# 5. Základní pojmy matematické statistiky

## 5.1. Motivace

V teorii pravděpodobnosti je matematickým modelem náhodného pokusu pravděpodobnostní prostor $(\Omega, A, P)$. Výsledky náhodného pokusu jsou popsány pomocí náhodné veličiny $X : \Omega \to R$. Pravděpodobnostní chování náhodné veličiny X je popsáno distribuční funkcí $\Phi(x)$. Pravděpodobnostní prostor a distribuční funkci považujeme za známé a hledáme pravděpodobnosti jevů určené náhodnou veličinou X.

V matematické statistice je situace odlišná. Máme číselné realizace n nezávislých pozorování náhodné veličiny X: $x_1 = X(\omega_1)$, ..., $x_n = X(\omega_n)$ a na jejich základě chceme učinit výpověď o distribuční funkci $\Phi(x)$ resp. o pravděpodobnostním prostoru $(\Omega, A, P)$.

Předpokládáme, že $X \sim \Phi_\vartheta(x)$, kde $\{\Phi_\vartheta(x); \vartheta \in \Xi\}$ je známá třída distribučních funkcí (její znalost plyne z úvah o podstatě náhodného pokusu, jímž jsme data získali nebo z dřívějších zkušeností s podobnými daty), $\vartheta$ je skalární nebo vektorový parametr a $\Xi$ je parametrický prostor, tj. množina všech přípustných hodnot parametru. Jedním z důležitých úkolů matematické statistiky je pomocí dat $x_1$, ..., $x_n$ odhadnout (bodově či intervalově) parametr $\vartheta$ (nebo nějakou jeho parametrickou funkci $h(\vartheta)$)a tím specifikovat distribuční funkci $\Phi_\vartheta(x)$, podle níž se řídí pravděpodobnostní chování náhodné veličiny X. Matematická statistika rovněž ověřuje pravdivost různých tvrzení o parametru $\vartheta$ či parametrické funkci $h(\vartheta)$.

## 5.2. Náhodný výběr a statistiky odvozené z náhodného výběru

### 5.2.1. Pojem náhodného výběru

Nechť $X_1$, ..., $X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\vartheta)$. Řekneme, že $X_1$, ..., $X_n$ je náhodný výběr rozsahu n z rozložení $L(\vartheta)$. (Číselné realizace $x_1$, ..., $x_n$ náhodného výběru $X_1$, ..., $X_n$ uspořádané do sloupcového vektoru představují datový soubor.)

Nechť $(X_1,Y_1)$, ..., $(X_n,Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory se stejným dvourozměrným rozložením $L_2(\vartheta)$. Řekneme, že $(X_1,Y_1)$, ..., $(X_n,Y_n)$ je dvourozměrný náhodný výběr rozsahu n z dvourozměrného rozložení $L_2(\vartheta)$. (Číselné realizace $(x_1,y_1)$, ..., $(x_n,y_n)$ náhodného výběru $(X_1,Y_1)$, ..., $(X_n,Y_n)$ uspořádané do matice typu $n \times 2$ představují dvourozměrný datový soubor.)

Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\vartheta)$.

### 5.2.2. Pojem statistiky, příklady důležitých statistik

Libovolná funkce $T = T(X_1, ..., X_n)$ náhodného výběru $X_1$, ..., $X_n$ (resp. p-rozměrného náhodného výběru) se nazývá statistika.

a) Nechť $X_1$, ..., $X_n$ je náhodný výběr, $n \geq 2$. Statistika $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ se nazývá výběrový průměr, $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ výběrový rozptyl, $S = \sqrt{S^2}$ výběrová směrodatná odchylka. Pro libovolné, ale pevně zvolené reálné číslo x je statistikou též hodnota výběrové distribuční funkce $F_n(x) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x\}$.

b) Nechť $X_{11},\ldots,X_{1n_1}$, ..., $X_{p1},\ldots,X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2$, ..., $n_p \geq 2$. Celkový rozsah je $n = \sum_{j=1}^{p} n_j$. Označme $M_1$, ..., $M_p$ výběrové průměry a $S_1^2$, ..., $S_p^2$ výběrové rozptyly jednotlivých výběrů. Statistika $\sum_{j=1}^{p} c_j M_j$, kde $c_1$, ..., $c_p$ jsou reálné konstanty, aspoň jedna nenulová, se nazývá lineární kombinace výběrových průměrů. Statistika $S_*^2 = \frac{\sum_{j=1}^{p}(n_j - 1)S_j^2}{n-p}$ se nazývá vážený průměr výběrových rozptylů.

c) Nechť $(X_1,Y_1)$, ..., $(X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení. Statistika $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$ je výběrová kovariance (přitom $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$, $M_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i$) a statistika $R_{12} = \begin{cases} \frac{1}{n-1}\sum_{i=1}^{n}\frac{X_i - M_1}{S_1}\cdot\frac{Y_i - M_2}{S_2} & \text{pro } S_1S_2 \neq 0 \\ 0 & \text{jinak} \end{cases}$ se nazývá výběrový koeficient korelace. Pro libovolnou, ale pevně zvolenou dvojici reálných čísel x,y je statistikou též hodnota výběrové simultánní distribuční funkce

$$F_n(x,y) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x \wedge Y_i \leq y\}.$$

(Číselné realizace m, $s^2$, s, $s_{12}$, $r_{12}$ statistik M, $S^2$, S, $S_{12}$, $R_{12}$ odpovídají číselným charakteristikám znaků v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n-1}$, nikoli $\frac{1}{n}$, jak tomu bylo v popisné statistice.)

## 5.3. Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru $X_1$, ..., $X_n$ z rozložení $L(\vartheta)$, které závisí na parametru $\vartheta$. Množinu všech přípustných hodnot tohoto parametru označíme $\Xi$. Parametr $\vartheta$ neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\vartheta)$).

Bodovým odhadem parametrické funkce $h(\vartheta)$ je statistika $T_n = T(X_1, ..., X_n)$, která nabývá hodnot blízkých $h(\vartheta)$, ať je hodnota parametru $\vartheta$ jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

Intervalovým odhadem parametrické funkce $h(\vartheta)$ rozumíme interval (D, H), jehož meze jsou statistiky $D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\vartheta)$, ať je hodnota parametru $\vartheta$ jakákoliv.

