KAPITOLA 1

Úvod do matematické statistiky

1. NÁHODNÝ VÝBĚR A VÝBĚROVÉ CHARAKTERISTIKY

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

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

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

Definujme nejprve základní pojmy matematické statistiky. Základním pojmem matematické statistiky je pojem náhodného výběru.

DEFINICE 1.1. Náhodný vektor $\mathbf{X}_n = (X_1, \ldots, X_n)'$ nazýváme **náhodným výběrem z rozdělení pravděpodobnosti** P, pokud

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

Číslo n nazýváme **rozsah náhodného výběru**. Libovolný bod $\mathbf{x}_n = (x_1, \ldots, x_n)'$, kde x_i je realizace náhodné veličiny X_i $(i = 1, \ldots, n)$, budeme nazývat **realizací náhodného výběru** $\mathbf{X}_n = (X_1, \ldots, X_n)'$. Množinu všech hodnot, kterých může náhodný výběr nabýt, nazýváme **výběrový prostor** a budeme jej značit $\mathcal{X}$.

Základní dělení matematické statistiky je dané strukturou množiny všech možných rozdělení (označme ji $\mathcal{P}$) náhodného výběru $\mathbf{X}$. Velmi často vybíráme do množiny $\mathcal{P}$ jen rozdělení, která jsou stejného typu a která závisí pouze na nějakém (skalárním či vícerozměrném) parametru. Tento parametr se většinou značí $\boldsymbol{\theta}$ a pravděpodobnostní míry z množiny $\mathcal{P}$ symbolem $P_{\boldsymbol{\theta}}$. Přitom předpokládáme, že parametr $\boldsymbol{\theta}$ nabývá hodnot z nějaké množiny $\boldsymbol{\Theta}$.

DEFINICE 1.2. Množinu $\mathcal{P}$ pravděpodobnostních měr tvaru

$$\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

nazýváme **parametrickou třídou rozdělení**. Vektor $\boldsymbol{\theta}$ nazýváme **parametrem rozdělení pravděpodobnosti** $P_{\boldsymbol{\theta}}$ a množinu $\boldsymbol{\Theta}$ možných hodnot parametru $\boldsymbol{\theta}$ **parametrický prostor**.

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

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x; \boldsymbol{\theta}).$$

Nyní se zmiňme o tzv. rodinách rozdělení.

DEFINICE 1.3. Nechť $g(x)$ je nějaká hustota. Definujme **rodiny rozdělení**

$$\mathcal{F}_1 = \{f(x;\theta) = g(x-\theta); \theta \in \mathbb{R}\}$$
$$\mathcal{F}_2 = \left\{f(x;\delta) = \tfrac{1}{\delta} g\left(\tfrac{x}{\delta}\right); \delta > 0\right\}$$
$$\mathcal{F}_3 = \left\{f(x;\theta,\delta) = \tfrac{1}{\delta} g\left(\tfrac{x-\theta}{\delta}\right); \theta \in \mathbb{R}, \delta > 0\right\}$$

Pak říkáme, že $\boxed{\mathcal{F}_1}$ je **rodina s parametrem polohy** (*location family*), $\boxed{\mathcal{F}_2}$ je **rodina s parametrem měřítka** (*scale family*) a $\boxed{\mathcal{F}_3}$ je **rodina s parametrem polohy a měřítka** (*location-scale family*).

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. $\gamma(\boldsymbol{\theta})$.

Funkci $\gamma(\boldsymbol{\theta})$ nazýváme **parametrickou funkcí**. V matematické statistice se pro funkce, pomocí kterých budeme odhady provádět, nazývají statistikou. (Tyto funkce jsou navíc měřitelné).

DEFINICE 1.4. Libovolnou náhodnou veličinu T_n, která vznikne jako funkce náhodného výběru $\mathbf{X}_n = (X_1, \ldots, X_n)'$, budeme nazývat **statistikou**, tj. $T_n = T(X_1, \ldots, X_n)'$.

Příklad 1.5. Výběrová (empirická) distribuční funkce.

Ukážeme, jakým způsobem lze například informaci obsaženou v náhodném výběru zužitkovat k popisu **distribuční funkce**. Mějme $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x;\boldsymbol{\theta})$.

Zaveďme tzv. **indikátor množiny** předpisem: $I_B(x) = \begin{cases} 1 & x \in B, \\ 0 & x \notin B \end{cases}$

a pro $x \in \mathbb{R}$ **indikátor jevu**: $I_i(x) = I_{(-\infty,x>}(X_i) = \begin{cases} 1 & X_i \le x, \\ 0 & X_i > x. \end{cases}$ pro $i = 1, \ldots, n.$

Potom $I_1(x), \ldots, I_n(x)$ jsou nezávislé náhodné veličiny se stejným alternativním rozdělením pravděpodobností s parametrem $\pi \in (0,1)$, tj. $\perp\!\!\!\perp\{I_1, \ldots, I_n\} \simeq A(\pi)$. Parametr π je roven pravděpodobnosti úspěchu, tj.

$$P(I_i(x) = 1) = P(X_i \le x) = F(x;\boldsymbol{\theta}) \quad \Rightarrow \quad \boxed{\perp\!\!\!\perp\{I_1, \ldots, I_n\} \simeq A(\pi = F(x;\boldsymbol{\theta}))}.$$

Položme

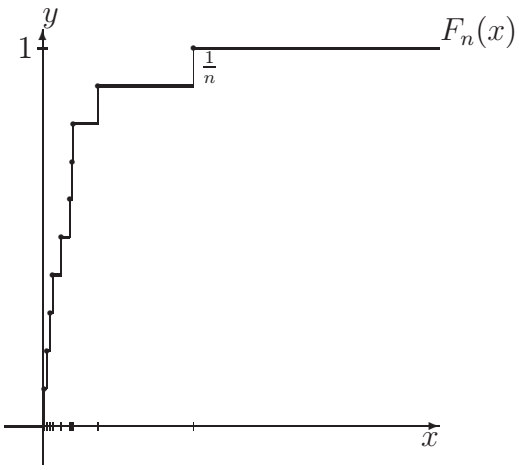
$$\begin{aligned} Y(x) &= \textstyle\sum_{i=1}^n I_i(x) \\ F_n(x) &= \frac{Y(x)}{n} \end{aligned}$$

a postupně počítejme

$$\boxed{EF_n(x)} = E\tfrac{Y(x)}{n} = \tfrac{1}{n} Y_n = \tfrac{1}{n} \sum_{i=1}^n I_i(x) = \tfrac{1}{n} \cdot n\, F(x;\boldsymbol{\theta}) = \boxed{F(x;\boldsymbol{\theta})}.$$

Protože posloupnost $\{F_n(x)\}_{n=1}^\infty$ splňuje jak slabý, tak silný zákon velkých čísel, tak platí

$$\boxed{\begin{aligned} \lim_{n\to\infty} P(|F_n(x) - F(x;\boldsymbol{\theta})| \ge \varepsilon) &= 0 \\ P(\lim_{n\to\infty} F_n(x) = F(x;\boldsymbol{\theta})) &= 1 \end{aligned}}.$$



Z uvedených vztahů je vidět, že pokud rozsah výběru bude dostatečně velký, lze **distribuční funkci rozdělení**, z něhož výběr pochází, **dostatečně přesně aproximovat** pomocí **výběrové (empirické) distribuční funkce**.

Předpokládejme, že rozdělení, z něhož výběr pochází, má konečné druhé momenty se střední hodnotou μ a rozptylem σ^2, což budeme dále značit

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu, \sigma^2).$$

Tedy pro každé $i = 1, \ldots, n$ platí

$$\begin{array}{rcl} EX_i & = & \mu \\ DX_i & = & \sigma^2 \end{array} .$$

Potom tyto charakteristiky zřejmě závisí na parametru $\boldsymbol{\theta}$, neboť

$$\begin{array}{rcl} \mu & = & \int_{-\infty}^{\infty} x dF(x;\boldsymbol{\theta}) \\ \sigma^2 & = & \int_{-\infty}^{\infty} (x-\mu)^2 dF(x;\boldsymbol{\theta}) \end{array} ,$$

proto bude lépe značit je $\boxed{\mu(\boldsymbol{\theta})}$ a $\boxed{\sigma^2(\boldsymbol{\theta})}$ místo μ a σ^2.

Všimněme si dále, že pro každé $x \in \mathbb{R}$ je $\boxed{F_n(x) = F_n(X_1, \ldots, X_n)}$ statistikou, tím také náhodnou veličinou (která nabývá hodnot mezi nulou a jedničkou) a tím i funkcí elementárního jevu $\omega \in \Omega$.

Zvolíme-li ω libovolně, ale pevně a uvažujeme-li $\boxed{F_n(x)}$ jako funkci proměnné x, pak lze snadno odvodit, že je tato funkce **distribuční funkcí** nějaké náhodné veličiny a lze zavést její střední hodnotu a rozptyl

$$\begin{array}{rcl} \mu_n & = & \int_{-\infty}^{\infty} x dF_n(x;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \sigma_n^2 & = & \int_{-\infty}^{\infty} (x-\mu)^2 dF(x;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_n)^2 \end{array} .$$

Zřejmě μ_n a σ_n^2 jsou borelovské funkce náhodného výběru a tedy statistiky a lze je považovat za odhady parametrických funkcí $\mu(\boldsymbol{\theta})$ a $\sigma^2(\boldsymbol{\theta})$. Lze očekávat, že čím bude rozsah náhodného výběru větší, tím bude odhad uvedených parametrických funkcí kvalitnější.

Poznámka 1.6.
Odhadem parametrické funkce $\boxed{\gamma(\boldsymbol{\theta})}$ budeme rozumět nějakou statistiku $\boxed{T_n = T(X_1, \ldots, X_n)'}$, která bude pro různé náhodné výběry kolísat kolem $\gamma(\boldsymbol{\theta})$.

Statistika $T_n = T(X_1, \ldots, X_n)'$ závisí na parametru $\boldsymbol{\theta}$ prostřednictvím distribuční funkce rozdělení, z něhož výběr pochází.

Také rozdělení této statistiky, tj. náhodné veličiny, závisí na parametru $\boldsymbol{\theta}$.

Proto střední hodnotu a rozptyl této statistiky budeme značit $\boxed{E_{\boldsymbol{\theta}} T_n}$ a $\boxed{D_{\boldsymbol{\theta}} T_n}$.

DEFINICE 1.7. VÝBĚROVÉ CHARAKTERISTIKY. Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr rozsahu n z rozdělení s distribuční funkcí $F(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Potom statistika

$\bar{X}_n = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$	se nazývá	**výběrový průměr**
$S_n^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$		**výběrový rozptyl**
$S_n = S = \sqrt{S_n^2} = \sqrt{S^2}$		**výběrová směrodatná odchylka**
$F_n(x) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,x>}(X_i)$		**výběrová (empirická) distribuční funkce**

2. NESTRANNOST, VÝCHÝLENÍ, KONZISTENCE ODHADŮ

Za lepší odhad se považuje ten, jehož rozdělení je více koncentrované okolo neznámé hodnoty parametru. Tento přirozený požadavek koncentrace rozdělení T_n okolo skutečné hodnoty parametru vyjadřujeme pomocí střední hodnoty a rozptylu.

DEFINICE 2.1. Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ 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 $T_n = T(X_1, \ldots, X_n)'$ je

nestranným (nevychýleným)	odhadem parametrické funkce $\gamma(\boldsymbol{\theta})$	pokud pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ 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		$\lim_{n\to\infty} E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$.
slabě konzistentním		pokud pro $\forall \varepsilon > 0$ platí $\lim_{n\to\infty} P_{\boldsymbol{\theta}}(\lvert T_n - \gamma(\boldsymbol{\theta})\rvert > \varepsilon) = 0$ tj. $T_n \xrightarrow{P_{\boldsymbol{\theta}}} \gamma(\boldsymbol{\theta})$
silně konzistentním		$P_{\boldsymbol{\theta}}(\lim_{n\to\infty} T_n = \gamma(\boldsymbol{\theta})) = 1$ tj. $T_n \xrightarrow{s.j.} \gamma(\boldsymbol{\theta})$

Poznámka 2.2.

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

Poznámka 2.3.

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**.

Příklad 2.4. Geometrické rozdělení.
Nechť náhodná veličina X má geometrické rozdělení,

$$f_X(x) = P(X = x) = (1-\theta)^x\theta \qquad 0 < \theta < 1 \;\; x = 0, 1, \ldots$$

Veličina X udává počet neúspěchů při výběru z alternativního rozdělení před výskytem prvního úspěchu. Hledejme nestranný odhad pro θ.

Je-li $T(X)$ takový nestranný odhad, musí pro něj platit

$$\boxed{E_\theta T(X)} = \sum_{x=0}^{\infty} T(x)(1-\theta)^x\theta = \boxed{\theta} \qquad 0 < \theta < 1,$$

Odtud dostáváme

$$\sum_{x=0}^{\infty} T(x)(1-\theta)^x = 1 \qquad 0 < \theta < 1,$$

takže musí platit

$$\begin{aligned} T(0) &= 1 \\ T(x) &= 0 \qquad \text{pro } x \geq 1. \end{aligned}$$

Tento odhad však není pokládán za vhodný, protože jen minimálně přihlíží k počtu neúspěchů před prvním úspěchem. Závisí jen na tom, zda úspěch nastal hned v prvním pokusu či nikoli.

Může se také stát, že nestranný odhad neexistuje.

Příklad 2.5. Parametrická funkce $\boxed{\frac{1}{\theta}}$ v případě binomického rozdělení.

Nechť náhodná veličina X má binomické rozdělení, tj. $X \sim Bi(n, \theta)$ a

$$f_X(x) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad n \geq 1, \;\; 0 < \theta < 1 \;\; x = 0, 1, \ldots, n.$$

Sporem ukážeme, že neexistuje nestranný odhad pro parametrickou funkci

$$\boxed{\gamma(\theta) = \frac{1}{\theta}}.$$

Nechť existuje taková funkce T, že pro každé $\theta \in (0, 1)$ platí

$$\boxed{E_\theta T(X)} = \sum_{x=0}^{n} T(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = \boxed{\frac{1}{\theta}} \qquad 0 < \theta < 1.$$

Na levé straně je však polynom proměnné θ nejvýše stupně n, který samozřejmě nemůže být identicky roven $\frac{1}{\theta}$ na intervalu $(0, 1)$.

Nyní vyšetříme případ, kdy odhadovanými parametry jsou **střední hodnota** a **rozptyl** rozdělení, ze kterého náhodný výběr pochází.

Věta 2.6. *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má střední hodnotu $\mu(\boldsymbol{\theta})$ pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Pak **výběrový průměr** je **nestranným odhadem** střední hodnoty, tj.*

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

Důkaz. Počítejme

$$E_{\boldsymbol{\theta}}\bar{X} = E_{\boldsymbol{\theta}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{\theta}}X_i = \frac{1}{n}\sum_{i=1}^{n}\mu(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta}).$$

□

VĚTA 2.7. *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má rozptyl $\sigma^2(\boldsymbol{\theta})$ pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$. Pak **výběrový rozptyl** je **nestranným odhadem** rozptylu, tj.*

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

Důkaz. Nejprve upravujme

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}\left[(X_i - \mu(\boldsymbol{\theta})) - (\bar{X} - \mu(\boldsymbol{\theta}))\right]^2 \\
&= \sum_{i=1}^{n}\left[(X_i - \mu(\boldsymbol{\theta}))^2 - 2(X_i - \mu(\boldsymbol{\theta}))(\bar{X} - \mu(\boldsymbol{\theta})) + (\bar{X} - \mu(\boldsymbol{\theta}))^2\right] \\
&= \sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 - 2(\bar{X} - \mu(\boldsymbol{\theta}))\underbrace{\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))}_{=n(\bar{X}-\mu(\boldsymbol{\theta}))} + n(\bar{X} - \mu(\boldsymbol{\theta}))^2 \\
&= \sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 - n(\bar{X} - \mu(\boldsymbol{\theta}))^2.
\end{aligned}$$

Pak počítejme

$$\begin{aligned}
E_{\boldsymbol{\theta}} S^2 = E_{\boldsymbol{\theta}}\left[\tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= \tfrac{1}{n-1} E_{\boldsymbol{\theta}}\left[\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 + n(\bar{X} - \mu(\boldsymbol{\theta}))^2\right] \\
= \tfrac{1}{n-1}\sum_{i=1}^{n}\Bigg[\underbrace{E_{\boldsymbol{\theta}}(X_i - \mu(\boldsymbol{\theta}))^2}_{=DX_i=\sigma^2(\boldsymbol{\theta})} &- n\underbrace{E_{\boldsymbol{\theta}}(\bar{X} - \mu(\boldsymbol{\theta}))^2}_{=D_{\boldsymbol{\theta}}\bar{X}}\Bigg]
\end{aligned}$$

Proto vypočtěme

$$D_{\boldsymbol{\theta}}\bar{X} = D_{\boldsymbol{\theta}}\left(\tfrac{1}{n}\sum_{i=1}^{n}X_i\right) \overset{nez.}{=} \tfrac{1}{n^2}\sum_{i=1}^{n}D_{\boldsymbol{\theta}}X_i = \frac{\sigma^2(\boldsymbol{\theta})}{n}$$

a celkově dostaneme

$$E_{\boldsymbol{\theta}} S^2 = \tfrac{1}{n-1}\left[n\sigma^2(\boldsymbol{\theta}) - \sigma^2(\boldsymbol{\theta})\right] = \sigma^2(\boldsymbol{\theta}).$$

□

Následující věta udává postačující podmínku pro konzistentní odhad.

VĚTA 2.8. *Nechť statistika $T_n = T(X_1, \ldots, X_n)'$ je nestranný nebo asymptoticky nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ a platí*

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

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

Důkaz. Nechť $\varepsilon > 0$. Z Čebyševovy nerovnosti plyne:

$$P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}} T_n| \geq \tfrac{\varepsilon}{2}) \leq \tfrac{4 D_{\boldsymbol{\theta}} T_n}{\varepsilon^2}.$$

Protože buď $E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$ nebo $\lim_{n\to\infty} E_{\boldsymbol{\theta}} T_n = \gamma(\boldsymbol{\theta})$, pak existuje přirozené číslo n_0 tak, že pro $\forall n > n_0$ platí:

$$-\tfrac{\varepsilon}{2} < \gamma(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}} T_n < \tfrac{\varepsilon}{2}.$$

Dále platí

$$\begin{aligned}
P_{\boldsymbol{\theta}}(|T_n - \gamma(\boldsymbol{\theta})| \geq \varepsilon) &= 1 - P_{\boldsymbol{\theta}}(|T_n - \gamma(\boldsymbol{\theta})| < \varepsilon) = 1 - P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}}T_n + ET_n - \gamma(\boldsymbol{\theta})| < \varepsilon) \\
&\leq 1 - P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}}T_n| + |ET_n - \gamma(\boldsymbol{\theta})| < \varepsilon) \\
&\leq 1 - P_{\boldsymbol{\theta}}(\{|T_n - E_{\boldsymbol{\theta}}T_n| < \tfrac{\varepsilon}{2}\} \cup \{|ET_n - \gamma(\boldsymbol{\theta})| < \tfrac{\varepsilon}{2}\}) \\
&\leq 1 - P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}}T_n| < \tfrac{\varepsilon}{2}) - P(|ET_n - \gamma(\boldsymbol{\theta})| < \tfrac{\varepsilon}{2}) \\
&\leq 1 - P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}}T_n| < \tfrac{\varepsilon}{2}) = P_{\boldsymbol{\theta}}(|T_n - E_{\boldsymbol{\theta}}T_n| \geq \tfrac{\varepsilon}{2}) \leq \tfrac{4D_{\boldsymbol{\theta}}T_n}{\varepsilon^2}
\end{aligned}$$

a tedy

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|T_n - \gamma(\boldsymbol{\theta})| \geq \varepsilon) \leq \tfrac{4}{\varepsilon^2} \lim_{n\to\infty} D_{\boldsymbol{\theta}}T_n = 0.$$

Tedy T_n je **slabě konzistentním odhadem** $\gamma(\boldsymbol{\theta})$. □

Důsledek 2.9. *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ střední hodnotu $\mu(\boldsymbol{\theta})$ a rozptyl $\sigma^2(\boldsymbol{\theta})$, tj.*

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

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

Důkaz. Vzhledem k tomu, že $\bar{X}$ je nestranným odhadem $\mu(\boldsymbol{\theta})$ a platí

$$\lim_{n\to\infty} D_{\boldsymbol{\theta}}\bar{X} = \lim_{n\to\infty} D_{\boldsymbol{\theta}} \left(\tfrac{1}{n} \sum_{i=1}^{n} X_i \right) \overset{nez.}{=} \lim_{n\to\infty} \tfrac{1}{n^2} \sum_{i=1}^{n} D_{\boldsymbol{\theta}}X_i = \lim_{n\to\infty} \frac{\sigma^2(\boldsymbol{\theta})}{n} = 0$$

tj. rozptyl konverguje k nule, jsou splněny předpoklady předchozí věty a platí tak tvrzení. □

Důsledek 2.10. *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ střední hodnotu $\mu(\boldsymbol{\theta})$ a rozptyl $\sigma^2(\boldsymbol{\theta})$, tj.*

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

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

Důkaz. Víme již, že statistika S^2 je nestranným odhadem $\sigma^2(\boldsymbol{\theta})$. Nyní budeme muset vypočítat rozptyl statistiky S^2, což není zdaleka tak triviální jako v případě výběrového průměru. Pro lepší přehlednost budeme psát místo $\mu(\boldsymbol{\theta})$ a $\sigma^2(\boldsymbol{\theta})$ pouze μ a σ^2, u středních hodnot $E_{\boldsymbol{\theta}}$ a rozptylu $D_{\boldsymbol{\theta}}$ také vynecháme parametr $\boldsymbol{\theta}$.

Položme

$$Y_i = (X_i - \mu)^2$$

$$S_0^2 = \tfrac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

a počítejme

$$S_0^2 = \tfrac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 = \tfrac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}.$$

Pak

$$EY_i = E(X_i - \mu)^2 = DX_i = \sigma^2$$
$$DY_i = EY_i^2 - (EY_i)^2 = E(X_i - \mu)^4 - \sigma^4 = \mu_4 - \sigma^4$$
$$ES_0^2 = E\bar{Y} = \tfrac{1}{n}\sum_{i=1}^{n} EY_i = \sigma^2 \tag{1}$$
$$DS_0^2 = D\left(\tfrac{1}{n}\sum_{i=1}^{n} Y_i\right) \overset{nez.}{=} \tfrac{1}{n^2}\sum_{i=1}^{n} DY_i = \frac{\mu_4 - \sigma^4}{n} \tag{2}$$

Označme

$$S_\star^2 = \tfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \tfrac{n-1}{n}S^2,$$

takže

$$S^2 = \tfrac{n}{n-1}S_\star^2. \tag{3}$$

Pak

$$\begin{aligned} S_\star^2 &= \tfrac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu) - (\bar{X} - \mu)\right]^2 \\ &= \tfrac{1}{n}\sum_{i=1}^{n}\left[(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right] \\ &= \underbrace{\tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}_{S_0^2} - \tfrac{2}{n}(\bar{X} - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \mu)}_{n\bar{X} - n\mu} + \tfrac{1}{n}n(\bar{X} - \mu)^2 \\ &= S_0^2 - (\bar{X} - \mu)^2 \end{aligned} \tag{4}$$

Počítejme nejprve

$$ES_\star^2 \overset{viz(4)}{=} E\left[S_0^2 - (\bar{X} - \mu)^2\right] = ES_0^2 - \underbrace{E(\bar{X} - \mu)^2}_{D\bar{X}} = \sigma^2 - \tfrac{\sigma^2}{n} = \tfrac{n-1}{n}\sigma^2$$
$$ES^2 \overset{viz(3)}{=} E\left[\tfrac{n}{n-1}S_\star^2\right] = \tfrac{n}{n-1}\tfrac{n-1}{n}\sigma^2 = \sigma^2.$$

Připomeňme, že rozptyl lze počítat pomocí vzorce

$$DS_\star^2 = ES_\star^4 - \left[ES_\star^2\right]^2,$$

a protože $ES_\star^2$ již známe, počítejme nyní

$$\begin{aligned} ES_\star^4 &\overset{viz(4)}{=} E[S_0^2 - (\bar{X} - \mu)^2]^2 = E[S_0^4 - 2S_0^2(\bar{X} - \mu)^2 + (\bar{X} - \mu)^4] \\ &= \underbrace{ES_0^4}_{(a)} - 2\underbrace{ES_0^2(\bar{X} - \mu)^2}_{(b)} + \underbrace{E(\bar{X} - \mu)^4}_{(c)}. \end{aligned} \tag{5}$$

Při výpočtu výrazu (a) ve vzorci (5) vyjdeme opět ze vztahu

$$DS_0^2 = ES_0^4 - \left(ES_0^2\right)^2,$$

takže

$$ES_0^4 = DS_0^2 + \left(ES_0^2\right)^2 = \tfrac{\mu_4 - \sigma^4}{n} + \sigma^4 = \tfrac{\mu_4}{n} + \tfrac{n-1}{n}\sigma^4.$$

Dále počítejme výraz (b) ve vzorci (5)

$$\begin{aligned}
E[S_0^2(\bar{X}-\mu)^2] &= \tfrac{1}{n^3} E\left\{\left[\sum_{i=1}^n (X_i-\mu)^2\right]\left[\sum_{i=1}^n (X_i-\mu)\right]^2\right\} \\
&= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n E[(X_i-\mu)^2(X_j-\mu)(X_k-\mu)] \\
&= \tfrac{1}{n^3}\sum_{i=1}^n \underbrace{E[(X_i-\mu)^4]}_{=\mu_4} \\
&\quad + \tfrac{1}{n^3}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\sum_{k=1,k\neq i,j}^n \underbrace{E[(X_i-\mu)^2(X_j-\mu)(X_k-\mu)]}_{=0 \quad \text{viz}^1} \\
&\quad + \underbrace{\tfrac{1}{n^3}\sum_{i=1}^n\sum_{j=1,i\neq j}^n E[(X_i-\mu)^2(X_j-\mu)^2]}_{=n(n-1)\sigma^4 \quad \text{viz}^2} \\
&= \tfrac{n\mu_4}{n^3} + \tfrac{n(n-1)\sigma^4}{n^3} \\
&= \tfrac{1}{n^2}\left[\mu_4 + (n-1)\sigma^4\right].
\end{aligned}$$

Ještě zbývá vypočítat poslední výraz (c) ve vzorci (5)

$$\begin{aligned}
E[(\bar{X}-\mu)^4] &= E\left[\frac{1}{n}\sum_{i=1}^n (X_i-\mu)\right]^4 \\
&= \tfrac{1}{n^4}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n\sum_{h=1}^n E[(X_i-\mu)(X_j-\mu)(X_k-\mu)(X_h-\mu)] \\
&= \tfrac{1}{n^4}\sum_{i=1}^n \underbrace{E[(X_i-\mu)^4]}_{=\mu_4} + \tfrac{1}{n^4}\underbrace{3\sum_{s=1}^n\sum_{t=1,t\neq s}^n E[(X_s-\mu)^2(X_t-\mu)^2]}_{=3n(n-1)\sigma^4 \quad \text{viz}^3} \\
&= \tfrac{1}{n^3}\left[\mu_4 + 3(n-1)\sigma^4\right]
\end{aligned}$$

Nyní předchozí tří výpočty můžeme shrnout a dostaneme

$$\begin{aligned}
ES_\star^4 &= \tfrac{\mu_4}{n} + \tfrac{n-1}{n}\sigma^4 - 2\left(\tfrac{\mu_4}{n^2} + \tfrac{n-1}{n^2}\sigma^4\right) + \tfrac{\mu_4}{n^3} + 3\tfrac{n-1}{n^3}\sigma^4 \\
&= \tfrac{(n-1)^2}{n^3}\mu_4 + \tfrac{(n-1)(n^2-2n+3)}{n^3}\sigma^4
\end{aligned}$$

[1] Díky nezávislosti náhodných veličin X_i, X_j a X_k máme: $E[(X_i-\mu)^2(X_j-\mu)(X_k-\mu)] = E(X_i-\mu)^2 E(X_j-\mu)E(X_k-\mu) = 0$, protože $E(X_i-\mu)^{2k+1} = 0$.

[2] Opět z nezávislosti náhodných veličin X_i a X_j plyne: $E[(X_i-\mu)^2(X_j-\mu)^2] = E(X_i-\mu)^2 E(X_j-\mu)^2 = \sigma^4$.

[3] Pouze v případech, kdy (1.) $s = i = j \wedge t = k = h \wedge s \neq t$, (2.) $s = i = k \wedge t = j = h \wedge s \neq t$ a (3.) $s = i = h \wedge t = j = k \wedge s \neq t$ dostaneme: $E[(X_s-\mu)^2(X_t-\mu)^2] = E(X_s-\mu)^2\ E(X_t-\mu)^2 = \sigma^4$, a to zase díky nezávislosti náhodných veličin X_t a X_s.

Nyní ještě spočtěme

$$\begin{aligned} DS_\star^2 &= \tfrac{(n-1)^2}{n^3}\mu_4 + \tfrac{(n-1)(n^2-2n+3)}{n^3}\sigma^4 - \left(\tfrac{n-1}{n}\sigma^2\right)^2 \\ &= \tfrac{(n-1)^2}{n^3}\mu_4 - \tfrac{(n-1)(n-3)}{n^3}\sigma^4 \end{aligned}$$

a konečně

$$DS^2 = (\tfrac{n}{n-1})^2 DS_\star^2 = \tfrac{\mu_4}{n} - \tfrac{n-3}{n(n-1)}\sigma^4.$$

Odtud snadno ukážeme, že rozptyl statistiky S^2 konverguje k nule, čímž je tvrzení dokázáno

$$\lim_{n\to\infty} DS^2 = \lim_{n\to\infty}\left(\tfrac{\mu_4}{n} - \tfrac{n-3}{n(n-1)}\sigma^4\right) = 0.$$

□

VĚTA 2.11. *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení, které má pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ střední hodnotu $\mu(\boldsymbol{\theta})$ a rozptyl $\sigma^2(\boldsymbol{\theta})$, tj.*

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta})).$$

Potom

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

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

Důkaz. Připomeňme nejprve, že náhodný výběr $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$ představuje nezávislé stejně rozdělené náhodné veličiny s konečnou střední hodnotou a rozptylem.

