

# M5VM05 Statistické modelování

## 2. Základní pojmy matematické statistiky

Jan Kolářek (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

podzim 2013



V teorii pravděpodobnosti se předpokládá, že

- je známý **pravděpodobnostní prostor**  $(\Omega, \mathcal{A}, P)$
- a že také známe **rozdělení pravděpodobnosti** náhodných veličin (resp. náhodných vektorů), které na tomto pravděpodobnostním prostoru uvažujeme.

V matematické statistice však

- máme k dispozici výsledky  $n$  nezávislých pozorování hodnot sledované náhodné veličiny  $X$ , které se ve statistice říká *statistický znak*, tj. máme

$$x_1 = X(\omega_1), \dots, x_n = X(\omega_n), \omega_1, \dots, \omega_n \in \Omega$$

- a na základě těchto pozorování chceme učinit výpověď o rozdělení zkoumané náhodné veličiny.

# Náhodný výběr

## Definice 1

Náhodný vektor  $\mathbf{X} = (X_1, \dots, X_n)'$  nazýváme **náhodným výběrem z rozdělení pravděpodobnosti**  $P$ , pokud

- (i)  $X_1, \dots, X_n$  jsou nezávislé náhodné veličiny,
- (ii)  $X_1, \dots, X_n$  mají stejné rozdělení pravděpodobnosti  $P$ .

Číslo  $n$  nazýváme **rozsah náhodného výběru**. Libovolný bod  $\mathbf{x} = (x_1, \dots, x_n)'$ , kde  $x_i$  je realizace náhodné veličiny  $X_i$  ( $i = 1, \dots, n$ ), budeme nazývat **realizací náhodného výběru**  $\mathbf{X} = (X_1, \dots, X_n)'$ .

Nechť náhodný výběr  $\mathbf{X} = (X_1, \dots, X_n)'$  je z rozdělení, které je dáno distribuční funkcí  $F(x, \boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \in \Theta$ . Zkráceně budeme značit:

$$\mathbb{P}\{X_1, \dots, X_n\} \simeq F(x; \boldsymbol{\theta}).$$

Cílem teorie odhadu je **na základě náhodného výběru** odhadnout

- rozdělení pravděpodobnosti,
- popřípadě některé parametry tohoto rozdělení,
- anebo nalézt odhad nějaké funkce parametrů  $\boldsymbol{\theta}$ , tj.  $\gamma(\boldsymbol{\theta})$ .

## Definice 2

Libovolnou náhodnou veličinou  $T_n$ , která vznikne jako funkce náhodného výběru  $\mathbf{X} = (X_1, \dots, X_n)'$ , budeme nazývat **statistikou**, tj.  $T_n = T(X_1, \dots, X_n)'$ .

## Definice 3

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr rozsahu  $n$  z rozdělení s distribuční funkcí  $F(x; \theta)$ ,  $\theta \in \Theta$ . Potom statistika

$$\bar{X}_n = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{se nazývá} \quad \text{výběrový průměr}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{výběrový rozptyl}$$

$$S = \sqrt{S^2} \quad \text{výběrová směrodatná odchylka}$$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x)}(X_i) \quad \text{výběrová (empirická) distribuční fce}$$

# Bodové odhady

**Bodovým odhadem** parametrické funkce  $\gamma(\boldsymbol{\theta})$  budeme rozumět nějakou statistiku  $T_n = T(X_1, \dots, X_n)'$ , která bude pro různé náhodné výběry kolísat kolem  $\gamma(\boldsymbol{\theta})$ .

## Definice 4

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr z rozdělení pravděpodobnosti  $P_{\boldsymbol{\theta}}$ , kde  $\boldsymbol{\theta}$  je vektor neznámých parametrů. Nechť  $\gamma(\boldsymbol{\theta})$  je daná parametrická funkce.

Řekneme, že statistika  $T_n = T(X_1, \dots, X_n)'$  je odhadem

**nestranným** (nevychýleným) pokud pro  $\forall \boldsymbol{\theta} \in \Theta$  platí  $E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$ .

**kladně vychýleným**  $E_{\boldsymbol{\theta}} T_n > \gamma(\boldsymbol{\theta})$ .

**záporně vychýleným**  $E_{\boldsymbol{\theta}} T_n < \gamma(\boldsymbol{\theta})$ .

**asymptoticky nestranným**  $\lim_{n \rightarrow \infty} E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$ .

**(slabě) konzistentním** pokud pro  $\forall \varepsilon > 0$  platí

$$\lim_{n \rightarrow \infty} P_{\boldsymbol{\theta}}(|T_n - \gamma(\boldsymbol{\theta})| > \varepsilon) = 0, \text{ tj. } T_n \xrightarrow{P_{\boldsymbol{\theta}}} \gamma(\boldsymbol{\theta})$$

## Poznámka 5

Vlastnost **nestrannosti** (tj. nevychýlenosti) ještě neposkytuje záruku dobrého odhadu, pouze vylučuje systematickou chybu.