### 5.3.1. Typy bodových odhadů

Nechť $X_1$, ..., $X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, T, $T_1$, $T_2$, ... jsou statistiky.

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\vartheta)$, jestliže $\forall\vartheta \in \Xi: E(T) = h(\vartheta)$.

(Význam nestrannosti spočívá v tom, že odhad T nesmí parametrickou funkci $h(\vartheta)$ syste-

maticky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li $T_1$, $T_2$ nestranné odhady téže parametrické funkce $h(\vartheta)$, pak řekneme, že $T_1$ je lepší odhad než $T_2$, jestliže
$\forall \vartheta \in \Xi : D(T_1) < D(T_2)$.

c) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost asymptoticky nestranných odhadů parametrické funkce $h(\vartheta)$, jestliže
$\forall \vartheta \in \Xi : \lim_{n\to\infty} E(T_n) = h(\vartheta)$.
(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu.)

d) Posloupnost $\{T_n\}_{n=1}^{\infty}$ se nazývá posloupnost konzistentních odhadů parametrické funkce $h(\vartheta)$, jestliže
$\forall \vartheta \in \Xi \forall \varepsilon > 0 : \lim_{n\to\infty} P(|T_n - h(\vartheta)| > \varepsilon) = 0$.
(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat „daleko“ od parametrické funkce $h(\vartheta)$.)

Lze dokázat, že z nestrannosti odhadu vyplývá jeho asymptotická nestrannost a z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů odhadu konverguje k nule.

### 5.3.2. Vlastnosti důležitých statistik

a) Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou $\mu$, rozptylem $\sigma^2$ a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$. Označme $M_n$ výběrový průměr, $S_n^2$ výběrový rozptyl a pro libovolné, ale pevně dané $x \in R$ označme $F_n(x)$ hodnotu výběrové distribuční funkce.

Pak $M_n$ je nestranným odhadem $\mu$ (tj. $E(M_n) = \mu$) s rozptylem $D(M) = \frac{\sigma^2}{n}$, $S_n^2$ je nestranným odhadem $\sigma^2$ (tj. $E(S_n^2) = \sigma^2$), ať jsou hodnoty parametrů $\mu$, $\sigma^2$ jakékoli. Dále platí, že pro libovolné, ale pevně dané $x \in R$ je výběrová distribuční funkce $F_n(x)$ nestranným odhadem $\Phi(x)$ (tj. $E(F_n(x)) = \Phi(x)$) s rozptylem $D(F_n(x)) = \Phi(x)(1- \Phi(x))/n$, ať je hodnota distribuční funkce $\Phi(x)$ jakákoliv.

Posloupnost $\{M_n\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů $\mu$, $\{S_n^2\}_{n=1}^{\infty}$ je posloupnost konzistentních odhadů $\sigma^2$ a pro libovolné, ale pevně dané $x \in R$ je $\{F_n(x)\}_{n=1}^{\infty}$ posloupnost konzistentních odhadů $\Phi(x)$.

b) Nechť $X_{11},\ldots,X_{1n_1}$, ..., $X_{p1},\ldots,X_{pn_p}$ je p stochasticky nezávislých náhodných výběrů o rozsazích $n_1 \geq 2$, ..., $n_p \geq 2$ z rozložení se středními hodnotami $\mu_1$, ..., $\mu_p$ a rozptylem $\sigma^2$. Celkový rozsah je $n = \sum_{j=1}^{p} n_j$. Nechť $c_1$, ..., $c_p$ jsou reálné konstanty, aspoň jedna nenulová.

Pak lineární kombinace výběrových průměrů $\sum_{j=1}^{p} c_j M_j$ je nestranným odhadem lineární kombinace středních hodnot $\sum_{j=1}^{p} c_j \mu_j$, ať jsou střední hodnoty $\mu_1$, ..., $\mu_p$ jakékoli a vážený průměr

výběrových rozptylů $S_*^2 = \frac{\sum_{j=1}^{p}(n_j - 1)S_j^2}{n - p}$ je nestranným odhadem rozptylu $\sigma^2$, ať je rozptyl $\sigma^2$ jakýkoliv.

c) Nechť $(X_1,Y_1)$, ..., $(X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí $\sigma_{12}$ a koeficientem korelace $\rho$. Pak výběrová kovariance $S_{12}$ je nestranným odhadem kovariance $\sigma_{12}$, ať je kovariance $\sigma_{12}$ jakákoli, avšak $E(R_{12})$ je rovno $\rho$ pouze přibližně (shoda je vyhovující pro $n > 30$), ať je korelační koeficient $\rho$ jakýkoli.

### 5.3.3. Pojem intervalu spolehlivosti

Nechť $X_1$, ..., $X_n$ je náhodný výběr z rozložení $L(\vartheta)$, $h(\vartheta)$ je parametrická funkce, $\alpha \in (0,1)$, $D = D(X_1, ..., X_n)$, $H = H(X_1, ..., X_n)$ jsou statistiky.

a) Interval (D, H) se nazývá 100(1-α)% (oboustranný) interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi : P(D < h(\vartheta) < H) \geq 1-\alpha$.

b) Interval $(D, \infty)$ se nazývá 100(1-α)% levostranný interval spolehlivosti pro parametrickou funkci $h(\upsilon)$, jestliže: $\forall \vartheta \in \Xi : P(D < h(\vartheta)) \geq 1-\alpha$.

c) Interval $(-\infty, H)$ se nazývá 100(1-α)% pravostranný interval spolehlivosti pro parametrickou funkci $h(\vartheta)$, jestliže: $\forall \vartheta \in \Xi : P(h(\vartheta) < H) \geq 1-\alpha$.

Číslo α se nazývá riziko (zpravidla $\alpha = 0{,}05$, méně často 0,1 či 0,01), číslo $1 - \alpha$ se nazývá spolehlivost.

