NMAI059 Pravděpodobnost a statistika 1
11. přednáška

Robert Šámal

Přehled

Statistika – bodové odhady (point estimation)

Statistika – intervalové odhady

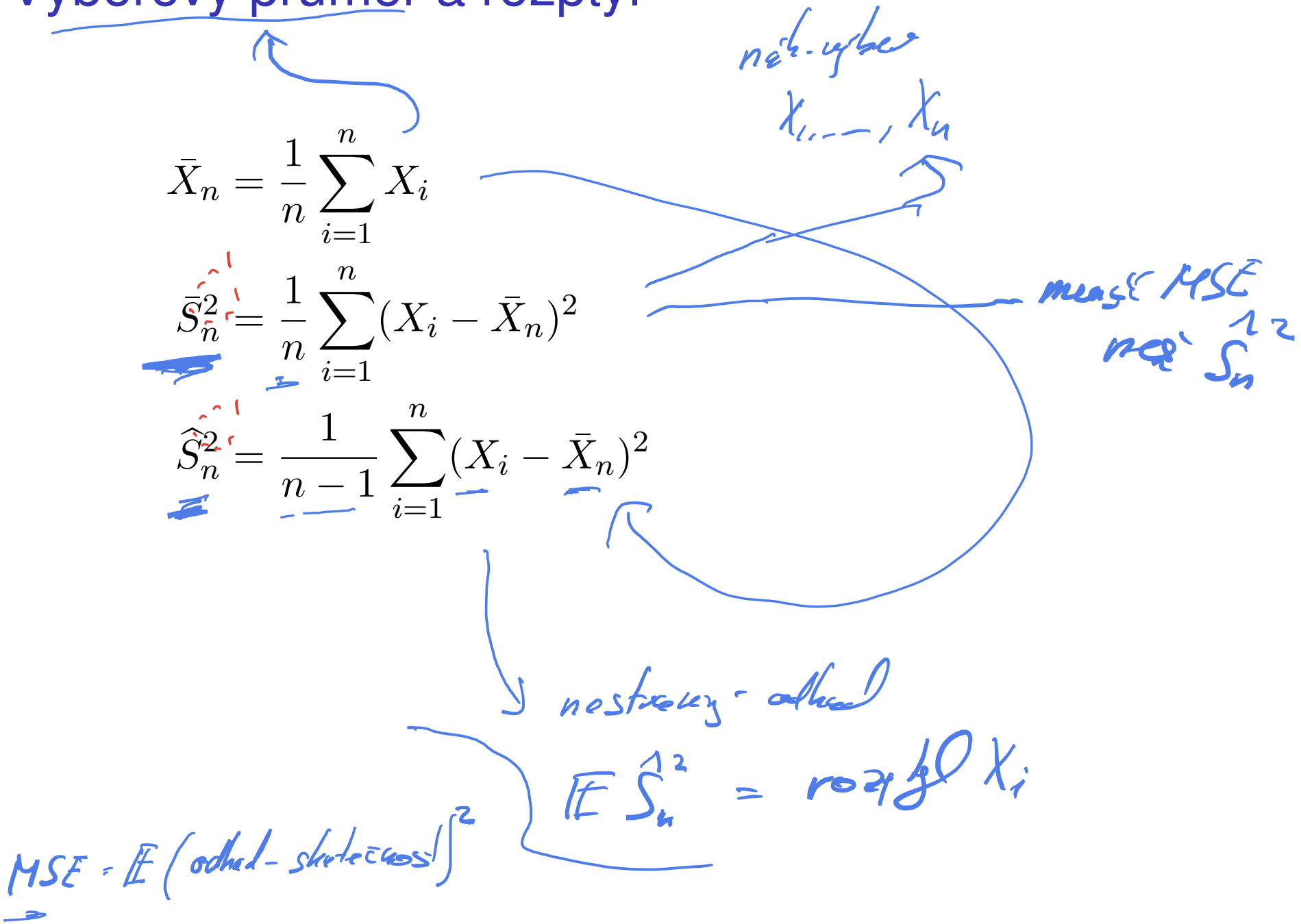
Testování hypotéz

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 ϑ, 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

