Pravděpodobnost a statistika

8. cvičení

(vytvořeno 15. listopadu 2020)

1 Centrální limitní věta a Čebyševova nerovnost

Centrální limitní věta (CLV):

- $\{X_i\}_{i\in\mathbb{N}}$ ... posloupnost i.i.d. náhodných veličin

  (i.i.d. = independent and identically distributed = nezávislé se stejným rozdělením)
- Nechť $\forall i \in \mathbb{N}$: $\mathrm{E}X_i = \mu$, $\mathrm{D}X_i = \sigma^2$.
- Potom:

$$
\begin{aligned}
\mathrm{E}\left(\sum_{i-1}^{n} X_i\right) &= \sum_{i-1}^{n} \mathrm{E}X_i = n\mu \\
\mathrm{D}\left(\sum_{i-1}^{n} X_i\right) &= \sum_{i-1}^{n} \mathrm{D}X_i = n\sigma^2 \\
\lim_{n\to\infty} P\left[\frac{\left(\sum_{i-1}^{n} X_i\right) - n\mu}{\sqrt{n\sigma^2}} \le x\right] &= \Phi(x)
\end{aligned}
$$

Čebyševova nerovnost:

- Pro každou náhodnou veličinu X (s jakýmkoliv rozdělením) platí:

$$\boxed{\forall\delta > 0: \qquad P\Big[|\mathrm{norm}X| < \delta\Big] \ \ge\ 1 - \frac{1}{\delta^2}}$$

$$
\begin{aligned}
\text{upravíme:} \qquad P\left[\left|\frac{X - \mathrm{E}X}{\sqrt{\mathrm{D}X}}\right| < \delta\right] &\ge 1 - \frac{1}{\delta^2} \\
P\left[|X - \mathrm{E}X| < \delta\sqrt{\mathrm{D}X}\right] &\ge 1 - \frac{1}{\delta^2}\cdot\frac{\sqrt{\mathrm{D}X}^2}{\sqrt{\mathrm{D}X}^2}
\end{aligned}
$$

$$\boxed{\forall\varepsilon > 0: \quad P\Big[|X - \mathrm{E}X| < \varepsilon\Big] \ \ge\ 1 - \frac{\mathrm{D}X}{\varepsilon^2}}$$

Úloha 1 (Pravidlo 3σ)**.** Ve vzdálenosti $3\cdot\sigma_X$ od EX se nachází méně než 3 promile pozorování. Ověřte toto tvrzení pro:

1. normální rozdělení,
2. libovolné rozdělení pomocí Čebyševovy nerovnosti.

Úloha 2.

1. Hodíme 100krát mincí. Jaká je pravděpodobnost, že celkový počet líců bude mezi 49% a 51% hodů?
2. Jaká bude tato pravděpodobnost, pokud hodíme 10 000krát?
3. Kolikrát musíme mincí hodit, aby tato pravděpodobnost byla alespoň 99%?

Úloha 3. Sečteme 300 čísel zaokrouhlených na jedno desetinné místo. Jaká je pravděpodobnost, že chyba výsledku bude menší než 1?

Úloha 4. Kolik respondentů je potřeba oslovit, abychom volební výsledek dané strany odhadli s přesností 1% a spolehlivostí 90%?

Řešení

Řešení 1:

1. Pokud má X normální rozdělení $N(\mu, \sigma^2)$:

$$\begin{aligned} P\left[|X-\mu| \le 3\sigma\right] &= P\left[-3\sigma \le X-\mu \le 3\sigma\right] = P\left[-3 \le \frac{X-\mu}{\sigma} \le 3\right] \\ &= \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 2 \cdot 0.99865 - 1 = 0.9973 \end{aligned}$$

2. Pokud rozdělení X neznáme:

$$P\left[|X - \mathrm{E}X| < 3\sigma\right] \ge 1 - \frac{\sigma^2}{(3\sigma)^2} = 1 - \frac{1}{9} = \frac{8}{9} \doteq 0.89$$

Řešení 2:

