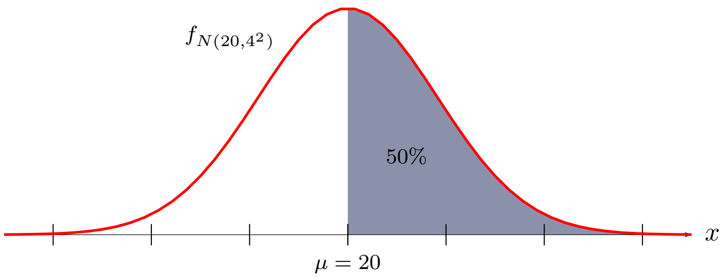
(a)

$$P(Z \geq 20) = P\left(\frac{Z-20}{\sqrt{16}} \geq \frac{20-20}{\sqrt{16}}\right) = P\big(norm(Z) \geq 0\big) =$$

$$= 1 - P\big(norm(Z) < 0\big) \overset{(CLV)}{\doteq} 1 - \Phi(0) = 1 - 0.5 = \mathbf{0.5} \ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.5398**.)

Když budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které ALE POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z, můžeme si představit hledanou pravděpodobnost (z HLEDISKA VÝPOČTU) takto:



(b)

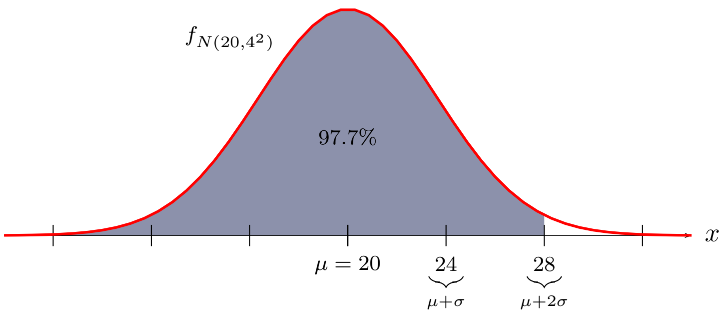
$$P(Z < 28) = P\left(\frac{Z-20}{\sqrt{16}} < \frac{28-20}{\sqrt{16}}\right) = P\Big(norm(Z) < 2\Big) \overset{(CLV)}{\doteq}$$

$$\overset{(CLV)}{\doteq} \Phi(2) \doteq \mathbf{0.977} \ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.9658**.)

Když opět budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z a vezmeme do úvahy pravidlo tří sigma, můžeme uvažovat (z HLEDISKA VÝPOČTU) takto:

Protože $28 = \mu + 2\sigma$, tak hodnota $P(Z < 28) \doteq P\Big(N(20, 4^2) < 28\Big)$ nám musí vyjít větší než 95% (viz náčrt).



(c)

$$P(14 \le Z \le 26) = P\left(\frac{14-20}{\sqrt{16}} \le \frac{Z-20}{\sqrt{16}} \le \frac{26-20}{\sqrt{16}}\right) = P\Big(-1.5 \le norm(Z) \le 1.5\Big) =$$

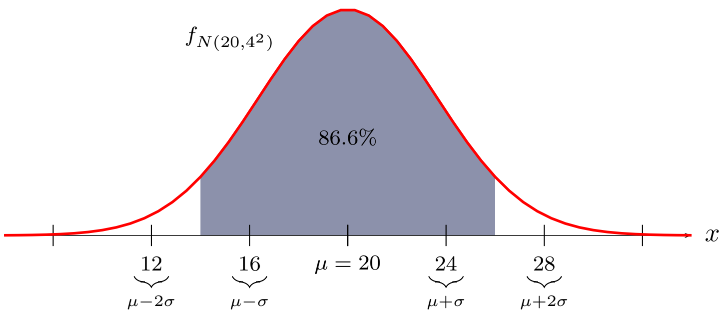
$$= P\Big(norm(Z) \le 1.5\Big) - P\Big(norm(Z) < -1.5\Big) \overset{(CLV)}{\doteq} \Phi(1.5) - \Phi(-1.5) =$$

$$= 2 \cdot \Phi(1.5) - 1 \doteq 2 \cdot 0.933 - 1 = \mathbf{0.866}\ .$$

(Pro srovnání: skutečná hodnota pro binomické rozdělení je **0.8973**.)

