

# NMAI059 Pravděpodobnost a statistika 1

## 11. přednáška

Robert Šámal

# Přehled

Statistika – bodové odhady (point estimation)

Statistika – intervalové odhady

Testování hypotéz

# Výběrový průměr a rozptyl

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

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

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

# Metoda maximální věrohodnosti (maximal likelihood, ML)

## ▶ Metoda MV (ML):

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

▶ definujeme také  $\ell(x; \vartheta) = \log(L(x; \vartheta))$

▶ díky nezávislosti je

$$L(x; \vartheta) = L(x_1; \vartheta) \dots L(x_n; \vartheta)$$

$$\ell(x; \vartheta) = \ell(x_1; \vartheta) + \dots + \ell(x_n; \vartheta)$$

Bin(20,p)	0.2	0.3	0.4	0.45	0.5	0.55	0.6
7	0.0545	0.1643	0.1659	0.1221	0.0739	0.0366	0.0146
8	0.0222	0.1144	0.1797	0.1623	0.1201	0.0727	0.0355
9	0.0074	0.0654	0.1597	0.1771	0.1602	0.1185	0.071
10	0.002	0.0308	0.1171	0.1593	0.1762	0.1593	0.1171
11	0.0005	0.012	0.071	0.1185	0.1602	0.1771	0.1597
12	0.0001	0.0039	0.0355	0.0727	0.1201	0.1623	0.1797
13	0	0.001	0.0146	0.0366	0.0739	0.1221	0.1659
14	0	0.0002	0.0049	0.015	0.037	0.0746	0.1244

## ML – další ilustrace

# Přehled

Statistika – bodové odhady (point estimation)

Statistika – intervalové odhady

Testování hypotéz

# 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  $X = (X_1, \dots, X_n)$  z distribuce  $F_{\vartheta}$ . Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti  $1 - \alpha$  ( $1 - \alpha$  confidence interval), pokud*

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

- ▶ tohle jsou tzv. oboustranné odhady
- ▶ jednostranný odhad:  $[\hat{\Theta}^-, \infty)$  nebo  $(-\infty, \hat{\Theta}^-]$

# Intervalové odhady normální náhodné veličiny

## Věta

$X_1, \dots, X_n$  je náhodný výběr z  $N(\vartheta, \sigma^2)$ .

$\sigma$  známe,  $\vartheta$  chceme určit,  $\alpha \in (0, 1)$ .

Nechť  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ . Zvolíme  $\hat{\Theta}_n := \bar{X}_n$ .

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

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  $\vartheta$ , rozptylem  $\sigma^2$ .

$\sigma$  známe,  $\vartheta$  chceme určit,  $\alpha \in (0, 1)$ .

Nechť  $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ . Zvolíme  $\hat{\Theta}_n := \bar{X}_n$ .

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

**Pak**  $\lim_{n \rightarrow \infty} P(C_n \ni \vartheta) = 1 - \alpha$ .

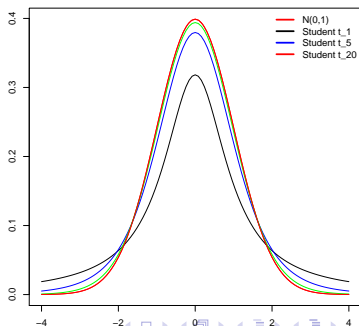
# Studentovo rozdělení

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

- ▶ Nechť  $X_1, \dots, X_n \sim N(\mu, \sigma^2)$
- ▶ Pak  $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
- ▶ Studentovo  $t$ -rozdělení s  $n - 1$  stupni volnosti je rozdělení

$$\text{n.v. } \frac{\bar{X}_n - \mu}{\hat{S}_n/\sqrt{n}}$$

- ▶ Distribuční funkci budeme značit  $\Psi_{n-1}$   
Je v tabulkách,  
v R: **pt**(x, n-1)



# Int. odhady normální n.v. pomocí Studentova t

## Věta

$X_1, \dots, X_n$  je náhodný výběr z  $N(\vartheta, \sigma^2)$ .