(i) Vzhledem k tomu, že $\bar{X} = \bar{X}_n$ je nestranným odhadem $\mu(\boldsymbol{\theta})$, tj. $E_{\boldsymbol{\theta}}\bar{X} = \mu(\boldsymbol{\theta})$, pak posloupnost $\{\bar{X}_n = \frac{1}{n}\sum_{n=1}^{\infty} X_i\}_{n=1}^{\infty}$ splňuje silný zákon velkých čísel, tj. platí

$$P_{\boldsymbol{\theta}}(\lim_{n\to\infty} \bar{X}_n = \mu(\boldsymbol{\theta})) = 1, \quad \text{pro } \forall \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

takže **výběrový průměr** $\bar{X}$ je **silně konzistentním odhadem** $\mu(\boldsymbol{\theta})$.

(ii) Připomeňme, že platí

$$\begin{aligned} S^2 = S_n^2 &= \tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}\left[(X_i - \mu(\boldsymbol{\theta})) - (\bar{X} - \mu(\boldsymbol{\theta}))\right]^2 \\ &= \tfrac{1}{n-1}\sum_{i=1}^{n}\left[(X_i - \mu(\boldsymbol{\theta}))^2 - 2(X_i - \mu(\boldsymbol{\theta}))(\bar{X} - \mu(\boldsymbol{\theta})) + (\bar{X} - \mu(\boldsymbol{\theta}))^2\right] \\ &= \tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 - 2(\bar{X} - \mu(\boldsymbol{\theta}))\tfrac{1}{n-1}\underbrace{\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))}_{=n(\bar{X}-\mu(\boldsymbol{\theta}))} + \tfrac{1}{n-1}n(\bar{X} - \mu(\boldsymbol{\theta}))^2 \\ &= \tfrac{n}{n-1}\left[\tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 - (\bar{X} - \mu(\boldsymbol{\theta}))^2\right]. \end{aligned} \tag{6}$$

Náhodné veličiny

$$Y_i = (X_i - \mu(\boldsymbol{\theta}))^2$$

jsou nezávislé stejně rozdělené se střední hodnotou $E_{\boldsymbol{\theta}}Y_i = E_{\boldsymbol{\theta}}(X_i - \mu(\boldsymbol{\theta}))^2 = \sigma^2(\boldsymbol{\theta})$, takže posloupnost

$$\left\{\tfrac{1}{n}\sum_{i=1}^{n} Y_i = \tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2\right\}_{i=1}^{n}$$

splňuje silný zákon velkých čísel, tj. platí

$$P_{\boldsymbol{\theta}}(\lim_{n\to\infty} \tfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu(\boldsymbol{\theta}))^2 = \sigma^2(\boldsymbol{\theta})) = 1.$$

Protože také platí

$$P_{\boldsymbol{\theta}}(\lim_{n\to\infty} \bar{X}_n = \mu(\boldsymbol{\theta})) = P_{\boldsymbol{\theta}}(\lim_{n\to\infty} \bar{X}_n - \mu(\boldsymbol{\theta})) = 0) = 1,$$

takže celkově, využijeme-li vztah (6), dostáváme

$$P_{\boldsymbol{\theta}}(\lim_{n\to\infty} S_n^2 = \sigma^2(\boldsymbol{\theta})) = 1, \quad \text{pro } \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

takže **výběrový rozptyl** S_n^2 je **silně konzistentním odhadem** $\sigma^2(\boldsymbol{\theta})$.

□

Poznámka 2.12. Více nestranných odhadů.

Obecně může existovat více nestranných odhadů. Například nejen výběrový průměr $\bar{X}$ je nestranným odhadem střední hodnoty $\mu(\boldsymbol{\theta})$, ale i každé jednotlivé pozorování X_i nebo každá jeho lineární kombinace $\sum_{i=1}^{n} c_i X_i$, pro kterou platí $\sum_{i=1}^{n} c_i = 1$.

Pokud tedy existuje více nestranných odhadů je přirozenou otázkou, který z nich je nejlepší.

Za nejlepší můžeme považovat ten, který má **nejmenší rozptyl** mezi všemi nestrannými odhady.

Rozdělení každé statistiky však závisí na parametru $\boldsymbol{\theta}$, z čehož vyplývá, že i rozptyl nestranné statistiky T_n závisí na parametru $\boldsymbol{\theta}$.

Může se stát, že odhad minimalizující rozptyl při určité hodnotě parametru není vhodný pro jinou hodnotu parametru - existuje jiný nestranný (nevychýlený) odhad, který má při této hodnotě parametru menší rozptyl.

Pokud taková situace nenastane, mluvíme o rovnoměrně nejlepším nestranném odhadu.

DEFINICE 2.13. Nechť T_n je nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ a pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ 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 2.14. Nejlepší nestranný lineární odhad střední hodnoty $\mu(\boldsymbol{\theta})$.

Jak jsme již dříve spočítali, pro náhodný výběr $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$ platí, že střední hodnota výběrového průměru $\bar{X}$ je rovna

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

a rozptyl výběrového průměru $\bar{X}$ je roven

$$D_{\boldsymbol{\theta}} \bar{X} = \frac{\sigma^2(\boldsymbol{\theta})}{n}.$$

Tedy variabilita této statistiky je n krát menší než variabilita jednotlivých pozorování $X_1, \ldots, X_n$ a tedy hodnoty statistiky $\bar{X}$ jsou více koncentrovány kolem odhadované střední hodnoty $\mu(\boldsymbol{\theta})$ než jednotlivá pozorování $X_1, \ldots, X_n$. Navíc je statistika $\bar{X}$ je lineární funkcí náhodných veličin $X_1, \ldots, X_n$.

Uvažujme všechny **lineární statistiky** tvaru $\sum_{i=1}^n c_i X_i$, kde $c_1, \dots, c_n \in \mathbb{R}$, které jsou nestrannými odhady střední hodnoty $\mu(\boldsymbol{\theta})$, tj. pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$ musí platit

$$\mu(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left(\sum_{i=1}^n c_i X_i \right) = \sum_{i=1}^n c_i \underbrace{E_{\boldsymbol{\theta}} X_i}_{=\mu(\boldsymbol{\theta})} = \mu(\boldsymbol{\theta}) \sum_{i=1}^n c_i \quad \Rightarrow \quad \sum_{i=1}^n c_i = 1.$$

Tím jsme dostali první podmínku, která se týká nestrannosti odhadu.

Nyní budeme hledat taková $c_1, \dots, c_n \in \mathbb{R}$, která minimalizují rozptyl

$$D_{\boldsymbol{\theta}} \left(\sum_{i=1}^n c_i X_i \right) \overset{nez.}{=} \sum_{i=1}^n c_i^2 D_{\boldsymbol{\theta}} X_i = \sigma^2(\boldsymbol{\theta}) \sum_{i=1}^n c_i^2$$

a pro něž platí $\sum_{i=1}^n c_i = 1$, tedy hledáme vázaný extrém, takže použijeme Lagrangeovu funkci s multiplikátorem λ, tj.

$$L(c_1, \dots, c_n, \lambda) = \sum_{i=1}^n c_i^2 - \lambda \left(\sum_{i=1}^n c_i - 1 \right).$$

Pak pro $j = 1, \dots, n$

$$\frac{\partial L}{\partial c_j} = 2c_j - \lambda = 0 \quad \Rightarrow \quad c_j = \tfrac{1}{2}\lambda$$

$$\frac{\partial L}{\partial \lambda} = -\sum_{i=1}^n c_i + 1 = 0 \quad \Rightarrow \quad \sum_{i=1}^n c_i = 1.$$

Prvních n rovnic implikuje, že

$$c_1 = c_2 = \cdots = c_n.$$

Označme společnou hodnotu symbolem c. Díky poslední rovnici dostaneme

$$1 = \sum_{i=1}^n c_i = nc \quad \Rightarrow \quad c = c_1 = c_2 = \cdots = c_n = \tfrac{1}{n},$$

tedy výběrový průměr $\bar{X}$ je **nejlepším nestranným lineárním odhadem** střední hodnoty $\mu(\boldsymbol{\theta})$.

Zkusme provést důkaz ještě jiným způsobem. Nechť $\sum_{i=1}^n c_i X_i$ je libovolný nestranný lineární odhad pro μ (tj. nutně musí platit $\sum_{i=1}^n c_i = 1$).

Položíme-li	$c_i = \frac{1}{n} + \delta_i$	pro	$i = 1, \dots, n$
je minimalizace výrazu	$\sum_{i=1}^n c_i^2$	za podmínky	$\sum_{i=1}^n c_i = 1$
ekvivalentní s úlohou minimalizovat	$\sum_{i=1}^n \left(\frac{1}{n} + \delta_i\right)^2$	za podmínky	$\sum_{i=1}^n \delta_i = 0.$

Za této podmínky je však

$$\sum_{i=1}^n \left(\tfrac{1}{n} + \delta_i\right)^2 = \underbrace{\sum_{i=1}^n \left(\tfrac{1}{n}\right)^2}_{=n\frac{1}{n^2}} + 2\tfrac{1}{n} \underbrace{\sum_{i=1}^n \delta_i}_{=0} + \sum_{i=1}^n \delta_i^2 = \boxed{\tfrac{1}{n} + \sum_{i=1}^n \delta_i^2},$$

což je minimální pro

$$\delta_i = 0 \qquad \text{pro} \qquad i = 1, \dots, n.$$

Tedy nejlepším nestranným lineárním odhadem je lineární kombinace X_i s koeficienty $c_i = \frac{1}{n}$.

3. POSTAČUJÍCÍ STATISTIKY

Nalezení rovnoměrně nejlepších nestranných odhadů není vždy jednoduché. Abychom nalezli odhad, který má nejmenší rozptyl, je vhodná jistá redukce výběru, tj. nahrazení celého výběru jedinou statistikou, takovou, která bude obsahovat „veškerou informaci o parametru $\boldsymbol{\theta}$“, která byla obsažená ve výběru. Takováto redukce výběrového prostoru se dosáhne pomocí postačujících statistik.

DEFINICE 3.1. Mějme náhodný výběr $\mathbf{X}_n = (X_1, \ldots, X_n)'$ z rozdělení pravděpodobnosti $P_{\boldsymbol{\theta}}$, kde $\boldsymbol{\theta}$ je neznámý parametr. Řekneme, že statistika $\mathbf{S}(\mathbf{X})$ je **postačující (sufici-entní) statistikou** (*sufficient statistic*), jestliže sdružené rozdělení náhodného výběru $\mathbf{X}_n = (X_1, \ldots, X_n)'$ podmíněné jevem $\mathbf{S}(\mathbf{X}) = \mathbf{s}$ je pro každé $\mathbf{s}$ nezávislé na $\boldsymbol{\theta}$.

Příklad 3.2. Nechť náhodný výběr $\mathbf{X}_n = (X_1, \ldots, X_n)'$ pochází z alternativního rozdělení s parametrem $\theta \in (0, 1)$, tj.

$$X_i \sim A(\theta) \sim p_x = \begin{cases} \theta^x (1-\theta)^{1-x} & n \in \mathbb{N},\ x = 0, \ldots, n, \\ 0 & \text{jinak.} \end{cases}$$

Nechť

$$S = \sum_{i=1}^{n} X_i \qquad \Rightarrow \qquad S \sim Bi(n, \theta).$$

Nechť $\mathbf{x}_n = (x_1, \ldots, x_n)'$ je realizace náhodného výběru. Uvažujme podmíněnou pravděpodobnost pro libovolně, ale pevně zvolené $s \in \mathbb{R}$

$$P_{\boldsymbol{\theta}}(X_1 = x_1, \ldots, X_n = x_n | S = s).$$

(a) Je-li $\sum_{i=1}^{n} x_i \neq s$, pak je tato podmíněná pravděpodobnost rovna nule.
(b) Nechť $\sum_{i=1}^{n} x_i = s$. Pak

$$P_{\boldsymbol{\theta}}(X_1 = x_1, \ldots, X_n = x_n | S = s) = \frac{P_{\boldsymbol{\theta}}(X_1 = x_1, \ldots, X_n = x_n)}{P_{\boldsymbol{\theta}}(S = s)}$$
$$= \frac{\prod_{i=1}^{n} P_{\boldsymbol{\theta}}(X_i = x_i)}{P_{\boldsymbol{\theta}}(S = s)} = \frac{\theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}}{\binom{n}{s} \theta^s (1-\theta)^{n-s}} = \frac{1}{\binom{n}{s}}.$$

Výsledek nezávisí na θ, takže statistika $S = \sum_{i=1}^{n} X_i$ je **postačující statistikou**.

Uvedeme větu, která se nazývá také větou o faktorizaci a která zjednodušuje hledání postačujících statistik. Kromě toho umožňuje rychle rozhodnout o tom, či je statistika postačující.

VĚTA 3.3. ***Neymanovo faktorizační kritérium.*** *Mějme náhodný výběr $\mathbf{X}_n = (X_1, \ldots, X_n)'$ z rozdělení s pravděpodobnostní funkcí (resp. hustotou) $f(\mathbf{x}; \boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Potom $S(\mathbf{X})$ je postačující statistika pro $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, právě když existují nezáporné měřitelné funkce g, h takové, že sdružené rozdělení náhodného výběru je součinem dvou faktorů:*

$$f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x})\ \ g\left(\mathbf{S}(\mathbf{x}), \boldsymbol{\theta}\right)$$

*(a říkáme, že hustota f se dá **faktorizovat**).*

Důkaz. Tvrzení ukážeme pouze pro diskrétní případ.

$\Rightarrow$ Nechť $\mathbf{S}$ je postačující statistika, pak podle definice

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} | \mathbf{S}(\mathbf{X}) = \mathbf{s}) = h(\mathbf{x})$$

a nezávisí na $\boldsymbol{\theta}$. Dále pro sdruženou pravděpodobnostní funkci platí

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}) = \underbrace{P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}|\mathbf{S}(\mathbf{X})=\mathbf{S}(\mathbf{x}))}_{h(\mathbf{x})}\underbrace{P_{\boldsymbol{\theta}}(\mathbf{S}(\mathbf{X})=\mathbf{S}(\mathbf{x}))}_{g(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta})}$$

$\Leftarrow$ Předpokládejme, že sdruženou pravděpodobnostní funkci lze vyjádřit ve tvaru

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})\; g\left(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta}\right),$$

tj. že ji lze faktorizovat. Označme

$$B_{\mathbf{s}} = \{\mathbf{x}\in\mathbb{R}^n; \mathbf{S}(\mathbf{x})=\mathbf{s}\}.$$

Nejprve spočtěme

$$\begin{aligned} P_{\boldsymbol{\theta}}(\mathbf{S}(\mathbf{X})=\mathbf{s}) &= \sum_{\mathbf{x}\in B_{\mathbf{s}}} P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}) = \sum_{\mathbf{x}\in B_{\mathbf{s}}} h(\mathbf{x})\; g\left(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta}\right) \\ &= g\left(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta}\right)\sum_{\mathbf{x}\in B_{\mathbf{s}}} h(\mathbf{x}). \end{aligned}$$

Je-li $P_{\boldsymbol{\theta}}(\mathbf{S}(\mathbf{X})=\mathbf{s})>0$ a $\mathbf{S}(\mathbf{x})\neq\mathbf{s}$, pak je podmíněná pravděpodobnost

$$P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}|\mathbf{S}(\mathbf{X})=\mathbf{s}) = 0.$$

Je-li $P_{\boldsymbol{\theta}}(\mathbf{S}(\mathbf{X})=\mathbf{s})>0$ a $\mathbf{S}(\mathbf{x})=\mathbf{s}$, pak

$$\begin{aligned} P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x}|\mathbf{S}(\mathbf{X})=\mathbf{s}) &= \frac{P_{\boldsymbol{\theta}}(\mathbf{X}=\mathbf{x})}{P_{\boldsymbol{\theta}}(\mathbf{S}(\mathbf{X})=\mathbf{s})} = \frac{h(\mathbf{x})\; g\left(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta}\right)}{g\left(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta}\right)\sum_{\mathbf{x}\in B_{\mathbf{s}}} h(\mathbf{x})} \\ &= \frac{h(\mathbf{x})}{\sum_{\mathbf{x}\in B_{\mathbf{s}}} h(\mathbf{x})} \end{aligned}$$

a tím je dokázáno, že podmíněné rozdělení vektoru $\mathbf{X}$ při dané hodnotě statistiky $\mathbf{S}$ nezávisí na $\boldsymbol{\theta}$ a $\mathbf{S}$ je postačující statistikou pro prametr $\boldsymbol{\theta}$.

□

Příklad 3.4. Nechť náhodný výběr $\mathbf{X}_n = (X_1,\ldots,X_n)'$ pochází z Poissonova rozdělení s parametrem $\theta>0$ s pravděpodobnostní funkcí

$$f_X(x) = P_{\boldsymbol{\theta}}(X=x) = \frac{e^{-\theta}\theta^x}{x!} \qquad x=0,1,2,\ldots.$$

Ukážeme, že statistika

$$S = \sum_{i=1}^n X_i$$

je postačující statistikou pro parametr θ, neboť sdružená hustota náhodného výběru je tvaru

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} = \underbrace{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}_{g(\mathbf{S}(\mathbf{x}),\boldsymbol{\theta})}\underbrace{\left(\prod_{i=1}^n x_i!\right)^{-1}}_{h(\mathbf{x})}.$$

Než uvedeme větu, která ukazuje praktický význam postačujících statistik pro konstrukci nejlepších nestranných odhadů, všimněme si podmíněných středních hodnot.

3.1. PODMÍNĚNÉ STŘEDNÍ HODNOTY.

Nechť $\mathbf{Z} = (X, Y)'$ je náhodný vektor, $F(x, y)$ je jeho sdružená distribuční funkce a $F_X(x)$ a $F_Y(y)$ odpovídající marginální distribuční funkce. Nechť vektor středních hodnot $E\mathbf{Z}$ existuje (a je konečný).

(1) Nechť pro každou borelovskou množinu $S \in \mathcal{B}$ a pro každé $x \in \mathbb{R}$ existuje funkce $F(x|y)$ taková, že platí

$$P(X \leq x, Y \in S) = \int_S F(x|y) dF_Y(y).$$

Potom funkci $F(x|y)$ nazveme **podmíněnou distribuční funkci náhodné veličiny X při daném $Y = y$** (podmíněnou jevem $Y = y$ nebo také vzhledem k Y).

(a) Diskrétní případ: $\mathbf{Z} = (X, Y)' \sim p(x, y)$, $M = \{(x, y) \in \mathbb{R}^2 : p(x, y) > 0\}$, $X \sim p_X(x)$, $M_X = \{x \in \mathbb{R} : p_X(x) > 0\}$, $Y \sim p_Y(y)$, $M_Y = \{y \in \mathbb{R} : p_Y(y) > 0\}$. Počítejme

$$\begin{aligned} P(X \leq x, Y \in S) &= \sum_{y \in S} \sum_{t \leq x} p(t, y) = \sum_{y \in S \cap M_Y} \sum_{t \leq x} p(t, y) + \sum_{y \in S \cap (\mathbb{R} - M_Y)} \sum_{t \leq x} \underbrace{p(t, y)}_{=0} \\ &= \sum_{y \in S \cap M_Y} \left(\sum_{t \leq x} \frac{p(t, y)}{p_Y(y)} \right) p_Y(y) = \int_{S \cap M_Y} \sum_{t \leq x} \frac{p(t, y)}{p_Y(y)} dF_Y(y). \end{aligned}$$

Takže **podmíněná distribuční funkce** je v **diskrétném** případě tvaru

$$F(x|y) = \begin{cases} \sum_{t \leq x} \frac{p(t,y)}{p_Y(y)} & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}$$

a **podmíněná pravděpodobnostní funkce** je rovna

$$p(x|y) = \begin{cases} \frac{p(x,y)}{p_Y(y)} & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}.$$

(b) Spojitý případ: $\mathbf{Z} = (X, Y)' \sim f(x, y)$, $X \sim f_X(x)$, $M_X = \{x \in \mathbb{R} : f_X(x) > 0\}$, $Y \sim f_Y(y)$, $M_Y = \{y \in \mathbb{R} : f_Y(y) > 0\}$. Počítejme

$$\begin{aligned} P(X \leq x, Y \in S) &= \int_S \int_{-\infty}^{x} f(t, y) dt dy \\ &= \int_{S \cap M_Y} \int_{-\infty}^{x} f(t, y) dt dy + \int_{S \cap (\mathbb{R} - M_Y)} \int_{-\infty}^{x} \underbrace{f(t, y)}_{=0} dt dy \\ &= \int_{S \cap M_Y} \left(\int_{-\infty}^{x} \frac{f(t, y)}{f_Y(y)} dt \right) f_Y(y) dy \\ &= \int_{S \cap M_Y} \left(\int_{-\infty}^{x} \frac{f(t, y)}{f_Y(y)} dt \right) dF_Y(y). \end{aligned}$$

Takže **podmíněná distribuční funkce** je v diskrétném případě tvaru

$$F(x|y) = \begin{cases} \int_{-\infty}^{x} \frac{f(t,y)}{f_Y(y)} dt & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}$$

a **podmíněná hustota** je rovna

$$f(x|y) = \begin{cases} \frac{f(x,y)}{f_Y(y)} & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}.$$

(2) Nechť $T = T(X, Y)$ je transformovaná náhodná veličina. Potom funkci

$$E(T(X,Y)|Y = y) = \int_{\mathbb{R}} T(x,y)dF(x|y) \qquad y \in \mathbb{R}$$

nazveme **podmíněnou střední hodnotou náhodné veličiny** X **za podmínky** $Y = y$ za předpokladu, že uvedený integrál pro všechna $y \in \mathbb{R}$ existuje (a je konečný).

Položme

$$E(T(X,Y)|Y = y) = h(y)$$

a definujme symbolem

$$E(T(X,Y)|Y) = h(Y)$$

náhodnou veličinu, kterou nazveme **(zobecněnou) podmíněnou střední hodnotou náhodné veličiny** $T(X, Y)$ **při daném** Y.

(a) Diskrétní případ:

$$E(T(X,Y)|Y = y) = \int_{\mathbb{R}} T(x,y)dF(x|y) = \sum_{x \in M_X} T(x,y)\, p(x|y)$$
$$= \begin{cases} \sum_{x \in M_X} T(x,y)\, \frac{p(x,y)}{p_Y(y)} & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}$$

a analogicky

$$E(T(X,Y)|Y) = \begin{cases} \sum_{x \in M_X} T(x,Y)\, \frac{p(x,Y)}{p_Y(Y)} & \text{pro } Y \in M_Y, \\ 0 & \text{pro } Y \in (\mathbb{R} - M_Y), \end{cases}.$$

(b) Spojitý případ:

$$E(T(X,Y)|Y = y) = \int_{\mathbb{R}} T(x,y)dF(x|y) = \int_{\mathbb{R}} T(x,y)\, f(x|y)dx$$
$$= \begin{cases} \int_{\mathbb{R}} T(x,y)\, \frac{f(x,y)}{f_Y(y)}dx & \text{pro } y \in M_Y, \\ 0 & \text{pro } y \in (\mathbb{R} - M_Y), \end{cases}$$

a analogicky

$$E(T(X,Y)|Y) = \begin{cases} \int_{\mathbb{R}} T(x,Y)\, \frac{f(x,Y)}{f_Y(Y)}dx & \text{pro } Y \in M_Y, \\ 0 & \text{pro } Y \in (\mathbb{R} - M_Y), \end{cases}.$$

Důležité vlastnosti podmíněných středních hodnot:

(i) Nechť X_1, X_2, Y jsou náhodné veličiny a a_0, a_1, a_2 jsou reálné konstanty, pak pokud střední hodnoty EX_1, EX_2 existují lze snadno dokázat, že platí

$$E(a_0 + a_1X_1 + a_2X_2|Y) = a_0 + a_1E(X_1|Y) + a_2E(X_2|Y), \tag{7}$$

(ii) Nechť X, Y jsou náhodné veličiny a střední hodnota EX existuje, pak

$$E\left[E(X|Y)\right] = EX. \tag{8}$$

Důkaz ukážeme pro spojitý případ:

$$EX = \int_{\mathbb{R}} xf_X(x)dx = \int_{\mathbb{R}} x\left(\int_{\mathbb{R}} f(x,y)dy\right)dx = \int_{\mathbb{R}} x\left(\int_{\mathbb{R}} f(x|y)f_Y(y)dy\right)dx$$
$$= \int_{\mathbb{R}} \underbrace{\left(\int_{\mathbb{R}} xf(x|y)dx\right)}_{h(y)=E(X|Y=y)} f_Y(y)dy = \int_{\mathbb{R}} h(y)f_Y(y)dy = E[h(Y)] = E\left[E(X|Y)\right].$$

(iii) Nechť $T_1 = T_1(X,Y)$ a $T_2 = T_2(Y)$ jsou transformované náhodné veličiny, pak

$$E(T_1T_2|Y) = T_2E(T_1|Y). \tag{9}$$

Důkaz ukážeme pro spojitý případ:

$$\begin{aligned} h(y) &= E(T_1T_2|Y=y) = E(T_1(X,Y)T_2(X)|Y=y) \\ &= \int_{\mathbb{R}} T_1(x,y)T_2(y)f(x|y)dx \\ &= T_2(y)\int_{\mathbb{R}} T_1(x,y)f(x|y)dx = T_2E(T_1|Y=y) \\ h(Y) &= E(T_1T_2|Y) = T_2E(T_1|Y). \end{aligned}$$

(3) Nechť $T = T(X,Y)$ je transformovaná náhodná veličina. **Podmíněný rozptyl** při daném $Y = y$ je definován vztahem

$$D(T(X,Y)|Y=y) = E\left\{[T - E(T|Y=y)]^2\,|Y=y\right\}$$

a **(zobecněný) podmíněný rozptyl** při daném Y je definován vztahem

$$D(T(X,Y)|Y) = E\left\{[T - E(T|Y)]^2\,|Y\right\}.$$

Platí

$$DT = E\left[D(T|Y)\right] + D\left[E(T|Y)\right], \tag{10}$$

neboť, spočítáme–li nejprve

$$\begin{aligned} D(T|Y) &= E\left\{[T - E(T|Y)]^2\,|Y\right\} \\ &= E\left\{[(T-ET) - (E(T|Y)-ET)]^2\,|Y\right\} \\ &= E\left\{(T-ET)^2 - 2(T-ET)[E(T|Y)-ET] + [E(T|Y)-ET]^2|Y\right\} \\ &= E[(T{-}ET)^2|Y] - 2[E(T|Y){-}ET]\underbrace{E[(T{-}ET)|Y]}_{\overset{viz(7)}{=}E(T|Y)-ET} + [E(T|Y){-}ET]^2 \\ &= E[(T-ET)^2|Y] - [E(T|Y)-ET]^2, \end{aligned}$$

tak odtud dostaneme

$$E[(T{-}ET)^2|Y] = D(T|Y) + [E(T|Y) - ET]^2$$

a nakonec

$$\begin{aligned} \underbrace{E\left\{E[(T{-}ET)^2|Y]\right\}}_{\overset{viz(7)}{=}E[(T-ET)^2=DT} &= E[D(T|Y)] + E[E(T|Y) - \underbrace{ET}_{\overset{viz(8)}{=}E[E(T|Y)]}\ ]^2 \\ &= E[D(T|Y)] + \underbrace{E[E(T|Y) - E[E(T|Y)]^2}_{=D[E(T|Y)]} \\ &= E[D(T|Y)] + D[E(T|Y)] \end{aligned}$$

Celkově tedy dostáváme

$$DT = E[D(T|Y)] + D[E(T|Y)].$$

VĚTA 3.5. ***Rao-Blackwellova***. *Nechť* $\mathbf{X}_n = (X_1, \ldots, X_n)'$ *je náhodný výběr z rozdělení pravděpodobnosti* $P_{\boldsymbol{\theta}}$, *kde* $\boldsymbol{\theta}$ *je vektor neznámých parametrů. Nechť existuje postačující statistika* $\mathbf{S}(\mathbf{X})$ *pro parametr* $\boldsymbol{\theta}$. *Nechť* $\gamma(\boldsymbol{\theta})$ *je daná parametrická funkce a statistika* $T(\mathbf{X})$ *je jejím nestranným odhadem, přičemž* $ET(\mathbf{X})^2 < \infty$ *pro každé* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Pak platí*

(i) Pro parametrickou funkci $\gamma(\boldsymbol{\theta})$ *existuje nestranný odhad*

$$S^*(\mathbf{X}) = S^*\left(\mathbf{S}(\mathbf{X})\right),$$

který je funkcí postačující statistiky $\mathbf{S}(\mathbf{X})$.

(ii) Pro rozptyl nestranného odhadu $S^*(\mathbf{X})$ *platí*

$$DS^*(\mathbf{X}) \leq DT(\mathbf{X}) \qquad \text{pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{11}$$

(iii) V nerovnosti (11) platí rovnost právě když

$$S^*(\mathbf{X}) = T(\mathbf{X}) \quad \text{s pravděpodobností 1 pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Důkaz. Nechť $T = T(\mathbf{X})$ je libovolný nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ a $\mathbf{S} = \mathbf{S}(\mathbf{X})$ je postačující statistika pro parametr $\boldsymbol{\theta}$.