## Poznámka 6

Používání **konzistentních** odhadů zaručuje

- malou pravděpodobnost velké chyby v odhadu parametru, pokud rozsah výběru dostatečně roste;
- volbou dostatečně velkého počtu pozorování lze učinit chybu odhadu libovolně malou.

## Věta 7

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr z rozdělení, které má střední hodnotu  $\mu(\boldsymbol{\theta})$  pro  $\forall \boldsymbol{\theta} \in \Theta$ . Pak **výběrový průměr je nestranným odhadem střední hodnoty**, tj.

$$E_{\boldsymbol{\theta}} \bar{X} = \mu(\boldsymbol{\theta}).$$

## Věta 8

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr z rozdělení, které má rozptyl  $\sigma^2(\boldsymbol{\theta})$  pro  $\forall \boldsymbol{\theta} \in \Theta$ . Pak **výběrový rozptyl je nestranným odhadem rozptylu**, tj.

$$E_{\boldsymbol{\theta}} S^2 = \sigma^2(\boldsymbol{\theta}).$$

## Věta 9

*Nechť statistika  $T_n = T(X_1, \dots, X_n)'$  je nestranný nebo asymptoticky nestranný odhad parametrické funkce  $\gamma(\boldsymbol{\theta})$  a platí*

$$\lim_{n \rightarrow \infty} D_{\boldsymbol{\theta}} T_n = 0.$$

*Pak je statistika  $T_n = T(X_1, \dots, X_n)$  konzistentním odhadem parametrické funkce  $\gamma(\boldsymbol{\theta})$ .*



## Důsledek 10

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr z rozdělení, které má pro  $\forall \boldsymbol{\theta} \in \Theta$  střední hodnotu  $\mu(\boldsymbol{\theta})$  a rozptyl  $\sigma^2(\boldsymbol{\theta})$ , tj.

$$\mathbb{L}\{X_1, \dots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

Potom je-li  $\mu(\boldsymbol{\theta}) < \infty$ , pak **výběrový průměr  $\bar{X}$  je slabě konzistentním odhadem  $\mu(\boldsymbol{\theta})$ .**

## Důsledek 11

Nechť  $\mathbf{X} = (X_1, \dots, X_n)'$  je náhodný výběr z rozdělení, které má pro  $\forall \boldsymbol{\theta} \in \Theta$  střední hodnotu  $\mu(\boldsymbol{\theta})$  a rozptyl  $\sigma^2(\boldsymbol{\theta})$ , tj.

$$\mathbb{L}\{X_1, \dots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

Potom je-li  $\sigma^2(\boldsymbol{\theta}) < \infty$ , pak **výběrový rozptyl  $S^2$  je slabě konzistentním odhadem  $\sigma^2(\boldsymbol{\theta})$ .**

# Více nestranných odhadů

## Definice 12

Nechť  $T_n$  je nestranný odhad parametrické funkce  $\gamma(\boldsymbol{\theta})$  a pro všechna  $\boldsymbol{\theta} \in \Theta$  platí

$$D_{\boldsymbol{\theta}}T_n \leq D_{\boldsymbol{\theta}}T_n^*,$$

kde  $T_n^*$  je libovolný nestranný odhad parametru  $\gamma(\boldsymbol{\theta})$ . Potom odhad  $T_n$  nazveme (**rovnoměrně**) **nejlepším nestranným odhadem** parametrické funkce  $\gamma(\boldsymbol{\theta})$ .

## Příklad 13

Nalezněte nejlepší nestranný lineární odhad střední hodnoty  $\mu(\boldsymbol{\theta})$ .

Odhady, jimiž jsme se doposud zabývali, se někdy nazývají **bodové odhady** parametrické funkce  $\gamma(\theta)$ . Je tomu tak proto, že pro danou realizaci náhodného výběru  $x_1, \dots, x_n$  představuje odhad daný statistikou  $T_n(x_1, \dots, x_n)$  **jediné číslo (bod)**, které je v jistém smyslu přiblížením ke skutečné hodnotě parametrické funkce  $\gamma(\theta)$ .