$\vartheta$  chceme určit,  $\sigma$  neznáme;  $\alpha \in (0, 1)$ . Necht'

$$\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2. \hat{\Theta}_n = \bar{X}_n, \hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

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

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

# Přehled

Statistika – bodové odhady (point estimation)

Statistika – intervalové odhady

Testování hypotéz

# Úvod do testování hypotéz

- ▶ Je naše mince spravedlivá?
  - ▶ Je naše kostka spravedlivá?
  - ▶ Má vylepšený program kratší dobu běhu než původní?
  - ▶ Je léčba nemoci metodou X dobrá? (Lepší než placebo, lepší než metoda Y, ...)
  - ▶ Jsou leváci lepší boxeři?
- 
- ▶ dvě hypotézy:  $H_0$ ,  $H_1$
  - ▶  $H_0$  – nulová hypotéza (*null hypothesis*) – značí defaultní, konzervativní model (léčba, mince je spravedlivá)
  - ▶  $H_1$  – alternativní hypotéza (*alternative hypothesis*) – značí alternativní model „pozoruhodnost“

# Testování hypotéz – ilustrace

- ▶ Chceme testovat, zda je mince spravedlivá.
- ▶ Hodíme  $n$ -krát mincí, orel padne  $S$ -krát.
- ▶ Pokud je  $|S - n/2|$  moc velké, tak mince není spravedlivá.

# Testování hypotéz – ilustrace

- ▶ Chceme testovat, zda je mince spravedlivá.
- ▶  $H_0$ : je spravedlivá
- ▶  $H_1$ : není spravedlivá („Vědci objevili, že v kasinu byla použita falešná mince.“ )
- ▶ Výsledky: zamítneme  $H_0$ /nezamítneme  $H_0$
- ▶ Chyba 1. druhu: chybné zamítnutí. Zamítneme  $H_0$ , i když platí. Trapas.
- ▶ Chyba 2. druhu: chybné přijetí. Nezamítneme  $H_0$ , ale ona neplatí. Promarněná příležitost.
- ▶ Potřebujeme určit  $k$  takové, že budeme zamítat  $H_0$  pokud  $|S - n/2| > k$ .

# Testování hypotéz – obecný postup

- ▶ Vybereme vhodný statistický model.
- ▶ Volíme *hladinu významnosti (significance level)*  $\alpha$ : pravd. chybného zamítnutí  $H_0$ . Typicky  $\alpha = 0.05$ .
- ▶ Určíme *testovou statistiku*  $S = h(X_1, \dots, X_n)$ , kterou budeme určovat z naměřených dat.
- ▶ Určíme *kritický obor (rejection region)* – množinu  $W$ .
- ▶ Naměříme hodnoty  $x_1, \dots, x_n$  náh. veličin  $X_1, \dots, X_n$ .
- ▶ Rozhodovací pravidlo: zamítneme  $H_0$  pokud  $h(x_1, \dots, x_n) \in W$ .
- ▶  $\alpha = P(h(X) \in W; H_0)$
- ▶  $\beta = P(h(X) \notin W; H_1) \dots$  *síla testu*
- ▶ často  $\alpha$  nevolíme předem, ale spočítáme *p-hodnotu*: minimální  $\alpha$ , pro které bychom  $H_0$  zamítlí.



## Testování hypotéz – příklad

- ▶  $X_1, \dots, X_n$  náhodný výběr z  $N(\vartheta, \sigma^2)$
- ▶  $\sigma^2$  známe
- ▶  $H_0 : \vartheta = 0$        $H_1 : \vartheta \neq 0$

## Testování hypotéz – příklad

- ▶  $X_1, \dots, X_{n_1}$  náhodný výběr z  $Ber(\vartheta_X)$
- ▶  $Y_1, \dots, Y_{n_2}$  náhodný výběr z  $Ber(\vartheta_Y)$
- ▶  $H_0 : \vartheta_X = \vartheta_Y$        $H_1 : \vartheta_X \neq \vartheta_Y$

# *p*-hacking

- ▶ napřed získáme data, pak v nich hledáme zajímavosti
- ▶ když máme dost dat, tak tam nějaké budou „shodou okolností“
- ▶ *reprodukovatelnost* – po explorační analýze dat uděláme nezávislý sběr dat a ten analyzujeme konfirmačně
- ▶ nebo dopředu náhodně rozdělíme data na část pro tvorbu hypotéz a část pro jejich potvrzení . . . jednoduchý případ křížové validace (cross validation)

## Letters in winning word of Scripps National Spelling Bee

correlates with

## Number of people killed by venomous spiders

