

NMAI059 Pravděpodobnost a statistika 1

9. přednáška

Robert Šámal

Nerovnosti, které známe z minula

- ▶ Markov:

$$X \geq 0 \Rightarrow P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$$

- ▶ Čebyšev/Chebyshev

$$P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq \frac{1}{a^2}$$

- ▶ Chernoff ($\sigma_X = \sqrt{n}$)

$$X = \sum_{i=1}^n X_i, X_i = \pm 1 \Rightarrow P(|X - \mathbb{E}(X)| > a\sigma_X) \leq 2e^{-a^2/2}$$

Přehled

Limitní věty

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť X_1, \dots, X_n jsou stejně rozdělené n.n.v. se stř. hodnotou μ a rozptylem σ^2 . Označme $S_n = (X_1 + \dots + X_n)/n$. Pak pro každé $\varepsilon > 0$ platí

$$\lim_{n \rightarrow \infty} P(|S_n - \mu| \geq \varepsilon) = 0.$$

Říkáme, že posloupnost S_n konverguje k μ v pravděpodobnosti (in probability), píšeme $S_n \xrightarrow{P} \mu$.

SZVČ → Centrální Limitní věta

Centrální Limitní věta

Věta

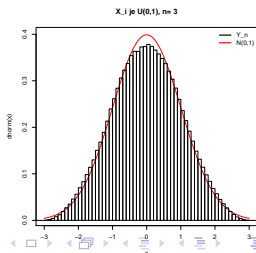
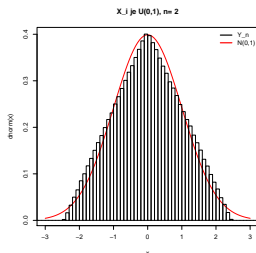
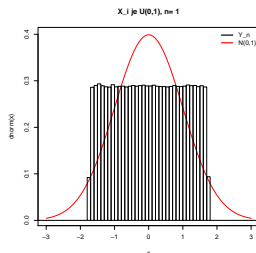
Nechť X_1, \dots, X_n jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2 . Označme

$$Y_n = ((X_1 + \dots + X_n) - n\mu) / (\sqrt{n} \cdot \sigma).$$

Pak $Y_n \xrightarrow{d} N(0, 1)$. Neboli, pokud F_n je distribuční funkce Y_n , tak

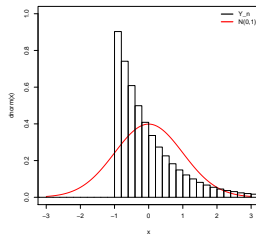
$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x) \quad \text{pro každé } x \in \mathbb{R}.$$

Říkáme, že posloupnost Y_n konverguje k $N(0, 1)$ v distribuci (in distribution).

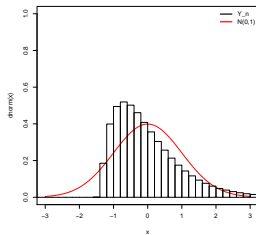


CLV další ukázka

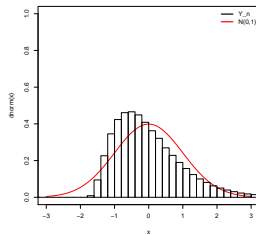
X_j je $\text{Exp}(1)$, $n=1$



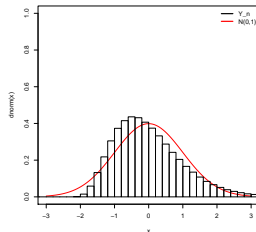
X_j je $\text{Exp}(1)$, $n=2$



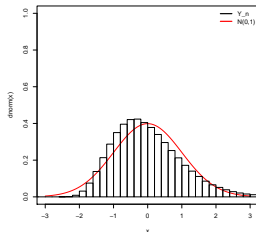
X_j je $\text{Exp}(1)$, $n=3$



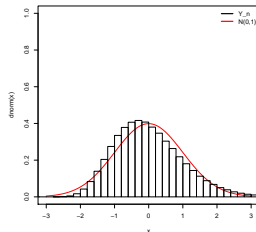
X_j je $\text{Exp}(1)$, $n=5$



X_j je $\text{Exp}(1)$, $n=7$



X_j je $\text{Exp}(1)$, $n=10$



Bonus: Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

$$M_X(t) = \mathbb{E}(e^{tX}).$$

Funkci $M_X(t)$ nazýváme momentová vytvořující funkce (moment generating function).