Úlohu odhadu však lze formulovat i jiným způsobem. Jde o to, sestrojít na základě daného náhodného výběru takový interval, jehož **hranice** jsou **statistiky**, a který se s dostatečně velkou přesností pokryje skutečnou hodnotu parametrické funkce  $\gamma(\theta)$ . V tomto případě mluvíme o **intervalovém odhadu** parametrické funkce  $\gamma(\theta)$ .

## Definice 14

Nechť  $\perp\!\!\!\perp\{X_1, \dots, X_n\} \simeq F(x; \theta)$  je náhodný výběr rozsahu  $n$  z rozdělení o distribuční funkci  $F(x; \theta)$ ,  $\theta \in \Theta$ . Dále mějme parametrickou funkci  $\gamma(\theta)$ ,  $\alpha \in (0, 1)$  a statistiky  $D = D(X_1, \dots, X_n)$  a  $H = H(X_1, \dots, X_n)$ . Potom intervaly  $\langle D, H \rangle$  nazveme  $100(1 - \alpha)$  % **intervalem spolehlivosti** pro parametrickou funkci  $\gamma(\theta)$  jestliže

$$P_{\theta}(D(X_1, \dots, X_n) \leq \gamma(\theta) \leq H(X_1, \dots, X_n)) = 1 - \alpha$$

Jestliže

$$P_{\theta}(D(X_1, \dots, X_n) \leq \gamma(\theta)) = 1 - \alpha,$$

pak statistiku  $D = D(X_1, \dots, X_n)$  nazýváme **dolním odhadem parametrické funkce**  $\gamma(\theta)$  se spolehlivostí  $1 - \alpha$  (nebo s rizikem  $\alpha$ ).

Jestliže

$$P_{\theta}(\gamma(\theta) \leq H(X_1, \dots, X_n)) = 1 - \alpha$$

pak statistiku  $H = H(X_1, \dots, X_n)$  nazýváme **horním odhadem parametrické funkce**  $\gamma(\theta)$  se spolehlivostí  $1 - \alpha$  (nebo s rizikem  $\alpha$ ).

# Konstrukce intervalových odhadů

- 1 Najdeme nějakou tzv. PIVOTOVOU STATISTIKU, tj. funkci  $h$  náhodného výběru  $\mathbf{X} = (X_1, \dots, X_n)'$  a parametrické funkce  $\gamma(\theta)$ , tedy náhodnou veličinu

$$h(\mathbf{X}, \gamma(\theta)),$$

tak aby její rozdělení již **nezáviselo na parametru**  $\theta$ .

- 2 Necht'  $q_{\alpha/2}$  a  $q_{1-\alpha/2}$  jsou kvantily rozdělení statistiky

$$h(\mathbf{X}, \gamma(\theta)).$$

Pak pro všechna  $\theta$  platí

$$P_{\theta}(q_{\alpha/2} < h(\mathbf{X}, \gamma(\theta)) \leq q_{1-\alpha/2}) = 1 - \alpha$$

- 3 Jestliže lze nerovnosti v závorce převést ekvivalentními úpravami na tvar, kde mezi nerovnostmi stojí jen  $\gamma(\theta)$ , pak jsme sestrojili intervalový odhad

$$D_n(\mathbf{X}) \leq \gamma(\theta) \leq H_n(\mathbf{X})$$

o spolehlivosti  $1 - \alpha$ .

Tedy, je-li  $h(\mathbf{X}, \gamma(\theta))$  **ryze monotónní funkce**, pak existuje inverzní funkce

$$h^{-1}(h(\mathbf{X}, \gamma(\theta))) = \gamma(\theta).$$

**(a)** Pokud je  $h(\mathbf{X}, \gamma(\theta))$  **rostoucí funkce**, pak platí

$$P_{\theta}(h^{-1}(q_{\alpha/2}) \leq \gamma(\theta) \leq h^{-1}(q_{1-\alpha/2})) = 1 - \alpha.$$

**(b)** Pokud je  $h(\mathbf{X}, \gamma(\theta))$  **klesající funkce**, pak platí

$$P_{\theta}(h^{-1}(q_{1-\alpha/2}) \leq \gamma(\theta) \leq h^{-1}(q_{\alpha/2})) = 1 - \alpha.$$

# Kvantily některých důležitých rozdělení

---

$\Phi$	distribuční funkce standardizovaného normálního rozdělení
$G_n$	distribuční funkce rozdělení $\chi^2$ o $n$ stupních volnosti
$H_n$	distribuční funkce Studentova rozdělení o $n$ stupních volnosti
$Q_{n,m}$	distribuční funkce Fisherova–Snedecorova rozdělení o $n$ a $m$ stupních volnosti