(i) Položme

$$\boxed{S^*(s) = E\left(T(\mathbf{X})|\mathbf{S}(\mathbf{X}) = \mathbf{s}\right)}.$$

Protože $S(\mathbf{X})$ je postačující statistikou, funkce $S^*(s)$ nezávisí na $\boldsymbol{\theta}$, tj.

$$S^* = S^*(\mathbf{S}) = S^*\left(\mathbf{S}(\mathbf{X})\right) = E\left[T(\mathbf{X})|\mathbf{S}(\mathbf{X})\right] = E(T|\mathbf{S})$$

je statistika. Ukážeme, že S^* je nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$. Pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí:

$$ES^* = E\left[E\left(T|\mathbf{S}\right)\right] = ET = \gamma(\boldsymbol{\theta}).$$

(ii) Počítejme a upravujme rozptyl statistiky T

$$\begin{aligned} DT &= E\left[T - \gamma(\boldsymbol{\theta})\right]^2 = E\left\{\left[T - S^*\right] + \left[S^* - \gamma(\boldsymbol{\theta})\right]\right\}^2 \\ &= \underbrace{E\left[T - S^*\right]^2}_{\geq 0} + 2\underbrace{E\left\{\left[T - S^*\right]\left[S^* - \gamma(\boldsymbol{\theta})\right]\right\}}_{=0} + \underbrace{E\left[S^* - \gamma(\boldsymbol{\theta})\right]^2}_{DS^*} \end{aligned}$$

tj.

$$DT \geq DS^*,$$

neboť střední hodnotu součinu dvou statistik lze vyjádřit takto

$$\begin{aligned} \underbrace{E\left\{\left[T - S^*\right]\left[S^* - \gamma(\boldsymbol{\theta})\right]\right\}}_{E(U \cdot V)} &= \underbrace{E\left\{E\left\{\left[T - S^*\right]\left[S^* - \gamma(\boldsymbol{\theta})\right]|S\right\}\right\}}_{E(E(U \cdot V|\mathbf{S}))} \\ &= E\left\{\left[S^* - \gamma(\boldsymbol{\theta})\right]\underbrace{E\left\{\left[T - S^*\right]|\mathbf{S}\right\}}_{=0}\right\} = 0. \end{aligned}$$

(iii) V nerovnosti (11) platí rovnost právě když

$$E\left[T - S^*\right]^2 = 0 \quad \text{pro všechna } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

tj. když pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí

$$S^*(\mathbf{X}) = T(\mathbf{X}) \quad \text{s pravděpodobností 1.} \quad \square$$

$\square$

Poznámka 3.6. Z uvedené věty vyplývá, že při hledání nejlepších nestranných odhadů se můžeme omezit na odhady, které jsou funkcemi postačujících statistik. Věta 3.5 dává návod, jak určit nestranný odhad, který je funkcí postačující statistiky, jestliže známe libovolný nestranný odhad.

Příklad 3.7. Uvažujme výběr z alternativního rozdělení s parametrem $\theta > 0$ s pravděpodobnostní funkcí

$$f_X(x) = P(X = x) = \theta^x(1-\theta)^{1-x} \qquad x = 0, 1$$

a odhad parametrické funkce $\boxed{\gamma(\theta) = \theta}$ počítejme pomocí podmíněné střední hodnoty $\boxed{S^* = E(T|S)}$, kde T je libovolný nestranný odhad $\gamma(\theta) = \theta$.

Je zřejmé, že nestranným odhadem parametru θ je i statistika

$$T = T(\mathbf{X}) = X_1,$$

tj. první člen výběru, neboť

$$EX_1 = \theta.$$

Jak jsme ukázali v příkladu 3.2, postačující statistikou pro parametr θ je statistika

$$S = S(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Statistika S je součtem nezávislých náhodných veličin s alternativním rozdělením a tedy má binomické rozdělení s parametry n a θ, tj.

$$S = \sum_{i=1}^{n} X_i \sim Bi(n, \theta).$$

Všimněme si, že pravděpodobnost

$$P\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right) = P\left(X_1 = x, \sum_{i=2}^{n} X_i = s - x\right).$$

Náhodné veličiny $X_1 \sim A(\theta) \equiv Bi(1, \theta)$ a $\sum_{i=2}^{n} X_i \sim Bi(n-1, \theta)$ jsou nezávislé, takže

$$\begin{aligned} P_{\boldsymbol{\theta}}\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right) &= P_{\boldsymbol{\theta}}\left(X_1 = x\right) P_{\boldsymbol{\theta}}\left(\sum_{i=2}^{n} X_i = s - x\right) \\ &= \theta^x(1-\theta)^{1-x}\binom{n-1}{s-x}\theta^{s-x}(1-\theta)^{n-1-s+x} \\ &= \binom{n-1}{s-x}\theta^s(1-\theta)^{n-s}. \end{aligned}$$

Počítejme podmíněnou střední hodnotu za podmínky, že $S = s$

$$S^*(s) = E(T|S=s) = E\left\{X_1|\sum_{i=1}^{n} X_i = s\right\} = \sum_{x=0,1} x\frac{P\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right)}{P_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} X_i = s\right)}$$

$$= \frac{\binom{n-1}{s-x}\theta^s(1-\theta)^{n-s}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}} = \frac{(n-1)!s!(n-s)!}{n!(s-1)!(n-s)!} = \frac{s}{n},$$

Tedy

$$S^*(S) = E(T|S) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

což je aritmetický průměr všech pozorování.

Podívejme se, jak to vypadá s rozptyly statistik $T = X_1$ a S^*.

$$DT = DX_1 = \theta(1-\theta)$$

$$DS^* = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{\theta(1-\theta)}{n},$$

tedy rozptyl druhého nestranného odhadu se n krát zmenšil.

Příklad 3.8. Uvažujme výběr z Poissonova rozdělení s parametrem $\theta > 0$ s pravděpodobnostní funkcí

$$f_X(x) = P(X = x) = \frac{e^{-\theta}\theta^x}{x!} \qquad x = 0, 1, 2, \ldots$$

a odhad parametrické funkce $\boxed{\gamma(\theta) = \theta}$ počítejme pomocí podmíněné střední hodnoty $\boxed{S^* = E(T|S)}$, kde T je libovolný nestranný odhad $\gamma(\theta) = \theta$.

Je zřejmé, že nestranným odhadem parametru θ je i statistika

$$T = T(\mathbf{X}) = X_1,$$

tj. první člen výběru, neboť

$$EX_1 = \theta.$$

Jak jsme ukázali v příkladu 3.4, postačující statistikou pro parametr θ je statistika

$$S = S(\mathbf{X}) = \sum_{i=1}^{n} X_i.$$

Dále je třeba si uvědomit, že statistika S je součtem nezávislých náhodných veličin s Poissonovým rozdělením a má také Poissonovo rozdělení s parametrem $n\theta$, tj.

$$S = \sum_{i=1}^{n} X_i \sim Po(n\theta).$$

Počítejme dále pravděpodobnost

$$P\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right) = P\left(X_1 = x, \sum_{i=2}^{n} X_i = s - x\right).$$

Náhodné veličiny $X_1 \sim Po(\theta)$ a $\sum_{i=2}^{n} X_i \sim Po((n-1)\theta)$ jsou nezávislé, takže

$$\begin{aligned} P\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right) &= P\left(X_1 = x\right) P\left(\sum_{i=2}^{n} X_i = s - x\right) \\ &= \frac{e^{-\theta}\theta^x}{x!} \frac{e^{-(n-1)\theta}\left[(n-1)\theta\right]^{s-x}}{(s-x)!}. \end{aligned}$$

Nyní již počítejme podmíněnou střední hodnotu za podmínky, že $S = s$

$$\begin{aligned} S^*(s) = E(T|S = s) = E\left\{X_1 | \sum_{i=1}^{n} X_i = s\right\} &= \sum_{x=0}^{s} x \frac{P\left(X_1 = x, \sum_{i=1}^{n} X_i = s\right)}{P\left(\sum_{i=1}^{n} X_i = s\right)} \\ = \sum_{x=0}^{s} x \frac{\frac{e^{-\theta}\theta^x}{x!} \frac{e^{-(n-1)\theta}\left[(n-1)\theta\right]^{s-x}}{(s-x)!}}{\frac{e^{-n\theta}(n\theta)^s}{s!}} &= \sum_{x=0}^{s} x \binom{s}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{s-x}. \end{aligned}$$

Protože výraz $\sum_{x=0}^{s} x \binom{s}{x} \left(\frac{1}{n}\right)^x \left(1 - \frac{1}{n}\right)^{s-x}$ je střední hodnotou náhodné veličiny s binomickým rozdělením $Bi(s, \frac{1}{n})$, ihned dostaneme

$$S^*(s) = E(T|S = s) = \frac{s}{n}.$$

Tedy

$$S^*(S) = E(T|S) = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

což je aritmetický průměr všech pozorování.

Stejně jak v předchozím případě, všimněme si rozptylů obou odhadů $T = X_1$ a S^*.

$$DT = DX_1 = \theta$$
$$DS^* = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{\theta}{n},$$

tedy rozptyl druhého nestranného odhadu se n krát zmenšil.

Poznámka 3.9. Nahrazení nestranného odhadu T odhadem $S^* = E(T|\mathbf{S})$ ještě neznamená, že jsme mezi všemi nestrannými odhady našli odhad s nejmenším rozptylem. Úplnost postačující statistiky je pro to dostatečnou podmínkou.

DEFINICE 3.10. Systém parametrických tříd rozdělení $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ nazveme **úplným**, pokud pro každou měřitelnou funkci $h(\mathbf{x})$ a náhodnou veličinu X s rozdělením z této třídy platí implikace: jestliže

$$E_{\boldsymbol{\theta}} h(X) = 0 \quad \text{pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

pak

$$h(X) = 0 \quad \text{s pravděpodobností 1 pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Příklad 3.11. Nechť $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ je třídou binomických rozdělení

$$X \sim P_{\boldsymbol{\theta}}(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad n \geq 1, \;\; 0 < \theta < 1 \;\; x = 0, 1, \ldots, n.$$

Ukážeme, že tento systém je **úplný**. Uvažujme funkci $h(x)$ na množině $\{0, 1, \ldots, n\}$, pro kterou platí

$$Eh(X) = 0 \quad \text{pro každé } \theta \in (0, 1).$$

Tato funkce musí splňovat podmínku

$$Eh(X) = \sum_{x=0}^{n} h(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = 0 \quad \text{pro každé } \theta \in (0, 1).$$

Tuto podmínku můžeme napsat takto

$$Eh(X) = \sum_{x=0}^{n} h(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = \underbrace{(1-\theta)^n}_{(1+z)^{-n}}\sum_{x=0}^{n} h(x)\binom{n}{x}\underbrace{\left(\frac{\theta}{1-\theta}\right)^x}_{z^x}$$

$$= (1+z)^{-n}\sum_{x=0}^{n}\binom{n}{x}h(x)z^x = 0 \quad \text{pro } z > 0$$

Na jedné straně máme polynom n-tého řadu v proměnné z. Pokud se má identicky rovnat nule, musí se všechny jeho koeficienty rovnat nule, tj.

$$h(x) = 0 \quad \text{pro } x = 0, 1, \ldots, n.$$

Proto

$$\boxed{P(h(X) = 0) = 1 \quad \text{pro každé } \theta \in (0, 1)}.$$

Příklad 3.12. Nechť $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ je třídou Poissonových rozdělení s pravděpodobnostní funkcí

$$f_X(x) = P(X = x) = \frac{e^{-\theta}\theta^x}{x!} \qquad x = 0, 1, 2, \ldots$$

Tento systém je opět **úplný**. Uvažujme funkci $h(x)$ na množině $\{0, 1, 2, \ldots\}$, pro kterou platí

$$Eh(X) = 0 \quad \text{pro každé } \theta > 0.$$

Tato funkce musí splňovat podmínku

$$Eh(X) = \sum_{x=0}^{\infty} h(x)\frac{e^{-\theta}\theta^x}{x!} = 0 \quad \text{pro každé } \theta > 0.$$

Takže

$$\sum_{x=0}^{\infty} h(x)\frac{\theta^x}{x!} = 0 \quad \text{pro každé } \theta > 0.$$

Tato mocninná řada je rovna nule pro všechna $\theta > 0$, takže všechny její koeficienty musí být rovnu nule, tj.

$$h(x) = 0 \quad \text{pro } x = 0, 1, 2, \ldots.$$

Proto

$$\boxed{P(h(X) = 0) = 1 \quad \text{pro každé } \theta > 0}.$$

Příklad 3.13. Nechť $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ je třídou normálních rozdělení

$$X \sim \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{1}{2}\left(\frac{x}{\theta}\right)^2} \quad x \in \mathbb{R}, ; \theta > 0$$

Tento systém není úplný. Definujme

$$h(x) = \begin{cases} -1 & x < 0, \\ 1 & x \geq 0. \end{cases}.$$

Pro libovolné $\theta > 0$ platí

$$\frac{1}{\sqrt{2\pi}\theta}\int_{-\infty}^{\infty} h(x) e^{-\frac{1}{2}\left(\frac{x}{\theta}\right)^2} dx = -\underbrace{\frac{1}{\sqrt{2\pi}\theta}\int_{-\infty}^{0} e^{-\frac{1}{2}\left(\frac{x}{\theta}\right)^2} dx}_{=\frac{1}{2}} + \underbrace{\frac{1}{\sqrt{2\pi}\theta}\int_{0}^{\infty} e^{-\frac{1}{2}\left(\frac{x}{\theta}\right)^2} dx}_{=\frac{1}{2}} = 0.$$

Tedy z vlastnosti, že $Eh(X) = 0$ neplyne, že $P(h(X) = 0) = 1$.

DEFINICE 3.14. Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rozdělení pravděpodobnosti $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Statistiku $T(\mathbf{X})$ nazveme **úplnou vzhledem k** $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, pokud její rozdělení pravděpodobností tvoří úplný systém.

Nyní vyslovíme větu o jednoznačnosti nestranných odhadů založených na postačujících statistikách.

VĚTA 3.15. ***První Lehmanova-Sheffého věta.*** *Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z rodělení pravděpodobnosti $\mathcal{P} = \{P_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. Předpokládejme, že $T = T(\mathbf{X})$ je* ***nestranný odhad parametrické funkce*** *$\gamma(\boldsymbol{\theta})$, přičemž $ET^2 < \infty$ pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Nechť $\mathbf{S} = \mathbf{S}(\mathbf{X})$ je* ***úplná postačující statistika.*** *Definujme*

$$S^* = E(T|\mathbf{S}).$$

Pak S^* je ***nejlepší nestranný odhad parametrické funkce*** *$\gamma(\boldsymbol{\theta})$ a je* ***jediný.***

Důkaz. Nechť $T = T(\mathbf{X})$ a $T_2 = T_2(\mathbf{X})$ jsou nestranné odhady parametrické funkce $\gamma(\boldsymbol{\theta})$ s konečnými druhými momenty. Označme $S_2^* = E(T_2|\mathbf{S})$. Pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí

$$ES^* = \gamma(\boldsymbol{\theta}) \qquad DS^* \leq DT$$
$$ES_2^* = \gamma(\boldsymbol{\theta}) \qquad DS_2^* \leq DT_2$$

Máme tedy

$$E(S^* - S_2^*) = E(E(T|\mathbf{S}) - E(T_2|\mathbf{S})) = 0 \quad \text{pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Z předpokladu o úplnosti plyne, že

$$P(S^* = S_2^*) = 1 \quad \text{pro každé } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Z toho plyne závěr, že pro nestranné odhady S^* a T_2 platí

$$DS^* \leq DT_2.$$

Proto S^* je nejlepší. Z Raovy-Blackwellovy věty plyne, že T_2 bude stejně dobrý odhad jako S_2^* právě tehdy, bude-li

$$T_2 = S_2^* \quad \text{skoro jistě při každém } \boldsymbol{\theta}.$$

Jelikož víme, že $S^* = S_2^*$, dostáváme odtud $T_2 = S^*$ skoro jistě. □

Poznámka 3.16. V tomto případě nejmenší možný rozptyl nestranného odhadu parametrické funkce $\gamma(\boldsymbol{\theta})$ je roven DS^*. Přitom jde o skutečné dosažitelné minimum.

VĚTA 3.17. ***Druhá Lehmanova-Sheffého věta.*** *Nechť* $\mathbf{S}$ *je úplná postačující statistika. Nechť*

$$W = g(\mathbf{S})$$

je nestranný odhad parametrické funkce $\gamma(\boldsymbol{\theta})$ *a nechť* $EW^2 < \infty$ *pro každé* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. *Pak* W *je* ***nejlepší nestranný odhad parametrické funkce*** $\gamma(\boldsymbol{\theta})$ *a* ***je jediný.***

Důkaz. Tvrzení je přímým důsledkem první Lehmannovy-Sheffého věty. □

Příklad 3.18. Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z alternativního rozdělení s pravděpodobnostní funkcí

$$f(x, \theta) = P_{\boldsymbol{\theta}}(X = x) = \theta^x (1-\theta)^{n-x} \qquad 0 < \theta < 1 \;\; x = 0, 1$$

s pravděpodobností úspěchu $\theta \in (0, 1)$, kde θ je neznámý parametr. Budeme hledat nejlepší nestranný odhad pro

- $\boxed{\theta}$, což je střední hodnota alternativního rozdělení
- a v případě, že $n \geq 2$ také pro $\boxed{\theta(1-\theta))}$, což je rozptyl alternativního rozdělení

$\boxed{\theta}$: Z příkladů 3.2 a 3.11 vyplývá, že statistika

$$S = \sum_{i=1}^{n} X_i \sim Bi(n, \theta)$$

je **úplnou postačující statistikou**, takže statistika

$$S^*(S) = E(T|S) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

odvozená pomocí Rao-Blackwellovy věty je podle první Lehmanovy-Sheffého věty **nejlepším nestranným odhadem parametru** θ.

$\boxed{\theta(1-\theta))}$: Pomocí Rao-Blackwellovy věty nejprve hledejme statistiku $S^* = E(T|S)$, kde T je nějaký nestranný odhad parametrické funkce $\gamma(\theta) = \theta(1-\theta)$ a S je postačující statistikou pro parametr θ.

Jako **nestranný odhad parametrické funkce** $\gamma(\theta) = \theta(1-\theta)$ vezměme například

$$T = X_1(1 - X_2),$$

neboť

$$ET = E[X_1(1 - X2)] = \underbrace{EX_1 \cdot E(1 - X_2)}_{\text{nezávislost } X_1, X_2} = \theta(1-\theta).$$

Pro $s = 0, 1, \ldots, n$ počítejme

$$S^*(s) = E(T|S = s) = E\left(X_1(1 - X_2)\middle|\sum_{i=1}^{n} X_i = s\right)$$
$$= \frac{P\left(X_1 = 1, 1 - X_2 = 1, \sum_{i=1}^{n} X_i = s\right)}{P(\sum_{i=1}^{n} X_i = s)}$$

Je-li $s = 0$, je zřejmé, že

$$E\left(X_1(1 - X_2)\middle|\sum_{i=1}^{n} X_i = s\right) = 0.$$

Nechť nyní $s > 0$. Pak

$$S^*(s) = \frac{P(X_1 = 1)P(X_2 = 0)P\left(\sum_{i=3}^{n} X_i = s - 1\right)}{P(\sum_{i=1}^{n} X_i = s)}$$
$$= \frac{\theta(1-\theta)\binom{n-2}{s-1}\theta^{s-1}(1-\theta)^{n-2-s+1}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}} = \frac{(n-2)!s!(n-s)!}{n!(s-1)!(n-s-1)!}$$
$$= \frac{s(n-s)}{n(n-1)} = \frac{n}{n-1} \cdot \frac{s}{n} \cdot \left(1 - \frac{s}{n}\right)$$

a

$$S^*(S) = \frac{n}{n-1}\bar{X}(1 - \bar{X}),$$

kde

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

Protože statistika

$$S = \sum_{i=1}^{n} X_i \sim Bi(n, \theta)$$

je **úplnou postačující statistikou**, pak podle první Lehmanovy-Sheffého věty je $S^*(S)$ **nejlepším nestranným odhadem parametrické funkce** $\theta(1 - \theta)$.

Veličiny $X_1, \ldots, X_n$ můžeme chápat jako výběr z $Bi(1, \theta)$. Toto rozdělení má rozptyl $\theta(1 - \theta)$. Všimněme si, že pro $i = 1, \ldots, n$ platí

$$X_i^2 = X_i,$$

neboť tyto veličiny nabývají pouze hodnot 0 a 1. Nestranný odhad rozptylu pořízený na základě daného výběru je

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

a odhad $\boxed{S^2}$ je tedy totožný s **nejlepším nestranným odhadem** parametrické funkce $\theta(1 - \theta)$.

Příklad 3.19. Nechť $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je náhodný výběr z Poissonova rozdělení s pravděpodobnostní funkcí

$$f_X(x) = P(X = x) = \frac{e^{-\theta}\theta^x}{x!} \qquad x = 0, 1, 2, \ldots$$

kde θ je neznámý parametr. Budeme hledat nejlepší nestranný odhad pro

- $\boxed{\theta}$, což je střední hodnota Poissonova rozdělení
- $\boxed{e^{-\theta} = P(X = 0)}$

$\boxed{\theta}$: Z příkladů 3.4 a 3.12 vyplývá, že statistika

$$S = \sum_{i=1}^{n} X_i \sim Po(n\theta)$$

je **úplnou postačující statistikou**, takže statistika

$$S^*(S) = E(T|S) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

odvozená pomocí Rao-Blackwellovy věty je podle první Lehmanovy-Sheffého věty **nejlepším nestranným odhadem parametru** θ.

$\boxed{e^{-\theta}}$: Pomocí Rao-Blackwellovy věty nejprve hledejme statistiku $S^* = E(T|S)$, kde T je nějaký nestranný odhad parametrické funkce $\gamma(\theta) = e^{-\theta}$ a S je postačující statistikou pro parametr θ.

Položme

$$T = I_{\{0\}}(X_1) = I(X_1 = 0) = \begin{cases} 1 & X_1 = 0, \\ 0 & \text{jinak.} \end{cases}$$

Protože

$$ET = 1 \cdot P_{\boldsymbol{\theta}}(T = 1) + 0 \cdot P_{\boldsymbol{\theta}}(T = 0) = P_{\boldsymbol{\theta}}(X_1 = 0) = e^{-\theta},$$

pak statistika T je **nestranným odhadem parametrické funkce** $\gamma(\theta) = e^{-\theta}$.

Je-li $n = 1$, pak statistika T je **nejlepším nestranným odhadem** parametrické funkce $\gamma(\theta) = e^{-\theta}$.

Pro $n > 1$ počítejme

$$\begin{aligned} S^*(s) = E(T|S = s) &= E\left(I(X_1 = 0) \,\middle|\, \sum_{i=1}^{n} X_i = s\right) \\ &= \frac{P\left(T = 1, \sum_{i=1}^{n} X_i = s\right)}{P(\sum_{i=1}^{n} X_i = s)} = \frac{P\left(X_1 = 0, \sum_{i=2}^{n} X_i = s\right)}{P(\sum_{i=1}^{n} X_i = s)} \\ &= \frac{P(X_1 = 0)P\left(\sum_{i=2}^{n} X_i = s\right)}{P(\sum_{i=1}^{n} X_i = s)} = \frac{\frac{e^{-\theta}e^{-(n-1)\theta}[(n-1)\theta]^s}{s!}}{\frac{e^{-n\theta}(n\theta)^s}{s!}} = \left(\frac{n-1}{n}\right)^s \end{aligned}$$

a

$$S^*(S) = \frac{n}{n-1}\bar{X}(1 - \bar{X}),$$

kde

$$\bar{X} = \left(\frac{n-1}{n}\right)^{\sum_{i=1}^{n} X_i}.$$

Protože statistika

$$S = \sum_{i=1}^{n} X_i \sim Po(n\theta)$$

je **úplnou postačující statistikou**, pak podle první Lehmanovy-Sheffého věty je $S^*(S)$ **nejlepším nestranným odhadem parametrické funkce** $e^{-\theta}$.

Spočítejme ještě

$$ES^* = ES^*(S) = E\left(\frac{n-1}{n}\right)^S = \sum_{s=0}^{\infty}\left(\frac{n-1}{n}\right)^s \frac{e^{-n\theta}(n\theta)^s}{s!}$$

$$= e^{-n\theta}\underbrace{\sum_{s=0}^{\infty}\frac{[(n-1)\theta]^s}{s!}}_{=e^{(n-1)\theta}} = e^{-\theta}$$

$$ES^{*2} = \sum_{s=0}^{\infty}\left(\frac{n-1}{n}\right)^{2s}\frac{e^{-n\theta}(n\theta)^s}{s!} = e^{-n\theta}\underbrace{\sum_{s=0}^{\infty}\frac{\left[\frac{(n-1)^2}{n}\theta\right]^s}{s!}}_{=e^{\frac{(n-1)^2}{n}\theta}} = e^{-2\theta+\frac{\theta}{n}}$$

$$DS^* = ES^{*2} - (ES^*)^2 = e^{-2\theta+\frac{\theta}{n}} - e^{-2\theta} = e^{-2\theta}\left(e^{\frac{\theta}{n}} - 1\right).$$

4. REGULÁRNÍ SYSTÉM HUSTOT A DOLNÍ MEZ ROZPTYLU REGULÁRNÍCH ODHADŮ

Je zcela zřejmé, že na základě konečně mnoho pozorování $\mathbf{X}_n = (X_1, \dots, X_n)'$ nelze odhadnout parametrickou funkce $\gamma(\boldsymbol{\theta})$ zcela bez chyby, tj. nelze najít nestranný odhad $T_n = T(X_1, \dots, X_n)'$ s nulovým rozptylem.

Existuje však **dolní mez**, pod kterou nemůže rozptyl žádného nestranného odhadu klesnout.

Tato dolní mez záleží ovšem, jak za chvíli ukážeme,

- na rozsahu náhodného výběru, tj. na n,
- na rodině rozdělení $F(x; \boldsymbol{\theta})$, ze kterého výběr pochází
- a na parametrické funkci $\gamma(\boldsymbol{\theta})$.

Při odvozování **dolní meze rozptylu nestranných odhadů** se omezíme

- na rodiny rozdělení $F(x; \theta)$, která splňují jisté podmínky, a to tzv. podmínky *regularity.*

V dalším budeme značit symbolem $f(x; \boldsymbol{\theta})$ jak *hustotu pravděpodobnost*i absolutně spojité náhodné veličiny, tak *pravděpodobnostní funkci* diskrétní náhodné veličiny, neboť obě jsou hustotami, v prvém případě vzhledem k Lebesgueově míře, v druhém případě vzhledem k čítací míře.

DEFINICE 4.1. Mějme parametrický prostor $\Theta \subset \mathbb{R}$. Řekneme, že **systém parametrických hustot**

$$\mathcal{F}_{\text{reg}} = \{f(x; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

je **regulární**, jestliže platí

(1) $\boldsymbol{\Theta} \subset \mathbb{R}^m$ je otevřená borelovská množina.

(2) Množina $M = \{x \in \mathbb{R} : f(x; \boldsymbol{\theta}) > 0\}$ nezávisí na parametru $\boldsymbol{\theta}$.

(3) Pro každé $x \in M$ existuje konečná parciální derivace

$$f'_i(x; \boldsymbol{\theta}) = \frac{\partial f(x; \boldsymbol{\theta})}{\partial \theta_i} \qquad (i = 1, \dots, m).$$

(4) Pro všechny $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)' \in \boldsymbol{\Theta}$ platí

$$\int_M \frac{f'_i(x; \boldsymbol{\theta})}{f(x; \boldsymbol{\theta})} dF(x; \boldsymbol{\theta}) = \int_M \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \theta_i} dF(x; \boldsymbol{\theta}) = 0 \qquad (i = 1, \dots, m),$$

kde $F(x; \boldsymbol{\theta})$ je odpovídající distribuční funkce.

(5) Pro všechny $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)' \in \boldsymbol{\Theta}$ je integrál

$$J_{ij} = J_{ij}(\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(x; \boldsymbol{\theta})}{\partial \theta_j} dF(x; \boldsymbol{\theta}) \qquad (i, j = 1, \dots, m)$$

konečný a matice $\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}) = (J_{ij}(\boldsymbol{\theta}))_{i,j=1}^m$ je pozitivně definitní. Matice $\mathbf{J}(\boldsymbol{\theta})$ se nazývá **Fisherova informační matice o parametru $\boldsymbol{\theta}$.**

Poznámka 4.2. Pro jednoduchost někdy hovoříme o **regulárnosti** $f(x; \boldsymbol{\theta})$, ne o **regulárnosti systému hustot**.

Poznámka 4.3. Ukážeme, že podmínka (4) souvisí s otázkou, zda při derivování rovnosti

$$1 = \int_M dF(x;\boldsymbol{\theta})$$

lze zaměnit pořadí derivace a integrálu, tj.

$$\boxed{0} = \frac{\partial}{\partial\theta_j}1 = \frac{\partial}{\partial\theta_j}\int_M dF(x;\boldsymbol{\theta}) \overset{?}{=} \underbrace{\int_M \frac{\partial}{\partial\theta_j} dF(x;\boldsymbol{\theta}) = \boxed{0}}_{(*)}.$$

Jestliže máme zaručeno, že platí vztah $(*)$, pak pořadí lze zaměnit. A nyní ukážeme, že podmínka (4) je ekvivaletní s podmínkou $(*)$. Nechť ν je čítací nebo Lebesgueova míra. Upravujme

$$\boxed{0} = \int_M \frac{\partial}{\partial\theta_j} dF(x;\boldsymbol{\theta}) = \int_M \frac{\partial}{\partial\theta_j} f(x;\boldsymbol{\theta})\, d\nu(x) = \boxed{\int_M f'_j(x;\boldsymbol{\theta})\, d\nu(x)}$$

někdy tato podmínka bývá v definici regularity

$$= \int_M f'_j(x;\boldsymbol{\theta}) \frac{f(x;\boldsymbol{\theta})}{f(x;\boldsymbol{\theta})}\, d\nu(x) = \boxed{\int_M \frac{f'_j(x;\boldsymbol{\theta})}{f(x;\boldsymbol{\theta})} dF(x;\boldsymbol{\theta})}$$

což je právě podmínka (4) v definici regularity.