### 5.3.4. Postup při konstrukci intervalu spolehlivosti

a) Vyjdeme ze statistiky V, která je nestranným bodovým odhadem parametrické funkce $h(\vartheta)$.
b) Najdeme tzv. pivotovou statistiku W, která vznikne transformací statistiky V, je monotónní funkcí $h(\vartheta)$ a přitom její rozložení je známé a na $h(\vartheta)$ nezávisí. Pomocí známého rozložení pivotové statistiky W najdeme kvantily $w_{\alpha/2}$, $w_{1-\alpha/2}$, takže platí: $\forall \vartheta \in \Xi : P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1 - \alpha$.
c) Nerovnost $w_{\alpha/2} < W < w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost $D < h(\vartheta) < H$.
d) Statistiky D, H nahradíme jejich číselnými realizacemi d, h a získáme tak 100(1-α)% empirický interval spolehlivosti, o němž prohlásíme, že pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1 - \alpha$. (Tvrzení, že (d,h) pokrývá $h(\vartheta)$ s pravděpodobností aspoň $1 - \alpha$ je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace $x_1$, ..., $x_n$ náhodného výběru $X_1$, ..., $X_n$ z rozložení $L(\vartheta)$ a pomocí každé této realizace sestrojíme 100(1-α)% empirický interval spolehlivosti pro $h(\vartheta)$, pak podíl počtu těch intervalů, které pokrývají $h(\vartheta)$ k počtu všech sestrojených intervalů bude přibližně $1 - \alpha$.)

**Příklad 1.**: Nechť $X_1$, ..., $X_n$ je náhodný výběr z $N(\mu,\sigma^2)$, kde $n \geq 2$ a rozptyl $\sigma^2$ známe. Sestrojte 100(1-α)% interval spolehlivosti pro neznámou střední hodnotu μ.

**Řešení**: V tomto případě parametrická funkce $h(\vartheta) = \mu$. Nestranným odhadem střední hodnoty je výběrový průměr (viz 5.3.(a)) $M = \frac{1}{n}\sum_{i=1}^{n} X_i$. Protože M je lineární kombinací normálně rozložených náhodných veličin, bude mít také normální rozložení se střední hodnotou

E(M) = μ a rozptylem D(M) = $\frac{\sigma^2}{n}$. Pivotovou statistikou W bude standardizovaná náhodná veličina $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. Kvantil $w_{\alpha/2} = u_{\alpha/2} = -u_{1-\alpha/2}$, $w_{1-\alpha/2} = u_{1-\alpha/2}$.

$\forall \vartheta \in \Xi : 1-\alpha \le P(-u_{1-\alpha/2} < U < u_{1-\alpha/2}) =$

$$P\left(-u_{1-\alpha/2} < \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\alpha/2}\right) = P\left(M - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} < \mu < M + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right).$$

Meze 100(1-α)% intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu $\sigma^2$ tedy jsou:

$D = M - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$, $H = M + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}$.

Při konstrukci jednostranných intervalů spolehlivosti se riziko nepůlí, tedy 100(1-α)% levostranný interval spolehlivosti pro μ je $\left(M - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}, \infty\right)$ a pravostranný je $\left(-\infty, M + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right)$.

Dosadíme-li do vzorců pro dolní a horní mez číselnou realizaci m výběrového průměru M, dostaneme 100(1-α)% empirický interval spolehlivosti.

### 5.3.5. Šířka intervalu spolehlivosti

Nechť (d, h) je 100(1-α)% empirický interval spolehlivosti pro $h(\vartheta)$ zkonstruovaný pomocí číselných realizací $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ z rozložení $L(\vartheta)$.

a) Při konstantním riziku klesá šířka h-d s rostoucím rozsahem náhodného výběru.
b) Při konstantním rozsahu náhodného výběru klesá šířka h-d s rostoucím rizikem.

Využití bodu (a) při stanovení minimálního rozsahu výběru y normálního rozložení: Nechť $X_1, ..., X_n$ je náhodný výběr z $N(\mu, \sigma^2)$, kde $\sigma^2$ známe. Jaký musí být minimální rozsah výběru n, aby šířka 100(1-α)% empirického intervalu spolehlivosti pro střední hodnotu μ nepřesáhla číslo Δ?

Řešení: Požadujeme, aby $\Delta \ge h - d = m + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} - (m - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}) = \frac{2\sigma}{\sqrt{n}}u_{1-\alpha/2}$. Z této podmínky dostaneme, že $n \ge \frac{4\sigma^2 u_{1-\alpha/2}^2}{\Delta^2}$. Za rozsah výběru zvolíme nejmenší přirozené číslo vyhovující této podmínce.

**Příklad 2.**: Hloubka moře se měří přístrojem, jehož systematická chyba je nulová a náhodné chyby měření mají normální rozložení se směrodatnou odchylkou σ = 1 m. Kolik měření je nutno provést, aby se hloubka stanovila s chybou nejvýše ± 0,25 m při spolehlivosti 0,95?

**Řešení**: Hledáme rozsah výběru tak, aby šířka 95% intervalu spolehlivosti pro střední hodnotu μ nepřesáhla 0,5 m. Přitom σ známe. Z 5.3.5. vyplývá, že

$n \geq \frac{4\sigma^2 u_{1-\alpha/2}^2}{\Delta^2} = \frac{4 \cdot 1{,}96^2}{0{,}5^2} = 61{,}4656$. Nejmenší počet měření je tedy 62.

## 5.4. Plánování pokusů

Abychom mohli správně vyhodnotit výsledky pokusu, musí být pokus dobře naplánován. Předpokládejme například, že zkoumáme vliv výkrmné diety (resp. několika výkrmných diet) na hmotnostní přírůstky selat. Selatům podáváme výkrmnou dietu po dobu půl roku a poté zjistíme průměrné denní přírůstky.

V závislosti na záměrech experimentátora rozeznáváme několik typů uspořádání pokusů.

### 5.4.1. Jednoduché pozorování

Náhodná veličina je pozorována za týchž podmínek. Situace je charakterizována jedním náhodným výběrem $X_1, ..., X_n$. (Náhodně vybereme n stejně starých selat téhož plemene, podrobíme je jediné výkrmné dietě a zjistíme hmotnostní přírůstky. Tak dostaneme realizaci jednoho náhodného výběru.)