---

$u_\alpha$	kvantily standardizovaného normálního rozdělení
$\chi_\alpha^2(\nu)$	kvantily rozdělení $\chi^2$ o $\nu$ stupních volnosti
$t_\alpha(\nu)$	kvantily Studentova rozdělení o $\nu$ stupních volnosti
$F_\alpha(\nu_1, \nu_2)$	kvantily Fisherova–Snedecorova rozdělení o $\nu_1$ a $\nu_2$ stupních volnosti

---

Je-li distribuční funkce  $F$  absolutně spojitá a ryze monotónní a je-li příslušná hustota  $f$  **sudá funkce**, pak platí

$$F(x) = 1 - F(-x) \quad x \in \mathbb{R}$$

a odtud

$$x_\alpha = -x_{1-\alpha} \quad \alpha \in (0, 1),$$

což speciálně platí pro **normální** a **Studentovo rozdělení**.



**Normální rozdělení** s hustotou

$$X \sim N(\mu, \sigma^2) \sim f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad x \in \mathbb{R}$$

má střední hodnotu  $EX = \mu$  a rozptyl  $DX = \sigma^2$ .

# Vlastnosti

Nechť  $k, n \in \mathbb{N}$ ,  $\nu, \nu_1, \nu_2, \dots, \nu_k \in \mathbb{N}$ ,  $b_0, b_1, \dots, b_n \in \mathbb{R}$ ,

$\exists i \in \{1, \dots, n\} : b_i \neq 0$

$$\perp \{X_1, \dots, X_n\} \wedge X_i \sim N(\mu_i, \sigma_i^2) \Rightarrow b_0 + \sum_{i=1}^n b_i X_i \sim N\left(b_0 + \sum_{i=1}^n b_i \mu_i, \sum_{i=1}^n b_i^2 \sigma_i^2\right)$$
$$X \sim N(\mu, \sigma^2) \Rightarrow U = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$\chi^2$  rozdělení:

$$\perp \{U_1, \dots, U_\nu\} \simeq N(0, 1) \Rightarrow K = U_1^2 + \dots + U_\nu^2 \sim \chi^2(\nu)$$

$$\perp \{K_1 \sim \chi^2(\nu_1), \dots, K_k \sim \chi^2(\nu_k)\} \Rightarrow K = K_1 + \dots + K_k \sim \chi^2(\nu_1 + \dots + \nu_k)$$

Studentovo t-rozdělení:

$$U \sim N(0, 1) \perp K \sim \chi^2(\nu) \Rightarrow T = \frac{U}{\sqrt{\frac{K}{\nu}}} \sim t(\nu)$$

Fisherovo F-rozdělení:

$$K_1 \sim \chi^2(\nu_1) \perp K_2 \sim \chi^2(\nu_2) \Rightarrow F = \frac{K_1/\nu_1}{K_2/\nu_2} \sim F(\nu_1, \nu_2)$$

## Věta 15

Mějme  $\perp\{X_1, \dots, X_n\} \simeq N(\mu, \sigma^2)$  a výběrový průměr  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  a výběrový

rozptyl  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . Pak platí

(1) Výběrový průměr  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

(2) Statistika  $U = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1)$

(3) Statistika  $K = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$

(4) Statistika  $T = \frac{\bar{X} - \mu}{S} \sqrt{n} \sim t(n-1)$

# Pivotové statistiky

Statistiky  $\boxed{U}$ ,  $\boxed{K}$  a  $\boxed{T}$  se nazývají PIVOTOVÉ STATISTIKY, přičemž

$U = \frac{\bar{X} - \mu}{\sigma} \sqrt{n}$	je pivotovou statistikou pro	$\mu$	při známém $\sigma$
	neznámý parametr		
$K = \frac{n-1}{\sigma^2} S^2$	- "-	$\sigma^2$	
$T = \frac{\bar{X} - \mu}{S} \sqrt{n}$	- "-	$\mu$	při neznámém $\sigma$

# Interval spolehlivosti pro střední hodnotu při známém rozptylu

## Důsledek 16

Mějme  $\perp\{X_1, \dots, X_n\} \simeq N(\mu, \sigma^2)$ , kde  $\mu$  je **neznámý parametr** a  $\sigma^2 \in \mathbb{R}$  je **známé** reálné číslo. Pak

$\langle \bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \rangle$  - je  $100(1 - \alpha)\%$  interval spolehlivosti pro střední hodnotu  $\mu$  při známém  $\sigma^2$

$\bar{X} - u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$  - je **dolní odhad** střední hodnoty  $\mu$  při známém  $\sigma^2$  se spolehlivostí  $1 - \alpha$

