10. cvičení z PST

22.- 26. dubna 2024

Odhad chyby v CLV pro binomické rozdělení: Pro binomické rozdělení $Z_n = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n,p)$, kde $X_i \sim Alt(p)$ jsou nezávislé, platí odhad

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

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

Příklad 10.1 *Letecká společnost prodává letenky a chce co nejvíce utržit. Letadlo má* 216 *míst, ale ví se, že průměrně* 5% *lidí se k odletu nedostaví. Jaká je pravděpodobnost, že pokud společnost prodá* 220 *letenek, nepřesáhne počet cestujících kapacitu letadla?*

Řešení:
Pro veličinu

$$Z = \text{“počet cestujících (z těch, co si koupili letenku), kteří se dostaví k odletu”}$$

nás zajímá $P(Z \leq 216)$.

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

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

(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

$$P(Z \leq 216) = P\left(\frac{Z-209}{\sqrt{10.45}} \leq \frac{216-209}{\sqrt{10.45}}\right) = P\left(norm(Z) \leq \frac{7}{\sqrt{10.45}}\right) \overset{(CLV)}{\doteq}$$
$$\overset{(CLV)}{\doteq} \Phi\left(\frac{7}{\sqrt{10.45}}\right) \doteq \Phi(2.165) \doteq \mathbf{0.985} \, .$$

CLV jsme použili, jelikož $\mathrm{var}(Z) = 10.45 \geq 9$. Jak je vidět, i když počet $n = 220$ se může zdát dostatečný, tak při odhadu záleží ve skutečnosti na součinu $\mathrm{var}(Z) = np(1-p) = 10.45$, který je kvůli malé hodnotě $1 - p = 0.05$ poměrně malý.

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:

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

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

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

Odhad a skutečná hodnota se tedy liší jen přibližně o 0.01, což je mnohem méně, než odhad chyby, který máme výše (zde k tomu přispívá i to, že obě hodnoty $P(Z \leq 216) = F_{norm(Z_n)}(2.165)$ i $\Phi(2.165)$ jsou blízké k 1.)

Příklad 10.2 *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 $norm(Z) = \frac{Z-E(Z)}{\sqrt{\text{var}(Z)}} = \frac{Z-20}{\sqrt{16}}$ má přibližně normované normální rozdělení $N(0,1)$.

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

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

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:

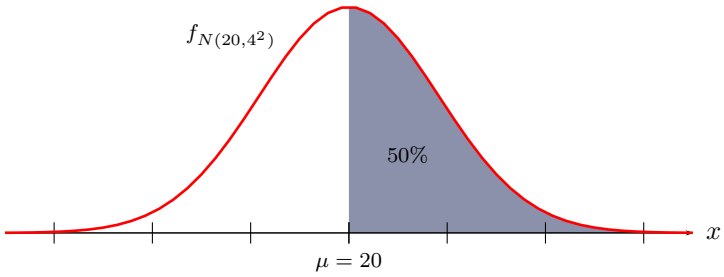
(a)

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

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

(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)

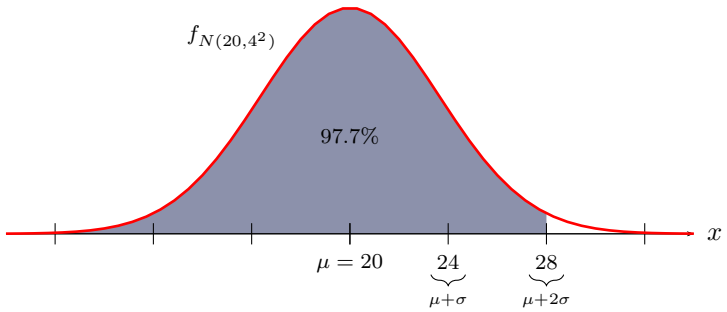
$$P(Z < 28) = P\left(\frac{Z-20}{\sqrt{16}} < \frac{28-20}{\sqrt{16}}\right) = P\Big(norm(Z) < 2\Big) \overset{(CLV)}{\doteq}$$

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

(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)

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

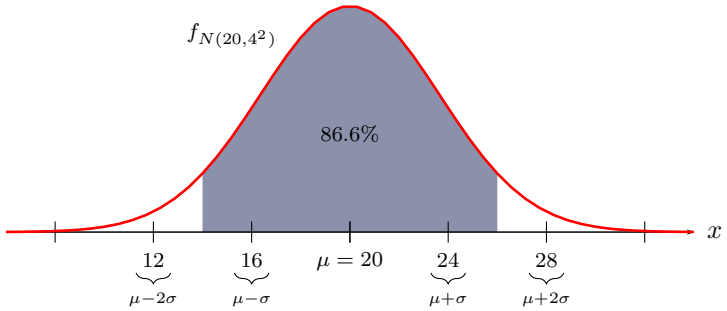
$$= P\Big(norm(Z) \le 1.5\Big) - P\Big(norm(Z) < -1.5\Big) \overset{(CLV)}{\doteq} \Phi(1.5) - \Phi(-1.5) =$$

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

(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).



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 $k = 1, 2, 3, \dots$)
  – *nevýhody:* obecně nemáme zaručeno, že dostaneme nějaký výsledek

Jak se používá metoda maximální věrohodnosti pro diskétně rozdělenou veličinu X:

Vyznačme závislost rozdělení X v závislosti na parametru ϑ dolním indexem takto $P_\vartheta(X = x)$, kde $x \in \mathbb{R}$.

Mějme konečnou množinu $K \subseteq \mathbb{R}$ naměřených hodnot veličiny X. Nechť každá z hodnot $i \in K$ má četnost při měřeních $n_i \in \mathbb{N}_0$. Pak věrohodnostní funkce má tvar

$$L(\vartheta) = \prod_{i \in K} \Big(P_\vartheta(X = i) \Big)^{n_i}$$

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 $(x_1, \dots, x_n)$ 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

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

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

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

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

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

kde X_j jsou jednotlivé nezávislé veličiny (v pokusech) a x_j naměřené hodnoty.

Funkci také můžeme jednoduše dostat použitím vzorce výše, tedy

$$L(\lambda) = \prod_{i \in \{0,1,2,3,5\}} \Big(P_\lambda(X = i) \Big)^{n_i} = \prod_{i \in \{0,1,2,3,5\}} \left(\frac{\lambda^i}{i!}\, \mathrm{e}^{-\lambda} \right)^{n_i} = \dots$$

Pro vyšetření maxima funkce L 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 rovnici $\lambda = \overline{\boldsymbol{x}}$, což vede 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 $\overline{\boldsymbol{x}}$.