### 5.4.2. Dvojné pozorování

Zkoumá se rozdílnost hodnot náhodné veličiny pozorované za dvojích různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

a) **Dvouvýběrové porovnávání**: Situace je charakterizována dvěma nezávislými výběry $X_{11},\ldots,X_{1n_1}$ a $X_{21},\ldots,X_{2n_2}$. (Z populace všech dostupných selat náhodně vybereme $n_1+n_2$ stejně starých selat téhož plemene. Náhodně je rozdělíme na dva soubory o rozsazích $n_1$ a $n_2$, první podrobíme výkrmné dietě č. 1 a druhý výkrmné dietě č. 2. Tak dostaneme realizace dvou nezávislých náhodných výběrů.)

b) **Párové porovnávání**: Situace je charakterizována jedním náhodným výběrem $(X_{11},X_{12}), ..., (X_{n1},X_{n2})$ z dvourozměrného rozložení. Párem se rozumí dvojice $(X_{i1},X_{i2})$, $i = 1, ..., n$. Úloha se zpravidla převádí na jednoduché pozorování náhodného výběru rozdílů $X_{i1} - X_{i2}$, kde $i = 1, ..., n$. (Náhodně vybereme n vrhů selat a z nich vždy dva sourozence a náhodně jim přiřadíme 1. a 2. výkrmnou dietu. Tak dostaneme realizaci náhodného výběru z dvourozměrného rozložení.)

### 5.4.3. Mnohonásobné pozorování

Zkoumá se rozdílnost hodnot náhodné veličiny pozorované za $r \geq 3$ různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

a) **Mnohovýběrové porovnávání**: Situace je charakterizována r nezávislými náhodnými výběry $X_{11},\ldots,X_{1n_1}, ..., X_{r1},\ldots,X_{rn_r}$. (Z populace všech dostupných selat náhodně vybereme $n_1+n_2+...+n_r$ stejně starých selat téhož plemene. Náhodně je rozdělíme na r souborů o rozsazích $n_1, n_2, ..., n_r$. Selata z prvního souboru podrobíme výkrmné dietě č. 1, ..., selata z r-tého souboru podrobíme výkrmné dietě č. r. Tak dostaneme realizace r nezávislých náhodných výběrů.)

b) **Blokové porovnávání**: Situace je charakterizována jedním náhodným výběrem $(X_{11},X_{12}, ..., X_{1r}), ..., (X_{n1},X_{n2}, ..., X_{nr})$ z r-rozměrného rozložení. Blokem se rozumí r-tice $(X_{i1},X_{i2}, ..., X_{ir})$, $i = 1, ..., n$. (Náhodně vybereme n vrhů selat a z nich vždy r sourozenců a náhodně jim přiřadíme 1. až r-tou výkrmnou dietu. Tak dostaneme realizaci náhodného výběru z r-rozměrného rozložení.)

## 5.5. Úvod do testování hypotéz

Testování hypotéz je důležitou úlohou statistické indukce. Pomocí statistické indukce se snažíme na základě znalosti náhodného výběru usuzovat na vlastnosti rozložení, z něhož tento výběr pochází.

Nulovou hypotézou rozumíme nějaké tvrzení o parametrech nebo typu rozložení, z něhož pochází náhodný výběr. Nulová hypotéza vyjadřuje nějaký teoretický předpoklad, často skeptického rázu a uživatel ji musí stanovit předem, bez přihlédnutí k datovému souboru. Proti nulové hypotéze stavíme alternativní hypotézu, která říká, co platí, když neplatí nulová hypotéza. Např. nulová hypotéza tvrdí, že střední hodnota hmotnosti balíčků cukru balených na automatické lince se nezměnila seřízením automatu, zatímco alternativní hypotéza tvrdí opak. Postup, který je založen na daném náhodném výběru a s jehož pomocí rozhodneme o zamítnutí či nezamítnutí nulové hypotézy, se nazývá testování hypotéz.

Rozlišujeme testy parametrické a neparametrické. Parametrické testy předpokládají, že daný náhodný výběr pochází z určitého typu rozložení (často normálního), které závisí na nějakých neznámých parametrech. Naproti tomu neparametrické testy nevyžadují předpoklad o určitém typu rozložení, ale stačí jim splnění jen velmi obecných podmínek, např. že distribuční funkce je spojitá.

### 5.5.1. Nulová a alternativní hypotéza

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení $L(\vartheta)$, kde parametr $\vartheta \in \Xi$ neznáme. Nechť $h(\vartheta)$ je parametrická funkce a c daná reálná konstanta.

a) Oboustranná alternativa: Tvrzení $H_0$: $h(\vartheta) = c$ se nazývá jednoduchá nulová hypotéza. Proti nulové hypotéze postavíme složenou alternativní hypotézu $H_1$: $h(\vartheta) \neq c$.

b) Levostranná alternativa: Tvrzení $H_0$: $h(\vartheta) \geq c$ se nazývá složená pravostranná nulová hypotéza. Proti jednoduché nebo složené pravostranné nulové hypotéze postavíme složenou levostrannou alternativní hypotézu $H_1$: $h(\vartheta) < c$.

c) Pravostranná alternativa: Tvrzení $H_0$: $h(\vartheta) \leq c$ se nazývá složená levostranná nulová hypotéza. Proti jednoduché nebo složené levostranné nulové hypotéze postavíme složenou pravostrannou alternativní hypotézu $H_1$: $h(\vartheta) > c$.

Testováním $H_0$ proti $H_1$ rozumíme rozhodovací postup založený na náhodném výběru $X_1, ..., X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

Volba alternativní hypotézy není libovolná, ale vyplývá z konkrétní situace. Např. při současné technologii je pravděpodobnost vyrobení zmetku $\vartheta = 0{,}01$.

a) Po rekonstrukci výrobní linky byla obnovena výroba, přičemž technologie zůstala stejná. Chceme ověřit, zda se změnila kvalita výrobků. Testujeme $H_0$: $\vartheta = 0{,}01$ proti $H_1$: $\vartheta \neq 0{,}01$.

b) Byly provedeny změny v technologii výroby s cílem zvýšit kvalitu. V tomto případě tedy testujeme $H_0$: $\vartheta = 0{,}01$ proti $H_1$: $\vartheta < 0{,}01$.

c) Byly provedeny změny v technologii výroby s cílem snížit náklady. V této situaci testujeme $H_0$: $\vartheta = 0{,}01$ proti $H_1$: $\vartheta > 0{,}01$.