$\bar{X} + u_{1-\alpha} \frac{\sigma}{\sqrt{n}}$  - je **horní odhad** střední hodnoty  $\mu$  při známém  $\sigma^2$  se spolehlivostí  $1 - \alpha$

Za pivotovou statistiku zvolíme statistiku

$$U = U_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N(0, 1).$$

Počítejme

$$\begin{aligned} 1 - \alpha &= P(u_{\frac{\alpha}{2}} \leq U \leq u_{1-\frac{\alpha}{2}}) \\ &= P(-u_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma} \sqrt{n} \leq u_{1-\frac{\alpha}{2}}) \\ &= P(\bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

## Příklad 17

Rychlost letadla byla určována v 5 zkouškách a z jejich výsledků byl vypočten odhad  $\bar{x} = 870,3$  m/s. Najděte 95% interval spolehlivosti pro  $\mu$ , je-li známo, že rozptýlení rychlosti letadla se řídí normálním rozdělením se směrodatnou odchylkou  $\sigma = 2,1$  m/s.

### Řešení

$$\bar{X} \pm u_{0,975} \frac{\sigma}{\sqrt{n}} = 870,3 \pm 1,959964 \frac{2,1}{\sqrt{5}} = (868,46; 872,14).$$

# Interval spolehlivosti pro střední hodnotu při neznámém rozptylu

## Důsledek 18

Mějme  $\mathbb{1}\{X_1, \dots, X_n\} \simeq N(\mu, \sigma^2)$ , kde  $\mu$  a  $\sigma^2$  jsou **neznámé parametry**. Pak pro střední hodnotu  $\boxed{\mu}$

$\langle \bar{X} - t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}} \rangle$  - je  $100(1 - \alpha)\%$  interval spolehlivosti pro střední hodnotu  $\mu$  při neznámém  $\sigma^2$

$\bar{X} - t_{1-\alpha}(n-1) \frac{S}{\sqrt{n}}$  - je **dolní odhad** střední hodnoty  $\mu$  při neznámém  $\sigma^2$  se spolehlivostí  $1 - \alpha$

$\bar{X} + t_{1-\alpha}(n-1) \frac{S}{\sqrt{n}}$  - je **horní odhad** střední hodnoty  $\mu$  při neznámém  $\sigma^2$  se spolehlivostí  $1 - \alpha$



# Interval spolehlivosti pro rozptyl

## Důsledek 19

Mějme  $\perp\{X_1, \dots, X_n\} \simeq N(\mu, \sigma^2)$ , kde  $\mu$  a  $\sigma^2$  jsou **neznámé parametry**. Pak pro rozptyl  $\sigma^2$

$\left\langle \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right\rangle$  - je  $100(1 - \alpha)\%$  interval spolehlivosti pro rozptyl  $\sigma^2$

$\frac{(n-1)S^2}{\chi_{1-\alpha}^2(n-1)}$  - je **dolní odhad** rozptylu  $\sigma^2$  se spolehlivostí  $1 - \alpha$

$\frac{(n-1)S^2}{\chi_{\alpha}^2(n-1)}$  - je **horní odhad** rozptylu  $\sigma^2$  se spolehlivostí  $1 - \alpha$

## Příklad 20

Deset balíčků mouky pocházejících z balícího stroje mělo hmotnosti v gramech: 987, 1 001, 993, 994, 993, 1 005, 1 007, 999, 995, 1 002. Sestrojte 95% interval spolehlivosti pro střední hodnotu a rozptyl hmotnosti (předpokládejte normální rozdělení).

**Řešení** Vypočteme průměr  $\bar{x} = 997,6$  a směrodatnou odchylku  $s = 6,2397$ .  
IS pro  $\mu$ :

$$\bar{x} \pm t_{0,975}(9) \frac{s}{\sqrt{10}} = 997,6 \pm 2,26 \frac{6,2397}{\sqrt{10}} = (993,14; 1002,06).$$

IS pro  $\sigma^2$ :

$$\left( \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right) = \left( \frac{9s^2}{\chi_{0,975}^2(9)}, \frac{9s^2}{\chi_{0,025}^2(9)} \right) = (18,42; 129,76).$$

# Dva výběry

## Věta 21

Nechť  $\underline{X} = \{X_1, \dots, X_{n_1}\} \sim N(\mu_1, \sigma_1^2)$ ,  $\bar{X}$  výběrový průměr a  $S_1^2$  výběrový rozptyl. Dále necht'  $\underline{Y} = \{Y_1, \dots, Y_{n_2}\} \sim N(\mu_2, \sigma_2^2)$ ,  $\bar{Y}$  výběrový průměr a  $S_2^2$  výběrový rozptyl. Předpokládejme  $\mathbf{X} \perp \mathbf{Y}$ . Pak