Když opět budeme uvažovat (spojité) rozdělení $N(20, 4^2)$, které POUZE aproximuje *původní diskrétní* binomické rozdělení veličiny Z a vezmeme do úvahy pravidlo tří sigma, můžeme uvažovat (z HLEDISKA VÝPOČTU) takto:

Protože $\mu - 2\sigma = 12 < 14 < 16 = \mu - \sigma$ a $\mu + \sigma = 24 < 16 < 28 = \mu + \sigma$, tak hodnota $P(14 \le Z \le 26) \doteq P\Big(14 \le N(20, 4^2) \le 26\Big)$ nám musí vyjít mezi 68% a 95% (viz náčrt).



Příklad 9.2 *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

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

nás zajímá $P(Z \ge 175)$. 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 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 $X_i \sim \text{Poiss}(4)$, 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 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**.)

Příklad 9.3 *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

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

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

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

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

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

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

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

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

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

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

Příklad 9.4 *Letecká společnost prodává letenky a chce co nejvíce utržit. Letadlo má* 216 *míst, ale ví se, že zhruba* 5% *lidí se k odletu nedostaví. Jaká je pravděpodobnost, že pokud společnost prodá* 220 *letenek, nepřesáhne počet cestujících kapacitu letadla?*

Řešení:
Pro veličinu

$Z =$ *"počet cestujících (z těch, co si koupili letenku), kteří se dostaví k odletu"*

nás zajímá $P(Z \le 216)$.

K řešení použijeme centrální limitní větu. Označme si tedy pro $i = 1, 2, ..., n$, kde $n = 220$, veličiny

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

Pokládáme je za nezávislé s alternativním rozdělením $X_i \sim \mathrm{Alt}(p) = \mathrm{Alt}(0.95)$. Protože platí $Z = \sum_{i=1}^{n} X_i$ dostaneme

$$E(Z) = n \cdot E(X_1) = n \cdot p = 220 \cdot 0.95 = 209$$
$$\mathrm{var}(Z) = n \cdot \mathrm{var}(X_1) = n \cdot p \cdot (1-p) = 209 \cdot 0.05 = 10.45$$
$$\Rightarrow \quad \sqrt{\mathrm{var}(Z)} = \sqrt{10.45} \doteq 3.233$$

(v případě rozptylu platí vztah díky nezávislosti veličin.)

Podle CLV bude mít veličina $norm(Z) = \frac{Z-E(Z)}{\sqrt{var(Z)}} = \frac{Z-209}{\sqrt{10.45}}$ přibližně rozdělení $N(0,1)$. Můžeme proto psát

$$P(Z \leq 216) = P\left(\frac{Z-209}{\sqrt{10.45}} \leq \frac{216-209}{\sqrt{10.45}}\right) = P\left(norm(Z) \leq \frac{7}{\sqrt{10.45}}\right) \overset{(CLV)}{\doteq}$$

$$\overset{(CLV)}{\doteq} \Phi\left(\frac{7}{\sqrt{10.45}}\right) \doteq \Phi(2.165) \doteq \mathbf{0.985}\ .$$

Doplnění: Rozdělení veličiny je zřejmě $Z \sim \mathrm{Bi}(n,p) = \mathrm{Bi}(220,\ 0.95)$ a její hodnoty jsou $Z \in \{0,1,\ldots,220\}$. Výpočet můžeme teď, vzhledem k malému počtu sčítanců, udělat i přímo:

$$P(Z \leq 216) = 1 - P(Z > 216) = 1 - \sum_{i=217}^{220} \binom{220}{i} 0.95^i \cdot 0.05^{220-i} =$$

$$= 1 - 0.95^{217}\Big(1750540 \cdot 0.05^3 + 24090 \cdot 0.95 \cdot 0.05^2 + 220 \cdot 0.95^2 \cdot 0.05 + 1 \cdot 0.95^3 \cdot 1\Big) \doteq$$

$$\doteq 1 - 0.95^{217} \cdot 286.82 \doteq 1 - 0.0042 = \mathbf{0.9958}\ .$$