NMAI059 Pravděpodobnost a statistika 1
10. přednáška

Robert Šámal

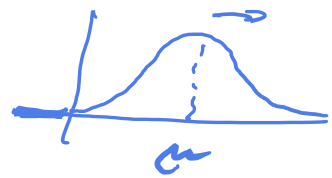
Přehled

Náhodný výběr

- bez vracení
  $\Omega = \{$všechny n-tice obyvatel ČR$\}$
  Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.
- s vracením
  $\Omega = \{$všechny n-tice obyvatel ČR, mohou se opakovat$\}$
  Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.
- varianty (stratifikovaný výběr)
  Chceme adekvátně reprezentovat různé podmnožiny (dané věkem, bydlištěm, ...).
  Nebudeme dále zkoumat.

Statistika – model

- nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$
  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. $T = T(X_1, \ldots, X_n)$.

Přehled

Výběrový průměr a rozptyl

↳ příklady statistik

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

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

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

↳ odhad rozptylu

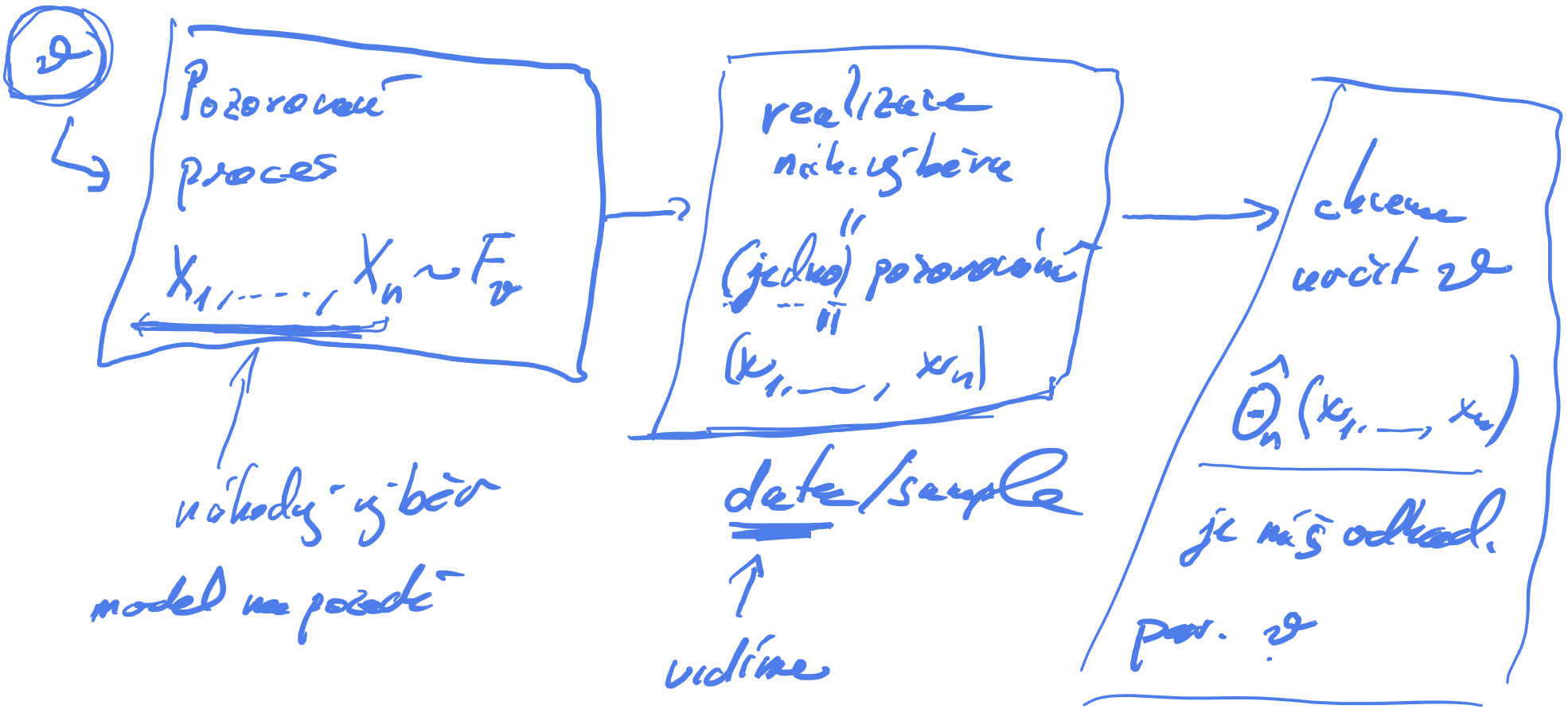
Odhad

ϑ velké theta

$\vartheta = t$ malé theta

Definice

Odhad je libovolná statistika.



Vlastnosti bodových odhadů

Definice

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

- *nestranný (unbiased) – pokud* $\vartheta = \mathbb{E}(\hat{\Theta}_n)$ *(pro každé* ϑ*)*
- *asymptoticky nestranný (asymptotically unbiased) – pokud* $\vartheta = \lim_{n\to\infty} \mathbb{E}(\hat{\Theta}_n)$
- *konzistentní (consistent) – pokud* $\hat{\Theta}_n \xrightarrow{P} \vartheta$.
- *vychýlení (bias)* $bias_\vartheta(\hat{\Theta}_n) := \mathbb{E}(\hat{\Theta}_n) - \vartheta$
- *střední kvadratická chyba (mean squared error, MSE) je* $MSE := \mathbb{E}((\hat{\Theta}_n - \vartheta)^2)$

Věta

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

Parametry výběrového momentu a rozptylu

$X_1, \dots, X_n$ náh. výběr $\sim F_\vartheta$ např. Bern(ϑ) chceme určit $\mu = \frac{1}{\vartheta}$ — nějaká fce ϑ