- ▶ $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}(X^n) \frac{t^n}{n!}$.
- ▶ $M_{Bern(p)}(t) = p \cdot e^t + (1 - p)$.
- ▶ $M_{X+Y}(t) = M_X(t)M_Y(t)$, jsou-li X, Y n.n.v.
- ▶ $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- ▶ $M_{N(0,1)} = e^{t^2/2}$
- ▶ $M_{Exp(\lambda)} = \frac{1}{1-t/\lambda}$
- ▶ Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

Přehled

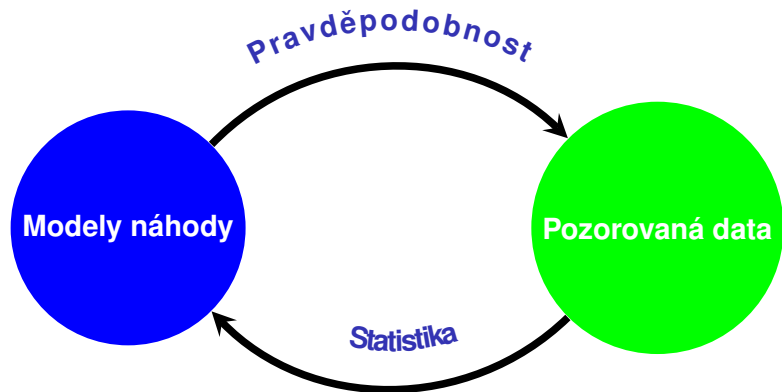
Limitní věty

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Plán přednášky



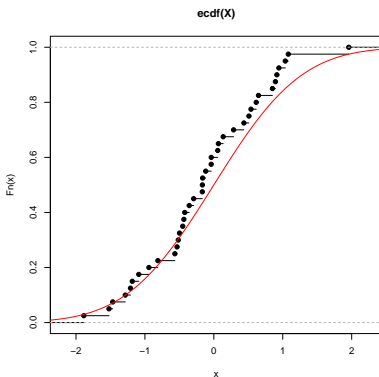
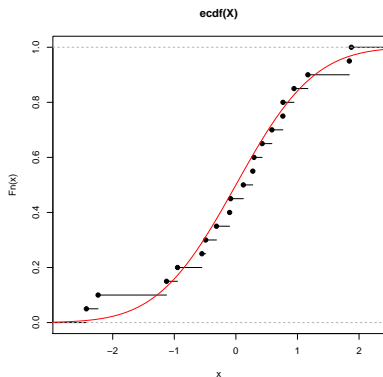
1. ilustrace – počet leváků

2. ilustrace – doba běhu programu

- ▶ $X_1, \dots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- ▶ **Definice:** Empirická distribuční funkce (empirical CDF) je definována

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.



Empirická distribuční funkce – vlastnosti

Věta

Pro pevné x platí

- ▶ $\mathbb{E}(\widehat{F}_n(x)) = F(x)$
- ▶ $\text{var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$
- ▶ $\widehat{F}_n(x)$ konverguje k $F(x)$ v pravděpodobnosti, píšeme $\widehat{F}_n(x) \xrightarrow{P} F(x)$.

Důkaz.

Slabý zákon velkých čísel.

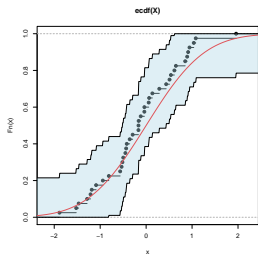
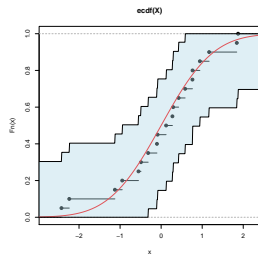


Empirická distribuční funkce – Dvoretzky-Kiefer-Wolfowitz (DKW)

Věta

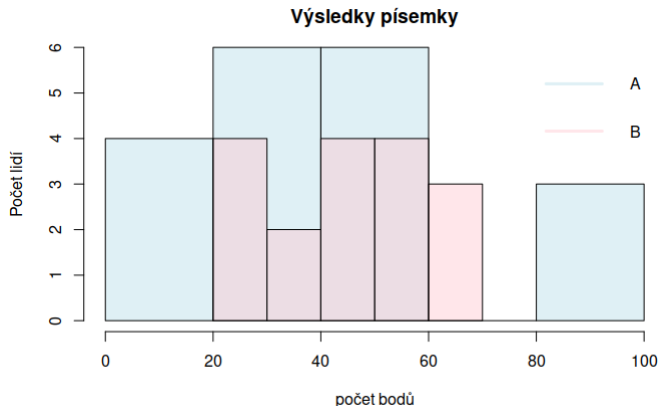
Nechť $X_1, \dots, X_n \sim F$ jsou n.n.v., \hat{F}_n jejich empirická distribuční funkce. Nechť $\mathbb{E}(X_i)$ je konečná. Zvolme $\alpha \in (0, 1)$ a označme $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. Pak platí

$$P(\hat{F}_n(x) - \varepsilon \leq F(x) \leq \hat{F}_n(x) + \varepsilon) \geq 1 - \alpha.$$