Poznámka 4.4. Označíme–li symbolem

$$U_i = U_i(\boldsymbol{\theta}) = \frac{f'_i(X;\boldsymbol{\theta})}{f(X;\boldsymbol{\theta})} = \frac{\partial \ln f(X;\boldsymbol{\theta})}{\partial\theta_i}$$

tzv. *i*–tý **skór** příslušný k hustotě $f(x;\boldsymbol{\theta})$ a

$$\mathbf{U} = \mathbf{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), \ldots, U_m(\boldsymbol{\theta}))'$$

tzv. **skórový vektor** příslušný k hustotě $f(x;\boldsymbol{\theta})$, pak podmínku (4) lze ekvivalentně napsat takto

$$\text{pro } \forall i \in \{1,\ldots,m\} \quad E_\theta U_i = 0, \qquad \text{tj.} \qquad E_\theta \mathbf{U} = (0,\ldots,0)' = \mathbf{0},$$

tj. skóry jsou centrované. V tomto značení podmínka (5) je ekvivalentní s existencí kovariancí

$$J_{ij} = \int_M \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_i} \frac{\partial \ln f(x;\boldsymbol{\theta})}{\partial\theta_j} dF(x;\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(U_i U_j) = C_{\boldsymbol{\theta}}(U_i, U_j) < \infty.$$

Pro sdruženou hustotu náhodného výběru $\mathbf{X}_n = (X_1,\ldots,X_n)'$ platí

$$f_{\mathbf{X}}(x_1,\ldots,x_n;\boldsymbol{\theta}) = \prod_{k=1}^n f(x_i;\boldsymbol{\theta}) \quad \Rightarrow \quad \frac{\partial \ln f_{\mathbf{X}}(x_1,\ldots,x_n;\boldsymbol{\theta})}{\partial\theta_j} = \sum_{k=1}^n \frac{\partial \ln f(x_k;\boldsymbol{\theta})}{\partial\theta_j}$$

a označíme–li pro k–tou složku náhodného výběru

$$\mathbf{U}_k = (U_{k,1},\ldots,U_{k,m})' = \left(\frac{\partial \ln f(X_k;\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \ln f(X_k;\boldsymbol{\theta})}{\partial\theta_m}\right)'$$

a pro celý náhodný výběr

$$\mathbf{U}^*_n = (U^*_1,\ldots,U^*_m)' = \left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \ln f_{\mathbf{X}}(\mathbf{X};\boldsymbol{\theta})}{\partial\theta_m}\right)',$$

dostaneme

pro skórový vektor náhodného výběru $\boxed{\mathbf{U}^*_n = \sum_{k=1}^n \mathbf{U}_k}$ a pro jednotlivé složky skórového vektoru $\boxed{U^*_j = \sum_{k=1}^n U_{k,j}}$.

VĚTA 4.5 (Raova-Cramerova nerovnost). *Nechť $T_n = T(X_1,\ldots,X_n)$ je **regulárním** odhaden parametrické funkce $\gamma(\boldsymbol{\theta})$, tj.*

(i) náhodný výběr $\mathbf{X}_n = (X_1,\ldots,X_n)'$ je z rozdělení s regulární hustotou $f \in \mathcal{F}_{reg}$,
(ii) $T_n(\mathbf{X})$ je nestranným odhadem parametrické funkce $\gamma(\boldsymbol{\theta})$,
*(iii) pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\forall j=1,\ldots,m$ **existují** parciální derivace $\frac{\partial\gamma(\boldsymbol{\theta})}{\partial\theta_j}$ a platí*

$$\frac{\partial}{\partial\theta_j}\int_M\ldots\int_M T_n(x_1,\ldots,x_n)\prod_{i=1}^n dF(x_i;\boldsymbol{\theta}) = \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\frac{\partial}{\partial\theta_j}\prod_{i=1}^n dF(x_i;\boldsymbol{\theta}).$$

*Pak existuje **dolní Rao–Cramerova hranice** C_n **rozptylu** odhadu T_n a platí*

$$C_n = C_n(\boldsymbol{\theta}) = \tfrac{1}{n}\boldsymbol{\gamma}'\mathbf{J}^{-1}\boldsymbol{\gamma} \le D_{\boldsymbol{\theta}}T_n, \qquad kde \qquad \boldsymbol{\gamma}' = \left(\tfrac{\partial\gamma(\boldsymbol{\theta})}{\partial\theta_1},\ldots,\tfrac{\partial\gamma(\boldsymbol{\theta})}{\partial\theta_m}\right)'.$$

Důkaz. Důkaz uděláme pro skalární parametr θ. Protože $T_n(\mathbf{Y})$ je nestranným odhadem parametrické funkce $\gamma(\theta)$, platí

$$\begin{aligned}\gamma(\theta) = E_\theta T_n(\mathbf{X}) &= \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\prod_{k=1}^n dF(x_k;\theta)\\ &= \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\prod_{k=1}^n f(x_k;\theta)d\nu(x_1)\cdots d\nu(x_n),\end{aligned}$$

kde ν je čítací nebo Lebesgueova míra. Díky předpokladům ve větě můžeme psát

$$\begin{aligned}\gamma'(\theta) = [E_\theta T_n(\mathbf{X})]' &= \tfrac{\partial}{\partial\theta}\int_M\ldots\int_M T_n(x_1,\ldots,x_n)\prod_{k=1}^n f(x_k;\theta)\ d\nu(x_1)\cdots d\nu(x_n)\\ &= \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\tfrac{\partial}{\partial\theta}\prod_{k=1}^n f(x_k;\theta)\ d\nu(x_1)\cdots d\nu(x_n)\\ &= \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\sum_{k=1}^n f'(x_k;\theta)\prod_{h=1,h\neq k}^n f(x_h;\theta)\ d\nu(x_1)\cdots d\nu(x_n)\\ &= \int_M\ldots\int_M T_n(x_1,\ldots,x_n)\sum_{k=1}^n \tfrac{f'(x_k;\theta)}{f(x_k;\theta)}\prod_{h=1}^n f(x_h;\theta)\ d\nu(x_1)\cdots d\nu(x_n)\\ &= E_\theta\left[T_n(\mathbf{X})\sum_{k=1}^n \tfrac{f'(X_k;\theta)}{f(X_k;\theta)}\right] = E_\theta\left[T_n(\mathbf{X})\sum_{k=1}^n U_{k,1}(\theta)\right] = E_\theta\left[T_n(\mathbf{X})\ \mathbf{U}_n^*\right]\end{aligned}$$

Protože $E_\theta\mathbf{U}_n^* = 0$, pak Fisherova informace pro skalární parametr θ, která se týká náhodného výběru, je rovna

$$J_n^* = E_\theta(\mathbf{U}_n^*)^2 = D_\theta\mathbf{U}_n^* = D_\theta\left(\sum_{k=1}^n U_{j,1}(\theta)\right) \overset{nez.}{=} \sum_{k=1}^n D_\theta U_{k,1}(\theta) = \sum_{k=1}^n \underbrace{E_\theta(U_{k,1}(\theta))^2}_{=J(\theta)} = nJ(\theta).$$

takže

$$|\gamma'(\theta)| = |E[\mathbf{U}_n^*T_n(\mathbf{X})]| = |\underbrace{C(\mathbf{U}_n^*(\theta),T_n(\mathbf{X}))}_{vizE\mathbf{U}_n^*=0}| \overset{Schwarz.ner.}{\le} \sqrt{DT_n(\mathbf{X})}\underbrace{\sqrt{D\mathbf{U}_n^*(\theta)}}_{=\sqrt{nJ(\theta)}}.$$

tj.

$$(\gamma'(\theta))^2 \le DT_n(\mathbf{X})nJ(\theta) \qquad \Rightarrow \qquad \frac{(\gamma'(\theta))^2}{nJ(\theta)} \le DT_n(\mathbf{X}),$$

čímž je tvrzení dokázáno. □

DEFINICE 4.6. Řekneme, že odhad $T_n(\mathbf{X})$ je
(a) **VYDATNÝM** (také EFICIENTNÍM) odhadem $\gamma(\boldsymbol{\theta})$, pokud

$$\varepsilon[T_n(\mathbf{X})] = \frac{C_n(\boldsymbol{\theta})}{DT_n(\mathbf{X})} = 1$$

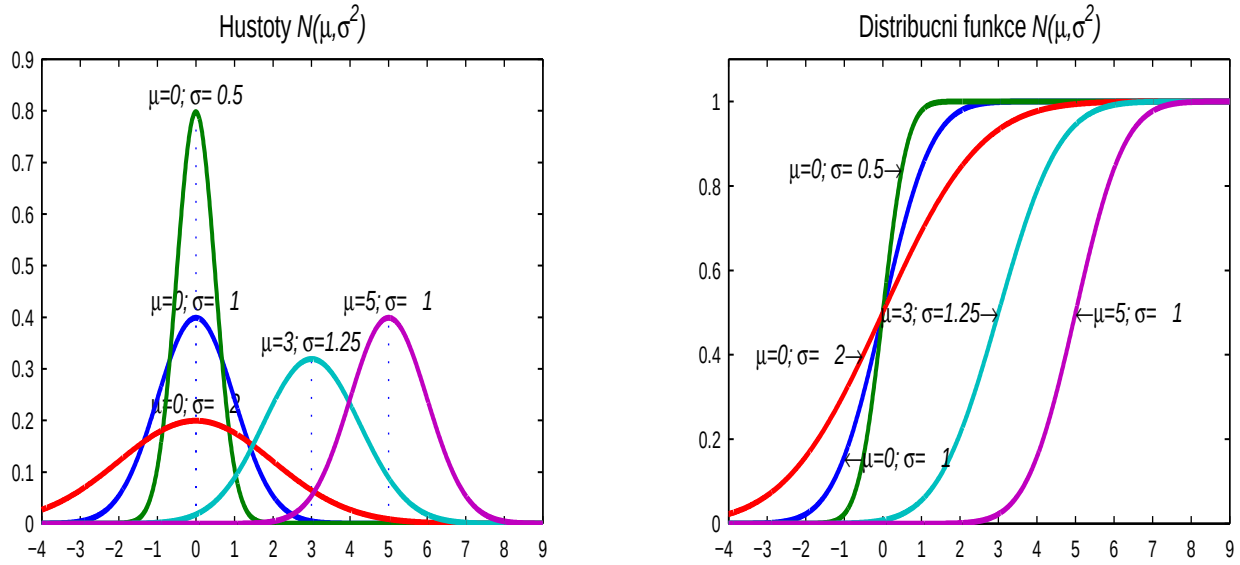
(b) **ASYMPTOTICKY VYDATNÝM** odhadem $\gamma(\boldsymbol{\theta})$, pokud

$$\lim_{n\to\infty} \varepsilon[T_n(\mathbf{X})] = 1$$

a číslo $\boxed{\varepsilon[T_n(\mathbf{X})]}$ se nazývá **vydatnost** (eficience) **odhadu** $T_n(\mathbf{X})$.

Příklad 4.7. NORMÁLNÍ ROZDĚLENÍ A REGULARITA.
Mějme náhodnou veličinu X s normálním rozdělením

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



Obrázek 1: Ukázky hustot a distribučních funkcí pro různé hodnoty parametrů μ a σ^2.

přičemž:

(a) σ^2 je **známé**, tj. $\theta_1 = \mu$. Pak **hustota** $f(x)$ **je regulární** (viz body (1) až (5)):

(1) Množina $\boldsymbol{\Theta}_1 = (-\infty, \infty)$ je neprázdná otevřená množina.

(2) Množina $M = \{x \in \mathbb{R} : f(x) > 0\}$ je $(-\infty, \infty)$ a nezávisí na $\mu \in \boldsymbol{\Theta}_1$.

(3) Pro každé $\mathbf{y} \in M$ existuje konečná derivace

$$f'_\mu(x) = \frac{d\, f(x)}{d\, \mu} = f(x)\frac{x-\mu}{\sigma^2} \quad \Rightarrow \quad U_1 = \frac{X-\mu}{\sigma^2}.$$

(4) Pro všechna $\mu \in \boldsymbol{\Theta}_1$ platí

$$EU_1 = \int_{-\infty}^{\infty} \frac{f'_\mu(x)}{f(x)} f(x)dx = \int_{-\infty}^{\infty} f'_\mu(x)dx = \frac{1}{\sigma^2} \underbrace{\int_{-\infty}^{\infty} (x-\mu) f(x)dx}_{0} = 0.$$

(5) Pro všechna $\mu \in \mathbb{R}$ je integrál J_{11} konečný a kladný

$$J(\mu) = J_{11} = EU_1^2 = \int_{-\infty}^{\infty} \left(\frac{f'_\mu(x)}{f(x)}\right)^2 f(x)dx = \int_{-\infty}^{\infty} \frac{\left(f'_\mu(x)\right)^2}{f(x)}dx$$
$$= \frac{1}{\sigma^4} \underbrace{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx}_{DX=\sigma^2}$$
$$= \frac{1}{\sigma^2} > 0.$$

(b) μ je **známé**, tj. $\theta_2 = \sigma^2$. Pak **hustota** $f(x)$ **je regulární** (viz body (1) až (5)):

(1) Množina $\mathbf{\Theta}_2 = (0, \infty)$ je neprázdná otevřená množina.

(2) Množina $M = \{x \in \mathbb{R} : f(x) > 0\}$ je $(-\infty, \infty)$ a nezávisí na $\sigma^2 \in \mathbf{\Theta}_2$.

(3) Pro každé $x \in M$ existuje konečná derivace

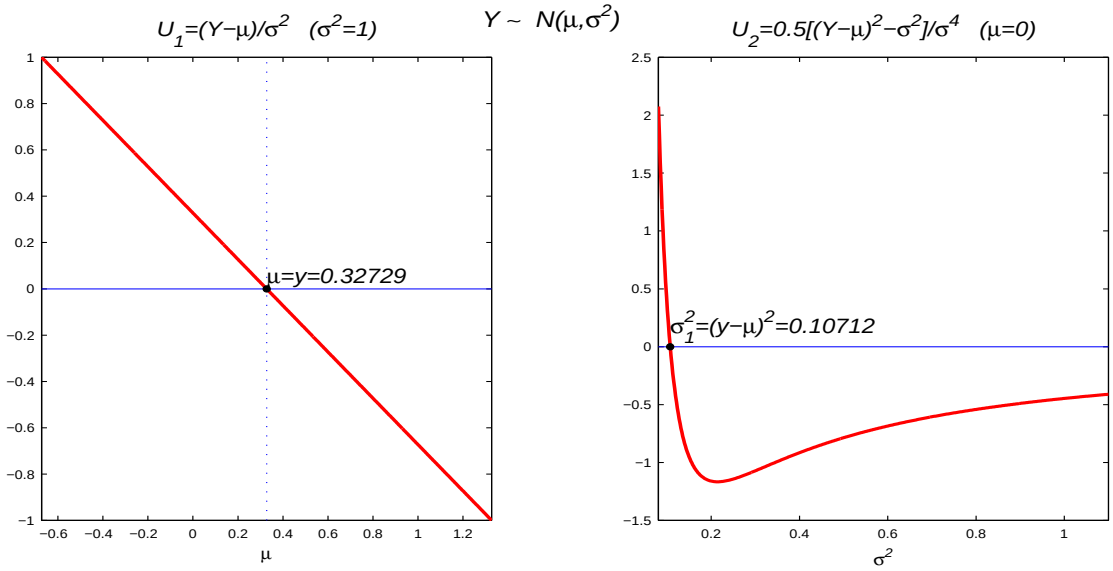
$$f'_{\sigma^2}(x) = \frac{d\, f(x)}{d\, \sigma^2} = f(x)\frac{(x-\mu)^2-\sigma^2}{2\sigma^4} \quad \Rightarrow \quad U_2 = \frac{(X-\mu)^2-\sigma^2}{2\sigma^4}.$$

(4) Pro všechna $\sigma^2 \in \mathbf{\Theta}_2$ platí

$$EU_2 = \int_{-\infty}^{\infty} \frac{f'_{\sigma^2}(x)}{f(x)} f(x)dx = \int_{-\infty}^{\infty} f'_{\sigma^2}(x)dx = \int_{-\infty}^{\infty} f(x)\frac{(x-\mu)^2-\sigma^2}{2\sigma^4}dx = 0.$$

(5) Pro všechna $\sigma^2 \in \mathbf{\Theta}_2$ je integrál J_{22} konečný a kladný

$$J(\sigma^2) = J_{22} = EU_2^2 = \int_{-\infty}^{\infty} \left(\frac{f'_{\sigma^2}(x)}{f(x)}\right)^2 f(x)dx = \frac{1}{4\sigma^8} \int_{-\infty}^{\infty} \left[(x-\mu)^2 - \sigma^2\right]^2 f(x)dx$$
$$= \frac{1}{4\sigma^8} \underbrace{\int_{-\infty}^{\infty} (x-\mu)^4 f(x)dx}_{\mu_4=3\sigma^4} - \frac{2\sigma^2}{4\sigma^8} \underbrace{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx}_{\sigma^2} + \frac{\sigma^4}{4\sigma^8} \underbrace{\int_{-\infty}^{\infty} f(x)dx}_{1}$$
$$= \frac{1}{2\sigma^4} > 0$$



Obrázek 2: Ukázky skórových funkcí U_1 (resp. U_2) pro $N(\mu, \sigma^2)$ při známém σ^2 (resp. μ).

(c) $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)'$. Pak **hustota** $f(x)$ **je regulární** (viz body (1) až (5)).

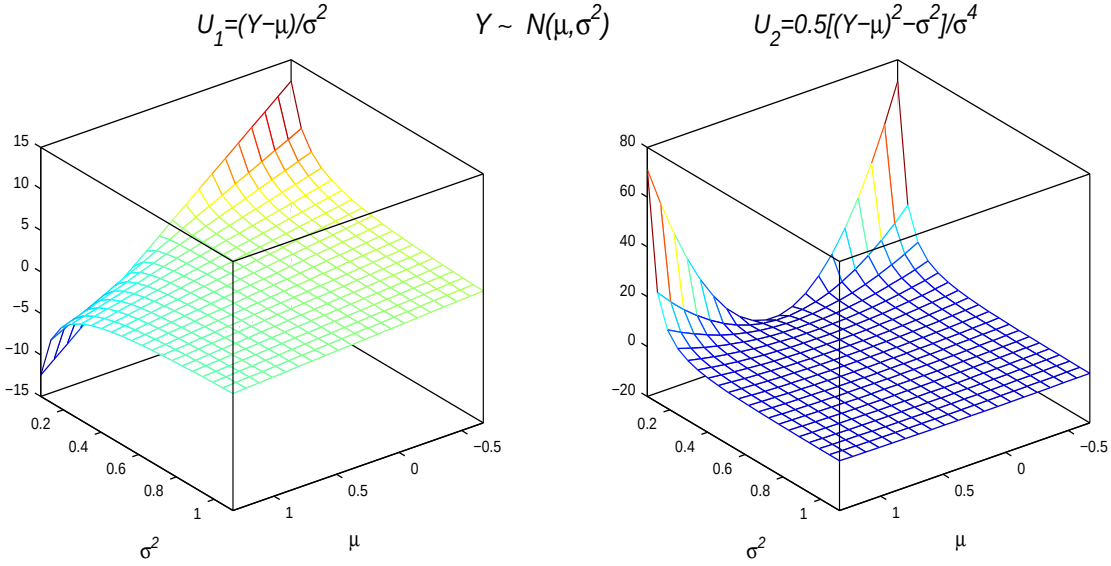
(1) Množina $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 = (-\infty, \infty) \times (0, \infty)$ je neprázdná otevřená množina.

(2) Množina $M = \{x \in \mathbb{R} : f(x) > 0\}$ je $(-\infty, \infty)$ a nezávisí na $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(3) Pro každé $x \in M$ existují konečné derivace $f'_\mu(x), f'_{\sigma^2}(x)$ (viz předchozí dva případy).

(4) Pro všechna $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)' \in \boldsymbol{\Theta}$ platí $EU_1 = EU_2 = 0$ (viz předchozí dva případy) a skórový vektor je roven

$$\mathbf{U} = \left(\tfrac{X-\mu}{\sigma^2}, \tfrac{(X-\mu)^2-\sigma^2}{2\sigma^4} \right)'.$$



Obrázek 3: Ukázky skórových funkcí U_1 a U_2 pro $N(\mu, \sigma^2)$ při neznámém σ^2 a μ.

(5) Pro všechna $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)' \in \boldsymbol{\Theta}$ jsou integrály J_{11}, J_{22} a $J_{12} = J_{21}$ konečné, přičemž

$$\begin{aligned} J(\mu, \sigma^2) = J_{12} &= \int_{-\infty}^{\infty} \tfrac{f'_\mu(x)}{f(x)} \tfrac{f'_{\sigma^2}(x)}{f(x)} f(x)dx \\ &= \tfrac{1}{2\sigma^6} \int_{-\infty}^{\infty} (x-\mu) \left[(x-\mu)^2 - \sigma^2\right] f(x)dx \\ &= \tfrac{1}{2\sigma^6} \underbrace{\int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx}_{\mu_3=0} - \tfrac{1}{2\sigma^4} \underbrace{\int_{-\infty}^{\infty} (x-\mu) f(x)dx}_{0} = 0 \end{aligned}$$

a Fisherova informační matice pro vektor parametrů $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)'$ je rovna

$$\mathbf{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

a je pozitivně definitní.

Příklad 4.8. WEIBULLOVO 3-PARAMETRICKÉ EXPONENCIÁLNÍ ROZDĚLENÍ $Wb(\gamma,\theta,\delta)$ **A REGULARITA**. Mějme náhodnou veličinu X s hustotou

$$f(x;\gamma,\theta,\delta)=\begin{cases}\frac{\gamma}{\delta}\left(\frac{x-\theta}{\delta}\right)^{\gamma-1}\exp\left\{-\left(\frac{x-\theta}{\delta}\right)^{\gamma}\right\} & x>\theta,\ \theta\in\mathbb{R},\ \gamma>0,\ \delta>0\\ 0 & \text{jinak.}\end{cases}$$

Zřejmě **nejde o regulární systém hustot**, neboť množina M, což je definiční obor náhodné veličiny, je **závislý na parametru** θ.

Příklad 4.9. NORMÁLNÍ ROZDĚLENÍ A VYDATNÉ ODHADY. Mějme náhodnou veličinu X s normálním rozdělením

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

a náhodný výběr $\mathbf{X}_n=(X_1,\ldots,X_n)'$ z téhož rozdělení, přičemž:

(a) σ^2 je **známé**, tj. $\theta_1=\mu$.

(1) Skórová funkce náhodného výběru (viz příklad 4.7):

$$U_1^*(\mu)=\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma^2}.$$

(2) Fisherova informace o parametru μ z náhodného výběru (viz příklad 4.7 a důkaz věty 4.5):

$$J_n^*(\mu)=nJ(\mu)=nJ_{11}=\frac{n}{\sigma^2}.$$

(3) Uvažujme parametrickou funkci

$$\gamma(\mu)=\mu$$

a výběrový průměr, tj. statistiku

$$T_n(\mathbf{X})=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

(i) Platí

$$E\bar{X}=\mu,$$

tj. $\bar{X}$ je **nestranným odhadem** parametru μ a

$$D\bar{X}=\frac{\sigma^2}{n}.$$

(ii) $\bar{X}$ je **regulárním odhadem** parametrické funkce $\gamma(\mu)=\mu$, přičemž

$$\gamma'_{\mu}(\mu)=1,$$

neboť $\bar{X}$ je nestranným odhadem parametru μ a platí

$$
\begin{aligned}
E\big(\bar{X}U_1^*(\mu)\big) &= \frac{1}{n\sigma^2}E\left[\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n X_i - n\mu\right)\right] \\
&= \frac{1}{n\sigma^2}\sum_{i=1}^n \underbrace{EX_i^2}_{\sigma^2+\mu^2} + \frac{2}{n\sigma^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\underbrace{E(X_iX_j)}_{\mu^2 (nez.)} - \frac{n\mu^2}{\sigma^2} \\
&= \frac{\sigma^2+\mu^2}{\sigma^2} + \frac{n(n-1)}{n\sigma^2}\mu^2 - \frac{n\mu^2}{\sigma^2} \\
&= 1 = \gamma'_\mu(\mu).
\end{aligned}
$$

(iii) $\bar{X}$ je **vydatným odhadem** μ, neboť dolní Raova-Cramerova hranice

$$
C_n(\mu) = \frac{\big[\gamma'_\mu(\mu)\big]^2}{J_n(\mu)} = \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} = D\bar{X}.
$$

(b) μ je **známé**, tj. $\theta_2 = \sigma^2$.

(1) Skórová funkce náhodného výběru (viz příklad 4.7):

$$
\begin{aligned}
U_2^*(\sigma^2) &= \sum_{i=1}^n \frac{(X_i-\mu)^2 - \sigma^2}{2\sigma^4} \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^n \left[\underbrace{(X_i-\mu)^2}_{\text{označme } Z_i} - \sigma^2\right] \\
&= \frac{1}{2\sigma^4}\sum_{i=1}^n Z_i - \frac{1}{2\sigma^2}.
\end{aligned}
$$

(2) Fisherova informace o parametru

$$
\gamma(\sigma^2) = \sigma^2
$$

z náhodného výběru (viz příklad 4.7 a důkaz věty 4.5):

$$
J_n^*(\sigma^2) = nJ(\sigma^2) = \frac{n}{2\sigma^4}.
$$

(3) Uvažujme parametrickou funkci

$$
\gamma(\sigma^2) = \sigma^2
$$

a výběrový rozptyl, tj. statistiku

$$
\begin{aligned}
T_n(\mathbf{Y}) = S^2 &= \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \\
&= \frac{1}{n-1}\left[\sum_{i=1}^n \underbrace{(X_i-\mu)^2}_{\text{označme } Z_i} - n(\bar{X}-\mu)^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^n Z_i - n(\bar{X}-\mu)^2\right].
\end{aligned}
$$

Počítejme

$$\begin{aligned} EZ_i &= DY_i = \sigma^2 \\ DZ_i &= EZ_i^2 - (EZ_i)^2 = \mu_4 - \sigma^4 = 2\sigma^4 \\ C(Z_i, Z_j) &= E(Z_iZ_j) - \underbrace{E(Z_i)E(Z_j)}_{\sigma^4} = 0 \qquad \Rightarrow \qquad E(Z_iZ_j) = \sigma^4 \quad \text{pro} \quad i \neq j. \end{aligned}$$

Pak

(i) Snadno lze ukázat, že platí

$$ES^2 = \sigma^2,$$

tj. S^2 je **nestranným odhadem** parametru σ^2. Dále obecně pro výběrový rozptyl platí:

$$DS^2 = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}\sigma^4$$

a protože v případě normálního rozdělení máme

$$\mu_4 = 3\sigma^4,$$

dostáváme

$$DS^2 = \frac{3\sigma^4}{n} - \frac{n-3}{n(n-1)}\sigma^4 = \frac{\sigma^4\left[3(n-1) - (n-3)\right]}{n(n-1)} = \frac{2\sigma^4}{n-1}.$$

(ii) S^2 je **regulárním odhadem** parametrické funkce $\gamma(\sigma^2) = \sigma^2$, přičemž

$$\gamma'_{\sigma^2}(\sigma^2) = 1,$$

neboť je nestranným odhadem a platí

$$\begin{aligned} E\big(S^2U_2^*(\sigma^2)\big) &= \frac{1}{2(n-1)\sigma^4} E\left[\Big(\sum_{i=1}^n Z_i - n(\bar{X}-\mu)^2\Big)\Big(\sum_{i=1}^n Z_i - n\sigma^2\Big)\right] \\ &= \frac{1}{2(n-1)\sigma^4}\left[\sum_{i=1}^n \underbrace{EZ_i^2}_{\mu^4=3\sigma^4} + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \underbrace{E(Z_iZ_j)}_{\sigma^4}\right. \\ &\qquad - n\sum_{i=1}^n \underbrace{E\big(Z_i(\bar{X}-\mu)^2\big)}_{\frac{(n+2)\sigma^4}{n^2}} \\ &\qquad \left. - n\sigma^2\sum_{i=1}^n \underbrace{EZ_i}_{\sigma^2} + n^2\sigma^2 \underbrace{E(\bar{X}-\mu)^2}_{D\bar{X}=\frac{\sigma^2}{n}}\right] \\ &= \frac{3n\sigma^4 + n(n-1)\sigma^4 - (n+2)\sigma^4 - n^2\sigma^4 + n\sigma^4}{2(n-1)\sigma^4} \\ &= 1 = \gamma'_{\sigma^2}(\sigma^2), \end{aligned}$$

přičemž platí

$$\begin{aligned} E\big(Z_i(\bar{Y}-\mu)^2\big) &= E\left\{ Z_i\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right]\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right]\right\} \\ &= \frac{1}{n^2}\Bigg[\underbrace{EZ_i^2}_{3\sigma^4} + \sum_{i\neq j=1}^{n}\underbrace{E(Z_iZ_j)}_{\sigma^4} \\ &\qquad + \sum_{i\neq j=1}^{n}\sum_{i\neq j\neq k=1}^{n}\underbrace{E\big(Z_i(X_j-\mu)(X_k-\mu)\big)}_{0}\Bigg] \\ &= \frac{1}{n^2}\big[3\sigma^4+(n-1)\sigma^4\big] = \frac{(n+2)\sigma^4}{n^2}. \end{aligned}$$

(iii) S^2 je **asymptoticky vydatným odhadem** σ^2, neboť dolní Raova-Cramerova hranice je rovna

$$\boxed{C_n(\sigma^2)} = \frac{\big[\gamma'_{\sigma^2}(\sigma^2)\big]^2}{J_n(\sigma^2)} = \frac{1}{\frac{n}{2\sigma^4}} = \boxed{\frac{2\sigma^4}{n}} < DS^2 = \frac{2\sigma^4}{n-1}$$

a

$$\lim_{n\to\infty}\frac{C_n(\sigma^2)}{DS^2} = 1.$$

5. KONSTRUKCE BODOVÝCH ODHADŮ

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ rozsahu n z rozdělení o distribuční funkci $F(x; \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Množina $\boldsymbol{\Theta}$ nechť je neprázdná a otevřená.

Budeme předpokládat, že distribuční funkci $F(x; \boldsymbol{\theta})$ lze vyjádřit ve tvaru

$$F(x; \boldsymbol{\theta}) = \int_{-\infty}^{x} f(x; \boldsymbol{\theta}) d\nu(t) \qquad x \in \mathbb{R} \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \boldsymbol{\Theta},$$

kde ν je $\sigma-$konečná míra na $(\mathbb{R}, \mathcal{B})$ (např. *Lebesgueova* nebo *čítací*) a $f(x; \boldsymbol{\theta})$ je nezáporná měřitelná funkce, tzv. **hustota pravděpodobnosti** (vzhledem k míře ν).

Pak **sdružená hustota** náhodného vektoru $\mathbf{X}_n = (X_1, \ldots, X_n)'$ je vzhledem k nezávislosti jednotlivých složek vektoru a jejich stejnému rozdělení rovna

$$f_{\mathbf{X}}(x_1, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

Mějme dále **parametrickou funkcí**

$$\gamma : \boldsymbol{\Theta} \to \mathbb{R}.$$

Předmětem našeho zájmu bude hodnota parametru $\boldsymbol{\theta}$ nebo, obecněji, hodnota některé parametrické funkce $\gamma(\boldsymbol{\theta})$.

