

# 10. cvičení

## Nerovnosti

**Úloha 1** (These two are the same picture)

Čebyševova nerovnost byla na přednášce v jiném znění:  $\Pr(|X - \mathbb{E}(X)| \geq a) \leq \text{var}(X)/a^2$ . Rozmyslete si, že znění v taháku je ekvivalentní.

### Řešení

Položme  $t := a/\sigma_X$  nebo  $a := t\sigma_X$ .

**Úloha 2** (Zase kostky?)

Házíme kostkou, za 1 a 2 dostaneme bod. Označme  $X$  počet bodů, které dostaneme po  $n$  (nezávislých) hodech. Odhadněte pravděpodobnost, že  $X \geq n/2$ .

- Pomocí Markovovy nerovnosti.
- Pomocí Čebyševovy nerovnosti.
- Pro konkrétní  $n$ , jak lze tuto hodnotu určit přesně?

**Řešení** a)  $(n/3)/(n/2) = 2/3$

b) Máme  $\text{Bin}(n, 1/3)$ , tedy  $\text{var}(X) = 2n/9$ , a  $P(X \geq n/2) \leq (2/9)n/(n/6)^2 = 8/n$

c) Stačí spočítat  $1 - F_X(49)$ , kde  $X \sim \text{Bin}(n, 1/3)$ .

**Úloha 3** (Statistika výšky)

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 možných lidí. Pro odhad použije výběrový průměr  $S_n = (X_1 + \dots + X_n)/n$ . Odhaduje, že směrodatná odchylka jednoho výběru je nejvýše 1 metr.

- Jak velké  $n$  má volit, aby směrodatná odchylka  $S_n$  byla nejvýše 1 cm?
- Pro jaké  $n$  zajistí Čebyševova nerovnost, že pravděpodobnost, že  $S_n$  se liší od  $h$  nejvýše o 5 cm s pravděpodobností alespoň 99 %?
- 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?

### Řešení

(Všechno počítáme v cm.)

- $X_i$  nezávislé, tedy rozptyl součtu je součet rozptylů, vynásobení konstantou znamená, že je to druhá mocnina konstanty rozptylu, tedy  $\text{var}(S_n) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{n}{n^2} 100^2 = \frac{10000}{n}$ , tedy chceme  $n \geq 10000$ .
- Máme tedy dle Čebyševa  $P(|S_n - h| \geq \frac{100a}{\sqrt{n}}) \leq \frac{1}{a^2}$ . Položíme-li  $100a/\sqrt{n} = 5$ , máme  $a = \frac{\sqrt{n}}{20}$ , a tedy  $1/a^2 = 400/n$ , což chceme rovno 0.01, a tedy  $n = 40000$ .
- Plyne z toho, že směrodatná odchylka je maximálně 0.35 (délka intervalu dělena dvěma, umím totiž odhadnout  $\mathbb{E}[(X - m)^2]$  při dosazení středu intervalu za  $m$ ).

## Zákony velkých čísel

### Úloha 4 (Počítání obsahu kruhu náhodným samplováním)

Vygenerujeme náhodný bod v jednotkovém čtverci (obě souřadnice budou mít rozdělení  $U(0, 1)$ ). Označíme  $X_i$  indikátor jevu „ $i$ -tý bod leží ve vepsaném kruhu“.

- Určete  $\mathbb{E}(X_i)$ ,  $\text{var}(X_i)$ .
- Položte  $\bar{X}_n = (X_1 + \dots + X_n)/n$ . Určete  $\mathbb{E}(\bar{X}_n)$  a  $\text{var}(\bar{X}_n)$ .
- Rozmyslete si, co říká slabý a silný zákon velkých čísel o aproximaci  $\pi$  pomocí  $X_n$ ?
- Pro jaké  $n$  čekáte, že dostaneme výsledek správně na jedno desetinné místo? Na dvě, tři, ...?
- Jiný výpočet obsahu kruhu:  $Y_i = \sqrt{1 - U_i^2}$ , kde  $U_i \sim U(0, 1)$ . Uvědomte si, že  $\mathbb{E}(Y_i)$  je obsah čtvrtkruhu, tedy  $\pi/4$ . Jaké je  $\text{var}(Y_i)$ ? Jaké je  $\text{var}(\bar{Y}_n)$ ?
- Která metoda je přesnější?

**Řešení** a)  $X_i \sim \text{Bern}(\pi/4) \rightsquigarrow \mathbb{E}[X_i] = \pi/4$ ,  $\text{var}(X_i) = (\pi/4)(1 - \pi/4)$

b)  $\mathbb{E}[\bar{X}_n] = \pi/4$  z linearity, z nezávislosti  $\text{var}(\bar{X}_n) = \frac{1}{n^2} \cdot (\sum \text{var}(X_i)) = \frac{(\pi/4)(1-\pi/4)}{n}$

c)  $S_n = ((n-1)S_{n-1} + X_n)/n$

d) **TODO**

## Centrální limitní věta

### Úloha 5 (Standardizace)

Připomeňme, že standardizací n.v.  $X$  myslíme  $\text{stand}(X) = (X - \mathbb{E}(X))/\sigma_X$ .