### 5.5.2. Chyba 1. a 2. druhu

Při testování $H_0$ proti $H_1$ se můžeme dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že $H_0$ zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že $H_0$ nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

| skutečnost | rozhodnutí | |
|---|---|---|
| | $H_0$ nezamítáme | $H_0$ zamítáme |
| $H_0$ platí | správné rozhodnutí | chyba 1. druhu |
| $H_0$ neplatí | chyba 2. druhu | správné rozhodnutí |

Pravděpodobnost chyby 1. druhu se značí $\alpha$ a nazývá se hladina významnosti testu (většinou bývá $\alpha = 0{,}05$, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí $\beta$. Číslo $1-\beta$ se nazývá síla testu a vyjadřuje pravděpodobnost, s jakou test vypoví, že $H_0$ neplatí.

### 5.5.3. Testování pomocí kritického oboru

Najdeme statistiku $T_0 = T_0(X_1, ..., X_n)$, kterou nazveme testovým kritériem. Množina všech hodnot, jichž může testové kritérium nabýt, se rozpadá na obor nezamítnutí nulové hypotézy (značí se V) a obor zamítnutí nulové hypotézy (značí se W a nazývá se též kritický obor). Tyto dva obory jsou odděleny kritickými hodnotami (pro danou hladinu významnosti $\alpha$ je lze najít ve statistických tabulkách).

Jestliže číselná realizace $t_0$ testového kritéria $T_0$ padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti $\alpha$ a znamená to skutečné vyvrácení testované hypotézy. Jestliže $t_0$ padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Pravděpodobnosti chyb 1. a 2. druhu nyní zapíšeme takto:
$P(T_0 \in W/H_0 \text{ platí}) = \alpha$, $P(T_0 \in V/H_1 \text{ platí}) = \beta$.

Stanovení kritického oboru pro danou hladinu významnosti $\alpha$:
Označme $t_{min}$ (resp. $t_{max}$) nejmenší (resp. největší) hodnotu testového kritéria.
Kritický obor v případě oboustranné alternativy má tvar
$W = \left(t_{min}, K_{\alpha/2}(T)\right\rangle \cup \left\langle K_{1-\alpha/2}(T), t_{max}\right)$, kde $K_{\alpha/2}(T)$ a $K_{1-\alpha/2}(T)$ jsou kvantily rozložení, jímž se řídí testové kritérium $T_0$, je-li nulová hypotéza pravdivá.
Kritický obor v případě levostranné alternativy má tvar:
$W = \left(t_{min}, K_{\alpha}(T)\right\rangle$.
Kritický obor v případě pravostranné alternativy má tvar:
$W = \left\langle K_{1-\alpha}(T), t_{max}\right)$.

Doporučuje se dodržovat následující postup:
- Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
- Zvolíme hladinu významnosti $\alpha$. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.
- Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
- Stanovíme kritický obor.
- Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti $\alpha$. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti $\alpha$.

### 5.5.4. Testování pomocí intervalu spolehlivosti

Sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro parametrickou funkci $h(\vartheta)$. Pokryje-li tento interval hodnotu c, pak $H_0$ nezamítáme na hladině významnosti $\alpha$, v opačném případě $H_0$ zamítáme na hladině významnosti $\alpha$.
Pro test $H_0$ proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.
Pro test $H_0$ proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.
Pro test $H_0$ proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.

### 5.5.5. Testování pomocí p-hodnoty

p-hodnota udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je-li p-hodnota $\leq \alpha$, pak $H_0$ zamítáme na hladině významnosti $\alpha$, je-li p-hodnota $> \alpha$, pak $H_0$ nezamítáme na hladině významnosti $\alpha$.
Způsob výpočtu p-hodnoty:

Pro oboustrannou alternativu $p = 2 \min\{P(T_0 \le t_0), P(T_0 \ge t_0)\}$.
Pro levostrannou alternativu $p = P(T_0 \le t_0)$.
Pro pravostrannou alternativu $p = P(T_0 \ge t_0)$.

p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ podporují $H_0$, je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium $T_0$, je-li $H_0$ pravdivá. Vzhledem k tomu, že v běžných statistických tabulkách jsou uvedeny pouze hodnoty distribuční funkce standardizovaného normálního rozložení, bez použití speciálního software jsme schopni vypočítat p-hodnotu pouze pro test hypotézy o střední hodnotě normálního rozložení při známém rozptylu.

**5.5.6. Příklad**

10 x nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, ..., X_{10}$ z rozložení $N(\mu, 0{,}04)$. Nějaká teorie tvrdí, že $\mu = 1{,}95$.

**a) Oboustranná alternativa**

Proti nulové hypotéze $H_0$: $\mu = 1{,}95$ postavíme oboustrannou alternativu $H_1$: $\mu \neq 1{,}95$. Na hladině významnosti 0,05 testujte $H_0$ proti $H_1$ všemi třemi popsanými způsoby.

**Řešení:**

$m = \frac{1}{10}(2 + ... + 2{,}2) = 2{,}06$, $\sigma^2 = 0{,}04$, $n = 10$, $\alpha = 0{,}05$, $c = 1{,}95$

a) Test provedeme pomocí kritického oboru.

Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku $U = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ (viz 5.3.4.). Testové kritérium tedy bude $T_0 = \frac{M - c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení $N(0, 1)$, pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria: $t_0 = \frac{2{,}06 - 1{,}95}{\frac{0{,}2}{\sqrt{10}}} = 1{,}74$. Stanovíme kritický obor:

$W = \left(t_{min}, K_{\alpha/2}(T)\right\rangle \cup \left\langle K_{1-\alpha/2}(T), t_{max}\right) = \left(-\infty, u_{\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) =$

$\left(-\infty, -u_{1-\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}, \infty\right) = \left(-\infty, -u_{0.975}\right\rangle \cup \left\langle u_{0{,}975}, \infty\right) = \left(-\infty, -1{,}96\right\rangle \cup \left\langle 1{,}96, \infty\right)$. Protože $1{,}74 \notin W$, $H_0$ nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.

Meze $100(1-\alpha)\%$ empirického intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu $\sigma^2$ jsou (viz 5.3.4.): $(d, h) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2})$.

V našem případě dostáváme: $d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} u_{0{,}975} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} . 1{,}96 = 1{,}936$, $h = 2{,}184$.