5.1. METODA MOMENTŮ. Předpokládejme, že pro náhodný výběr existují obecné momenty:

$$\mu'_k = \mu'_k(\boldsymbol{\theta}) = EX_i^k \qquad i = 1, \ldots, n \qquad k = 1, \ldots, m.$$

Výběrové obecné momenty jsou definovány vzorcem

$$M'_k = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} X_i^k \qquad k = 1, 2, \ldots$$

Momentová metoda odhadu parametru $\boldsymbol{\theta}$ spočívá v tom, že za odhad $\tilde{\boldsymbol{\theta}}$ vezmeme řešení rovnic

$$M'_k = \mu'_k(\boldsymbol{\theta}) \qquad k = 1, \ldots, m.$$

a nazveme je **odhadem metodou momentů**.

Někdy se může stát, že m rovnic nepostačuje k jednoznačnému určení $\tilde{\boldsymbol{\theta}}$, pak se většinou připojují další rovnice

$$M'_k = \mu'_k(\boldsymbol{\theta}) \qquad \text{pro} \qquad k = m+1, m+2$$

atd., až se získá potřebný počet rovnic. To samozřejmě lze provádět jen za předpokladu, že existují příslušné momenty μ'_k.

Odhadem dané **parametrické funkce** $\gamma(\boldsymbol{\theta})$ **metodou momentů** rozumíme statistiku

$$\widetilde{\gamma} = \gamma(\widetilde{\boldsymbol{\theta}}).$$

Odhady získané metodou momentů obvykle nejsou dostatečně kvalitní, v jednotlivých konrétních případech zpravidla lze dokázat konzistenci odhadů.

Příklad 5.1. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ rozsahu n z normálního rozdělení o parametrech μ a σ^2, které odhadneme momentovou metodou.

Pak

$$\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)',$$

tj. $m = 2$ a $\boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$.

Snadno lze spočítat, že

$$\mu_1' = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu$$
$$\mu_2' = \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu^2 + \sigma^2.$$

Výběrové obecné momenty jsou rovny

$$M_1' = \tfrac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$
$$M_2' = \tfrac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Chceme-li najít odhady momentovou metodou, musíme řešit soustavu rovnic:

$$M_1' = \mu$$
$$M_2' = \mu^2 + \sigma^2$$

Z první rovnice ihned dostaneme

$$\widetilde{\mu} = \bar{X},$$

což dosadíme do druhé rovnice a počítáme

$$\widetilde{\sigma}^2 = M_2' - \bar{X}^2 = \tfrac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \underbrace{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)}_{=(n-1)S^2} = \tfrac{n-1}{n} S^2,$$

kde

$$S^2 = \tfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad \text{je výběrový rozptyl.}$$

Protože

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

vidíme, že že odhad $\widetilde{\mu}$ **je nestranný**, avšak

$$E_{\boldsymbol{\theta}}(\tilde{\sigma}^2) = E \tfrac{n-1}{n} S^2 = \tfrac{n-1}{n} \sigma^2,$$

takže $\widetilde{\sigma}^2$ **není nestranný**, avšak je **asymptoticky nestranný**.

Lze ukázat, že oba odhady jsou **konzistentní** (slabě i silně).

5.2. METODA MAXIMÁLNÍ VĚROHODNOSTI. Označme sdruženou hustotu pravděpodobnosti náhodného vektoru $\mathbf{X}$ takto

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$$

a nazveme ji **věrohodnostní funkcí náhodného výběru**.

Odhad $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ nazveme **maximálně věrohodným**, jestliže pro každé $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ platí

$$L(\widehat{\boldsymbol{\theta}}_{\text{MLE}}; x_1, \ldots, x_n) \geq L(\boldsymbol{\theta}; x_1, \ldots, x_n).$$

Zpravidla je vhodnější pracovat s logaritmem funkce L. Pak za předpokladů známých z diferenciálního počtu vede hledání maximálně věrohodného odhadu $\hat{\boldsymbol{\theta}}$ k řešení rovnic

$$\frac{\partial}{\partial \theta_j} \ln L(\theta_1, \ldots, \theta_m; x_1, \ldots, x_n) = \frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}; \mathbf{x}) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \ln f(x_i; \theta_1, \ldots, \theta_m) = 0 \qquad j = 1, \ldots, m$$

které jsou ve statistické literatuře známé pod názvem **soustava věrohodnostních rovnic**.

Příklad 5.2. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ rozsahu n z binomického rozdělení o parametrech m a π. Parametr π odhadneme metodou maximální věrohodnosti.
Pro náhodný výběr z binomického rozdělení platí

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq Bi(m, \pi) \quad \sim \quad p(x) = \begin{cases} \binom{m}{x} \pi^x (1-\pi)^{m-x} & x = 0, 1, \ldots, m, \\ 0 & \text{jinak.} \end{cases}$$

Věrohodnostní funkce:

$$\begin{aligned} L(\pi; X_1, \ldots, X_n) &= \prod_{i=1}^{n} \binom{m}{X_i} \pi^{X_i} (1-\pi)^{m-X_i} \\ &= \pi^{\sum_{i=1}^n X_i} (1-\pi)^{nm - \sum_{i=1}^n X_i} \prod_{i=1}^{n} \binom{m}{X_i} = \pi^{n\bar{X}} (1-\pi)^{n(m-\bar{X})} \prod_{i=1}^{n} \binom{m}{X_i}. \end{aligned}$$

Logaritmus věrohodnostní funkce:

$$\boxed{l(\pi; X_1, \ldots, X_n) = \sum_{i=1}^{n} \ln \binom{m}{X_i} + n\bar{X} \ln \pi + n(m - \bar{X}) \ln(1-\pi)}$$

Věrohodnostní rovnice:

$$\frac{\partial l}{\partial \pi} = \frac{1}{\pi} n\bar{X} - \frac{1}{1-\pi} n(m - \bar{X}) = 0 \quad \Rightarrow \quad \boxed{\widehat{\pi}_{\text{MLE}} = \frac{\bar{X}}{m}}.$$

Vzhledem k tomu, že nepředpokládáme degenerované binomické rozdělení s nulovým rozptylem, takže s pravděpodobností 1 musí platit

$$0 < \bar{X} < m,$$

pak snadno ověříme, že jde o maximum, neboť pokud spočítáme druhé parciální derivace

$$\frac{\partial^2}{\partial^2 \pi} l(\pi; X_1, \ldots, X_n) = -\frac{1}{\pi^2} n\bar{X} - \frac{1}{(1-\pi)^2} n(m - \bar{X}) = -n \left[\frac{\bar{X}}{\pi^2} + \frac{m - \bar{X}}{(1-\pi)^2} \right] < 0.$$

Příklad 5.3. Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ rozsahu n z normálního rozdělení o parametrech μ a σ^2. Tyto parametry odhadneme metodou maximální věrohodnosti.

Opět $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\mu, \sigma^2)'$, tj. $m = 2$ a $\boldsymbol{\Theta} = \mathbb{R} \times (0, \infty)$.

Pak

$$L(\boldsymbol{\theta}; X_1, \ldots, X_n) = L(\mu, \sigma^2; X_1, \ldots, X_n) = \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{X_i-\mu}{\sigma}\right)^2} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

$$\boxed{\ln L(\mu, \sigma^2; X_1, \ldots, X_n) = l(\mu, \sigma^2; X_1, \ldots, X_n) = -\tfrac{n}{2}\ln(2\pi\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2}.$$

Vyjádřeme věrohodnostní rovnice

$$\tfrac{\partial \ln L}{\partial \sigma^2} = -\tfrac{n}{2}\tfrac{1}{2\pi\sigma^2}2\pi + \tfrac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

$$\tfrac{\partial \ln L}{\partial \mu} = \tfrac{1}{2\sigma^2}\sum_{i=1}^{n} 2(X_i - \mu) = 0$$

Z druhé rovnice plyne, že

$$\boxed{\widehat{\mu}_{\mathrm{MLE}}} = \tfrac{1}{n}\sum_{i=1}^{n} X_i = \boxed{\bar{X}} \quad \ldots \quad \text{výběrový průměr}$$

Po dosazení do první věrohodnostní rovnice dostaneme

$$-n\sigma^2 + \sum_{i=1}^{n}(X_i - \bar{X})^2 = 0 \qquad \Rightarrow \qquad \boxed{\widehat{\sigma}^2_{\mathrm{MLE}}} = \tfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \boxed{\tfrac{n-1}{n}S^2 = S^{*2}},$$

kde

$$S^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{je výběrový rozptyl.}$$

Upravme nejprve logaritmus věrohodnostní funkce takto:

$$\begin{aligned} l(\mu, \sigma^2; X_1, \ldots, X_n) &= -\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln(\sigma^2) - \tfrac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(X_i - \bar{X}) + (\bar{X} - \mu)\right]^2 \\ &= -\tfrac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \tfrac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(X_i - \bar{x})^2 + n(\bar{X} - \mu)^2\right\} \\ &= -\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln(\sigma^2) - \tfrac{1}{2\sigma^2}\left[nS^{*2} + n(\bar{X} - \mu)^2\right]. \end{aligned}$$

Nyní dokažme, že funkce $l(\mu, \sigma^2; X_1, \ldots, X_n)$ nabývá pro jakoukoliv realizaci

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega) \quad \text{pro každé} \quad \omega \in \Omega$$

v bodě $(\widehat{\mu}_{\mathrm{MLE}}, \widehat{\sigma}^2_{\mathrm{MLE}}) = (\bar{x}, s^{*2})$ svého maxima, takže po dosazení dostáváme

$$l(\bar{x}, s^{*2}; x_1, \ldots, x_n) = -\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln(s^{*2}) - \tfrac{n}{2}.$$

Ověřme, zda platí

$$\begin{aligned} l(\mu,\sigma^2;x_1,\dots,x_n) &\overset{?}{\leq} l(\bar{x},s^{*2};x_1,\dots,x_n) \\ -\tfrac{n}{2}\ln(2\pi)-\tfrac{n}{2}\ln(\sigma^2)-\tfrac{ns^{*2}+n(\bar{x}-\mu)^2}{2\sigma^2} &\overset{?}{\leq} -\tfrac{n}{2}\ln(2\pi)-\tfrac{n}{2}\ln(s^{*2})-\tfrac{n}{2} \\ 0 &\overset{?}{\leq} \underbrace{\left(\tfrac{s^{*2}}{2\sigma^2}-\tfrac{1}{2}\right)-\ln\tfrac{s^*}{\sigma}}_{\text{1. člen}}+\underbrace{\tfrac{(\bar{x}-\mu)^2}{2\sigma^2}}_{\geq 0} \end{aligned}$$

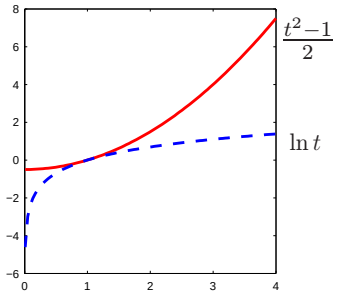
Protože pro všechna kladná

$$t=\tfrac{s^*}{\sigma}>0$$

platí

$$\ln t<\tfrac{t^2-1}{2},$$

je první i druhý člen nezáporný a nerovnost platí.



Protože

$$E_{\boldsymbol{\theta}}(\widehat{\mu}_{\text{MLE}})=E_{\boldsymbol{\theta}}\bar{X}=\mu,$$

ale

$$E_{\boldsymbol{\theta}}(\hat{\sigma}^2_{\text{MLE}})=E_{\boldsymbol{\theta}}\tfrac{n-1}{n}S^2=\tfrac{n-1}{n}\sigma^2,$$

vidíme že odhad $\widehat{\mu}_{\text{MLE}}$ **je nestranný**, avšak $\widehat{\sigma}^2_{\text{MLE}}$ již **nestranný není** (ale asymptoticky nestranný).

V tomto případě jsme došli ke stejnému výsledku jako u momentové metody.

Poznámka 5.4. Maximálně věrohodné odhady mají řadu výhodných vlastností:

(1) Existuje-li **vydatný** (eficientní) **odhad**, má soustava věrohodnostních rovnic **jediné řešení** a to je rovné **vydatnému** (eficientnímu) **odhadu**.
(2) Existuje-li **postačující** (suficientní) **odhad**, **je každé řešení věrohodnostních rovnic** funkcí **postačujícího** (suficientního) **odhadu**.
(3) Pochází-li náhodný výběr z **regulárního rozdělení**, pak existuje maximálně věrohodný odhad, který je **konzistentní** a **asymptoticky normální**, tj. v jednorozměrném případě

$$\widehat{\theta}_{\text{MLE}}\overset{A}{\sim}N(\theta,nJ(\theta)).$$

5.3. Srovnání metody momentů s metodou maximální věrohodnosti.

Obecně se dá říci, že momentová metoda je poměrně jednoduchá. Používá se zejména v těch případech, kdy jiné metody odhadu jsou numericky či z jiných důvodů těžko zvládnutelné. Na druhé straně pokud jde o rozdělení, která nemají konečné momenty, pak se tato metoda nedá aplikovat vůbec. Někdy se odhady pořízené momentovou metodou berou alespoň jako počáteční aproximace pro řešení věrohodnostních rovnic, pokud je pro jejich řešení nutný iterační postup.

5.4. METODA MINIMÁLNÍHO χ^2.

Nejprve si připomeňme jedno velmi důležité vícerozměrné diskrétní rozdělení, a to multinomické.

Multinomické rozdělení popisuje situaci, kdy máme k neslučitelných jevů, které mohou nastat v každém z n nezávislých pokusů s pravděpodobnostmi

$$\pi_1, \dots, \pi_k \qquad \text{přičemž} \qquad \sum_{j=1}^{k} \pi_j = 1.$$

Nechť náhodná veličina Y_j značí počet případů, kdy nastal j-tý jev, takže Y_j může nabývat hodnot od nuly do n a musí platit

$$\sum_{j=1}^{k} Y_j = n.$$

Náhodný vektor $\mathbf{Y} = (Y_1, \dots, Y_k)'$ pak má multinomické rozdělení s pravděpodobnostní funkcí

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} n! \prod_{j=1}^{k} \frac{\pi_j^{y_j}}{y_j!} & \text{pro } y_j = 0, 1, \dots, n;\ \sum_{j=1}^{k} y_j = n\ \sum_{j=1}^{k} \pi_j = 1 \\ 0 & \text{jinak} \end{cases},$$

což lze ekvivalentně napsat i takto

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} n! \frac{\pi_1^{y_1} \cdot \dots \cdot \pi_{k-1}^{y_{k-1}} (1-\pi_1-\dots-\pi_{k-1})^{(n-y_1-\dots-y_{k-1})}}{y_1! \cdot \dots \cdot y_{k-1}! (n-y_1-\dots-y_{k-1})!} & \text{pro } y_j = 0, 1, \dots, n \\ 0 & \text{jinak.} \end{cases}$$

a značíme

$$\boxed{\mathbf{Y} \quad \sim \quad Mn(n, \pi_1, \dots, \pi_k)},$$

přičemž platí pro $j, h = 1, \dots, k$

$$\begin{aligned} EY_j &= n\pi_j \\ DY_j &= n\pi_j(1-\pi_j) \\ C(Y_j, Y_h) &= -n\pi_j\pi_h. \end{aligned}$$

Multinomické rozdělení je zobecněním binomického rozdělení a je patrně nejdůležitějším diskrétním mnohorozměrným rozdělením. Svým významem by se dalo přirovnat k mnohorozměrnému normálnímu rozdělení, jemuž se podobá především díky dvěma vlastnostem: podmíněná i marginální rozdělení jsou opět multinomická.

Nyní se opět vrátíme k náhodnému výběru $\mathbf{X} = (X_1, \dots, X_n)'$ rozsahu n z rozdělení o distribuční funkci $F(x; \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)' \in \boldsymbol{\Theta} \subset \mathbb{R}^m$.

Při odhadu neznámého parametru $\boldsymbol{\theta}$ metodou minimálního χ^2 na základě náhodného výběru $\mathbf{X} = (X_1, \dots, X_n)'$ postupujeme tak, že

(1) rozdělí se interval $(-\infty, \infty)$ na konečný počet pod dvou disjunktních podmnožin $B_1, \dots, B_k$ (pokud nejde o výběr z diskrétního rozdělení, které nabývá pouze konečného počtu hodnot)

(2) určí se pravděpodobnosti

$$p_j(\boldsymbol{\theta}) = \int_{B_j} dF(x;\boldsymbol{\theta})$$

jako funkce parametru $\boldsymbol{\theta}$

(3) pro danou realizaci náhodného výběru se určí bod $\widehat{\boldsymbol{\theta}}$, v němž funkce

$$\chi^2(\boldsymbol{\theta}) = \sum_{j=1}^{k} \left(\frac{Y_j - np_j(\boldsymbol{\theta})}{\sqrt{np_j(\boldsymbol{\theta})}} \right)^2$$

nabývá minima, přičemž

$$Y_j = \sum_{i=1}^{n} I(X_i \in B_j)$$

je počet bodů $X_1, \ldots, X_n$ ležících v B_j (samozřejmě musí platit $\sum_{j=1}^{k} Y_j = n$).

Pokud je tato funkce diferencovatelná, hledání minima vede na řešení soustavy rovnic

$$-\frac{1}{2}\frac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_h} = \sum_{j=1}^{k} \left(\frac{Y_j - np_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} + \frac{[Y_j - np_j(\boldsymbol{\theta})]^2}{2np_j^2(\boldsymbol{\theta})} \right) \frac{\partial p_j(\boldsymbol{\theta})}{\partial \theta_h} = 0 \quad (h = 1, \ldots, k) \quad (12)$$

vzhledem k neznámým $\theta_1, \ldots, \theta_k$. Avšak i v nejjedodušších případech je velmi obtížné řešit systém rovnice (12). Potíže způsobuje člen

$$\frac{[Y_j - np_j(\boldsymbol{\theta})]^2}{2np_j^2(\boldsymbol{\theta})}.$$

Pro velká n je však vliv tohoto členu zanedbatelný, a proto se řešení soustavy (12) nahrazuje řešením soustavy

$$\sum_{j=1}^{k} \frac{Y_j - np_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} \frac{\partial p_j(\boldsymbol{\theta})}{\partial \theta_h} = 0 \quad (h = 1, \ldots, k) \quad (13)$$

Tento postup se nazývá **modifikovanou metodou minimálního** χ^2.

Odhady získané oběma metodami jsou při dosti obecných podmínkách **konzistentními odhady**.

6. INTERVALOVÉ ODHADY

6.1. Definice intervalového odhadu. Odhady, jimiž jsme se doposud zabývali, se někdy nazývají **bodové odhady** parametrické funkce $\gamma(\boldsymbol{\theta})$.

Je tomu tak proto, že pro danou realizaci náhodného výběru $x_1, \ldots, x_n$ představuje odhad daný statistikou $T_n(x_1, \ldots, x_n)$ **jediné číslo** (**bod**), které je v jistém smyslu přiblížením ke skutečné hodnotě parametrické funkce $\gamma(\boldsymbol{\theta})$.

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

Podobná je úloha zkonstruovat na základě náhodného výběru statistiku, o níž lze s dostatečně velkou spolehlivostí prohlásit, že skutečná hodnota parametrické funkce je větší než tato statistika. V tomto případě mluvíme o **dolním odhadu** parametrické funkce $\gamma(\boldsymbol{\theta})$. Analogicky lze zavést pomocí opačné nerovnosti pojem **horního odhadu** $\gamma(\boldsymbol{\theta})$.

DEFINICE 6.1. Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x; \boldsymbol{\theta})$ je náhodný výběr rozsahu n z rozdělení o distribuční funkci $F(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Dále mějme parametrickou funkci $\gamma(\boldsymbol{\theta})$, $\alpha \in (0, 1)$ a statistiky $D = D(X_1, \ldots, X_n)$ a $H = H(X_1, \ldots, X_n)$.

Potom intervaly $\langle D, H\rangle$ nazveme $100(1-\alpha)$ % **intervalem spolehlivosti** pro parametrickou funkci $\gamma(\boldsymbol{\theta})$ jestliže

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

Jestliže

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

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

Jestliže

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

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

Poznámka 6.2. Vysvětleme si nyní smysl pojmu **spolehlivost intervalových odhadů**.

Konkrétní data $x_1, \ldots, x_n$ (tj. realizace náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$) nejsou náhodnými veličinami, nýbrž jsou to výsledky určitého pokusu ω, tj.

$$x_1 = X_1(\omega), \ldots, x_n = X_n(\omega).$$

Sestrojíme-li tedy na jejich základě intervalový odhad, řekněme (a, b), parametrické funkce $\gamma(\theta)$, pak nemá smysl mluvit o pravděpodobnosti $P(a < \gamma(\theta) < b)$, protože všechny tři symboly jsou reálná čísla (třebaže $\gamma(\theta)$ neznáme) a nerovnost $a < \gamma(\theta) < b$ buď platí nebo neplatí, tj. náš intervalový odhad je buď správný nebo nesprávný.

Budeme-li však sestrojovat intervalové odhady vícekrát po sobě, pak poměrná četnost případů, kdy intervalový odhad bude správný, bude přibližně rovna $1 - \alpha$.

Číslo $\boxed{\alpha}$ se volí poměrně malé, nejčastěji	0.05	spolehlivost je pak	0.95	tj. 95%
	0.01		0.99	tj. 99%

Kromě dostatečné spolehlivosti bychom chtěli, aby interval $\langle D_n(\mathbf{X}), T_n(\mathbf{X})\rangle$ byl co možná nejkratší.

Tyto požadavky jsou však (při pevném rozsahu výběru n) protichůdné. Žádáme-li větší spolehlivost, musíme se smířit s delším intervalem; žádáme-li naopak kratší interval, musíme se smířit s nižší spolelivostí.

6.2. Kvantily. Nyní definujme kvantilovou funkci a kvantil.

Definice 6.3. Nechť F je distribuční funkcí a $\alpha \in (0,1)$. Potom funkci

$$F^{-1}(\alpha) = Q(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$$

se nazývá **kvantilová funkce** a číslo

$$x_\alpha = Q(\alpha)$$

se nazývá α-**kvantilem** rozdělení s distribuční funkcí $F(x)$, přičemž

$x_{0.25}$	$=$	$Q(0.25)$	se nazývá	**dolní kvartil**
$x_{0.5}$	$=$	$Q(0.5)$		**medián**
$x_{0.75}$	$=$	$Q(0.75)$		**horní kvartil**
$x_{0.75} - x_{0.25}$	$=$	IQR		**interkvartilé rozpětí**

Z definice kvantilů vyplývá následující vztah. Je-li X absolutně spojitá náhodná veličina, pak platí

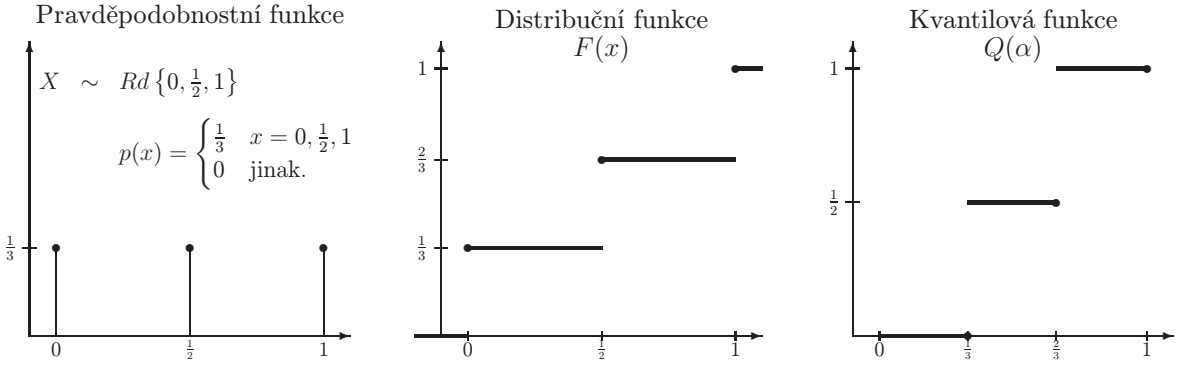
$$P(x_{\alpha/2} < X \leq x_{1-\alpha/2}) = F(x_{1-\alpha/2}) - F(x_{\alpha/2})$$

Příklad 6.4.

Kvantilová funkce diskrétního rozdělení

Uvažujme diskrétní rozdělení, ve kterém náhodná veličina X nabývá pouze tří hodnot $0, \frac{1}{2}$ a 1 se stejnými pravděpodobnostmi.

Toto rozdělení nazveme **rovnoměrně diskrétní** a budeme značit $Rd\left\{0, \frac{1}{2}, 1\right\}$, takže pravděpodobnostní, distribuční a kvantilová funkce jsou tvaru



Příklad 6.5.

KVANTILOVÁ FUNKCE SPOJITÉHO ROZDĚLENÍ

Uvažujme spojité exponenciální rozdělení s parametrem $\lambda > 0$, značíme $Ex(\lambda)$. Náhodná veličina X nabývá pouze nezáporných hodnot a její hustota je tvaru

$$X \sim Ex(\lambda) \quad \sim \quad f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0,\ \lambda > 0 \\ 0 & \text{jinak.} \end{cases}$$

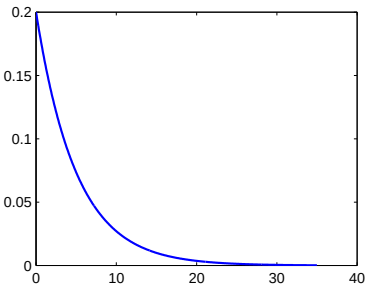
Odvodíme distribuční funkci

$$F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & x < 0, \\ \int_0^x \lambda e^{-\lambda t}dt = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x} & x \geq 0. \end{cases}$$

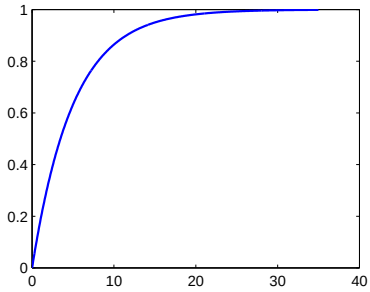
a kvantilovou funkci pro $0 \leq \alpha \leq 1$

$$\begin{array}{rcl} \alpha & = & 1 - e^{-\lambda x} \\ e^{-\lambda x} & = & 1 - \alpha \\ -\lambda x & = & \ln(1-\alpha) \\ x & = & \frac{-\ln(1-\alpha)}{\lambda} \end{array} \quad \Rightarrow \quad \boxed{Q(\alpha) = \frac{-\ln(1-\alpha)}{\lambda}} \quad \text{pro} \quad 0 \leq \alpha \leq 1.$$

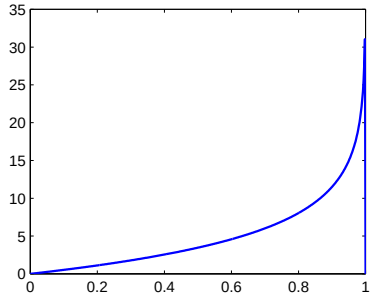
Hustota $f(x)$ pro $\lambda = 0.2$



Distribuční funkce $F(x)$



Kvantilová funkce $Q(\alpha)$



Na závěr tohoto příkladu ještě nalezneme dolní, horní kvartil a medián.

Medián: $x_{0.5} = \frac{-\ln\left(1-\frac{1}{2}\right)}{\lambda} = \frac{\ln 2}{\lambda}$

Dolní kvartil: $x_{0.5} = \frac{-\ln\left(1-\frac{1}{2}\right)}{\lambda} = \frac{\ln \frac{4}{3}}{\lambda}$

Horní kvartil: $x_{0.5} = \frac{-\ln\left(1-\frac{1}{2}\right)}{\lambda} = \frac{\ln 2}{\lambda}$

6.3. Kvantily některých důležitých rozdělení. Zaveďme následující značení:

Φ	distribuční funkce standardizovaného normálního rozdělení
G_n	distribuční funkce rozdělení χ^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_α	kvantily standardizovaného normálního rozdělení
$\chi^2_\alpha(\nu)$	kvantily rozdělení χ^2 o ν stupních volnosti
$t_\alpha(\nu)$	kvantily Studentova rozdělení o ν stupních volnosti
$F_\alpha(\nu_1, \nu_2)$	kvantily Fisherova-Snedecorova rozdělení o ν_1 a ν_2 stupních volnosti

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

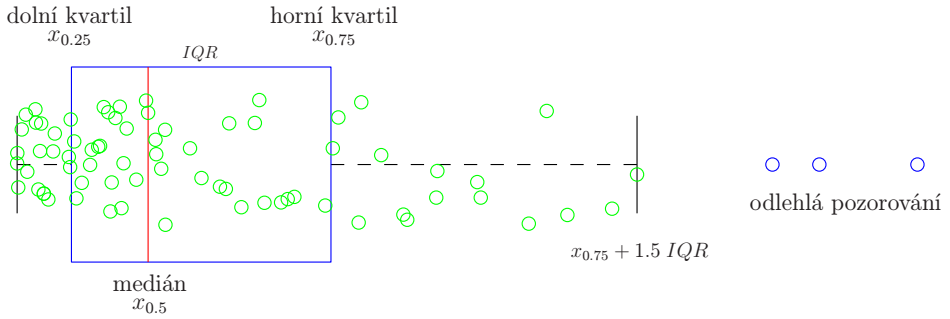
$$F(x) = 1 - F(-x) \qquad x \in \mathbb{R}$$

a odtud

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

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

6.4. Krabicový graf (*box plot, box and whisker plot*). Velmi často užívaným grafem, který se řadí k metodám **průzkumové analýzy dat** (*EDA - Exploratory Data Analysis*)



6.5. Empirická (výběrová) kvantilová funkce.

Je definována pomocí náhodného výběru

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\}$$

takto

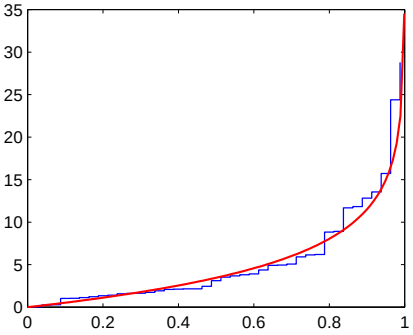
$$Q_{emp}(p_i) = X_{(i)} \qquad \text{pro} \qquad p_i = \frac{i-\frac{1}{2}}{n},$$

kde

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

jsou tzv. **pořádkové statistiky**, tj. uspořádaný náhodný výběr.