(1) Statistika

$$U_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

(2) Pokud  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , pak statistika

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_{12}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sim t(n_1 + n_2 - 2), S_{12}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(3) Statistika

$$F = \frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \sim F(n_1 - 1, n_2 - 1).$$

# IS pro $\mu_1 - \mu_2$

## Důsledek 22

Nechť  $\mathbb{I}\{X_1, \dots, X_{n_1}\} \sim N(\mu_1, \sigma_1^2)$ ,  $\bar{X}$  výběrový průměr a  $S_1^2$  výběrový rozptyl. Dále necht'  $\mathbb{I}\{Y_1, \dots, Y_{n_2}\} \sim N(\mu_2, \sigma_2^2)$ ,  $\bar{Y}$  výběrový průměr a  $S_2^2$  výběrový rozptyl. Předpokládejme  $\mathbf{X} \perp \mathbf{Y}$ . Pak

- jsou-li  $\sigma_1^2$  a  $\sigma_2^2$  známé, pak  $100(1 - \alpha)\%$  IS pro  $\mu_1 - \mu_2$

$$\left\langle \bar{X} - \bar{Y} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X} - \bar{Y} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\rangle.$$

- Jestliže  $\sigma_1^2$  a  $\sigma_2^2$  neznáme a platí  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , pak  $100(1 - \alpha)\%$  IS

$$\left\langle \bar{X} - \bar{Y} - t_{1-\frac{\alpha}{2}}(n_1+n_2-2) S_{12} \sqrt{\frac{n_1+n_2}{n_1 n_2}}, \bar{X} - \bar{Y} + t_{1-\frac{\alpha}{2}}(n_1+n_2-2) S_{12} \sqrt{\frac{n_1+n_2}{n_1 n_2}} \right\rangle,$$

kde

$$S_{12}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}.$$

## Příklad 23

Bylo vylosováno 11 stejně starých selat téhož plemene, šest z nich bylo krmeno směsí A, zbývajících pět bylo krmeno směsí B. Denní přírůstky váhy selat (v dkg) byly při krmení směsí A : 62, 54, 55, 60, 53, 58, u směsi B : 52, 56, 50, 49, 51. Předpokládáme, že neznámé směrodatné odchylky si budou rovny u obou skupin. Sestrojte interval spolehlivosti pro rozdíl neznámých středních hodnot  $\mu_1 - \mu_2$  při riziku  $\alpha = 0,05$ .

**Řešení** Vypočteme průměry  $\bar{x} = 57$ ,  $\bar{y} = 51,6$  a směrodatné odchylky  $s_1 = 3,58$ ,  $s_2 = 2,7$ ,  $s_{12} = 3,22$ .

IS pro  $\mu_1 - \mu_2$ :

$$\bar{x} - \bar{y} \pm t_{0,975}(6 + 5 - 2)s_{12}\sqrt{\frac{5+6}{5 \cdot 6}} = 5,4 \pm 2,26 \cdot 3,22 \cdot \sqrt{\frac{11}{30}} = (0,99; 9,81).$$

## Důsledek 24

Nechť  $\underline{X} \{X_1, \dots, X_{n_1}\} \sim N(\mu_1, \sigma_1^2)$ ,  $\bar{X}$  výběrový průměr a  $S_1^2$  výběrový rozptyl. Dále necht'  $\underline{Y} \{Y_1, \dots, Y_{n_2}\} \sim N(\mu_2, \sigma_2^2)$ ,  $\bar{Y}$  výběrový průměr a  $S_2^2$  výběrový rozptyl. Předpokládejme  $\underline{X} \perp \underline{Y}$ . Pak

- Při  $\text{neznámých } \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$  je  $100(1 - \alpha)\%$  IS roven

$$\left\langle \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)}, \frac{S_1^2}{S_2^2} \frac{1}{F_{\frac{\alpha}{2}}(n_1-1, n_2-1)} \right\rangle.$$

## Alternativně

$$\left\langle \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1)}, \frac{S_1^2}{S_2^2} \frac{1}{F_{1-\frac{\alpha}{2}}(n_2-1, n_1-1)} \right\rangle$$

# Příklad

## Příklad 25

V tabulce jsou uvedeny výsledky analýz niklu získané dvěma analytickými metodami. Stanovte interval spolehlivosti pro podíl směrodatných odchylek obou metod při riziku  $\alpha = 0,05$ , jestliže tyto výsledky považujeme za realizace náhodných výběrů z normálního rozdělení.