- $\text{stand}(X)$  má střední hodnotu 0 a rozptyl 1
- $Y_n$  v CLV je rovna  $\text{stand}(\bar{X}_n)$  a také  $\text{stand}(X_1 + \dots + X_n)$ .

**Řešení** a) Plyne z vlastností střední hodnoty a rozptylu.

b)  $Y_n$  v CLV je rovna  $\text{stand}(\bar{X}_n)$  a také  $\text{stand}(X_1 + \dots + X_n)$ .

### Úloha 6 (Downloading...)

Měříme rychlost stahování souborů z cloudového úložiště. Každý čas stahování jednoho souboru je náhodná veličina s průměrem  $\mu = 5$  minut a standardní odchylkou  $\sigma = 2$  minuty. Předpokládejme, že časy stahování jednotlivých souborů jsou na sobě nezávislé, stahování probíhá jedno po druhém (tj. vždy se stahuje jen jeden soubor, hned po jeho dokončení začneme stahovat další).

- Pokud stáhneme 50 souborů, jaká je přibližná pravděpodobnost, že celková doba stahování přesáhne 270 minut?
- Jaká je přibližná pravděpodobnost, že průměrná doba stahování na soubor je kratší než 4,5 minuty?

Použijte Centrální limitní větu. Napište přesnou formuli pomocí funkce  $\Phi$  a použijte tabulku na předchozí straně pro odhad.

**Řešení** a) Chceme  $P(S_n \geq 270)$ , tak označíme  $Y_n = (S_n - 50 \cdot 5)/(\sqrt{50} \cdot 2) \rightsquigarrow P(Y_n \geq \sqrt{2}) = 1 - \Phi(\sqrt{2})$ .

b) Chceme  $P(\bar{X}_n \leq 4.5)$ , označme tedy  $Z_n = (\bar{X}_n \cdot 50 - 50 \cdot 5)/(\sqrt{50} \cdot 2) \rightsquigarrow P(Z_n \leq -2.5/\sqrt{2}) \approx \Phi(-1.77)$ , kde v obou případech předpokládáme pro použití CLV  $50 = \infty$ .

### Úloha 7 (Rozdíl počtu hodů mincí)

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

- Vyjádřete  $S$  pomocí distribuční funkce  $X$ .

- b) Použijte CLV na odhad této pravděpodobnosti.  
 c) Zkuste  $S$  vyčíslit vhodným softwarem a srovnajte.

**Řešení** a)  $S = 2^{100} \cdot P[X \leq 30] = 2^{100} \cdot F_X(30)$ .

- b) Protože evidentně  $\mathbb{E}[X] = 50$ ,  $\text{var}(X) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25$  (bo  $S \sim \text{Bin}(100, 1/2)$ ), podle CLV máme  $\mathbb{E}[X_i] = 0$ ,  $\text{var}(X_i) = p(1-p) = 1/4$ , když budeme mít jednotlivé samplly  $X_1, \dots, X_n$ , zajímá nás  $P[\sum X_i \leq 30] = P[\frac{\sum X_i - 50}{5} \leq -4] = \Phi(-4)$  (řekněme, že 100 je dost velké)  
 c) **TODO**

## Kvantilová funkce

### Úloha 8 (Kvantilová funkce)

Kvantilovou funkci jsme definovali předpisem  $Q_X(p) = F_X^{-1}(p)$ .

- a) Jaký je obor hodnot  $Q_X$ ? Kdy dává definice smysl?  
 b) Nahlédněte, že taková funkce má (zjevně) tu vlastnost, že  $Q_X(p) = x \iff p = F_X(x)$ .  
 c) Rozmyslete si, jak je rozumné definovat  $Q_X$  pro diskrétní n.v. (a v čem je vlastně problém).

**Řešení** a) Interval  $(0, 1)$ , občas i s krajními hodnotami; dává smysl, když je  $F_X$  bijekce.

- b) Ano, nahlédněte.  
 c)  $Q_X(p) := \inf\{x \in \mathbb{R} : p \leq F_X(x)\}$

## Tahák

- **Markovova nerovnost:**  $P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$  pro  $X \geq 0$ .
- **Čebyševova nerovnost:**  $P(|X - \mathbb{E}(X)| \geq t\sigma_X) \leq \frac{1}{t^2}$ .
- Nechtě  $X_1, \dots, X_n$  jsou stejně rozdělené n.n.v. se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ . Definujeme  $\bar{X}_n := (X_1 + \dots + X_n)/n$ .

- **Silný zákon velkých čísel**  $\bar{X}_n \xrightarrow{s.j.} \mu$
- **Slabý zákon velkých čísel**  $\bar{X}_n \xrightarrow{P} \mu$ , neboli  $(\forall \varepsilon > 0) \Pr(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0$ .  
 Ukazovali jsme si, že dokonce  $\Pr(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$ .
- **Centrální limitní věta:** Označme  $Y_n = ((X_1 + \dots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$ . Pak  $Y_n \xrightarrow{d} N(0, 1)$ . Neboli

$$\lim_{n \rightarrow \infty} F_{Y_n}(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$

$x$	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5
$\Phi(x)$	0.01	0.02	0.07	0.16	0.31	0.5	0.69	0.84	0.93	0.98	0.99