Protože $1{,}95 \in (1{,}936; 2{,}184)$, $H_0$ nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.

Protože proti nulové hypotéze stavíme oboustrannou alternativu, použijeme vzorec
$p = 2 \min\{P(T_0 \le t_0), P(T_0 \ge t_0)\} = 2 \min \{P(T_0 \le 1{,}74), P(T_0 \ge 1{,}74)\} =$
$= 2 \min \{ \Phi(1{,}74), 1 - \Phi(1{,}74) \} = 2 \min \{ 0{,}95907, 1 - 0{,}95907 \} = 0{,}08186$. Jelikož $0{,}08186 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

**b) Levostranná alternativa**
Proti nulové hypotéze $H_0$: $\mu = 1{,}95$ postavíme levostrannou alternativu $H_1$: $\mu < 1{,}95$. Na hladině významnosti 0,05 testujte $H_0$ proti $H_1$ všemi třemi popsanými způsoby.

**Řešení:**
a) Test provedeme pomocí kritického oboru.
Na rozdíl od oboustranné alternativy bude mít kritický obor tvar $W = \langle -\infty, u_\alpha ) = \langle -\infty, u_{0,05} ) = \langle -\infty, -1{,}645 )$.
Protože $1{,}74 \notin W$, $H_0$ nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ empirického pravostranného intervalu spolehlivosti pro střední hodnotu $\mu$ při známém rozptylu $\sigma^2$ jsou (viz 5.3.4.): $(-\infty, h) = (-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha})$.
V našem případě dostáváme: $h = 2{,}06 + \frac{0{,}2}{\sqrt{10}} u_{0,95} = 2{,}06 + \frac{0{,}2}{\sqrt{10}} . 1{,}645 = 2{,}164$.
Protože $1{,}95 \in (-\infty; 2{,}164)$, $H_0$ nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.
Protože proti nulové hypotéze stavíme levostrannou alternativu, použijeme vzorec
$p = P(T_0 \le t_0) = \Phi(1{,}74) = 0{,}95907$. Jelikož $0{,}95907 > 0{,}05$, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

**c) Pravostranná alternativa**
Proti nulové hypotéze $H_0$: $\mu = 1{,}95$ postavíme pravostrannou alternativu $H_1$: $\mu > 1{,}95$. Na hladině významnosti 0,05 testujte $H_0$ proti $H_1$ všemi třemi popsanými způsoby.

**Řešení:**
a) Test provedeme pomocí kritického oboru.
Na rozdíl od oboustranné alternativy bude mít kritický obor tvar $W = \langle u_{1-\alpha}, \infty) = \langle u_{0,95}, \infty) = \langle 1{,}645, \infty)$.
Protože $1{,}74 \in W$, $H_0$ zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ empirického levostranného intervalu spolehlivosti pro střední hodnotu $\mu$ při známém rozptylu $\sigma^2$ jsou (viz 5.3.4.): $(d, \infty) = (m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty)$.
V našem případě dostáváme: $d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} u_{0,95} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} . 1{,}645 = 1{,}956$.
Protože $1{,}95 \notin (1{,}956, \infty)$, $H_0$ zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

c) Test provedeme pomocí p-hodnoty.
Protože proti nulové hypotéze stavíme pravostrannou alternativu, použijeme vzorec $p = P(T_0 \geq t_0) = 1 - \Phi(1,74) = 1 - 0,95907 = 0,04093$. Jelikož $0,04093 \leq 0,05$, nulovou hypotézu zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

## Příklady k 5. kapitole

**Příklad 1. :** Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou $\mu$, rozptylem $\sigma^2$ a distribuční funkcí $\Phi(x)$. Nechť $n \geq 2$.

a) Vypočtěte střední hodnotu a rozptyl výběrového průměru $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ .

b) Vypočtěte střední hodnotu výběrového rozptylu $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$

c) Pro libovolné, ale pevně zvolené reálné x vypočtěte střední hodnotu a rozptyl výběrové distribuční funkce $F_n(x) = \frac{1}{n}\mathrm{card}\{i; X_i \leq x\}$.

**Řešení:**

ad a) $E(M) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$,

$D(M) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$.

ad b)

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2\right) = \frac{1}{n-1}\sum_{i=1}^{n} E\left([(X_i - \mu) - (M - \mu)]^2\right) =$$

$$\frac{1}{n-1}\sum_{i=1}^{n} E\left((X_i - \mu)^2 - 2(X_i - \mu)(M - \mu) + (M - \mu)^2\right) =$$

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

$$\frac{1}{n-1}\left[\sum_{i=1}^{n} D(X_i) - 2E\left((nM - n\mu)(M - \mu)\right) + nE\left((M - \mu)^2\right)\right] =$$

$$\frac{1}{n-1}\left[\sum_{i=1}^{n} \sigma^2 - 2nE\left((M - \mu)^2\right) + nE\left((M - \mu)^2\right)\right] = \frac{1}{n-1}\left(n\sigma^2 - nD(M)\right) = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \sigma^2$$

(Pro informaci, bez důkazu: $D(S^2) = \frac{\gamma_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, kde $\gamma_4$ je 4. centrální moment.

ad c) Pro libovolné, ale pevně zvolené reálné číslo x zavedeme transformovanou náhodnou veličinu $Y_n$, která udává počet těch náhodných veličin $X_1, ..., X_n$, které nabývají hodnoty menší nebo rovné x, tedy $F_n(x) = Y_n/n$. Odvodíme rozložení náhodné veličiny $Y_n$. Je to diskrétní náhodná veličina. Její pravděpodobnostní funkce $\pi(y) = P(Y_n = y)$. Má-li veličina $Y_n$ nabýt hodnoty y, musí právě y veličin z n veličin $X_1, ..., X_n$ nabýt hodnoty menší nebo rovné x. Přitom $P(X_i \leq x) = \Phi(x)$. Odtud tedy plyne, že $Y_n \sim Bi(n, \Phi(x))$. Je nám známo, že $E(Y_n) = = n \Phi(x)$ a $D(Y_n) = n \Phi(x)(1- \Phi(x))$. Využitím vlastností střední hodnoty a rozptylu snadno odvodíme, že $E(F_n(x)) = \Phi(x)$ a $D(F_n(x)) = \Phi(x)(1- \Phi(x))/n$.