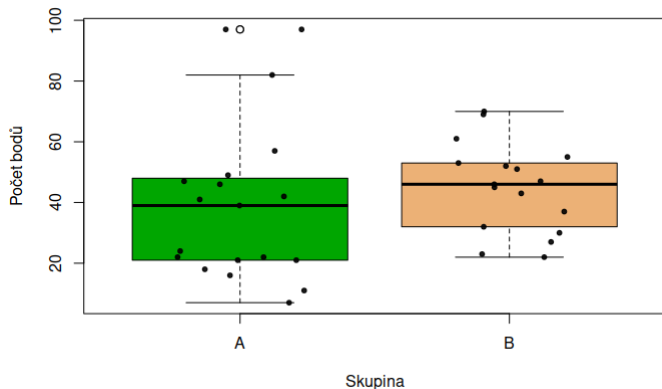
Intro – explorační analýza dat (exploratory data analysis)

- ▶ posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, ...)
- ▶ různé tabulky (třeba v Excelu a spol.)
- ▶ vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



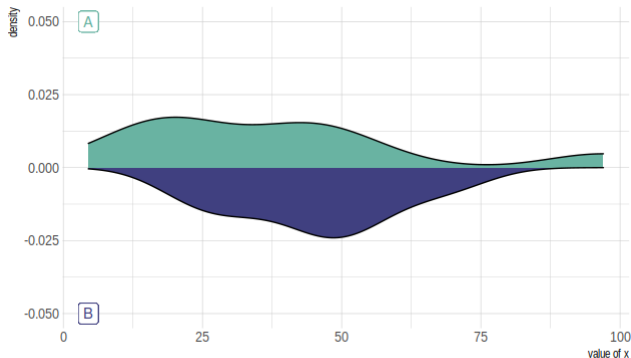
Intro – explorační analýza dat (exploratory data analysis)

- ▶ posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, ...)
- ▶ různé tabulky (třeba v Excelu a spol.)
- ▶ vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



Intro – explorační analýza dat (exploratory data analysis)

- ▶ posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, . . .)
- ▶ různé tabulky (třeba v Excelu a spol.)
- ▶ vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



Náhodný výběr

- ▶ s vracením
- ▶ bez vracení

$\Omega = \{\text{všechny } n\text{-tice obyvatel ČR}\}$

Pro $\omega = (\omega_1, \dots, \omega_n)$ zvolíme $X_i = I(\omega_i \text{ je levák})$.

Statistika – přehled

- ▶ nezávislá měření – hodnoty n.n.v. $X_1, \dots, X_n \sim F$
náhodný výběr s distribuční funkcí F s rozsahem n
- ▶ neparametrické modely: povolujeme velkou třídu F
- ▶ parametrické modely: $F \in \{F_\vartheta : \vartheta \in \Theta\}$
- ▶ příklady:
 - ▶ $Pois(\lambda)$ (parametr $\vartheta = \lambda, \Theta = \mathbb{R}^+$)
 - ▶ $U(a, b)$ (parametr $\vartheta = (a, b), \Theta = \mathbb{R}^2$)
 - ▶ $N(\mu, \sigma^2)$ (parametr $\vartheta = (\mu, \sigma), \Theta = \mathbb{R} \times \mathbb{R}^+$)
- ▶ „Všechny modely jsou špatné, ale některé jsou užitečné.“
(George Box)

Zkoumané úlohy – cíle konfirmační analýzy (confirmatory data analysis)

- ▶ bodové odhady
 - ▶ intervalové odhady
 - ▶ testování hypotéz
 - ▶ (lineární) regrese
-
- ▶ *statistika* – libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum, atd.
Tj. $T = T(X_1, \dots, X_n)$.

Další typy zkoumaných problémů

- ▶ Je zkoumaný lék účinný?
- ▶ Je naše mince, kostka spravedlivá?
- ▶ Předpokládáme, že výška člověka je normálně rozdělená. Jaké je μ , σ ?
- ▶ Předpokládáme, že výška člověka je normálně rozdělená. V jakém vztahu je průměrná výška mužů a žen? Praváků a leváků?
- ▶ Jak závisí náklon šikmé věže v Pise na čase?