1. X ... počet líců ... Binomické rozdělení s parametry $m = 100$, $q = \frac{1}{2}$

$$\begin{aligned} P\left[X = x\right] &= \binom{100}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{100-x} = \binom{100}{x}\frac{1}{2^{100}} \\ \mathrm{E}X &= mq = 50 \\ \mathrm{D}X &= mq(1-q) = 25 \end{aligned}$$

$$P\left[49 \le X \le 51\right] = \sum_{x=49}^{51}\binom{100}{x}\frac{1}{2^{100}} = \left(\binom{100}{49} + \binom{100}{50} + \binom{100}{51}\right) \cdot \frac{1}{2^{100}}$$

$$= \frac{98913082887808032681188722800+100891344545564193334812497256+98913082887808032681188722800}{1267650600228229401496703205376} \doteq 0.23564656559733319$$

- Pomocí CLV:

$$\begin{aligned} P\left[49 \le X \le 51\right] &= P\left[\frac{49-50}{\sqrt{25}} \le \frac{X - \mathrm{E}X}{\sqrt{\mathrm{D}X}} \le \frac{51-50}{\sqrt{25}}\right] = P\left[-\frac{1}{5} \le \mathrm{norm}X \le \frac{1}{5}\right] \\ &= \Phi(0.2) - \Phi(-0.2) = 2\,\Phi(0.2) - 1 \doteq 2 \cdot 0.5793 - 1 = 0.1586 \end{aligned}$$

- Pomocí Čebyševovy nerovnosti:

$$P\left[|X - 50| < 1\right] \ge 1 - \frac{25}{1^2} = 1 - 25 = -24 \ \ldots \ \text{žádná užitečná informace}$$

2. X ... počet líců ... Binomické rozdělení s parametry $m = 10\,000$, $q = \frac{1}{2}$

$$\begin{aligned} P\left[X = x\right] &= \binom{10\,000}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10\,000-x} = \binom{10\,000}{x}\frac{1}{2^{10\,000}} \\ \mathrm{E}X &= mq = 5\,000 \\ \mathrm{D}X &= mq(1-q) = 2\,500 \\ P\left[4\,900 \le X \le 5\,100\right] &= \sum_{x=4\,900}^{5\,100}\binom{10\,000}{x}\frac{1}{2^{10\,000}} \doteq 0.9555742009539193 \end{aligned}$$

- Pomocí CLV:

$$
\begin{aligned}
P\left[4\,900 \le X \le 5\,100\right] &= P\left[\frac{4\,900-5\,000}{\sqrt{2\,500}} \le \frac{X-\mathrm{E}X}{\sqrt{\mathrm{D}X}} \le \frac{5\,100-5\,000}{\sqrt{2\,500}}\right] \\
&= P\left[\frac{-100}{50} \le \mathrm{norm}X \le \frac{100}{50}\right] \\
&= \Phi(2)-\Phi(-2) \;=\; 2\Phi(2)-1 \;\doteq\; 2\cdot 0.9772-1 \;=\; 0.9544
\end{aligned}
$$

- Pomocí Čebyševovy nerovnosti:

$$
P\left[|X-5\,000| < 100\right] \;\ge\; 1-\frac{2\,500}{100^2} \;=\; 1-\frac{1}{4} \;=\; \frac{3}{4} \;=\; 0.75
$$

3. X … počet líců … Binomické rozdělení s parametry $m \in \mathbb{N}$, $q = \frac{1}{2}$

$$
\begin{aligned}
P\left[X = x\right] &= \binom{m}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{m-x} \;=\; \binom{m}{x}\frac{1}{2^m} \\
\mathrm{E}X &= mq \;=\; \frac{m}{2} \\
\mathrm{D}X &= mq(1-q) \;=\; \frac{m}{4} \\
P\left[0.49m \le X \le 0.51m\right] &= \sum_{x=0.49m}^{0.51m}\binom{m}{x}\frac{1}{2^m} \;\ge\; 0.99 \;\;\ldots\;\text{hledáme } m\text{, těžké}
\end{aligned}
$$