σ^2 = jiná fce ϑ

Věta

1. $\bar{X}_n$ *je konzistentní nestranný odhad* μ $= \mathbb{E}X_1 = \mathbb{E}X_2 = \dots$
2. $\bar{S}_n$ *je konzistentní asymptoticky nestranný odhad* σ^2
3. $\hat{S}_n$ *je konzistentní nestranný odhad* σ^2

(1) $\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$

$\bar{X}_n$ je nestranný, tj. $\mathbb{E}\bar{X}_n = \mu$ $\Rightarrow$ i asympt. nestr.

$\frac{\mathbb{E}X_1 + \mathbb{E}X_2 + \dots + \mathbb{E}X_n}{n} = \frac{\mu + \mu + \dots + \mu}{n} = \mu$

$\bar{X}_n$ je konsist., tj. $\bar{X}_n \xrightarrow{P} \mu$ Slabý z.v.č. (Čebyšev.)

$\left(\mathrm{var}(\bar{X}_n) = \frac{\sigma^2}{n}\right)$

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

$\mu = \mathbb{E}X_i$

$\sigma^2 = \operatorname{var}(X_i)$

$\mu = \mathbb{E}\bar{X}_n$

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

$= \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)^2 - 2(X_i-\mu)(\bar{X}_n-\mu) + (\bar{X}_n-\mu)^2\right]$

$\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - \mathbb{E}\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\bar{X}_n-\mu) + \mathbb{E}(\bar{X}_n-\mu)^2$

$\mathbb{E}\frac{1}{n}\sum(X_i-\mu)^2 - 2\mathbb{E}(\bar{X}_n-\mu)(\bar{X}_n-\mu) + \mathbb{E}(\bar{X}_n-\mu)^2$

$\frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}(X_i-\mu)^2}_{\operatorname{var}X_i=\sigma^2} - \underbrace{\mathbb{E}(\bar{X}_n-\mu)^2}$

asymptotyczny

$\sigma^2 - \operatorname{var}\bar{X}_n = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2$

$$\mathbb{E}\,\bar{S}_n = \frac{n-1}{n}\sigma^2 \longrightarrow \sigma^2$$

konzistence vychýlení

Bessel korekce

(3) $$\hat{S}_n = \frac{1}{n-1}\sum(\dots)^2 = \frac{n}{n-1}\bar{S}_n$$

$$\mathbb{E}\,\hat{S}_n = \frac{n}{n-1}\mathbb{E}\,\bar{S}_n = \sigma^2 \longrightarrow \hat{S}_n \text{ je nestranný odhad}$$

Je lepší $\hat{S}_n$ nebo $\bar{S}_n$?