Zkoumané úlohy – předpoklady

- ▶ Vždy předpokládáme, že máme nezávislá měření – hodnoty n.n.v. $X_1, \dots, X_n \sim F$
- ▶ O F předpokládáme, že patří do nějakého *modelu* – množiny vhodných distr. funkcí.
- ▶ parametrické/neparametrické modely

Zkoumané úlohy – cíle

- ▶ bodové odhady
- ▶ intervalové odhady
- ▶ testování hypotéz
- ▶ (lineární) regrese

Přehled

Limitní věty

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Výběrový průměr a rozptyl

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\bar{S}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

$$\hat{S}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

Cíle

Definice

Odhad $\hat{\Theta}_n = \hat{\Theta}_n(X_1, \dots, X_n)$ parametru ϑ je

- ▶ *neustranný (unbiased)* – pokud $\vartheta = \mathbb{E}(\hat{\Theta}_n)$
- ▶ *asymptoticky neustranný (asymptotically unbiased)* – pokud $\vartheta = \lim_{n \rightarrow \infty} \mathbb{E}(\hat{\Theta}_n)$
- ▶ *vychýlení (bias)* $bias_{\vartheta}(\hat{\Theta}_n) = \mathbb{E}(\hat{\Theta}_n) - \vartheta$
- ▶ *střední kvadratická chyba (mean squared error, MSE)* je $\mathbb{E}((\hat{\Theta} - \vartheta)^2)$

Věta

$$MSE = bias_{\vartheta}(\hat{\Theta}_n)^2 + var_{\vartheta}(\hat{\Theta}_n)$$

Parametry výběrového momentu a rozptylu

Věta

1. \bar{X}_n je konzistentní nestranný odhad μ
2. \bar{S}_n je konzistentní asymptoticky nestranný odhad μ
3. \hat{S}_n je konzistentní nestranný odhad μ

Metoda momentů

- ▶ $m_r(\vartheta) := \mathbb{E}(X^r)$ pro $X \sim F_\vartheta$... r -tý moment
- ▶ $\widehat{m}_r(\vartheta) := \frac{1}{n} \sum_{i=1}^n X_i^r$ pro náhodný výběr X_1, \dots, X_n z F_ϑ
... r -tý výběrový moment

Věta

$\widehat{m}_r(\vartheta)$ je nestranný konzistentní odhad pro $m_r(\vartheta)$

- ▶ Odhad metodou momentů je řešení soustavy rovnic

$$m_r(\vartheta) = \widehat{m}_r(\vartheta) \quad r = 1, \dots, k.$$

Metoda momentů – příklady

Metoda maximální věrohodnosti (maximal likelihood, ML)

- ▶ náh. výběr $X = (X_1, \dots, X_n)$ z modelu s parametrem ϑ
- ▶ možný výsledek $x = (x_1, \dots, x_n)$
- ▶ ... sdružená pravděpodobnostní funkce $p_X(x; \vartheta)$
- ▶ ... sdružená hustota $f_X(x; \vartheta)$
- ▶ *věrohodnost (likelihood)* $L(x; \vartheta)$ značí p_X nebo f_X
- ▶ normálně: máme pevné ϑ , a $L(x; \vartheta)$ je funkce x
- ▶ teď: máme pevné x a $L(x; \vartheta)$ je funkce ϑ

Metoda MV (ML):

volíme takové ϑ , pro které je $L(x; \vartheta)$ maximální

Metoda maximální věrohodnosti (maximal likelihood, ML)

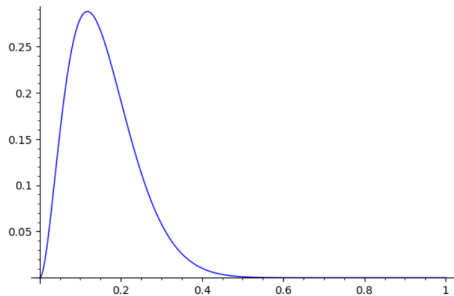
- ▶ **Metoda MV (ML):**
volíme takové ϑ , pro které je $L(x; \vartheta)$ maximální
- ▶ definujeme také $\ell(x; \vartheta) = \log(L(x; \vartheta))$
- ▶ díky nezávislosti je

$$L(x; \vartheta) =$$

$$\ell(x; \vartheta) =$$

ML – leváci

```
plot(binomial(17,2)*p^2*(1-p)^15, [0,1])
```



Přehled

Limitní věty

Statistika – úvod

Statistika – bodové odhady

Statistika – intervalové odhady

Intervalové odhady

- ▶ místo jednoho čísla s nejistým významem vypočítáme z dat interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

Definice

Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ jsou n.v. které závisí na náhodném výběru $X = (X_1, \dots, X_n)$. Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti $1 - \alpha$ ($1 - \alpha$ confidence interval), pokud

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha.$$

Intervalové odhady normální náhodné veličiny