- Pomocí CLV:

$$
\begin{aligned}
P\left[0.49m \le X \le 0.51m\right] &\ge 0.99 \\
P\left[\frac{0.49m-0.5m}{\sqrt{\frac{m}{4}}} \le \frac{X-\mathrm{E}X}{\sqrt{\mathrm{D}X}} \le \frac{0.51m-0.5m}{\sqrt{\frac{m}{4}}}\right] &\ge 0.99 \\
P\left[-\frac{0.01m}{\sqrt{\frac{m}{4}}} \le \mathrm{norm}X \le \frac{0.01m}{\sqrt{\frac{m}{4}}}\right] &\ge 0.99 \\
P\left[-0.02\sqrt{m} \le \mathrm{norm}X \le 0.02\sqrt{m}\right] &\ge 0.99 \\
2\,\Phi\Big(0.02\sqrt{m}\Big)-1 &\ge 0.99 \\
\Phi(0.02\sqrt{m}) &\ge \frac{1.99}{2} \\
0.02\sqrt{m} &\ge \Phi^{-1}\left(\frac{1.99}{2}\right) \\
\sqrt{m} &\ge \frac{1}{0.02}\cdot\Phi^{-1}(0.995) \\
m &\ge \left(\frac{2.576}{0.02}\right)^2 \;=\; 16\,589.44
\end{aligned}
$$

  Mincí musíme hodit alespoň 16 590krát.

- Pomocí Čebyševovy nerovnosti:

$$
\begin{aligned}
P\left[\left|X-\frac{m}{2}\right| < 0.01m\right] \;\ge\; 1-\frac{\frac{m}{4}}{(0.01m)^2} &\ge 0.99 \\
1-\frac{m}{0.0004\cdot m^2} &\ge 0.99 \\
\frac{1}{0.0004\cdot m} &\le 0.01 \\
\frac{1}{0.0004\cdot 0.01} &\le m \\
m &\ge 250\,000 \\
m &\ge \frac{1}{0.0004\cdot 0.01} \;=\; 250\,000
\end{aligned}
$$

  Mincí musíme hodit alespoň 250 000krát.

Řešení 3:

$$P\left[\left|\sum_{i=1}^{300} X_i\right| < 1\right] \quad = \quad ?$$

- X_i chyba při zaokrouhlení i-tého čísla
  - rovnoměrné rozdělení na $\langle -0.05, 0.05\rangle$
  - $\mathrm{E}X_i = 0$
  - $\mathrm{D}X_i = \mathrm{E}(X_i^2) = \int_{-0.05}^{0.05} x^2 \cdot 10 \,\mathrm{d}x = \frac{10}{3}\left[x^3\right]_{-0.05}^{0.05} = \frac{1}{1200}$
  - $\mathrm{E}\left(\sum_{i=1}^{300} X_i\right) = \sum_{i=1}^{300} \mathrm{E}X_i = 0$
  - $\mathrm{D}\left(\sum_{i=1}^{300} X_i\right) = \sum_{i=1}^{300} \mathrm{D}X_i = 300 \cdot \frac{1}{1200} = \frac{1}{4}$
- Pomocí CLV:

$$\begin{aligned}
P\left[-1 \leq \sum_{i=1}^{300} X_i \leq 1\right] \quad &= \quad P\left[-\frac{1}{\sqrt{\frac{1}{4}}} \leq \frac{\sum_{i=1}^{300} X_i}{\sqrt{\mathrm{D}\left(\sum_{i=1}^{300} X_i\right)}} \leq \frac{1}{\sqrt{\frac{1}{4}}}\right] \\
&= \quad P\left[-2 \leq \mathrm{norm} \sum_{i=1}^{300} X_i \leq 2\right] \\
&= \quad \Phi(2) - \Phi(-2) \quad = \quad 2\Phi(2) - 1 \quad \doteq \quad 2 \cdot 0.9772 - 1 \quad = \quad 0.9544
\end{aligned}$$

- Pomocí Čebyševovy nerovnosti:

$$P\left[\left|\sum_{i=1}^{300} X_i\right| < 1\right] \quad \geq \quad 1 - \frac{\mathrm{D}\left(\sum_{i=1}^{300} X_i\right)}{1^2} \quad = \quad 1 - \frac{1}{4} \quad = \quad \frac{3}{4}$$