**Příklad 2. :** Nechť $(X_1,Y_1), ..., (X_n,Y_n)$ je náhodný výběr z dvourozměrného rozložení s vektorem středních hodnot $(\mu_1, \mu_2)$ a kovariancí $\sigma_{12}$. Vypočtěte střední hodnotu výběrové kovariance $S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)$.

**Řešení:**

$$E(S_{12}) = E\Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2)\Big) =$$

$$\frac{1}{n-1}\sum_{i=1}^{n} E\big([(X_i-\mu_1)-(M_1-\mu_1)][(Y_i-\mu_2)-(M_2-\mu_2)]\big) =$$

$$\frac{1}{n-1}\sum_{i=1}^{n} E\big((X_i-\mu_1)(Y_i-\mu_2)\big) - (M_1-\mu_1)(Y_i-\mu_2) - (M_2-\mu_2)(X_i-\mu_1) + (M_1-\mu_1)(M_2-\mu_2)\Big) =$$

$$\frac{1}{n-1}\left[\sum_{i=1}^{n} C(X_i, Y_i) - E\left((M_1-\mu_1)(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n}\mu_2)\right) - E\left((M_2-\mu_2)(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_1)\right) + \sum_{i=1}^{n} E\big((M_1-\mu_1)(M_2-\mu_2)\big)\right] =$$

$$\frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma_{12} - nE\big((M_1-\mu_1)(M_2-\mu_2)\big) - nE\big((M_1-\mu_1)(M_2-\mu_2)\big) + nE\big((M_1-\mu_1)(M_2-\mu_2)\big)\right] =$$

$$\frac{1}{n-1}\big[n\sigma_{12} - nC(M_1, M_2)\big] = \frac{1}{n-1}(n\sigma_{12} - \sigma_{12}) = \sigma_{12}$$

$$C(M_1, M_2) = C\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{j=1}^{n} Y_j\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C(X_i, Y_j) = \frac{1}{n^2}\sum_{i=1}^{n} C(X_i, Y_i) = \frac{1}{n^2} n\sigma_{12} = \frac{1}{n}\sigma_{12}$$

(Kovariance $C(X_i, Y_j) = 0$ pro $i \neq j$, protože se jedná o náhodný výběr.)

**Příklad 3:** Nechť $X_{11},\ldots,X_{1n_1}$ a $X_{21},\ldots,X_{2n_2}$ jsou stochasticky nezávislé náhodné výběry, první z rozložení se střední hodnotou $\mu_1$ a rozptylem $\sigma^2$, druhý z rozložení se střední hodnotou $\mu_2$ a rozptylem $\sigma^2$. Označme $M_1$, $M_2$ výběrové průměry, $S_1^2$, $S_2^2$ výběrové rozptyly a

$S_*^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ vážený průměr výběrových rozptylů

a) Dokažte, že statistika $M_1 - M_2$ je nestranným odhadem parametrické funkce $\mu_1 - \mu_2$.
b) Dokažte, že $S_*^2$ je nestranným odhadem $\sigma^2$.

**Řešení:**

ad a) $E(M_1 - M_2) = E(M_1) - E(M_2) = \mu_1 - \mu_2$.

ad b)

$$E(S_*^2) = E\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right) = \frac{1}{n_1+n_2-2}\big[(n_1-1)E(S_1^2) + (n_2-1)E(S_2^2)\big] =$$

$$\frac{1}{n_1+n_2-2}\big[(n_1-1)\sigma^2 + (n_2-1)\sigma^2\big] = \sigma^2$$

**Příklad 4:** Nezávisle opakovaná laboratorní měření určité konstanty jsou charakterizována náhodným výběrem $X_1, \ldots, X_n$, $E(X_i) = \mu$, $D(X_i) = \sigma^2$, $i = 1, \ldots, n$. Uvažme statistiky

$M_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, $L_n = \dfrac{X_1 + X_n}{2}$.

a) Dokažte, že $M_n$ a $L_n$ jsou nestranné odhady konstanty $\mu$ a zjistěte, který z nich je lepší.
b) Dokažte, že $\{M_n\}_{n=1}^{\infty}$ a $\{L_n\}_{n=1}^{\infty}$ tvoří posloupnosti asymptoticky nestranných odhadů konstanty $\mu$.
c) Zjistěte, zda $\{M_n\}_{n=1}^{\infty}$ a $\{L_n\}_{n=1}^{\infty}$ tvoří posloupnosti konzistentních odhadů konstanty $\mu$.

**Řešení:**

ad a) $E(M_n) = \mu$ – viz příklad 1, bod (a).

$$E(L_n) = E\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}\left[E(X_1) + E(X_n)\right] = \frac{\mu + \mu}{2} = \mu$$

$D(M_n) = \frac{\sigma^2}{n}$ – viz příklad 1, bod (a).

$$D(L_n) = D\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4}\left[D(X_1) + D(X_n)\right] = \frac{\sigma^2 + \sigma^2}{4} = \frac{\sigma^2}{2}$$. Pro $n \geq 3$ je tedy lepším odhadem konstanty $\mu$ výběrový průměr $M_n$.

ad b) Asymptotická nestrannost plyne z nestrannosti.

ad c) Posloupnost asymptoticky nestranných odhadů je posloupností konzistentních odhadů, pokud posloupnost rozptylů konverguje k nule.

$\lim_{n\to\infty} D(M_n) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$, tedy posloupnost výběrových rozptylů je posloupností konzistentních odhadů konstanty $\mu$.

$\lim_{n\to\infty} D(L_n) = \lim_{n\to\infty} \frac{\sigma^2}{2} > 0$, tedy posloupnost $\{L_n\}_{n=1}^{\infty}$ není posloupností konzistentních odhadů konstanty $\mu$.

# Práce se systémem STATISTICA

**Téma: Ilustrace základních pojmů matematické statistiky**

**a) Průzkum chování výběrového průměru a výběrového rozptylu**

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení se střední hodnotou $\mu$ a rozptylem $\sigma^2$. Pak pro výběrový průměr $M = \frac{1}{n}\sum_{i=1}^{n} X_i$ a výběrový rozptyl $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2$ platí: $E(M) = \mu$, $D(M) = \sigma^2/n$, $E(S^2) = \sigma^2$. Tyto vlastnosti budeme ilustrovat na náhodném výběru rozsahu 100 z rozložení Rs(0,1). V tomto případě $E(X_i) = 1/2$, $D(X_i) = 1/12$.