Metoda I	3,26	3,26	3,27	3,27
Metoda II	3,23	3,27	3,29	3,29

**Řešení** Vypočteme směrodatné odchylky  $s_1 = 0,0058$ ,  $s_2 = 0,028$ .

IS pro  $\frac{\sigma_1^2}{\sigma_2^2}$ :

$$\left( \frac{s_1^2}{s_2^2} \frac{1}{F_{0,975}(4-1,4-1)}, \frac{s_1^2}{s_2^2} F_{0,975}(4-1,4-1) \right) = (0,0027; 0,643),$$

odtud dostáváme IS pro  $\frac{\sigma_1}{\sigma_2}$ :

$$\sqrt{(0,0027; 0,643)} = (0,052; 0,802).$$

## Věta 26

Nechť  $\mathbf{X}_1 = (X_1, Y_1)', \dots, \mathbf{X}_n = (X_n, Y_n)'$  je náhodný výběr z dvourozměrného normálního rozdělení  $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  s parametry  $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$  a  $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ , kde  $\mu_1, \mu_2 \in \mathbb{R}$ ,  $\sigma_1^2 > 0$ ,  $\sigma_2^2 > 0$  a  $\rho \in (0, 1)$ .

$$\begin{aligned} \text{Pro } i = 1, \dots, n \text{ označme} \quad Z_i &= X_i - Y_i \\ \bar{Z} &= \frac{1}{n} \sum_{i=1}^n Z_i \\ S_Z^2 &= \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z})^2. \end{aligned}$$

Pak

$$\left\langle \bar{Z} - t_{1-\frac{\alpha}{2}}(n-1) \frac{S_Z}{\sqrt{n}}, \bar{Z} + t_{1-\frac{\alpha}{2}}(n-1) \frac{S_Z}{\sqrt{n}} \right\rangle$$

je intervalový odhad parametrické funkce  $\mu_1 - \mu_2$  o spolehlivosti  $1 - \alpha$ .



## Příklad 27

U 6 aut bylo zjištěno ojetí předních pneumatik (v mm)

L	1,8	1,0	2,2	0,9	1,5	1,6
P	1,5	1,1	2,0	1,1	1,4	1,4

Určete 95 % interval spolehlivosti pro rozdíl středních hodnot ojetí levé a pravé pneumatiky.

**Řešení** Vypočteme rozdíl ojetí na každém autě

$z = (0,3; -0,1; 0,2; -0,2; 0,1; 0,2)$  a průměr  $\bar{z} = 0,083$  a směrodatnou odchylku  $s = 0,194$ .

IS pro  $\mu_1 - \mu_2$ :

$$\bar{z} \pm t_{0,975}(6-1) \frac{s}{\sqrt{n}} = 0,083 \pm 2,57 \cdot \frac{0,194}{\sqrt{6}} = (-0,120; 0,287).$$

# Odhady založené na centrální limitní větě

Často lze najít takovou **transformaci**  $h$ , že náhodná veličina  $h(\mathbf{X}, \gamma(\boldsymbol{\theta}))$  má pro  $n \rightarrow \infty$  **asymptoticky** standardizované normální rozdělení  $N(0, 1)$ , tj.

$$h(\mathbf{X}, \gamma(\boldsymbol{\theta})) \stackrel{A}{\sim} N(0, 1)$$

Přitom rozdělení, z něhož výběr pochází

- nemusí splňovat požadavky **spojitosti** a **ryzí monotonie** distribuční funkce,
- může být i diskrétní.

## Věta 28

Mějme  $\perp\{X_1, \dots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$  a výběrový průměr  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . Necht'  $S_*^2 = S_*^2(\mathbf{X})$  je (**slabě**) **konzistentním odhadem** rozptylu  $\sigma^2(\boldsymbol{\theta})$ . Pak statistika

$$U_* = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{S_*} \sqrt{n} \stackrel{A}{\sim} N(0, 1).$$

## Důsledek 29 (Binární náhodné výběry)

Nechť  $\mathbb{1}\{X_1, \dots, X_n\} \simeq A(p)$  je náhodný výběr s alternativním (binárním) rozdělením. Potom intervalovým odhadem parametru  $\boxed{p}$  o asymptotické spolehlivosti  $1 - \alpha$  je interval

$$\left\langle \bar{X} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right\rangle.$$

## Důsledek 30 (Poissonovské náhodné výběry)

Nechť  $\mathbb{1}\{X_1, \dots, X_n\} \simeq Po(\lambda)$  je náhodný výběr s Poissonovým rozdělením. Potom intervalovým odhadem parametru  $\boxed{\lambda}$  ( $0 < \lambda < \infty$ ) o asymptotické spolehlivosti  $1 - \alpha$  je interval

$$\left\langle \bar{X} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}}{n}}, \bar{X} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\bar{X}}{n}} \right\rangle.$$

## Příklad 31

Z 42 náhodně vybraných účastníků sportovního odpoledne bylo 16 dívek a 26 chlapců. Určete 95 % interval spolehlivosti pro podíl dívek mezi účastníky.