$X_1, \dots, X_n \sim N(\mu, \sigma^2)$ $\quad \vartheta = (\mu, \sigma)$

naměříme v cíli $x_1 = 1{,}5$, $x_2 = 2{,}7$, $x_3 = \dots$ $\dots$ x_n

$x = (x_1, \dots, x_n)$

chceme: μ, σ

$$L(x_1; \vartheta) = f(x_1; \vartheta) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\left(\frac{x-\mu}{\sigma}\right)^2/2}$$

$$\ell(x_1; \vartheta) = -\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2 - \log\sigma - \log\sqrt{2\pi}$$

vstup — naměříme — statistika $\pi = 3{,}14$

$$\ell(x; \vartheta) = \sum_{i=1}^{n} \ell(x_i; \vartheta) = -\frac{1}{2}\sum\left(\frac{x_i-\mu}{\sigma}\right)^2 - n\log\sigma - n\log\sqrt{2\pi}$$

cíl: μ, σ

odhad $\ell(x; \vartheta)$ bude max.

→ zderivuj: podle σ } položíme $= 0$ ⇒ 2 rovnice pro

→ — " — μ 2 neznámé, μ a σ

$$\frac{\partial \ell}{\partial \mu} = +\frac{1}{2}\sum_{i=1}^{n} 2\left(\frac{x_i-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma^2}\sum(x_i-\mu) = 0 \Rightarrow \sum x_i = n\cdot\mu$$

→ odhad $\mu = \bar{x}_n$

$$\frac{\partial \ell}{\partial \sigma} = +\frac{1}{2}\sum \frac{(x_i-\mu)^2}{\sigma^3}\cdot(2) - n\cdot\frac{1}{\sigma} = 0$$

$$\sum(x_i-\mu)^2 = n\sigma^2 \Rightarrow$$

$$\hat{\mu} = \bar{x}_n$$

$$\sigma^2 = \frac{1}{n}\sum(x_i-\mu)^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i-\bar{x}_n)^2$$

Met. mom. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

} porovnání 1, 2, ..., r-tý moment < výběrový / teoretický

$$\mathbb{E}X_i = \mu = m_1(\vartheta)$$

$$\mathbb{E}X_i^2 = \underbrace{\mathbb{E}X_i^2 - (\mathbb{E}X_i)^2}_{var(X_i) = \sigma^2} + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2 = m_2(\vartheta)$$

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

$$\widehat{m_2(\vartheta)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = m_2(\vartheta) = \sigma^2 + \mu^2$$

$U(0, \vartheta)$

$$\hat{\mu} = \overline{X}_n$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \left(\sum X_i\right)^2 = \frac{1}{n}\sum (X_i - \overline{X}_n)^2$$

Přehled

Intervalové odhady (interval estimation)

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

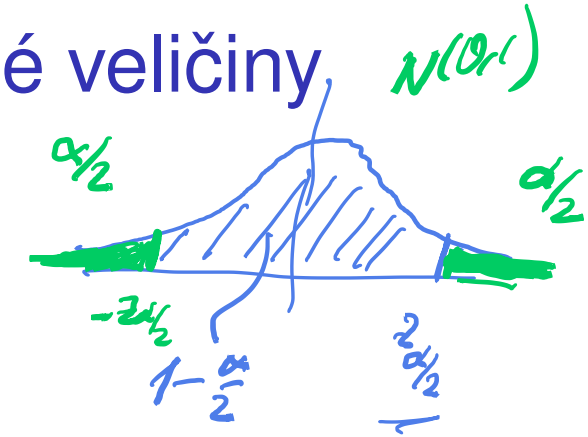
Definice

Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ jsou n.v. které závisí na náhodném výběru $X = (X_1, \ldots, X_n)$ z distribuce F_ϑ. 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