1. Vytvořte nový datový soubor o 103 proměnných a 100 případech. Pomocí programu gener.svb, který si stáhnete z webové stránky, se naplní prvních 100 proměnných 100 realizacemi náhodných veličin $X_i \sim Rs(0,1)$, i=1, …, 100, do proměnné v101 se uloží pořadová čísla 1 až 100, do proměnné v102 (resp. v103) se uloží průměry (resp. rozptyly) proměnných v1 až v100. Proměnnou v101 přejmenujte na PORADI, v102 na PRUMER a v103 na ROZPTYL. Vzniklý datový soubor uložte pod názvem uniform.sta.
2. Graficky znázorněte hodnoty některé z proměnných v1, …, v100 (např. v1) a hodnoty proměnné PRUMER.
   Návod: Graphs – Scatterplots – Graph type Multiple – vypneme Linear fit – Variables X PORADI, Y v1, PRUMER, OK, OK.Vidíme, že hodnoty proměnné v1 se nacházejí mezi 0 a 1, zatímco hodnoty proměnné PRUMER se koncentrují v úzkém pásu kolem 0,5.
3. Pomocí Descriptive Statistics vypočtěte průměr a rozptyl např. proměnné v1 a proměnné PRUMER. Průměr proměnné v1 by měl být blízký 0,5, rozptyl 1/12 = 0,083. Průměr proměnné PRUMER by se měl blížit 0,5, zatímco rozptyl by měl být 100 x menší než 1/12, tj. 0,00083. Dále vypočtěte průměr proměnné ROZPTYL. Měl by se blížit 1/12 = 0,083.

**b) Ověření nestrannosti výběrové distribuční funkce**

Nechť $X_1, ..., X_n$ je náhodný výběr z rozložení s distribuční funkcí $\Phi(x)$. Pro libovolné, ale pevně zvolené reálné číslo x označme $F_n(x) = \frac{1}{n}\text{card}\{i; X_i \le x\}$ hodnotu výběrové distribuční funkce v bodě x. Pak $E(F_n(x)) = \Phi(x)$, tj. pro libovolné. Ale pevně zvolené reálné x je výběrová distribuční funkce nestranným odhadem distribuční funkce $\Phi(x)$. Tuto vlastnost budeme ilustrovat na náhodném výběru rozsahu 1000 z rozložení N(0,1).

1. Vytvořte nový datový soubor o dvou proměnných a 1000 případech.
2. Do proměnné v1 uložte 1000 realizací náhodné veličiny s rozložením N(0,1) tak, že v Long Name použijte příkaz =vnormal(rnd(1);0;1).
3. Hodnoty proměnné v1 setřiďte podle velikosti: Data - Sort.
4. Proměnnou v2 transformujte tak, že v Long Name použijte příkaz =v0/1000.
   Nakreslete dvourozměrný tečkový diagram, kde na osu x vyneste v1 a na osu y v2. Graf této výběrové distribuční funkce porovnejte s grafem distribuční funkce N(0,1) na intervalu -3 až 3. Ten získáte tak, že k datovému souboru přidáte proměnnou x a proměnnou y. Do Long Name proměnné x napište příkaz =6*(v0-1)/1000-3 a do Long Name proměnné y napište příkaz =INormal(x;0;1). Pomocí Scatterplot vykreslíte graf distribuční funkce rozložení N(0,1).

**c) Šířka intervalu spolehlivosti v závislosti na rozsahu výběru a v závislosti na riziku**
Nechť (d, h) je 100(1-α)% empirický interval spolehlivosti pro $h(\vartheta)$ zkonstruovaný pomocí číselných realizací $x_1, ..., x_n$ náhodného výběru $X_1, ..., X_n$ z rozložení $L(\vartheta)$.
- Při konstantním riziku klesá šířka h-d s rostoucím rozsahem náhodného výběru.
- Při konstantním rozsahu náhodného výběru klesá šířka h-d s rostoucím rizikem.

Tyto vlastnosti budeme ilustrovat na intervalu spolehlivosti sestrojeném pro střední hodnotu μ rozložení N(0,1).

**Sledování vlivu rozsahu výběru na šířku intervalu spolehlivosti (při α=0,05)**
Pro hypotetické náhodné výběry rozsahu n (n = 5, 7, 9, …, 85) z rozložení N(0,1), jejichž výběrové průměry se vždy realizovaly hodnotou 0, vypočtěte dolní a horní meze 95% intervalů spolehlivosti pro μ a graficky znázorněte závislost těchto mezí na rozsahu n.
**Návod:** Program intsp1.svb otevřete v programovacím okně. Vytvořte nový datový soubor o 3 proměnných a 41 případech. Po spuštění programu intsp1 se do proměnné v1 uloží rozsahy výběrů 5, 7, ..., 85, do v2 (resp. v3) dolní (resp. horní) meze 95% intervalů spolehlivosti pro μ. Vytvoření grafu: Graphs – Scatterplots – Graph type Multiple – vypnout Linear fit – Variables X v1, Y v2, v3, OK, OK.

**Sledování vlivu rizika na šířku intervalu spolehlivosti (při konstatním rozsahu výběru)**
Pro hypotetický náhodný výběr rozsahu n=25 z rozložení N(0,1), jehož výběrový průměr se realizoval hodnotou 0, vypočtěte dolní a horní meze 100(1-α)% intervalů spolehlivosti (α=0,20, 0,19, …, 0,01) pro μ a graficky znázorněte závislost těchto mezí na riziku α.
**Návod:** Program intsp2.svb otevřete v programovacím okně. Vytvořte nový datový soubor o 3 proměnných a 20 případech. Po spuštění programu intsp2 se do proměnné v1 uloží rizika 0,20, 0,19, ..., 0,01, do v2 (resp. v3) dolní (resp. horní) meze 100(1-α)% intervalů spolehlivosti pro μ.Vytvoření grafu: stejným způsobem jako v předešlém případě.