Teoritická a empirická kvantilová funkce exponenciálního rozdělení



6.6. Q–Q grafy (*Q–Q plots, Quantile–quantile plots*). Velmi užitečný graf, pomocí kterého můžeme např. porovnávat

- teoretické a výběrové kvantily
- kvantily dvou výběrů

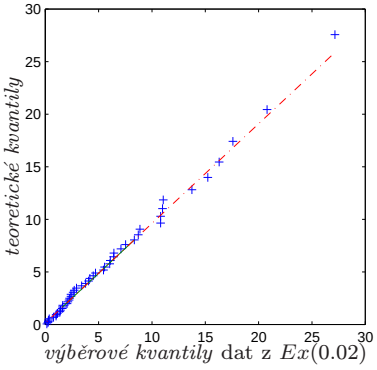
Na následujících třech obrázcích budeme demostrovat použití Q–Q grafů pro simulovaná data z exponenciálního, Poissonova a normálního rozdělení.

Pokud jsou generovaná data ze stejné rodiny rozdělení, body leží zhruba na přímce a platí

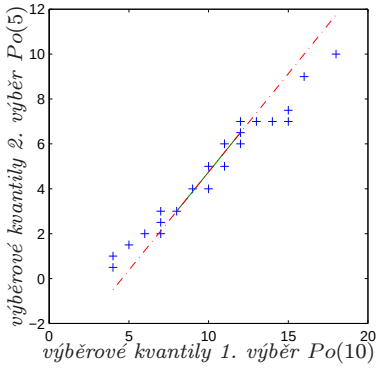
$$X_{(i)} \approx Q(p_i) = F^{-1}(p_i) \quad \text{pro} \quad X \sim F(x) \qquad \text{a} \qquad Y_{(i)} \approx a + bQ(p_i) \quad \text{pro} \quad Y \sim F\left(\tfrac{x-a}{b}\right).$$

Pocházejí-li z různých rozdělení, část bodů leží výrazně mimo přímku.

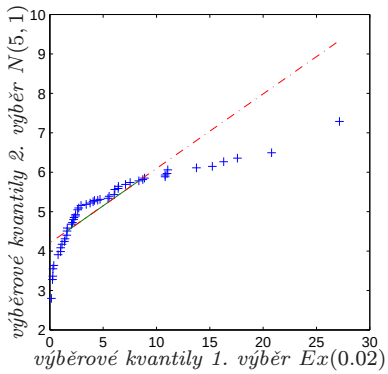
Exponenciální rozdělení



Poissonovo rozdělení



Exponenciální a normální rozdělení



6.7. Konstrukce intervalových odhadů. Popíšeme nyní jednu metodu konstrukce intervalových odhadů, která je použitelná ve většině případů.

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

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

tak aby její rozdělení již **nezáviselo na parametru** θ.

(2) Nechť $q_{\alpha/2}$ a $q_{1-\alpha/2}$ jsou kvantily rozdělení statistiky

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

Pak pro všechna θ 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 monotonní funkce**, pak existuje inverzní funkce

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

(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

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

7. BODOVÉ A INTERVALOVÉ ODHADY PARAMETRŮ NORMÁLNÍHO ROZDĚLENÍ

Nechť $k, n \in \mathbb{N}$, $\nu, \nu_1, \nu_2, \ldots, \nu_k \in \mathbb{N}$, $b_0, b_1, \ldots, b_n \in \mathbb{R}$, $\exists\ i \in \{1, \ldots, n\} : b_i \neq 0$ Připomeňme, že platí:

Normální rozdělení:

s hustotou

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

má střední hodnotu $EX = \mu$ a rozptyl $DX = \sigma^2$. Toto rozdělení má následující vlastnosti:

$$\begin{aligned} \perp\!\!\!\perp \{X_1, \ldots, X_n\} \ \wedge\ X_i \sim N(\mu_i, \sigma_i^2) \quad &\Rightarrow \quad b_0 + \sum_{i=1}^{n} b_i X_i \sim N\left(b_0 + \sum_{i=1}^{n} b_i \mu_i, \sum_{i=1}^{n} b_i^2 \sigma_i^2\right) \\ X \sim N(\mu, \sigma^2) \quad &\Rightarrow \quad U = \tfrac{X-\mu}{\sigma} \sim N(0,1) \end{aligned}$$

χ^2 **rozdělení**:

$$\begin{aligned} \perp\!\!\!\perp \{U_1, \ldots, U_\nu\} \simeq N(0,1) \quad &\Rightarrow \quad K = U_1^2 + \cdots + U_\nu^2 \sim \chi^2(\nu) \\ \perp\!\!\!\perp \{K_1 \sim \chi^2(\nu_1), \ldots, K_k \sim \chi^2(\nu_k)\} \quad &\Rightarrow \quad K = K_1 + \cdots + K_k \sim \chi^2(\nu_1 + \cdots + \nu_k) \end{aligned}$$

Studentovo t-rozdělení:

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

Fisherovo-Snedecorovo F-rozdělení:

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

Ještě než začneme odvozovat rozdělení výběrových statistik, připomeňme si, že platí věty:

VĚTA 7.1. *Nechť náhodný vektor*

$$\mathbf{X} = (X_1, \ldots, X_n)' \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

má $n-$rozměrné normální rozdělení a $\mathbf{B}$ *je* ***regulární*** *matice reálných čísel typu $n \times n$ a $\mathbf{a} \in \mathbb{R}^n$. Potom náhodný vektor*

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X} \sim N_n(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}).$$

DŮKAZ. Hustota pravděpodobnosti náhodného vektoru $\mathbf{X}$ je tvaru

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})}.$$

Inverzní transformace k transformaci

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$$

je rovna

$$\mathbf{X} = \mathbf{B}^{-1}(\mathbf{Y} - \mathbf{a})$$

a jakobián této inverzní transformace je tvaru

$$|\mathbf{J}| = \left|\mathbf{B}^{-1}\right| = |\mathbf{B}|^{-1}.$$

Pak hustotu pravděpodobnosti transformované náhodného vektoru

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$$

lze vyjádřit takto

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= f_{\mathbf{X}}(\mathbf{B}^{-1}(\mathbf{Y}-\mathbf{a}))|\mathbf{B}|^{-1} \\ &= (2\pi)^{-\frac{n}{2}}\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}}|\mathbf{B}|^{-1}e^{-\frac{1}{2}[\mathbf{B}^{-1}(\mathbf{y}-\mathbf{a})-\boldsymbol{\mu}]'\boldsymbol{\Sigma}^{-1}[\mathbf{B}^{-1}(\mathbf{y}-\mathbf{a})-\boldsymbol{\mu}]} \\ &= (2\pi)^{-\frac{n}{2}}\,|\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}|^{-\frac{1}{2}}e^{-\frac{1}{2}(\mathbf{y}-\mathbf{a}-\mathbf{B}\boldsymbol{\mu})'|\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}|^{-1}(\mathbf{y}-\mathbf{a}-\mathbf{B}\boldsymbol{\mu})} \quad \sim \quad N_n(\mathbf{a}+\mathbf{B}\boldsymbol{\mu},\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}) \end{aligned}$$

□

VĚTA 7.2. *Nechť $X_1,\dots,X_n$ jsou **nezávislé** náhodné veličiny takové, že*

$$X_i \sim N(\mu_i,\sigma^2) \qquad i=1,\dots,n.$$

*a $\mathbf{B}$ je **ortonormální** matice typu $n\times n$. Položme $\mathbf{X}=(X_1,\dots,X_n)'$ a*

$$\mathbf{Y}=(Y_1,\dots,Y_n)'=\mathbf{B}'(\mathbf{X}-\boldsymbol{\mu}),$$

*kde $\boldsymbol{\mu}=(\mu_1,\dots,\mu_n)'$. Potom Y_j jsou **nezávislé** náhodné veličiny a*

$$Y_j \sim N(0,\sigma^2).$$

DŮKAZ. Protože $X_1,\dots,X_n$ jsou nezávislé náhodné veličiny s rozdělením $X_i\sim N(\mu_i,\sigma^2)$, má náhodný vektor $\mathbf{X}$ hustotu pravděpodobnosti

$$f_{\mathbf{X}}(\mathbf{x})=\prod_{i=1}^{n}\left[\tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma}\right)^2}\right]=(2\pi)^{-\frac{n}{2}}\,e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma}\right)^2}\sim N_n(\boldsymbol{\mu},\boldsymbol{\Sigma}),\quad \text{kde } \boldsymbol{\Sigma}=\sigma^2\mathbf{I}_n.$$

Je-li $\mathbf{B}$ ortonormální matice (tj. $\mathbf{B}^{-1}=\mathbf{B}'$), pak z věty 7.1 plyne, že náhodný vektor

$$\mathbf{Y}=\mathbf{B}'(\mathbf{X}-\boldsymbol{\mu})\sim N_n(\mathbf{O},\mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}),\qquad \text{kde } \mathbf{B}'\boldsymbol{\Sigma}\mathbf{B}=\sigma^2\mathbf{B}'\mathbf{B}=\sigma^2\mathbf{I}_n$$

s hustotou pravděpodobnosti

$$f_{\mathbf{Y}}(\mathbf{Y})=\prod_{j=1}^{n}\left[\tfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y_j}{\sigma}\right)^2}\right]=\prod_{j=1}^{n}f_{Y_j}(y_j).$$

Odtud plyne tvrzení věty. □

Na základě těchto vlastností můžeme odvodit rozdělení výběrových statistik v případě náhodných výběrů z normálního rozdělení.

VĚTA 7.3. *Mějme $\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq N(\mu,\sigma^2)$ 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* $T=\frac{\bar X-\mu}{S}\sqrt{n}\sim t(n-1)$

Důkaz. Mějme ortonormální matici typu $n\times n$, jejíž první řádek je $\left(\frac{1}{\sqrt{n}},\dots,\frac{1}{\sqrt{n}}\right)'$, tj. např.

$$\mathbf{B}=\begin{pmatrix}\mathbf{b}_1'\\ \mathbf{b}_2'\\ \mathbf{b}_3'\\ \vdots\\ \mathbf{b}_{n-1}'\\ \mathbf{b}_n'\end{pmatrix}=\begin{pmatrix}\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots & \frac{1}{\sqrt{n}}\\ \frac{1}{\sqrt{1\cdot2}} & -\frac{1}{\sqrt{1\cdot2}} & 0 & \cdots & \cdots & 0\\ \frac{1}{\sqrt{2\cdot3}} & \frac{1}{\sqrt{2\cdot3}} & -\frac{2}{\sqrt{2\cdot3}} & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \ddots & 0\\ \frac{1}{\sqrt{(n-2)(n-1)}} & \frac{1}{\sqrt{(n-2)(n-1)}} & \cdots & \frac{1}{\sqrt{(n-2)(n-1)}} & -\frac{n-2}{\sqrt{(n-2)(n-1)}} & 0\\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \cdots & \cdots & \frac{1}{\sqrt{(n-1)n}} & -\frac{n-1}{\sqrt{(n-1)n}}\end{pmatrix}.$$

Podle věty 7.2

$$\mathbf{Y} = (Y_1, \ldots, Y_n)' = \mathbf{B}(\mathbf{X} - \boldsymbol{\mu}) \ \sim \ N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

a Y_i jsou nezávislé normálně rozdělené náhodné veličiny s nulou střední hodnotou a se stejným rozptylem σ^2.

Nejprve dokážeme důležité vztahy

(a) Počítejme: $\boxed{\mathbf{Y}'\mathbf{Y}} = (\mathbf{X} - \boldsymbol{\mu})' \underbrace{\mathbf{B}'\mathbf{B}}_{=\mathbf{I}_n} (\mathbf{X} - \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})'(\mathbf{X} - \boldsymbol{\mu}) = \boxed{\sum_{i=1}^{n} (X_i - \mu)^2}$.

(b) Vyjádřeme $\boxed{Y_1} = \mathbf{b}_1'(\mathbf{X} - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) = \frac{1}{\sqrt{n}} (n\bar{X} - n\mu) = \boxed{\sqrt{n}(\bar{X} - \mu)}$.

(c) Nakonec spočítejme

$$\begin{aligned}
\boxed{\sum_{i=1}^{n} (X_i - \bar{X})^2} &= \sum_{i=1}^{n} [(X_i - \mu) - (\bar{X} - \mu)]^2 \\
&= \underbrace{\sum_{i=1}^{n} (X_i - \mu)^2}_{\mathbf{Y}'\mathbf{Y}} - 2(\bar{X} - \mu) \underbrace{\sum_{i=1}^{n} (X_i - \mu)}_{n(\bar{X} - \mu)} + n(\bar{X} - \mu)^2 \\
&= \mathbf{Y}'\mathbf{Y} - \underbrace{n(\bar{X} - \mu)^2}_{Y_1^2} = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \boxed{\sum_{i=2}^{n} Y_i^2}.
\end{aligned}$$

Nyní budeme dokazovat jednotlivá tvrzení věty:

(1) Ze vztahu (b) dostaneme

$$Y_1 = \sqrt{n}(\bar{X} - \mu) = \mathbf{b}_1'(\mathbf{X} - \boldsymbol{\mu}) \sim N(\mu_{Y_1}, \sigma_{Y_1}^2),$$

přičemž

$$\mu_{Y_1} = \mathbf{b}_1' E(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{b}_1'(\mu - \mu) = 0$$

$$\sigma_{Y_1}^2 = \mathbf{b}_1' D\mathbf{X} \mathbf{b}_1 = \sigma^2 \mathbf{b}_1' \mathbf{b}_1 = \sigma^2.$$

Odtud ihned dostaneme, že

$$\bar{X} = \mu + \frac{Y_1}{\sqrt{n}} \ \sim \ N\left(\mu, \frac{\sigma^2}{n}\right).$$

Provedeme-li standardizaci, tj. takovou lineární transformaci, která zajišťuje nulovou střední hodnotu a jednotkový rozptyl, dostaneme první tvrzení věty:

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

(2) Náhodné veličiny Y_i jsou nezávislé normálně rozdělené náhodné veličiny s nulou střední hodnotou a se stejným rozptylem σ^2, tj.

$$\perp\!\!\!\perp \{Y_1, \ldots, Y_n\} \simeq N(0, \sigma^2).$$

Provedeme-li opět jejich standardizaci, dostaneme posloupnost nezávislých standardizovaných normálních náhodných veličin

$$\perp\!\!\!\perp \{\tfrac{Y_1}{\sigma}, \ldots, \tfrac{Y_n}{\sigma}\} \simeq N(0, 1),$$

jejichž kvadráty $K_i = \left(\frac{Y_i}{\sigma}\right)^2$ mají χ^2 rozdělení o jednom stupni volnosti, tj.

$$\perp\!\!\!\perp \{K_2 = \left(\tfrac{Y_2}{\sigma}\right)^2, \ldots, K_n = \left(\tfrac{Y_n}{\sigma}\right)^2\} \simeq \chi^2(1).$$

Protože náhodná veličina, která je součtem několika nezávislých náhodných veličin s χ^2 rozdělením, má opět χ^2 rozdělení, přitom její stupeň volnosti je roven součtu jednotlivých stupňů volnosti, dostáváme druhé tvrzení věty:

$$K = K_2 + \cdots + K_n = \sum_{i=2}^{n} \left(\tfrac{Y_i}{\sigma}\right)^2 = \tfrac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

(3) Protože $Y_1, \ldots, Y_n$ jsou nezávislé náhodné veličiny a nám se již dříve podařilo vyjádřit výběrový průměr a výběrový rozptyl takto

$$\bar{X} = \mu + \frac{Y_1}{\sqrt{n}} \qquad \text{a} \qquad S^2 = \frac{1}{n-1} \sum_{i=2}^{n} Y_i^2,$$

je vidět, že statistiky $\bar{X}$ a S^2 jsou **stochasticky nezávislé**, značíme $\boxed{\bar{X} \perp S^2}$.

Abychom dostali náhodnou veličinu, která má Studentovo rozdělení, potřebujeme mít dvě nezávislé náhodné veličiny, z nichž jedna, označme ji jako U^*, má standardizované normální rozdělení, a druhá, označme ji jako K^*, má χ^2 rozdělení s ν stupni volnosti. Pak náhodná veličina $T^* = \frac{U^*}{\sqrt{\frac{K^*}{\nu}}}$ má Studentovo rozdělení s ν stupni volnosti, tj.

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

Položíme-li

$$U^* = U = U_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \ \sim \ N(0,1) \qquad \text{a} \qquad K^* = K = \tfrac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

pak statistika

$$T^* = \frac{U^*}{\sqrt{\frac{K^*}{\nu}}} = \frac{\frac{\bar{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}} = \frac{\bar{X} - \mu}{S}\sqrt{n} \ \sim \ t(n-1),$$

čímž jsme dokázali poslední tvrzení věty.

□

Poznámka 7.4. Statistiky $\boxed{U}$, $\boxed{K}$ a $\boxed{T}$ se nazývají PIVOTOVÉ STATISTIKY, přičemž

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

DŮSLEDEK 7.5. *Mějme* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$, *kde* μ *je* ***neznámý parametr*** *a* $\sigma^2 \in \mathbb{R}$ *je* ***známé*** *reálné číslo. Pak*

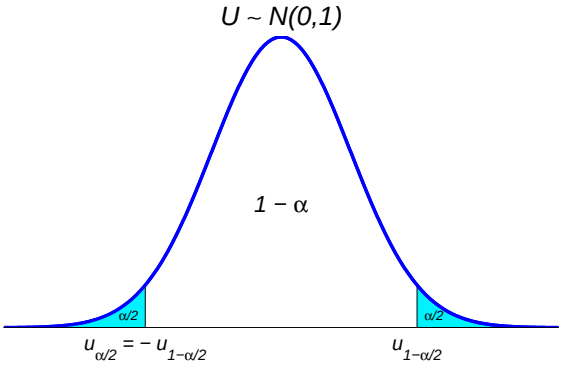
$\langle \bar{X} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\rangle$ - *je* $100(1-\alpha)\%$ *interval spolehlivosti pro střední hodnotu* μ *při známém* σ^2

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

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

Důkaz. Za pivotovou statistiku zvolíme statistiku

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



Pro lepší čitelnost místo $P_\theta = P_\mu$ budeme psát pouze P.

Počítejme

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

□

DŮSLEDEK 7.6. *Mějme* $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$, *kde* μ *a* σ^2 *jsou* ***neznámé parametry***. *Pak*

(1) 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* μ *při neznámém* σ^2

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

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

(2) pro rozptyl $\boxed{\sigma^2}$

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

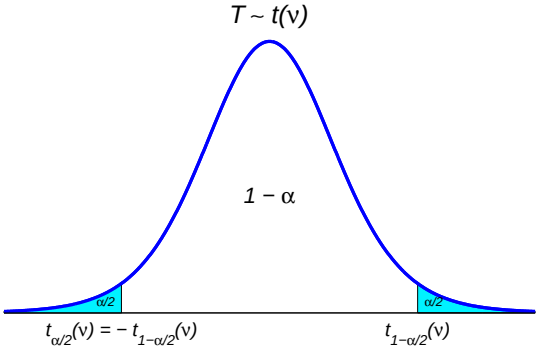
$\frac{(n-1)S^2}{\chi^2_{1-\alpha}(n-1)}$ - *je* ***dolní odhad*** *rozptylu* σ^2 *se spolehlivostí* $1-\alpha$

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

Důkaz.

(1) V případě hledání intervalu spolehlivosti pro střední hodnotu při neznámém rozptylu za pivotovou statistiku zvolíme statistiku

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



Pro lepší čitelnost místo $P_{\boldsymbol{\theta}} = P_{\mu,\sigma^2}$ budeme psát pouze P.

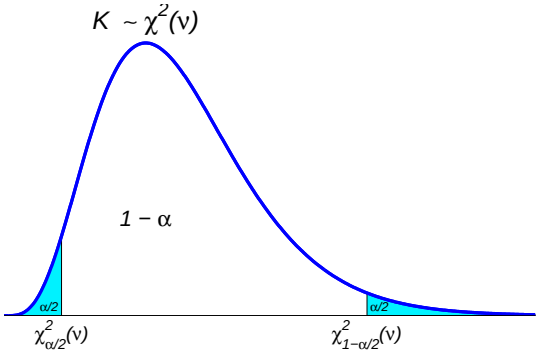
$$\begin{aligned} 1-\alpha &= P(t_{\alpha/2}(n-1) \le T \le t_{1-\alpha/2}(n-1)) \\ &= P(t_{\alpha/2}(n-1) \le \tfrac{\bar{X}-\mu}{S}\sqrt{n} \le t_{1-\alpha/2}(n-1)) \\ &= P(\bar{X} - t_{1-\alpha/2}(n-1)\tfrac{S}{\sqrt{n}} \le \mu \\ &\qquad \le \bar{X} + t_{1-\alpha/2}(n-1)\tfrac{S}{\sqrt{n}}) \end{aligned}$$

(2) V případě hledání intervalu spolehlivosti pro rozptyl za pivotovou statistiku zvolíme statistiku

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

Počítejme

$$\begin{aligned} 1-\alpha &= P(\chi^2_{\frac{\alpha}{2}}(n-1)) \le K \le \chi^2_{1-\frac{\alpha}{2}}(n-1) \\ &= P(\chi^2_{\frac{\alpha}{2}}(n-1) \le \tfrac{n-1}{\sigma^2}S^2 \le \chi^2_{\frac{1-\alpha}{2}}(n-1)) \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right) \end{aligned}$$



□

V dalším si budeme všímat intervalů spolehlivosti pro <u>DVA NEZÁVISLÉ VÝBĚRY</u>.

VĚTA 7.7. *Nechť* $\perp\!\!\!\perp\{X_1, \dots, X_{n_X}\} \sim N(\mu_X, \sigma_X^2)$ *je náhodný výběr rozsahu* n_X *z normálního rozdělení* $N(\mu_X, \sigma_X^2)$, $\bar{X}$ *je jeho výběrový průměr a* S_X^2 *jeho výběrový rozptyl.*

Dále nechť $\perp\!\!\!\perp\{Y_1, \dots, Y_{n_Y}\} \sim N(\mu_Y, \sigma_Y^2)$ *je náhodný výběr rozsahu* n_Y *z normálního rozdělení* $N(\mu_Y, \sigma_Y^2)$, $\bar{Y}$ *je jeho výběrový průměr a* S_Y^2 *jeho výběrový rozptyl.*

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$. *Pak*

(1) Statistika

$$U_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}}} \ \sim \ N(0,1).$$

(2) Pokud $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, *pak statistika*

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_{XY}}\sqrt{\frac{n_X n_Y}{n_X + n_Y}} \ \sim \ t(n_X + n_Y - 2), \text{ kde } S_{XY}^2 = \tfrac{(n_X-1)S_X^2 + (n_Y-1)S_Y^2}{n_X+n_Y-2}.$$

(3) Statistika

$$F = \frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2} \ \sim \ F(n_X - 1, n_Y - 1).$$

Důkaz. Z nezávislosti náhodných výběrů vyplývá, že všechny statistiky $\bar{X}$, $\bar{Y}$, S_X^2 a S_Y^2 jsou nezávislé, tj.

$$\perp\!\!\!\perp\{\bar{X}, \bar{Y}, S_X^2, S_Y^2\}.$$

(1) Protože výběrové průměry normálních náhodných výběrů mají opět normální rozdělení, tj.

$$\bar{X} \ \sim \ N\left(\mu_X, \tfrac{\sigma_X^2}{n_X}\right)$$

a

$$\bar{Y} \ \sim \ N\left(\mu_Y, \tfrac{\sigma_Y^2}{n_Y}\right),$$

tak i jejich rozdíl je opět normální, tj.

$$Z = \bar{X} - \bar{Y} \ \sim \ N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \tfrac{\sigma_Y^2}{n_Y}\right).$$

Potom standardizovaná náhodná veličina U_Z má standardní normální rozdělení, tj.

$$U_Z = U_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \ \sim \ N(0,1),$$

tím jsme dokázali první tvrzení věty.

(2) Je-li $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, pak statistika U_Z je tvaru

$$\begin{aligned} U_Z = U_{\bar{X}-\bar{Y}} &= \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_Y}}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \\ &= \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma}\sqrt{\frac{n_X n_Y}{n_X + n_Y}} \ \sim \ N(0,1). \end{aligned}$$

Označíme-li dvě nezávislé statistiky s χ^2 rozdělením

$$K_X = \frac{n_X - 1}{\sigma^2} S_X^2 \ \sim \ \chi^2(n_X - 1) \qquad \text{a} \qquad K_Y = \frac{n_Y - 1}{\sigma^2} S_Y^2 \ \sim \ \chi^2(n_Y - 1),$$

pak statistika $K = K_X + K_Y$ má opět χ^2 rozdělení se stupni volnosti, které jsou součtem stupňů volnosti statistik K_X a K_Y, tj.

$$\begin{aligned} K = K_X + K_Y &= \frac{n_X - 1}{\sigma^2} S_X^2 + \frac{n_Y - 1}{\sigma^2} S_Y^2 \\ &= \frac{1}{\sigma^2}\left[(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2\right] \ \sim \ \chi^2(n_X + n_Y - 2). \end{aligned}$$

Položme

$$S_{XY}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2},$$

pak

$$K = \frac{n_X + n_Y - 2}{\sigma^2} S_{XY}^2.$$

Abychom dostali náhodnou veličinu, která má Studentovo rozdělení, potřebujeme mít dvě nezávislé náhodné veličiny, z nichž jedna, označme ji jako U^*, má standardizované normální rozdělení, a druhá, označme ji jako K^*, má χ^2 rozdělení s ν stupni volnosti. Pak náhodná veličina $T^* = \frac{U^*}{\sqrt{\frac{K^*}{\nu}}}$ má Studentovo rozdělení s ν stupni volnosti, tj.

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

Položíme-li

$$U^* = U = U_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sigma} \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \ \sim \ N(0,1)$$

a

$$K^* = K = \frac{n_X + n_Y - 2}{\sigma^2} S_{XY}^2 \sim \chi^2(n_X + n_Y - 2)$$

pak statistika

$$\begin{aligned} T^* &= \frac{U^*}{\sqrt{\frac{K^*}{\nu}}} = \frac{\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sigma}\sqrt{\frac{n_X n_Y}{n_X+n_Y}}}{\sqrt{\frac{\frac{n_X+n_Y-2}{\sigma^2}S_{XY}^2}{n_X+n_Y-2}}} \\ &= \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_{XY}} \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \ \sim \ t(n_X + n_Y - 2), \end{aligned}$$

čímž jsme dokázali druhé tvrzení věty.

(3) Chceme-li dokázat třetí tvrzení, musíme najít dvě nezávislé náhodné veličiny, které mají χ^2 rozdělení. Označme je $K_1^* \ \sim \ \chi^2(\nu_1)$ a $K_2^* \ \sim \ \chi^2(\nu_2)$. Pak náhodná veličina

$$F^* = \frac{K_1^*/\nu_1}{K_2^*/\nu_2} \ \sim \ F(\nu_1, \nu_2).$$

Položíme-li

$$K_1^* = K_X = \frac{n_X - 1}{\sigma_X^2} S_X^2 \qquad \text{a} \qquad K_2^* = K_Y = \frac{n_Y - 1}{\sigma_Y^2} S_Y^2,$$

dostáváme

$$F^* = \frac{K_1^*/\nu_1}{K_2^*/\nu_2} = \frac{\frac{n_X-1}{\sigma_X^2}S_X^2/(n_X-1)}{\frac{n_Y-1}{\sigma_Y^2}S_Y^2/(n_Y-1)} = \frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2} \ \sim \ F(n_X - 1, n_Y - 1)$$

a tím jsme dokázali i poslední tvrzení věty.

□

DŮSLEDEK 7.8. *Nechť* $\perp\!\!\!\perp\{X_1,\dots,X_{n_X}\} \sim N(\mu_X,\sigma_X^2)$ *je náhodný výběr rozsahu* n_X *z normálního rozdělení* $N(\mu_X,\sigma_X^2)$, $\bar{X}$ *je jeho výběrový průměr a* S_X^2 *jeho výběrový rozptyl.*

Dále nechť $\perp\!\!\!\perp\{Y_1,\dots,Y_{n_Y}\} \sim N(\mu_Y,\sigma_Y^2)$ *je náhodný výběr rozsahu* n_Y *z normálního rozdělení* $N(\mu_Y,\sigma_Y^2)$, $\bar{Y}$ *je jeho výběrový průměr a* S_Y^2 *jeho výběrový rozptyl.*

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$. *Pak*

(1) jsou-li $\boxed{\sigma_Y^2 \text{ a } \sigma_X^2 \text{ známé}}$, *pak* $100(1-\alpha)\%$ *interval spolehlivosti pro rozdíl středních hodnot* $\mu_X - \mu_Y$ *je tvaru*

$$\left\langle \bar{X}-\bar{Y}-u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\sigma_X^2}{n_X}+\tfrac{\sigma_Y^2}{n_Y}}, \bar{X}-\bar{Y}+u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\sigma_X^2}{n_X}+\tfrac{\sigma_Y^2}{n_Y}} \right\rangle.$$

(2) Jestliže $\boxed{\sigma_Y^2 \text{ a } \sigma_X^2 \text{ nejsou známé}}$ *a platí* $\boxed{\sigma_Y^2=\sigma_X^2=\sigma^2}$, *pak* $100(1-\alpha)\%$ *interval spolehlivosti pro rozdíl středních hodnot* $\mu_X-\mu_Y$ *je tvaru*

$$\left\langle \bar{X}-\bar{Y}-t_{1-\frac{\alpha}{2}}(n_X+n_Y-2)\,S_{XY}\sqrt{\tfrac{n_X+n_Y}{n_Xn_Y}}, \bar{X}-\bar{Y}+t_{1-\frac{\alpha}{2}}(n_X+n_Y-2)\,S_{XY}\sqrt{\tfrac{n_X+n_Y}{n_Xn_Y}} \right\rangle,$$

kde

$$S_{XY}^2=\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}.$$

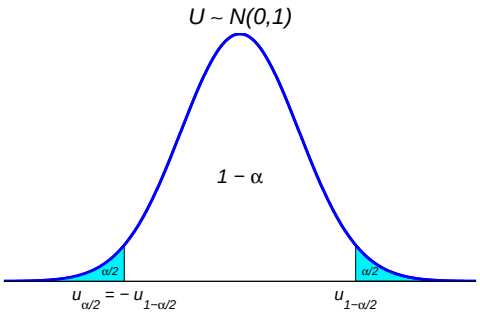
(3) Při $\boxed{\text{neznámých } \mu_x,\mu_Y,\sigma_X^2,\sigma_Y^2}$ *je* $100(1-\alpha)\%$ *interval spolehlivosti pro podíl rozptylů* $\frac{\sigma_X^2}{\sigma_Y^2}$ *roven*

$$\left\langle \frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_X-1,n_Y-1)}, \frac{S_X^2}{S_Y^2}\frac{1}{F_{\frac{\alpha}{2}}(n_X-1,n_Y-1)} \right\rangle.$$

Důkaz. Obdobně jako v předchozí větě

(1) jako pivotovou statistiku použijeme

$$U_{\bar{X}-\bar{Y}}=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_Y}}} \sim N(0,1).$$



Počítejme

$$\begin{aligned}1-\alpha &= P\left(u_{\frac{\alpha}{2}} \le U_{\bar{X}-\bar{Y}} \le u_{1-\frac{\alpha}{2}}\right)\\ &= P\left(u_{\frac{\alpha}{2}} \le \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_x}+\frac{\sigma_Y^2}{n_Y}}} \le u_{1-\frac{\alpha}{2}}\right)\\ &= P\left(\bar{X}-\bar{Y}-u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\sigma_X^2}{n_X}+\tfrac{\sigma_Y^2}{n_Y}} \le \mu_X-\mu_Y \right.\\ &\qquad \left. \le \bar{X}-\bar{Y}+u_{1-\frac{\alpha}{2}}\sqrt{\tfrac{\sigma_X^2}{n_X}+\tfrac{\sigma_Y^2}{n_Y}}\right)\end{aligned}$$

Tím jsme dokázali první tvrzení.