**Řešení** Označme  $X_i$ ,  $i = 1, \dots, 42$  náhodnou veličinu nabývající hodnoty 1, pokud vybraný účastník je dívka a hodnoty 0, pokud vybraný účastník je chlapec. Zřejmě  $X_i \sim A(p)$ . Vypočteme průměr  $\bar{x} = \frac{16}{42} = 0,38$  a směrodatnou odchylku  $s = \sqrt{\bar{x}(1 - \bar{x})} = 0,4856$ .  
IS pro  $p$ :

$$\bar{x} \pm u_{0,975} \frac{s}{\sqrt{n}} = 0,38 \pm 1,96 \cdot \frac{0,4856}{\sqrt{42}} = (0,234; 0,527).$$

# Úlohy k procvičení

## Příklad 1.1

Při zjišťování přesnosti nově zaváděné metody pro stanovení obsahu manganu v oceli bylo rozhodnuto provést 4 nezávislá měření. Stanovte dolní odhad pro  $\sigma$  s rizikem 0,05, když výsledky měření byly: 0,31%; 0,30%; 0,29%; 0,32%.

[0,00799]

## Příklad 1.2

Ze základního souboru byl proveden náhodný výběr s naměřenými intervalovými hodnotami a jejich četnostmi sledovaného znaku

$x_i$	(15, 17)	(17, 19)	(19, 21)	(21, 23)	(23, 25)	(25, 27)
$n_i$	10	30	50	70	60	30

Určete

- interval ve kterém se nachází střední hodnota  $\mu$  s pravděpodobností 0,95
- interval ve kterém se nachází rozptyl  $\sigma^2$  s pravděpodobností 0,95.

[a) (21,5094; 22,1706), b) (5,952; 8,464)]

# Úlohy k procvičení

## Příklad 1.3

V tabulce jsou uvedeny hodnoty odporu (v ohmech) vzorků drátů A a B. Je známo, že výsledky takových zkoušek mají normální rozdělení s rozptyly  $\sigma_1^2 = 4 \cdot 10^{-6}$ ,  $\sigma_2^2 = 9 \cdot 10^{-6}$ . Stanovte dolní odhad pro rozdíl středních hodnot odporu drátů při riziku  $\alpha = 0,05$ .

A	0,140	0,138	0,143	0,142	0,144	0,137
B	0,135	0,140	0,142	0,136	0,138	

[−0,000116]

## Příklad 1.4

Bylo vylosováno 6 vrhů selat a z nich vždy dva sourozenci. Jeden z nich vždy dostal náhodně dietu č. 1 a druhý dietu č. 2. Přírůstky (v dkg) jsou následující: (62;52), (54;56), (55;49), (60;50), (53;51), (58;50). Sestrojte 95% interval spolehlivosti pro  $\mu_1 - \mu_2$ .

[(0,626;10,707)]

# Úlohy k procvičení

## Příklad 1.5

V tabulce jsou uvedeny výsledky analýz niklu získané dvěma analytickými metodami. Stanovte horní odhad pro podíl směrodatných odchylek obou metod při riziku  $\alpha = 0,05$ , jestliže tyto výsledky považujeme za realizace náhodných výběrů z normálního rozdělení.

Metoda I	3,26	3,26	3,27	3,27
Metoda II	3,23	3,27	3,29	3,29

[0,622]

## Příklad 1.6

Mezi 160 pracovníky (náhodně vybranými z 8 000 pracujících v závodě) 48 cestuje do práce vlakem. Napište bodový odhad a 95% interval spolehlivosti pro podíl a počet zaměstnanců dopravujících se vlakem.

[podíl: 0,3; (0,229; 0,371), počet: 2 400; (1 832; 2 968)]

## Příklad 1.7

*Naprogramujte funkci `ukol.R`, která pro jediný vstupní parametr  $n$  vygeneruje  $n$ -rozměrný datový soubor z normálního rozdělení  $N(1/2, 1)$  a na základě vygenerovaných dat sestojí 95% interval spolehlivosti pro střední hodnotu  $\mu$ . Sledujte, pro jak velká  $n$  tento interval obsahuje nulu a jak se mění šířka intervalu. Dokážete interpretovat pozorované jevy?*