11. cvičení z PSt — 2.–5.5.2023

Připomenutí teorie

- Čebyševova nerovnost: Nechť X má konečnou střední hodnotu μ a rozptyl σ^2. Pak
$$P(|X-\mu| \geq a \cdot \sigma) \leq \frac{1}{a^2}.$$
- Centrální limitní věta: Nechť $X_1, \dots, X_n$ jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2. Označme $Y_n = ((X_1 + \cdots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$. Pak $Y_n \xrightarrow{d} N(0,1)$. Neboli
$$\lim_{n\to\infty} F_{Y_n}(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$
Říkáme, že posloupnost Y_n konverguje k $N(0,1)$ *v distribuci (in distribution)*.
- De Moivre–Laplaceho věta říká o něco silnější věc: v okolí pn je i pravděpodobnostní funkce $Bin(n,p)$ dobře aproximována hustotou $N(np, np(1-p))$, tedy vhodně přeškálovanou Gaussovou funkcí φ. Přesněji:
$$\binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}$$
pro k blízké pn. Ještě přesněji: pokud pro nějaké c je $|k - pn| < c\sqrt{np(1-p)}$ a n se blíží nekonečnu, tak poměr dvou výrazů nahoře se blíží k jedné.

Aplikace nerovností a Centrální Limitní Věty

1. Statistik chce odhadnout průměrnou výšku h (v metrech) lidí v nějaké populaci, pomocí n nezávislých vzorků $X_1, \dots, X_n$, které vybíráme uniformně náhodně ze všech možnych lidí. Pro odhad použije výběrový průměr $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Odhaduje, že směrodatná odchylka jednoho měření je nejvýše 1 metr.

(a) Jak velké n má volit, aby směrodatná odchylka $\bar{X}_n$ byla nejvýše 1 cm?

(b) Pro jaké n zajistí Čebyševova nerovnost, že $\bar{X}_n$ se liší od h nejvýše o 5 cm s pravděpodobností alespoň 99 %?

(c) Statistik si všimne, že všichni měření lidé mají výšku v intervalu $(1.4, 2.1)$. Jak má upravit odhad směrodatné odchylky? Jak se změní odpovědi na předchozí otázky?

2. Označme $S = \sum_{k=0}^{30} \binom{100}{k}$. Označme dále $X = \sum_{i=1}^{100} X_i$, kde X_i je 0 nebo 1, obojí s pravděpodobností 1/2 a veličiny $X_1, \dots, X_n$ jsou nezávislé. Je tedy $X \sim Bin(100, 1/2)$.

(a) Vyjádřete S pomocí distribuční funkce F_X.

(b) Použijte CLV na odhad této pravděpodobnosti.

(c) Případně vyčíslete S vhodným softwarem a srovnejte.

3. Odhadněte $\binom{100}{30}$ pomocí CLV. Nápověda: použijte CLV pro odhad $P(29.5 < X < 30.5)$ pro vhodnou n.v. X. Na druhou stranu pro $P(X = 30)$ máme vzorec $\binom{100}{30}/2^{100}$ z binomického rozdělení. Alternativně, můžete použít Moivre–Laplaceho větu.

4. Chceme odhadnout, zda naše mince (a způsob jak s ní házíme) je spravedlivá. Pokud ze sta hodů padne orel více než 55-krát, řekneme, že spravedlivá není. Jaká je pravděpodobnost, že se zmýlíme?

Intervalové odhady

5. Máme jedno měření $X \sim N(\mu, 1)$. (Tj. parametr $\vartheta = \mu$.)

(a) Najděte intervalový odhad pro μ se spolehlivostí 95 %.

(b) Místo jednoho měření jich provedeme n (pochopitelně nezávislých). Jaký bude teď intervalový odhad pro μ?

(c) Nechť X má stále střední hodnotu μ a rozptyl 1, ale není už nutně normální. Co se změní?

1)

a) $\sqrt{var(\bar{X}_n)} \leq 0{,}01$

$$\sqrt{var(\bar{X}_n)} = \sqrt{\frac{var(X_1) + \dots + Var(X_n)}{n^2}} \leq \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$$

$$\sqrt{var(\bar{X}_n)} \leq \frac{1}{\sqrt{n}} \leq 0{,}01 \implies n \geq 1000$$

b)

$$0{,}01 > \frac{\sigma^2}{n \cdot \varepsilon^2} =$$

$$P(|S_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n \cdot \varepsilon^2}$$

$$x \leq \frac{\sigma^2}{n \cdot c^2}$$

$$\frac{1}{100} > \frac{100^2}{25n}$$

$$25n > 100^3$$

$$n > \frac{100^3}{25}$$

c) $\sigma = 35$

pak a) $= 1225$

b) $= 900$

2) $S = \sum_{k=0}^{30} \binom{100}{k}$, $X = \sum_{i=0}^{100} X_i$, kde $X_i = \begin{cases} 0 & p = \frac{1}{2} \\ 1 & p = \frac{1}{2} \end{cases}$

$$F_X = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

$$F_X(30) = \sum_{k=0}^{30} \binom{100}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{100-k} = \left(\frac{1}{2}\right)^{100} \cdot \sum_{k}^{30} \binom{100}{k}$$

$$S = 2^{100} \cdot F_X(30)$$