(2) V případě hledání intervalu spolehlivosti pro rozdíl středních hodnot při neznámém rozptylu $\sigma^2=\sigma_X^2=\sigma_Y^2$ za pivotovou statistiku zvolíme statistiku

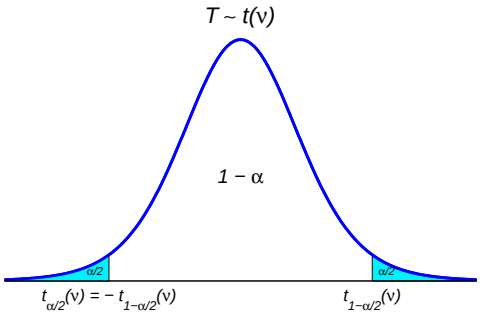
$$T_{\bar{X}-\bar{Y}}=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_{XY}}\sqrt{\frac{n_Xn_Y}{n_X+n_Y}} \sim t(n_X+n_Y-2),$$

kde

$$S_{XY}^2=\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}.$$

Označme $\nu = n_X + n_Y - 2$ a počítejme

$$\begin{aligned}1-\alpha &= P(t_{\alpha/2}(\nu) \le T_{\bar{X}-\bar{Y}} \le t_{1-\alpha/2}(\nu)) \\ &= P\left(t_{\alpha/2}(\nu) \le \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_{XY}}\sqrt{\frac{n_X n_Y}{n_X+n_Y}}\right. \\ &\qquad \left. \le t_{1-\alpha/2}(\nu)\right) \\ &= P\left(\bar{X}-\bar{Y}-t_{1-\frac{\alpha}{2}}(\nu)\, S\sqrt{\tfrac{n_X+n_Y}{n_X n_Y}} \le \mu_X-\mu_Y\right. \\ &\qquad \left. \le \bar{X}-\bar{Y}+t_{1-\frac{\alpha}{2}}(\nu)\, S\sqrt{\tfrac{n_X+n_Y}{n_X n_Y}}\right),\end{aligned}$$

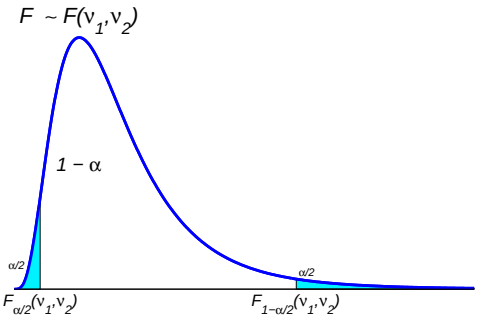


čímž jsme dokázali druhé tvrzení.

(3) V případě hledání intervalu spolehlivosti pro podíl rozptylů za pivotovou statistiku zvolíme statistiku

$$F = \frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2} \;\sim\; F(n_X-1, n_Y-1).$$

Položme $\nu_1 = n_X - 1$ a $\nu_2 = n_Y - 1$ a počítejme



$$\begin{aligned}1-\alpha &= P(F_{\frac{\alpha}{2}}(\nu_1,\nu_2)) \le F \le F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2)) \\ &= P\left(F_{\frac{\alpha}{2}}(\nu_1,\nu_2)) \le \frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2} \le F_{1-\frac{\alpha}{2}}(\nu_1,\nu_2))\right) \\ &= P\left(\frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_X-1,n_Y-1)} \le \frac{\sigma_X^2}{\sigma_Y^2}\right. \\ &\qquad \left. \le \frac{S_X^2}{S_Y^2}\frac{1}{F_{\frac{\alpha}{2}}(n_X-1,n_Y-1)}\right)\end{aligned}$$

a tím jsme dokázali i poslední tvrzení.

□

Poznámka 7.9. Ve statistických tabulkách bývají uváděny kvantily F-rozdělení pouze pro hodnoty $\alpha \ge 0.5$. Ukážeme, proč není třeba uvádět hodnoty kvantilů pro $\alpha < 0.5$. Uvažujme místo pivotové statistiky F statistiku

$$F^* = \frac{S_Y^2}{S_X^2}\frac{\sigma_X^2}{\sigma_Y^2} = \frac{1}{F} \;\sim\; F(n_Y-1, n_x-1).$$

Opět označme $\nu_1 = n_X - 1$ a $\nu_2 = n_Y - 1$ a počítejme interval spolehlivosti pro takto navrženou pivotovou statistiku

$$\begin{aligned}1-\alpha &= P(F_{\frac{\alpha}{2}}(\nu_2,\nu_1)) \le F^* \le F_{1-\frac{\alpha}{2}}(\nu_2,\nu_1)) = P\left(F_{\frac{\alpha}{2}}(\nu_2,\nu_1)) \le \tfrac{S_Y^2}{S_X^2}\tfrac{\sigma_X^2}{\sigma_Y^2} \le F_{1-\frac{\alpha}{2}}(\nu_2,\nu_1))\right) \\ &= P\left(\tfrac{S_Y^2}{S_X^2}F_{\frac{\alpha}{2}}(n_Y-1,n_X-1) \le \tfrac{\sigma_X^2}{\sigma_Y^2} \le \tfrac{S_Y^2}{S_X^2}F_{1-\frac{\alpha}{2}}(n_Y-1,n_X-1)\right)\end{aligned}$$

Takže $\boxed{F_{1-\frac{\alpha}{2}}(n_Y-1,n_X-1) = \dfrac{1}{F_{\frac{\alpha}{2}}(n_X-1,n_Y-1)}}$ a interval spolehlivosti pro $\boxed{\frac{\sigma_X^2}{\sigma_Y^2}}$ lze vyjádřit i takto

$$\left\langle \frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_X-1,n_Y-1)}, \frac{S_X^2}{S_Y^2}F_{1-\frac{\alpha}{2}}(n_Y-1,n_X-1)\right\rangle.$$

V dalším se zaměříme na interval spolehlivosti pro rozdíl středních hodnot u tzv. PÁROVÝCH VÝBĚRŮ.

VĚTA 7.10. *Nechť* $\mathbf{X}_1 = (X_1, Y_1)', \ldots, \mathbf{X}_n = (X_n, Y_n)'$ *je náhodný výběr z dvourozměrného normálního rozdělení* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *s parametry* $\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ *a* $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$, *kde* $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X^2 > 0$, $\sigma_Y^2 > 0$ *a* $\rho \in (0,1)$.

Pro $i = 1, \ldots, n$ *označme*

$$\begin{aligned} Z_i &= X_i - Y_i \\ \bar{Z} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i \\ S_Z^2 &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^n (Z_i - \bar{Z})^2. \end{aligned}$$

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 $\mu_X - \mu_Y$ *o spolehlivosti* $1 - \alpha$.

Důkaz. Připomeňme, že marginální náhodné veličiny vícerozměrného náhodného vektoru jsou opět normální náhodné veličiny, tj.

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq N(\mu_X, \sigma_X^2)$$

a

$$\perp\!\!\!\perp\{Y_1, \ldots, Y_n\} \simeq N(\mu_Y, \sigma_Y^2).$$

Takže pro jejich rozdíl

$$Z_i = X_i - Y_i \qquad i = 1, \ldots, n$$

platí, že mají také normální rozdělení

$$\perp\!\!\!\perp\{Z_1, \ldots, Z_n\} \simeq N(\mu_Z D, \sigma_Z^2),$$

kde

$$EZ_i = E(X_i - Y_i) = \mu_x - \mu_Y$$

$$\begin{aligned} DZ_i &= D(X_i - Y_i) = C(X_i - Y_i, X_i - Y_i) \\ &= C(X_i, X_i) - C(X_i, Y_i) - C(Y_i, X_i) + C(Y_i, Y_i) \\ &= DX_i - 2\underbrace{C(X_i, Y_i)}_{=\rho\sigma_X\sigma_Y} + DY_i = \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \sigma_Y^2. \end{aligned}$$

Budeme-li aplikovat důsledek 7.6 na

$$Z_1, \ldots, Z_n,$$

dostaneme tvrzení věty. □

8. BODOVÉ A INTERVALOVÉ ODHADY ZALOŽENÉ NA CENTRÁLNÍ LIMITNÍ VĚTĚ

Odhady parametrů normálního rozdělení, které jsme doposud zkoumali, mají díky **centrální limitní větě** (CLV) širší použití.

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

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

Přitom rozdělení, z něhož výběr pochází

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

Bodové i intervalové odhady lze pak sestrojit stejným způsobem jako v případě normálních náhodných výběrů, jejich **spolehlivost** bude $\boxed{1-\alpha}$ jen přibližně, tj. **asymptoticky**.

VĚTA 8.1. *Mějme* $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$ *a výběrový průměr* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. *Nechť* $S_*^2 = S_*^2(\mathbf{X})$ *je* ***(slabě) konzistentním odhadem*** *rozptylu* $\sigma^2(\boldsymbol{\theta})$. *Pak statistika*

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

Důkaz. Podle Lindebergovy-Levyho CLV mají standardizované průměry asymptoticky standardizované normální rozdělení, tj.

$$U_{\bar{X}} = \frac{\bar{X} - E\bar{X}}{\sqrt{D\bar{X}}} = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{\sqrt{\frac{\sigma^2(\boldsymbol{\theta})}{n}}} = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta})}\sqrt{n} \overset{A}{\sim} N(0,1),$$

což lze ekvivalentně napsat také takto

$$U_{\bar{X}} \xrightarrow{\mathcal{L}} U \sim N(0,1).$$

Abychom dokázali, že také $U_* = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{S_*}\sqrt{n} \overset{A}{\sim} N(0,1)$, budeme potřebovat následující tvrzení, které uvedeme bez důkazu (lze najít např. v knize Rao, R. C.: *Lineární metody statistické indukce a jejich aplikace*. Academia Praha, 1978)

Jestliže $\quad Z_n \xrightarrow{\mathcal{L}} Z \quad \wedge \quad Y_n \xrightarrow{P} c \quad \Rightarrow \quad Z_n \cdot Y_n \xrightarrow{\mathcal{L}} cZ$

Pokud položíme

$$Z_n = U_{\bar{X}} \xrightarrow{\mathcal{L}} Z = U$$

a

$$Y_n = \frac{\sigma(\boldsymbol{\theta})}{S_*} \xrightarrow{P} 1,$$

neboť S_*^2 je (slabě) konzistentním odhadem rozptylu $\sigma^2(\boldsymbol{\theta})$, pak již dostaneme tvrzení věty, tj.

$$U_* = Z_n Y_n = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{S_*}\sqrt{n} \xrightarrow{\mathcal{L}} cZ = 1 \cdot U \sim N(0,1).$$

Jako transformaci jsme zvolili funkci

$$h(\mathbf{X}, \mu(\boldsymbol{\theta})) = U_{\bar{X}} \cdot \frac{\sigma(\boldsymbol{\theta})}{S_*} = \frac{\bar{X} - \mu(\boldsymbol{\theta})}{S_*}\sqrt{n}.$$

□

Důsledek 8.2. *Nechť* $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu(\boldsymbol{\theta}), \sigma^2(\boldsymbol{\theta}))$ *je náhodný výběr s konečnými druhými momenty. Potom intervalovým odhadem střední hodnoty* $\boxed{\mu(\boldsymbol{\theta})}$ *o asymptotické spolehlivosti* $1-\alpha$ *je interval*

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

kde S^2 *je výběrový rozptyl, tj.*

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

Důkaz. Důkaz je zřejmý, neboť $S_*^2 = S^2$ je konzistentním odhadem rozptylu a jako pivotovou statistiku jsme při tvorbě intervalového odhadu použili U_* s asymptoticky standardizovaným normálním rozdělením. □

Důsledek 8.3. *(**Binární náhodné výběry**). Nechť* $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq A(p)$ *je náhodný výběr s alternativním (binárním) rozdělením. Potom intervalovým odhadem parametru* $\boxed{p}$ *o asymptotické spolehlivosti* $1-\alpha$ *je interval*

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

Důkaz. Nejprve připomeňme, že pro náhodné veličiny s alternativním (binárním) rozdělením platí

$$EX_i = p \qquad \text{a} \qquad DX_i = p(1-p).$$

Protože $\bar{X}$ je konzistentním odhadem střední hodnoty, což je parametr p, pak statistika

$$S_*^2 = \bar{X}(1-\bar{X})$$

je konzistentním odhadem rozptylu $p(1-p)$.

Při tvorbě intervalového odhadu jako pivotovou statistiku jsme opět použili U_* s asymptoticky standardizovaným normálním rozdělením. □

Důsledek 8.4. *(**Poissonovské náhodné výběry**). Nechť* $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq Po(\lambda)$ *je náhodný výběr s Poisonovým rozdělením. Potom intervalovým odhadem parametru* $\boxed{\lambda}$ *(*$0 < \lambda < \infty$*) o asymptotické spolehlivosti* $1-\alpha$ *je interval*

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

Důkaz. Připomeňme, že pro náhodné veličiny s Poissonovým rozdělením platí

$$EX_i = DX_i = \lambda.$$

Protože $\bar{X}$ je konzistentním odhadem střední hodnoty, což je parametr λ, pak statistika

$$S_*^2 = \bar{X}$$

je konzistentním odhadem rozptylu λ.

Při tvorbě intervalového odhadu jako pivotovou statistiku jsme opět použili U_* s asymptoticky standardizovaným normálním rozdělením. □

9. TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

Mějme náhodný výběr $\mathbf{X} = (X_1, \ldots, X_n)'$ rozsahu n z rozdělení o distribuční funkci $F(x; \boldsymbol{\theta})$, kde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \boldsymbol{\Theta} \subset \mathbb{R}^m$. Množina $\boldsymbol{\Theta}$ nechť je neprázdná a otevřená.

Předpokládejme, že o parametru $\boldsymbol{\theta}$ existují dvě konkurující si hypotézy: H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$, H_1: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$

Tvrzení $\boxed{H_0}$ / $\boxed{H_1}$ se nazývá **nulovou hypotézou** / **alternativní hypotézou**.

Je-li $\boxed{\boldsymbol{\Theta}_0}$ / $\boxed{\boldsymbol{\Theta}_1}$ jednobodová, nazývá se **jednoduchou**, v opačném případě **složenou hypotézou**.

O platnosti této hypotézy se má rozhodnout na základě náhodného výběru $\mathbf{X} = (X_1, \ldots, X_n)'$, a to tak, že ↗ **zamítneme** *nebo* ↘ **nezamítneme** platnost hypotézy H_0.

Na testování použijeme statistiku $T_n = T(\mathbf{X})$, kterou nazýváme **testovací statistikou**. Množinu hodnot, které může testovací statistika nabýt, rozdělíme na dvě disjunktní oblasti. Jednu označíme $\boxed{W_\alpha}$, a nazveme ji **kritickou oblastí** (nebo také *oblastí zamítnutí hypotézy*) a druhá je doplňkovou oblastí (*oblast nezamítnutí testované hypotézy*).

Na základě realizace náhodného výběru $\mathbf{x} = (x_1, \ldots, x_n)'$ vypočítáme hodnotu testovací statistiky $t_n = T(\mathbf{x})$.

- Pokud hodnota testovací statistiky t_n nabude hodnoty z kritické oblasti, tj. $\boxed{t_n = T(\mathbf{x}) \in W_\alpha}$, pak **nulovou hypotézu zamítáme**.
- Pokud hodnota testovací statistiky nabude hodnoty z oblasti nezamítnutí, tj. $\boxed{t_n = T(\mathbf{x}) \notin W_\alpha}$, tak **nulovou hypotézu nezamítáme**, což ovšem neznamená že přijímáme alternativu.

Toto rozhodnutí nemusí však být správné. V následující tabulce jsou uvedeny možné situace

H_0	PLATÍ	NEPLATÍ
ZAMÍTÁME $t_n = T(\mathbf{x}) \in W_\alpha$	***chyba 1. druhu*** (α_0 je ***hladina testu***) $\alpha_0 = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_0) \leq \alpha$	O.K. (tzv. ***síla testu*** či ***silofunkce***) $1 - \beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_1)$ pro $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$
NEZAMÍTÁME $t_n = T(\mathbf{x}) \notin W_\alpha$	O.K.	***chyba 2. druhu*** $\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \notin W_\alpha \mid H_1)$ pro $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$

Volba kritického oboru W_α se řídí požadavky:

(1) Chceme, aby pravděpodobnost chyby 1. druhu byla menší nebo rovna předem zvolenému malému $\alpha \in (0, 1)$ (obvykle se volí $\alpha = 0.01$ nebo $\alpha = 0.05$), tj. aby platilo pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\alpha_0 = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha \mid H_0) \leq \alpha.$$

Pro spojitá rozdělení je vždy možné (i když ne nutné) zvolit test, jehož hladina je právě rovna α. U diskrétních rozdělení jsou možnými hladinami testu jen některé diskrétní hodnoty. Není-li zvolená hladina mezi nimi, rozhodneme se pro hladinu, která je nejbližší nižší (nebo nejbližší vyšší).

(2) Mezi testy na hladině $\boxed{\alpha}$ se pak snažíme zvolit test s co **nejmenší pravděpodobností chyby druhého druhu**, tj. **co nejsilnější test**.

Vidíme, že postavení obou hypotéz je nesymetrické. Za **nulovou hypotézu** volíme tu, jejíž **neoprávněné zamítnutí** (chyba 1. druhu) **je závažnější**.

DEFINICE 9.1. Chybu, která spočívá **v nesprávném zamítnutí nulové hypotézy, i když je správná**, budeme nazývat **chybou prvého druhu**, pravděpodobnost

$$\alpha_0 = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(T(\mathbf{X}) \in W_\alpha | H_0)$$

nazveme **hladinou významnosti** (též **hladinou testu**).

Chybu, která spočívá **v nesprávném přijetí nulové hypotézy, i když neplatí**, budeme nazývat **chybou druhého druhu** a její pravděpodobnost pro $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ označíme

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(T(\mathbf{X}) \notin W_\alpha | H_1).$$

Pravděpodobnost $1-\beta(\boldsymbol{\theta})$ nazýváme **silou testu** (též **silou kritické oblasti** W_α) a jakožto funkci $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$ ji také nazveme **silofunkcí testu**.

9.1. JEDNODUCHÁ HYPOTÉZA A JEDNODUCHÁ ALTERNATIVA.

Nejprve rozebereme nejjednodušší případ, kdy $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1\}$.

V dalším budeme značit symbolem ν $\sigma-$konečnou míru na $(\mathbb{R}^n, \mathcal{B}^n)$ (např. *Lebesgueova* nebo *čítací*) a $f(\mathbf{x};\boldsymbol{\theta})$ nezápornou měřitelnou funkci, tzv. **hustotu pravděpodobnosti vzhledem k míře** ν. Tedy $f(\mathbf{x};\boldsymbol{\theta})$ jsou jak hustoty absolutně spojitých náhodných veličin, tak pravděpodobnostní funkce.

Budeme předpokládat, že pravděpodobnostní míry $P_{\boldsymbol{\theta}_0}$ a $P_{\boldsymbol{\theta}_1}$ jsou absolutně spojité vzhledem k σ-konečné míře ν.

Označme hustoty $p_0(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_0)$, $p_1(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}_1)$.

LEMMA 9.2 (Neymanovo–Pearsonovo). *Nechť k danému $\alpha \in (0,1)$ existuje takové kladné číslo $c > 0$, že pro množinu $W_0 = \{\mathbf{x} \in \mathbb{R}^n : p_1(\mathbf{x}) \geq c p_0(\mathbf{x})\}$ platí $\int_{W_0} p_0(\mathbf{x})\, d\nu(\mathbf{x}) = \alpha$.*

Pak pro libovolnou množinu $W \in \mathcal{B}^n$ splňující podmínku $\int_W p_0(\mathbf{x})\, d\nu(\mathbf{x}) \leq \alpha$ platí

$$\int_{W_0} p_1(\mathbf{x})\, d\nu(\mathbf{x}) \geq \int_W p_1(\mathbf{x})\, d\nu(\mathbf{x}).$$

Důkaz. Pro jednoduchost pro $j = 0, 1$ místo $\int_{W_0} p_j(\mathbf{x})\, d\nu(\mathbf{x})$ pišme $\int_{W_0} p_j\, d\nu$. Vzhledem k tomu, že množiny W a W_0 lze psát jako disjunktní sjednocení, tj.

$$W = (W - W_0) \cup (W \cap W_0) \qquad \text{a} \qquad W_0 = (W_0 - W) \cup (W \cap W_0),$$

pak platí

$$\begin{aligned}\int_{W_0} p_1\, d\nu - \int_W p_1\, d\nu &= \int_{W_0-W} p_1\, d\nu + \int_{W\cap W_0} p_1\, d\nu - \int_{W-W_0} p_1\, d\nu - \int_{W\cap W_0} p_1\, d\nu \\ &= \int_{W_0-W} p_1\, d\nu - \int_{W-W_0} p_1\, d\nu. \qquad (14)\end{aligned}$$

Integrační obor prvního integrálu v (14) je částí množiny W_0, takže vzhledem k definici této množiny můžeme ho odhadnout zdola. Obdobně integrační obor druhého integrálu v (14) není částí W_0, takže ho můžeme opět díky definici W_0 odhadnout shora, tj.

$$\int_{W_0} p_1\,d\nu - \int_{W} p_1\,d\nu = \int_{W_0-W\in W_0} \underbrace{p_1}_{\geq cp_0}\,d\nu \quad - \int_{W-W_0\notin W_0} \underbrace{p_1}_{<cp_0}\,d\nu$$

$$\geq \int_{W_0-W} c\,p_0\,d\nu \quad - \int_{W-W_0} c\,p_0\,d\nu = c\left(\underbrace{\int_{W_0} p_0\,d\nu}_{=\alpha} - \underbrace{\int_{W} p_0\,d\nu}_{\leq\alpha}\right) \geq 0.$$

Předpoklady lemmatu požadují, aby kritické obory W_0 a W měly za platnosti nulové hypotéz v prvém případě pravděpodobnost α a v druhém případě pravděpodobnost nejvýše α. Tvrzení lemmatu porovnává pro dva kritické obory W_0 a W pravděpodobnost, s jakou zamítnou nulovou hypotézu, když platí hypotéza alternativní, tj. **porovnává sílu testu** obou kritických oborů. Pro kritický obor W_0 je síla testu stejná nebo větší než pro libovolný kritický obor W, to znamená, že kritický obor W_0 je mezi kritickými obory s danou hladinou α **nejsilnější možný**. □

Poznámka 9.3. Předchozí lemma lze vyslovit takto:
Test s kritickým oborem $\boxed{W_0 = \{\mathbf{x}\in\mathbb{R}^n : p_1(\mathbf{x}) \geq cp_0(\mathbf{x})\}}$ (pro $c>0$) určuje nejsilnější test hypotézy H_0 proti H_1 na dané hladině α.

Příklad 9.4 (Jednoduchá hypotéza i alternativa pro náhodný výběr z normálního rozdělení při známém rozptylu)**.** Mějme $\perp\!\!\!\perp\{X_1,\ldots,X_n\} \simeq N(\mu,\sigma^2)$, kde σ^2 je známé. Nechť $\mu_0,\mu_1\in\mathbb{R}$. Je třeba najít kritický obor $\boxed{W_0}$ nejsilnějšího testu

$$\boxed{H_0}: \mu=\mu_0 \qquad \text{proti} \qquad \boxed{H_1}: \mu=\mu_1 \quad \text{na hladině} \quad \alpha\in(0,1).$$

Platí

$$\mathbf{X}\sim\boxed{f_{\mathbf{X}}(\mathbf{x};\mu)} = \prod_{i=1}^n f_{X_i}(x_i;\mu) = \prod_{i=1}^n \tfrac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} = \boxed{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\}}.$$

Dále si připomeňme, že položíme-li $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, resp. pro realizace $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, pak za platnosti nulové hypotézy $\boxed{H_0}$

$$\bar{X} \sim N\left(\mu_0, \tfrac{\sigma^2}{n}\right) \qquad \Rightarrow \qquad U_{\bar{X}} = \frac{\bar{X}-E_{\mu_0}(\bar{X})}{\sqrt{D_{\mu_0}(\bar{X})}} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1). \tag{15}$$

Dále využijeme vztah

$$\sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n (x_i-\mu)^2 - n(\bar{x}-\mu)^2 \quad \Rightarrow \quad \sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2. \tag{16}$$

Označme $p_0(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x};\mu=\mu_0)$ a $p_1(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x};\mu=\mu_1)$.

Podmínku $\boxed{p_1(\mathbf{x}) \geq cp_0(\mathbf{x})}$ lze napsat také takto $\boxed{\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \geq c > 0}$.

Počítejme s využitím vztahu (16)

$$\tfrac{p_1(\mathbf{x})}{p_0(\mathbf{x})} = \exp\left\{\tfrac{n}{2\sigma^2}\left[(\bar{x}-\mu_0)^2 - (\bar{x}-\mu_1)^2\right]\right\} \geq c.$$

Po zlogaritmování dostaneme

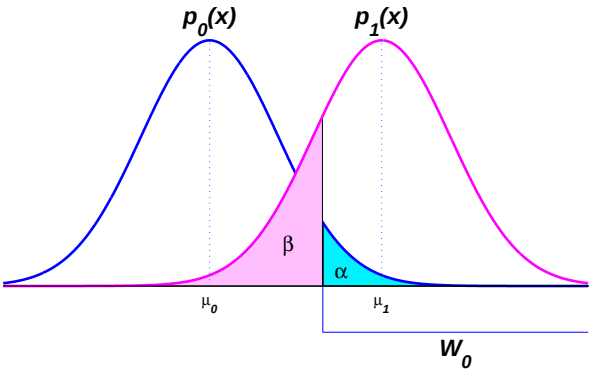
$$\frac{n}{2\sigma^2}\left[(\bar{x}-\mu_0)^2-(\bar{x}-\mu_1)^2\right]=\frac{n}{2\sigma^2}\left[2\bar{x}(\mu_1-\mu_0)-(\mu_1^2-\mu_0^2)\right]\geq \ln c \qquad (17)$$

(1) Předpokládejme, že $\mu_0 < \mu_1$.

Pak nerovnost (17) dále upravujme takto

$$\bar{x} \geq \underbrace{\frac{\mu_1+\mu_0}{2}+\frac{\sigma^2 \ln c}{n(\mu_1-\mu_0)}}_{\text{označme } k_1}$$

Dokážeme najít takové k_1, aby platilo $P_{\mu_0}(\bar{X}\geq k_1)=\alpha$?



Díky normalitě výběrového průměru (viz (15)) však můžeme počítat a upravovat

$$\alpha = P_{\mu_0}(\bar{X}\geq k_1)=P_{\mu_0}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\geq\frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right)=1-\Phi\left(\frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right)$$

takže

$$\Phi\left(\frac{k_1-\mu_0}{\sigma/\sqrt{n}}\right)=1-\alpha \quad \Rightarrow \quad u_{1-\alpha}=\frac{k_1-\mu_0}{\sigma/\sqrt{n}} \quad \Rightarrow \quad \boxed{k_1=\mu_0+\frac{\sigma}{\sqrt{n}}u_{1-\alpha}}$$

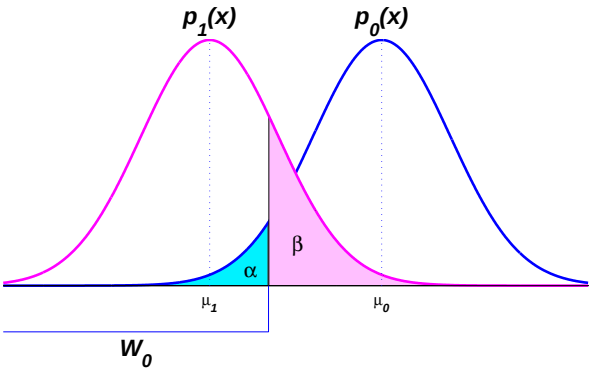
a kritický obor lze vyjádřit takto

$$W_0=\{\mathbf{x}\in\mathbb{R}^n:\bar{x}\geq k_1\}=\left\{\mathbf{x}\in\mathbb{R}^n:\bar{x}\geq\mu_0+\frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right\}.$$

(2) Nyní předpokládejme, že $\mu_0 > \mu_1$.