Řešení 4:

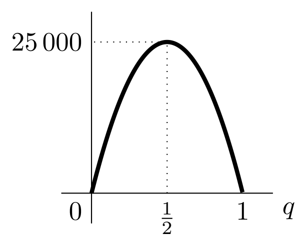
- $\varepsilon = 0.01$ ... přesnost
- $\alpha = 0.9$ ... spolehlivost
- n ... počet respondentů
- q ... (neznámá) pravděpodobnost
- S_n ... počet respondentů preferujících danou stranu ... Binomické rozdělení
- $\mathrm{E}S_n = nq$
- $\mathrm{E}S_n = nq(1-q)$
- $\overline{X} = \frac{S_n}{n}$ ... výběrový průměr (bodový odhad parametru p)
- jaké musí být n, aby:

$$\begin{aligned}
P\left[\left|\overline{X} - p\right| < \varepsilon\right] \quad &\geq \quad \alpha \\
P\left[\left|\frac{S_n}{n} - p\right| < \varepsilon\right] \quad &\geq \quad \alpha \\
P\left[\left|S_n - np\right| < n\varepsilon\right] \quad &\geq \quad \alpha
\end{aligned}$$

- Pomocí Čebyševovy nerovnosti:

$$\begin{aligned}
P\left[|S_n - np| < n\varepsilon\right] \;\geq\; 1 - \frac{\mathrm{D}S_n}{(n\varepsilon)^2} &\geq \alpha \\
1 - \frac{nq(1-q)}{n^2\varepsilon^2} &\geq \alpha \\
\frac{q(1-q)}{n\varepsilon^2} &\leq 1-\alpha \\
n &\geq \frac{q(1-q)}{\varepsilon^2(1-\alpha)} \\
\text{pro } \varepsilon = 0.01, \alpha = 0.9: \qquad n &\geq \frac{q(1-q)}{0.0001 \cdot 0.1} \;\geq\; 100\,000 \cdot q(1-q)
\end{aligned}$$

- Výsledek závisí na neznámém parametru $q \in \langle 0, 1\rangle$. Pro kterou jeho hodnotu to bude vycházet nejméně příznivě, tj., kde bude mít funkce $q \mapsto 100\,000 \cdot q(1-q)$ maximum?



- Maximum je v bodě $q = \frac{1}{2}$ a výsledkem je $n \geq \underline{25\,000}$.
- Pomocí CLV:

$$\begin{aligned}
P\left[-\varepsilon \leq \frac{S_n}{n} - q \leq \varepsilon\right] &\geq \alpha \\
P\left[-n\varepsilon \leq S_n - nq \leq n\varepsilon\right] &\geq \alpha \\
P\left[-\frac{n\varepsilon}{\sqrt{nq(1-q)}} \leq \frac{S_n - nq}{\sqrt{nq(1-q)}} \leq \frac{n\varepsilon}{\sqrt{nq(1-q)}}\right] &\geq \alpha \\
2\Phi\left(\frac{n\varepsilon}{\sqrt{nq(1-q)}}\right) - 1 &\geq \alpha
\end{aligned}$$

$$\begin{aligned}
\frac{n\varepsilon}{\sqrt{n}} &\geq \sqrt{q(1-q)} \cdot \Phi^{-1}\left(\frac{\alpha+1}{2}\right) \\
n &\geq \left(\frac{\sqrt{q(1-q)}}{\varepsilon} \cdot \Phi^{-1}\left(\frac{\alpha+1}{2}\right)\right)^2 \\
\text{pro } \varepsilon = 0.01, \alpha = 0.9, q = \frac{1}{4}: \qquad n &\geq \left(\frac{\sqrt{\frac{1}{4}}}{0.01} \cdot \Phi^{-1}(0.95)\right)^2 \\
n &\geq \left(\frac{1.645}{0.02}\right)^2 \;=\; \underline{6765.1}
\end{aligned}$$