$\longrightarrow$ $\hat{S}_n$ je nestranný, $\bar{S}_n$ ne.

$\longrightarrow$ Který má lepší M.S.E. ?

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 $X_1, \ldots, X_n$ 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

$$m_r(\vartheta) = \widehat{m_r(\vartheta)} \qquad r = 1, \ldots, k.$$

Metoda momentů – příklady

(1) $X_1, \dots, X_n \sim Bern(p)$ --- $X_i = $ "i-tý člověk je levák"

$\vartheta = p \in [0,1]$

naměřená data

$x_1, \dots, x_n$

$m_1(\vartheta) = \mathbb{E}X_1 = \vartheta$

$\widehat{m_1(\vartheta)} = \frac{1}{n}(X_1 + \dots + X_n) = \overline{X_n}$

$\frac{x_1 + \dots + x_n}{n} = \vartheta$

$\hat{\vartheta}_n(x_1, \dots, x_n)$

(2) $X_1, \dots, X_n \sim U(0, \vartheta)$

$m_1(\vartheta) = \mathbb{E}X_1 = \frac{\vartheta}{2}$ $\longrightarrow$ $\frac{\vartheta}{2} = \frac{x_1 + \dots + x_n}{n}$

$\widehat{m_1(\vartheta)} = \overline{X_n}$

bod. odhad $\frac{x_1 + \dots + x_n}{n} \cdot 2$

statistika $\hat{\vartheta}_n(X_1, \dots, X_n)$

Metoda maximální věrohodnosti (maximal likelihood, ML)

- náh. výběr $X = (X_1, \ldots, X_n)$ z modelu s parametrem ϑ
- možný výsledek $x = (x_1, \ldots, x_n)$
- … 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

6 leváků z 43 měření

$X_1, \dots, X_{43}$ $\quad X_i$ = "i-tý člověk je levák"

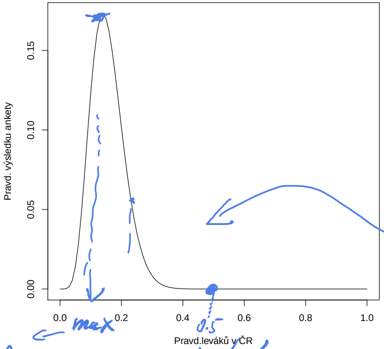
$EX_i = \vartheta$ = procento leváků v ČR

realizace $X_1 = 0, X_2 = 0, X_3 = 1, \dots$

$\sum_{i=1}^{43} X_i = 6$

$X = X_1 + \dots + X_{43} \sim Bin(43, \vartheta)$

$P(X=6) = \binom{43}{6}\vartheta^6(1-\vartheta)^{37} = f(\vartheta)$



odkud je pomocí max. věroh.

X_i = i-tý člověk, kterého potkám, je levák

náh. člověk z ČR

= i-tý člověk na přednášce je levák

$f'(\vartheta) = c \cdot 6\vartheta^5(1-\vartheta)^{37} - c \cdot 37\vartheta^6(1-\vartheta)^{36} = c\,\vartheta^6(1-\vartheta)^{37}\left(\frac{6}{\vartheta} - \frac{37}{1-\vartheta}\right) = 0 \Rightarrow \vartheta = \frac{6}{43}$

Přehled

Intervalové odhady (interval estimation)

- 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* $X = (X_1, \ldots, X_n)$. *Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti* $1 - \alpha$ *(*$1 - \alpha$ *confidence interval), pokud*

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