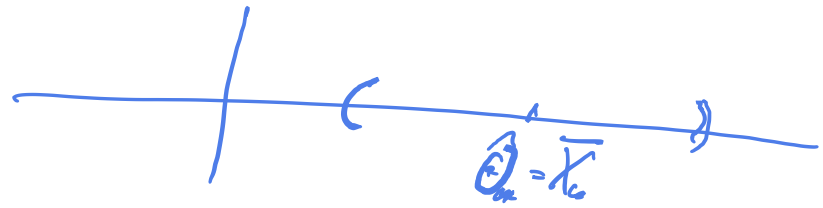


Věta

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

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

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



Důkaz.

$C_n \ni \vartheta \iff |\hat{\Theta}_n - \vartheta| \le z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

$\iff \left| \frac{\hat{\Theta}_n - \vartheta}{\sigma/\sqrt{n}} \right| \le z_{\alpha/2}$

$Z \sim N(0,1)$

$P(C_n \ni \vartheta) = P(|Z| \le z_{\alpha/2})$
$= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})$
$= (1 - \alpha/2) - (+\alpha/2) = 1 - \alpha.$

Intervalové odhady pomocí CLV

Věta

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

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

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

Dk $Z_n = \frac{\bar{\Theta}_n - \vartheta}{\sigma/\sqrt{n}}$ je podíl. $N(0,1)$ dle CLV

$Z_n \xrightarrow{d} N(0,1)$

$P(C_n \ni \vartheta) = P(-z_{\alpha/2} \le Z \le z_{\alpha/2})$
$= F_{Z_n}(z_{\alpha/2}) - F_{Z_n}(-z_{\alpha/2}) \to \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2})$
$= 1 - \alpha$

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

$S_n = \sqrt{S_n^2}$ … výběr. směr. odch.

- Nechť $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- Pak $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

víme z minula

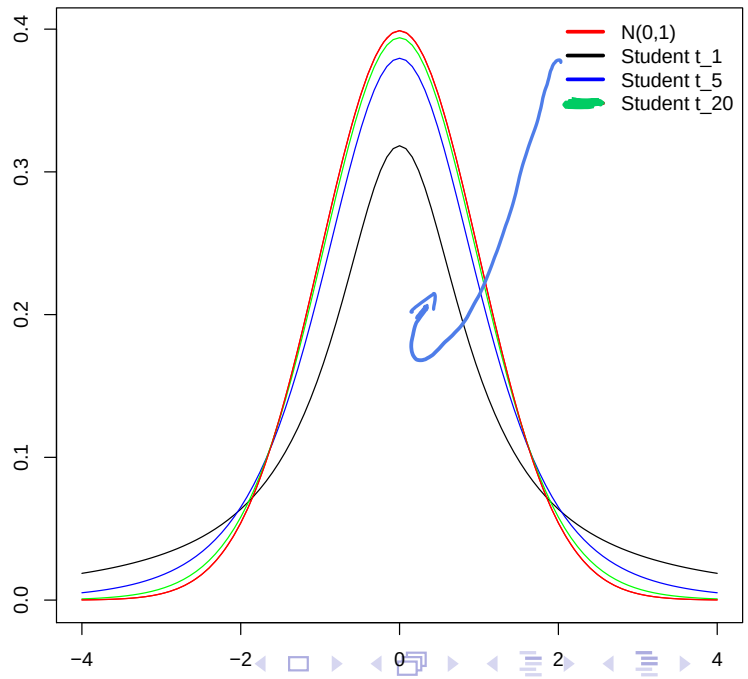
potřebujeme znát σ

- *Studentovo t-rozdělení s $n-1$ stupni volnosti* je rozdělení n.v. $\dfrac{\bar{X}_n - \mu}{\widehat{S}_n/\sqrt{n}}$

$\xrightarrow{d} N(0,1)$

- Distribuční funkci budeme značit Ψ_{n-1}
  Je v tabulkách,
  v R: **pt**(x,n−1)

Důležité — je to stejné rozdělení pro všechna μ, σ



Int. odhady normální n.v. pomocí Studentova t

Věta

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

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

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

Dk $P(C_n \ni \vartheta) = P(|Z| \le z_{\alpha/2}) = \Psi_{n-1}(z_{\alpha/2}) - \Psi_{n-1}(-z_{\alpha/2})$

$= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$

$Z = \frac{\bar{X}_n - \vartheta}{\hat{S}_n/\sqrt{n}}$ — St. t-rozd. s $(n-1)$ st-vole.

