NMAI059 Pravděpodobnost a statistika 1
10. přednáška

Robert Šámal

Přehled

Statistika – model situace

Statistika – bodové odhady

Statistika – intervalové odhady

Náhodný výběr

- bez vracení
  $\Omega = \{$všechny n-tice obyvatel ČR$\}$
  Pro $\omega = (\omega_1, \dots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák).
- s vracením
  $\Omega = \{$všechny n-tice obyvatel ČR, mohou se opakovat$\}$
  Pro $\omega = (\omega_1, \dots, \omega_n)$ 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. $T = T(X_1, \dots, X_n)$.

Přehled

Statistika – model situace

Statistika – bodové odhady

Statistika – intervalové odhady

Výběrový průměr a rozptyl

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

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

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

Odhad

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

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. $\widehat{S}_n$ *je konzistentní nestranný odhad* μ

Metoda momentů

- $m_r(\vartheta) := \mathbb{E}\big(X^r\big)$ 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

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

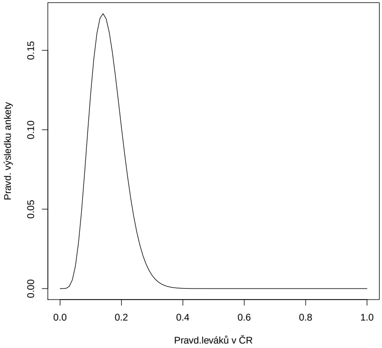
Metoda maximální věrohodnosti (maximal likelihood, ML)

- **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



0.00
0.05
0.10
0.15
Pravd. výsledku ankety
0.0
0.2
0.4
0.6
0.8
1.0
Pravd.leváků v ČR

Přehled

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

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

Intervalové odhady normální náhodné veličiny

Věta

$X_1, \dots, X_n$ je náhodný výběr z $N(\vartheta, \sigma^2)$.
σ známe, ϑ chceme určit, $\alpha \in (0,1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. $\hat{\Theta}_n = \bar{X}_n$.

$$C_n := [\hat{\Theta}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\Theta}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.

Důkaz.

Intervalové odhady pomocí CLV

Věta

$X_1, \dots, X_n$ je náhodný výběr z rozdělení se střední hodnotou ϑ, rozptylem σ^2.
σ známe, ϑ chceme určit, $\alpha \in (0, 1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

$$C_n := [\hat{\Theta}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\Theta}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta)$ se limitně blíží $1 - \alpha$.

Studentovo rozdělení

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

- Nechť $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- Pak $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- *Studentovo t-rozdělení s $n-1$ stupni volnosti* je rozdělení n.v.
$$\frac{\bar{X}_n - \mu}{\widehat{S}_n/\sqrt{n}}$$
- Distribuční funkce Ψ_{n-1} (v tabulkách ...)

Int. odhady normální n.v. pomocí Studentova t

Věta

$X_1, \dots, X_n$ *je náhodný výběr z* $N(\vartheta, \sigma^2)$.
ϑ *chceme určit,* σ *neznáme;* $\alpha \in (0,1)$. *Nechť*
$\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2$. $\hat{\Theta}_n = \bar{X}_n$, $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.