$\Phi^{-1}(0.975) = 1.96$

$\Psi_4^{-1}(0.975) = 12.7$

$\Psi_5^{-1}(0.975) = 2.57$

$\Psi_{20}^{-1}(0.975) = 2.09$

Přehled

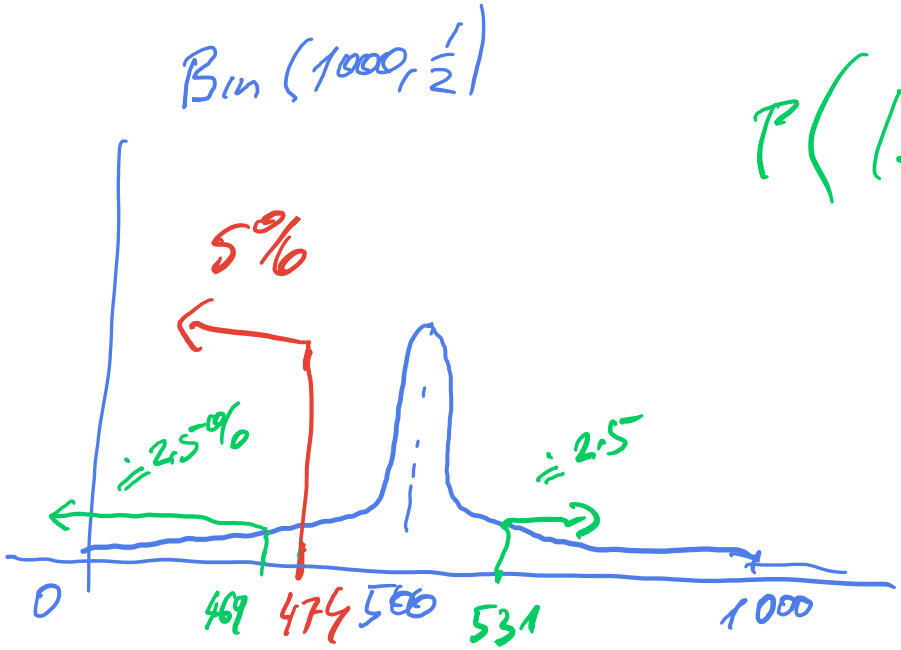
Úvod do testování hypotéz

- 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.



$P\left(\left|S - \frac{n}{2}\right| > 31\right) \doteq 5\%$

Nezamítáme, že je mince nespravedlivé.

(nezamítáme H_0) na hladině významnosti 5%

psychol. ~ 5%

fyzika .... $< 3 \cdot 10^{-3}$ evidence

$< 3 \cdot 10^{-7}$ objev.

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)* α: pravd. chybného zamítnutí H_0. Typicky $\alpha = 0.05$.
- Určíme *testovou statistiku* $S = h(X_1, \ldots, X_n)$, kterou budeme určovat z naměřených dat.
- Určíme *kritický obor (rejection region)* – množinu W.
- Naměříme hodnoty $x_1, \ldots, x_n$ náh. veličin $X_1, \ldots, X_n$.
- Rozhodovací pravidlo: zamítneme H_0 pokud $h(x_1, \ldots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ … *síla testu*

- často α nevolíme předem, ale spočítáme p*-hodnotu:* minimální α, pro které bychom H_0 zamítli.

Testování hypotéz – příklad

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

Testování hypotéz – příklad

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

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)