Pak nerovnost (17) dále upravujme takto

$$\bar{x} \leq \underbrace{\frac{\mu_1+\mu_0}{2}-\frac{\sigma^2 \ln c}{n(\mu_0-\mu_1)}}_{\text{označme } k_2}$$



Díky normalitě výběrového průměru (viz (15)) však můžeme počítat a upravovat

$$\alpha = P_{\mu_0}(\bar{X}\leq k_2)=P_{\mu_0}\left(\frac{\bar{X}-\theta_0}{\sigma/\sqrt{n}}\leq\frac{k_2-\mu_0}{\sigma/\sqrt{n}}\right)=\Phi\left(\frac{k_2-\mu_0}{\sigma/\sqrt{n}}\right)$$

takže

$$\Phi\left(\frac{k_2-\mu_0}{\sigma/\sqrt{n}}\right)=\alpha \quad \Rightarrow \quad u_{\alpha}=-u_{1-\alpha}=\frac{k_2-\mu_0}{\sigma/\sqrt{n}} \quad \Rightarrow \quad \boxed{k_2=\mu_0-\frac{\sigma}{\sqrt{n}}u_{1-\alpha}}$$

a kritický obor lze vyjádřit takto

$$W_0=\{\mathbf{x}\in\mathbb{R}^n:\bar{x}\leq k_2\}=\left\{\mathbf{x}\in\mathbb{R}^n:\bar{x}\leq\mu_0-\frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right\}.$$

Všimněme si, že při jednoduché hypotéze i alternativě

$$\boxed{H_0} : \mu = \mu_0 \qquad \text{proti} \qquad \boxed{H_1} : \mu = \mu_1 \quad \text{na hladině} \quad \alpha \in (0,1)$$

při (1) $\mu_0 < \underbrace{\mu_1}_{\text{libovolné}}$ má W_0 stejný tvar nezávislý na μ_1

(2) $\mu_0 > \underbrace{\mu_1}_{\text{libovolné}}$ má W_0 stejný tvar nezávislý na μ_1

Říkáme, že test je **stejnoměrně nejsilnější** vůči všem alternativám typu $\begin{matrix}(1)\ \mu_0 < \mu_1 \\ (2)\ \mu_0 > \mu_1\end{matrix}$.

Příklad 9.5. Mějme pro jednoduchost náhodný výběr rozsahu $n = 1$, tj. jedinou náhodnou veličinu X z rozdělení s hustotou

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & x \in (0,1), \\ 0 & \text{jinak.} \end{cases}$$

Najdeme nejsilnější test hypotézy

$$H_0 : \theta = 1 \qquad \text{proti} \qquad H_1 : \theta = 2 \qquad \text{na dané hladině } \alpha = 0.05.$$

Je třeba najít kritický obor $\boxed{W_0} = \{x \in \mathbb{R} : p_1(x) \geq c p_0(x)\}$ (pro $c > 0$), přičemž

$$p_j(x) = f(x;\theta_j) = \begin{cases} \theta_j x^{\theta_j - 1} & x \in (0,1), j = 0, 1 \\ 0 & \text{jinak.} \end{cases}$$

Podmínku $\boxed{p_1(x) \geq c p_0(x)}$ lze napsat také takto $\boxed{\frac{p_1(x)}{p_0(x)} \geq c > 0}$, takže

$$\frac{p_1(x)}{p_0(x)} = 2x^{2-1} \geq c \quad \Rightarrow \quad x \geq \underbrace{\frac{c}{2}}_{=k}$$

a k určíme z požadavku na hladinu významnosti, tj.

$$\alpha = 0.05 = \int_k^1 p_0 dx = \int_k^1 dx = 1 - k \quad \Rightarrow \quad k = 1 - 0.05 = 0.95$$

a

$$\boxed{W_0} = \{x \in \mathbb{R} : x \geq 0.95\}$$

Všimněme si dále, že pokud bychom zvolili alternativní hypotézu trochu jinak, např.

$$H_1 : \theta = 3 \quad \Rightarrow \quad \frac{p_1(x)}{p_0(x)} = 3x^{3-1} \geq c \quad \Rightarrow \quad x^2 \geq \underbrace{\frac{c}{3}}_{=k^*},$$

pak zřejmě dostaneme jinou kritickou oblast, neboť tvar kritické oblasti závisí jak na nulové hypotéze, tak na alternativní.

Poznámka 9.6. V současné době běžný statistický software (Statistika, SPSS, S$^+$, SAS) udává **dosaženou hladinu** (v anglicky psané literatuře *P–value*, *significance value*). Je to **nejmenší hladina testu**, při které bychom ještě **hypotézu** H_0 **zamítli**.

9.2. JEDNODUCHÁ HYPOTÉZA A SLOŽENÁ ALTERNATIVA.

Nechť parametrický prostor $\boldsymbol{\Theta}$ má nejméně 3 různé body, z nichž jeden je $\boldsymbol{\theta}_0$. Položme $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_0\}$. Je třeba otestovat hypotézu

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{proti} \qquad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta} - \boldsymbol{\Theta}_0 .$$

Nejprve si představme, že bychom se snažili najít pomocí N-P lemmatu nejsilnější test hypotézy H_0 proti alternativě

$$H_1' : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \in \boldsymbol{\Theta} - \boldsymbol{\Theta}_0 .$$

Obecně je třeba počítat s tím, že každý takovýto dílčí test bude mít jiný kritický obor. Může se však stát, že **kritické obory budou stejné pro všechny zmíněné dílčí testy**. Pak je rozumné test H_0 proti složené alternativě H_1 založit právě na tomto společném kritickém oboru. V tomto případě říkáme, že jde o

stejnoměrně nejsilnější test H_0 proti H_1.

Pokud však tato situace nenastane, vzniká otázka, jak postupovat v tomto případě. Zaveďme si proto nejprve pojem **zkreslený** (**vychýlený**) test.

DEFINICE 9.7. Testujme jednoduchou hypotézu $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ proti alternativě $H_0 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ na základě náhodného výběru s hustotou $f(\mathbf{x}; \boldsymbol{\theta})$. Nechť W_α je kritický obor testu. Řekneme, že test je ZKRESLENÝ (VYCHÝLENÝ), jestliže existuje taková hodnota parametru $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$, pro kterou platí nerovnost

$$\underbrace{\int_{W_\alpha} p_1(\mathbf{x}) d\nu}_{\text{síla testu}} < \underbrace{\int_{W_\alpha} p_0(\mathbf{x}) d\nu}_{\text{chyba 1. druhu}},$$

kde $p_0(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}_0)$ a $p_1(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}_1)$.

Tato podmínka říká, že existuje parametr $\boldsymbol{\theta}_1$, pro který je síla testu menší než chyba 1. druhu, tedy

pravděpodobnost zamítnutí správné hypotézy	$>$	pravděpodobnost zamítnutí nesprávné hypotézy

což je naprosto nežádoucí vlastnost.

Tedy v případech, kdy nebude existovat rovnoměrně nejsilnější test, budeme se snažit vytvořit alespoň nezkreslený test.

Příklad 9.8 (JEDNODUCHÁ HYPOTÉZA A SLOŽENÁ ALTERNATIVA PRO NÁHODNÝ VÝBĚR Z NORMÁLNÍHO ROZDĚLENÍ PŘI ZNÁMÉM ROZPTYLU). Mějme $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq N(\mu, \sigma^2)$, kde σ^2 je známé. Nechť $\mu_0, \mu_1 \in \mathbb{R}$.

Jak jsme již ukázali v příkladě 9.4, kritický obor je jiný pro $\mu_1 < \mu_0$ a $\mu_2 > \mu_0$, takže nenajdeme kritický obor stejnoměrně nejsilnějšího testu

$$H_0 : \mu = \mu_0 \qquad \text{proti} \qquad H_1 : \mu \neq \mu_0 \quad \text{na hladině} \quad \alpha \in (0, 1),$$

proto se budeme snažit najít kritický obor alespoň **nezkresleného testu**.

(A) Zvolíme-li kritický obor typu

$$W_\alpha = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \geq k_1\} = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \geq \mu_0 + \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\}.$$

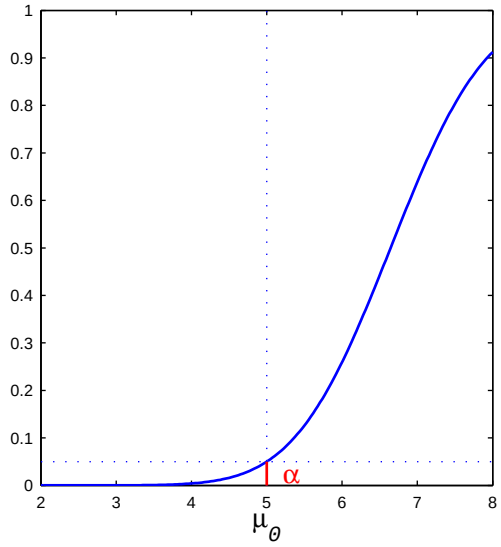
Pak **silofunkce** (což je síla testu jakožto funkce parametru $\boldsymbol{\theta} \in \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$) je tvaru

$$\begin{aligned}\beta^*(\boldsymbol{\theta}) = 1 - \beta(\boldsymbol{\theta}) = \beta^*(\mu) &= \int_{W_\alpha} p_1 \, d\nu \\ &= P_{\mu,\sigma}(\bar{X} \geq k_1) \\ &= P_{\mu,\sigma}(\bar{X} \geq \mu_0 + \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}) \\ &= P_{\mu,\sigma}\left(\tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq \tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} + u_{1-\alpha}\right) \\ &= 1 - \Phi\left(\tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} + u_{1-\alpha}\right)\end{aligned}$$

Zřejmě platí

$$\beta^*(\mu_0) = \alpha$$

Silofunkce $\beta^*(\mu)$

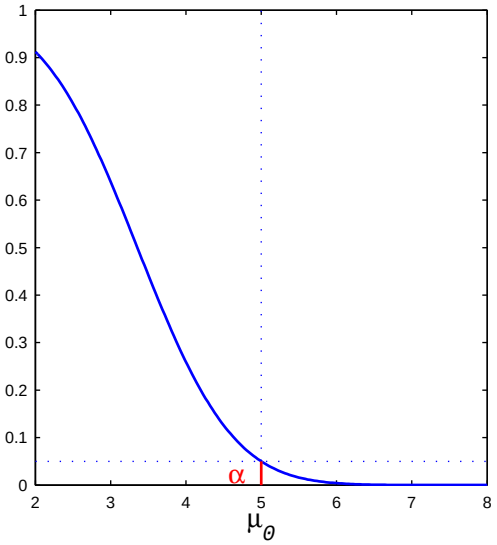


a pro $\mu_1 < \mu_0$ je síla testu < chyba 1. druhu.

(B) Zvolíme-li kritický obor typu

$$W_\alpha = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \leq k_2\} = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \leq \mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\}.$$

Silofunkce $\beta^*(\mu)$



Pak **silofunkce** je tvaru

$$\begin{aligned}\beta^*(\boldsymbol{\theta}) = 1 - \beta(\boldsymbol{\theta}) = \beta^*(\mu) &= \int_{W_\alpha} p_1 \, d\nu \\ &= P_{\mu,\sigma}(\bar{X} \leq k_2) \\ &= P_{\mu,\sigma}(\bar{X} \leq \mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}) \\ &= P_{\mu,\sigma}\left(\tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\alpha}\right) \\ &= \Phi\left(\tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\alpha}\right)\end{aligned}$$

Zřejmě opět platí

$$\beta^*(\mu_0) = \alpha$$

a pro $\mu_1 > \mu_0$

je síla testu < chyba 1. druhu.

(C) Abychom se vyvarovali předchozích obtíží, zvolme nyní kritický obor takto

$$W_\alpha = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \notin (k_1, k_2), \text{ kde } k_1 < k_2\} = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \notin \left(\mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \mu_0 + \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right)\right\}.$$

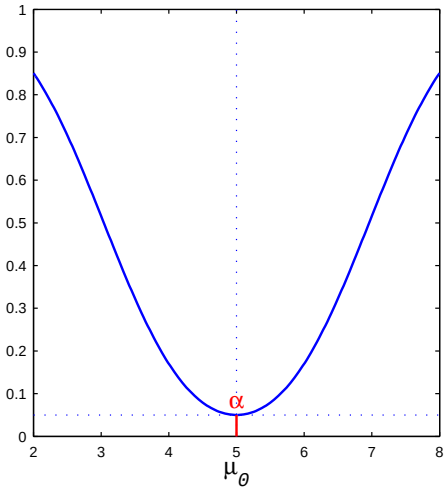
Pak **silofunkce** je tvaru

$$\begin{aligned}\beta^*(\boldsymbol{\theta}) &= 1-\beta(\boldsymbol{\theta}) = \beta^*(\mu) = \int_{W_\alpha} p_1 \, d\nu \\ &= P_{\mu,\sigma}(\bar{X} \le k_1 \wedge \bar{X} \ge k_2) \\ &= 1 - P_{\mu,\sigma}(\mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \le \bar{X} \le \mu_0 + \tfrac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}) \\ &= 1 - P_{\mu,\sigma}\left(\tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\frac{\alpha}{2}} \le \tfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} + u_{1-\frac{\alpha}{2}}\right) \\ &= 1 - \Phi\left(\tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} + u_{1-\alpha}\right) + \Phi\left(\tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} - u_{1-\alpha}\right)\end{aligned}$$

Zřejmě platí

$$\beta^*(\mu_0) = \alpha$$

Silofunkce $\beta^*(\mu)$



a neexistuje žádné $\mu \neq \mu_0$, pro které je síla testu menší než chyba 1. druhu, takže jde o **nezkreslený test**.

9.3. TESTY PODÍLEM VĚROHODNOSTÍ A TESTY ZALOŽENÉ NA INTERVALOVÝCH ODHADECH. Neymanovu-Pearsonovu větu nelze bezprostředně aplikovat na případ, kdy množiny $\boldsymbol{\Theta}_0$, $\boldsymbol{\Theta}_1$ **nejsou obě jednobodové**. Její princip konstrukce kritického oboru lze však použít s tím, že na místě $\boxed{p_j(\mathbf{x})}$, $j = 0, 1$, píšeme $\boxed{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_j} p(x;\boldsymbol{\theta})}$.

Dostáváme tedy kritický obor tvaru

$$W_0^* = \left\{\mathbf{x} \in \mathbb{R}^n : \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1} p(x;\boldsymbol{\theta}) \ge c \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} p(x;\boldsymbol{\theta})\right\}.$$

Pokud $c > 1$ (což je pravidlem) je ekvivalentně

$$W_0^* = \left\{\mathbf{x} \in \mathbb{R}^n : \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(x;\boldsymbol{\theta}) \ge c \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} p(x;\boldsymbol{\theta})\right\} = \left\{\mathbf{x} \in \mathbb{R}^n : p(x;\widehat{\boldsymbol{\theta}}_{\text{MLE}}) \ge c p(x;\widehat{\boldsymbol{\theta}}_{0,\text{MLE}})\right\},$$

kde $\widehat{\boldsymbol{\theta}}_{\text{MLE}}$ je maximálně věrohodný odhad $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ a $\widehat{\boldsymbol{\theta}}_{0,\text{MLE}}$ je maximálně věrohodný odhad za hypotézy H_0.

Příklad 9.9 (Náhodný výběr z normálního rozdělení při neznámém rozptylu a oboustranné alternativě)**.** Mějme $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq N(\mu, \sigma^2)$, kde μ a σ^2 jsou neznámé parametry. Máme testovat hypotézu

$\boxed{H_0} : \mu = \mu_0$ proti alternativě $\boxed{H_1} : \mu \neq \mu_0$ na hladině významnosti $\alpha \in (0, 1)$

Parametr $\boldsymbol{\theta} = (\mu, \sigma^2)$ je zde dvourozměrný, množina $\boldsymbol{\Theta} = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, 0 < \sigma^2 < \infty\}$. Maximálně věrohodné odhady jsou

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \left(\bar{X} = \tfrac{1}{n}\sum_{i=1}^n X_i, \tfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right) \quad \text{a} \quad \widehat{\boldsymbol{\theta}}_{0,\text{MLE}} = \left(\mu_0, \tfrac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2\right)$$

Dosadíme-li tyto odhady za $\boldsymbol{\theta} = (\mu, \sigma^2)$ do výrazu

$$p(\mathbf{x};\boldsymbol{\theta}) = \textstyle\prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x_i-\mu)^2}{2\sigma^2}\right\}\right) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\},$$

dostaneme pro W_0^* nerovnost

$$\left(\tfrac{2\pi}{n}\textstyle\sum_{i=1}^n (x_i - \bar{x})^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\} \ge c \left(\tfrac{2\pi}{n}\textstyle\sum_{i=1}^n (x_i - \mu_0)^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\},$$

což je

$$\sum_{i=1}^n (x_i - \bar{x})^2 \le c_1 \sum_{i=1}^n (x_i - \mu_0)^2.$$

Dále využijeme vztah

$$\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})^2}_{=(n-1)s_n^2}=\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu_0)^2 \quad \Rightarrow \quad \sum_{i=1}^{n}(x_i-\mu_0)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu_0)^2,$$

$$\text{takže} \quad \sum_{i=1}^{n}(x_i-\bar{x})^2 \le c_1\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right]$$

což nakonec můžeme vyjádřit takto

$$|\bar{x}-\mu_0|\sqrt{n} \ge c_2\sqrt{\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2}=c_2 s_n \quad \Rightarrow \quad \tfrac{|\bar{x}-\mu_0|}{s_n}\sqrt{n} \ge c_2 .$$

Protože veličina $\boxed{T_n=\frac{\bar{X}-\mu_0}{S_n}\sqrt{n}}$ má za platnosti nulové hypotézy **Studentovo** t–rozdělení o $n-1$ stupních volností, pak na základě tohoto rozdělení můžeme určit **kritickou hodnotu**

$$c_2=t_{1-\frac{\alpha}{2}}(n-1),$$

neboť

$$\alpha=P_{(\mu_0,\sigma^2)}(|T_n|\ge c_2)=P_{(\mu_0,\sigma^2)}\left(\tfrac{|\bar{X}-\mu_0|}{S_n}\sqrt{n}\ge t_{1-\frac{\alpha}{2}}(n-1)\right)$$

nebo ekvivalentně

$$1-\alpha=P_{(\mu_0,\sigma^2)}\left(\bar{X}-\tfrac{S_n}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\le\mu_0\le\bar{X}+\tfrac{S_n}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right)$$

Hypotézu $\boxed{H_0}: \mu=\mu_0$ tedy **zamítáme** ve prospěch alternativy $\boxed{H_1}: \mu\ne\mu_0$ na hladině významnosti $\boxed{\alpha}$, pokud realizace

$$t_n=\tfrac{|\bar{x}-\mu_0|}{s_n}\sqrt{n}\ge t_{1-\frac{\alpha}{2}}(n-1).$$

Výsledky příkladů 9.4 a 9.9 naznačují, že existuje určitý **VZTAH MEZI TESTY A INTERVALOVÝMI ODHADY**, který lze popsat následovně.

Mějme náhodný výběr $\mathbf{X}=(X_1,\dots,X_n)'$ rozsahu n z rozdělení, které závisí na parametru $\boldsymbol{\theta}=(\theta_1,\dots,\theta_m)'\in\boldsymbol{\Theta}$ a parametrickou funkci $\gamma(\boldsymbol{\theta})$.

(A) Hypotéza $\boxed{H_0}: \gamma(\boldsymbol{\theta})=\gamma(\boldsymbol{\theta}_0)$ proti (tzv. *oboustranné*) alternativě $\boxed{H_1}: \gamma(\boldsymbol{\theta})\ne\gamma(\boldsymbol{\theta}_0)$:

Mějme **intervalový odhad** $(D_n(\mathbf{X}),H_n(\mathbf{X}))$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $1-\alpha$. Pokud platí nulová hypotéza, pak

$$1-\alpha=P_{\boldsymbol{\theta}}\left(D_n(\mathbf{X})\le\gamma(\boldsymbol{\theta}_0)\le H_n(\mathbf{X})\right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha=\{\mathbf{X}\in\mathbb{R}^n:\gamma(\boldsymbol{\theta}_0)\notin(D_n(\mathbf{X}),H_n(\mathbf{X}))\}.$$

Zjistíme-li v konkrétní situaci, že

$$\gamma(\boldsymbol{\theta}_0) \notin (d_n(\mathbf{x}), h_n(\mathbf{x})) \text{ tj. realizace } \mathbf{x} \in W_\alpha ,$$

potom

- buď nastal jev, který má pravděpodobnost α (volí se blízká nule),
- nebo neplatí nulová hypotéza.

Protože při obvyklé volbě $\alpha = 0.05$ nebo $\alpha = 0.01$ je tento jev „prakticky nemožný“, proto nulovou hypotézu $\boxed{H_0}$ **zamítáme ve prospěch alternativy** $\boxed{H_1}$.

V opačném případě, tj. pokud

$$\gamma(\boldsymbol{\theta}_0) \in (d_n(\mathbf{x}), h_n(\mathbf{x})) \text{ tj. realizace } \mathbf{x} \notin W_\alpha ,$$

nulovou hypotézu $\boxed{H_0}$ **nezamítáme**.

(B) Hypotéza $\boxed{H_0} : \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti (tzv. *jednostranné*) alternativě $\boxed{H_1} : \gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$:

V tomto případě využijeme **dolní odhad** $D_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $\boxed{1-\alpha}$. Pokud platí nulová hypotéza, pak

$$1 - \alpha = P_{\boldsymbol{\theta}} \left(D_n(\mathbf{X}) \leq \gamma(\boldsymbol{\theta}_0) \right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha = \left\{ \mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0) \right\}.$$

(C) Hypotéza $\boxed{H_0} : \gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$ proti (tzv. *jednostranné*) alternativě $\boxed{H_1} : \gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$

V tomto případě využijeme **horní odhad** $H_n(\mathbf{X})$ parametrické funkce $\gamma(\boldsymbol{\theta})$ o spolehlivosti $\boxed{1-\alpha}$. Pokud platí nulová hypotéza, pak

$$1 - \alpha = P_{\boldsymbol{\theta}} \left(\gamma(\boldsymbol{\theta}_0) \leq H_n(\mathbf{X}) \right),$$

takže **kritický obor** tohoto testu má tvar:

$$W_\alpha = \left\{ \mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0) \right\}.$$

Předchozí úvahy shrňme do následující tabulky:

H_0	H_1	Hypotézu H_0 zamítáme, pomocí	
		intervalu spolehlivosti	kritické oblasti, tj. pokud $\mathbf{x} \in W_\alpha$, kde $W_\alpha =$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) \neq \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) \notin (d_n(\mathbf{x}), h_n(\mathbf{x}))$	$\{\mathbf{X} \in \mathbb{R}^n : \gamma(\boldsymbol{\theta}_0) \notin (D_n(\mathbf{X}), H_n(\mathbf{X}))\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) > \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) < d_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : D_n(\mathbf{X}) > \gamma(\boldsymbol{\theta}_0)\}$
$\gamma(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}) < \gamma(\boldsymbol{\theta}_0)$	$\gamma(\boldsymbol{\theta}_0) > h_n(\mathbf{x})$	$\{\mathbf{X} \in \mathbb{R}^n : H_n(\mathbf{X}) < \gamma(\boldsymbol{\theta}_0)\}$

9.4. TESTY O PARAMETRECH NORMÁLNÍHO ROZDĚLENÍ. TESTY ZALOŽENÉ NA CENTRÁLNÍ LIMITNÍ VĚTĚ.

Pomocí intervalových (dolních, horních) odhadů, které jsme již dříve odvodili v sekci 7, dostáváme celou řadu kritických oblastí testů o parametrech normálního rozdělení. Poznamenejme, že se shodují s testy podílem věrohodností.

Přehled takto získaných testů pro JEDEN NÁHODNÝ VÝBĚR $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$ podáváme v následující tabulce:

H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{X}\in W_\alpha$, tj.	Předpoklady
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert\bar X-\mu_0\rvert\sqrt{n}\geq \sigma u_{1-\frac{\alpha}{2}}$	σ^2 známé
$\mu=\mu_0$	$\mu>\mu_0$	$(\bar X-\mu_0)\sqrt{n}\geq \sigma u_{1-\alpha}$	σ^2 známé
$\mu=\mu_0$	$\mu<\mu_0$	$(\bar X-\mu_0)\sqrt{n}\leq -\sigma u_{1-\alpha}$	σ^2 známé
$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert\bar X-\mu_0\rvert\sqrt{n}\geq S_n t_{1-\frac{\alpha}{2}}(n-1)$	σ^2 neznámé
$\mu=\mu_0$	$\mu>\mu_0$	$(\bar X-\mu_0)\sqrt{n}\geq S_n t_{1-\alpha}(n-1)$	σ^2 neznámé
$\mu=\mu_0$	$\mu<\mu_0$	$(\bar X-\mu_0)\sqrt{n}\leq -S_n t_{1-\alpha}(n-1)$	σ^2 neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$\frac{(n-1)S_n^2}{\sigma_0^2}\notin\left(\chi^2_{\frac{\alpha}{2}}(n-1),\chi^2_{1-\frac{\alpha}{2}}(n-1)\right)$	μ neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$	$\frac{(n-1)S_n^2}{\sigma_0^2}\leq \chi^2_{\alpha}(n-1)$	μ neznámé
$\sigma^2=\sigma_0^2$	$\sigma^2<\sigma_0^2$	$\frac{(n-1)S_n^2}{\sigma_0^2}\geq \chi^2_{1-\alpha}(n-1)$	μ neznámé

V případě DVOU NEZÁVISLÝCH VÝBĚRŮ

- první náhodný výběr $\perp\!\!\!\perp\{X_1,\dots,X_{n_X}\}\sim N(\mu_X,\sigma_X^2)$ (s výběrovým průměrem $\bar X$ a výběrový rozptylem S_X^2),
- druhý náhodný výběr $\perp\!\!\!\perp\{Y_1,\dots,Y_{n_Y}\}\sim N(\mu_Y,\sigma_Y^2)$ (s výběrovým průměrem $\bar Y$ a výběrový rozptylem S_Y^2),
- a pokud označíme

$$S_{XY}^2=\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2},$$

pak následující tabulka se týká testů rovnosti středních hodnot a rozptylů:

H_0	H_1	Hypotézu H_0 zamítáme, pokud $(\mathbf{X}',\mathbf{Y}')'\in W_\alpha$, tj.	Předpoklady
$\mu_X=\mu_Y$	$\mu_X\neq\mu_Y$	$\lvert\bar X-\bar Y\rvert\geq u_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}}$	σ^2 známé
$\mu_X=\mu_Y$	$\mu_X\neq\mu_Y$	$\lvert\bar X-\bar Y\rvert\geq t_{1-\frac{\alpha}{2}}(n_X+n_Y-2)\ S_{XY}\sqrt{\frac{n_X+n_Y}{n_Xn_Y}}$	σ^2 neznámé
$\sigma_X^2=\sigma_Y^2$	$\sigma_X^2\neq\sigma_Y^2$	$\frac{S_X^2}{S_Y^2}\notin\left(F_{\frac{\alpha}{2}}(n_X-1,n_Y-1),F_{1-\frac{\alpha}{2}}(n_X-1,n_Y-1)\right)$	μ_X,μ_Y neznámé

Následující tabulka nabízí ASYMPTOTICKÉ TESTY pro náhodné výběry $\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq\mathcal{L}(\mu(\boldsymbol\theta),\sigma^2(\boldsymbol\theta))$ s konečnými druhými momenty (s výběrovým průměrem $\bar X=\frac{1}{n}\sum_{i=1}^n X_i$ a se $S_*^2=S_*^2(\mathbf{X})$, což je (**slabě**) **konzistentní odhad** rozptylu $\sigma^2(\boldsymbol\theta)$):

H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{X}\in W_\alpha$, tj.	Předpoklady
$\mu=\mu_0$	$\mu\neq\mu_0$	$\frac{\lvert\bar X-\mu_0\rvert}{S_*}\sqrt{n}\geq\ u_{1-\frac{\alpha}{2}}$	$0<\sigma^2(\boldsymbol\theta)<\infty$
$\mu=\mu_0$	$\mu\neq\mu_0$	$\frac{\lvert\bar X-\mu_0\rvert}{\sqrt{\bar X}}\sqrt{n}\geq u_{1-\frac{\alpha}{2}}$	$\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq Po(\mu)$
$p=p_0$	$p\neq p_0$	$\frac{\lvert\bar X-p_0\rvert}{\sqrt{p_0(1-p_0)}}\sqrt{n}\geq u_{1-\frac{\alpha}{2}}$	$\perp\!\!\!\perp\{X_1,\dots,X_n\}\simeq A(p)$

9.5. Vztah mezi pravděpodobností chyby prvního, druhého druhu a počtem pozorování. Abychom si uvědomili vztah mezi oběma chybami, ukážeme jednoduchý příklad.

Příklad 9.10 (Jednoduchá hypotéza i alternativa pro binomické rozdělení)**.** Dva chlapci, Honzík a František, mají každý svůj pytlík s barevnými kuličkami. Honzík má 80 bílých a 20 modrých kuliček, František 30 bílých a 70 modrých kuliček. Oba pytlíky jsou k nerozeznání. Vybereme náhodně jeden z pytlíků a chceme rozhodnout, kterému z chlapců patří. Za tím účelem provedeme následující test:

Výchozí test A: Vybereme z pytlíku 10 kuliček. Pokud mezi nimi bude méně než $k = 8$ bílých kuliček, zamítneme hypotézu, že patří Honzíkovi.

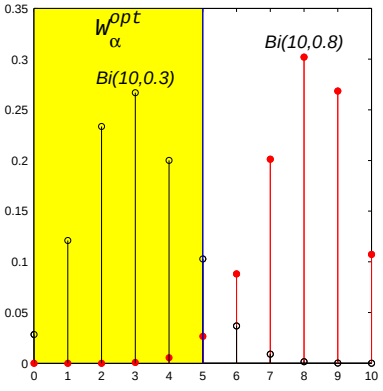
Vypočítejme chybu prvního i druhého druhu a pokusme se najít takový test, který by zajistil, aby chyby prvního i druhého druhu byly vůči chlapcům co nejvíce spravedlivé.

Označme jako $\boxed{Y}$ náhodnou veličinu, která značí počet bílých kuliček mezi deseti vybranými. Náhodná veličina $Y \in \{0, 1, \ldots, n\}$, $n = 10$. Zřejmě má binomické rozdělení, což pro $j = 0, 1$ značíme

$Y \sim Bi(n, \theta)$ s pravděpodobnostní funkcí $p_j(x) = \begin{cases} \binom{n}{y}\theta_j^y(1-\theta_j)^{n-y} & y = 0, \ldots, n, \\ 0 & \text{jinak.} \end{cases}$

Budeme testovat hypotézu $H_0 : \theta = \theta_0 = 0.8$ proti alternativě $H_1 : \theta = \theta_1 = 0.3$, kde kritický obor je $W_\alpha = \{0, 1, \ldots, k-1\}$. „Spravedlivý“ test budeme hledat pomocí procedury v Matlabu s využitím příkazů „`binocdf(y,n,theta)`“

$p_1(y)$ (vlevo) a $p_0(y)$ (vpravo)

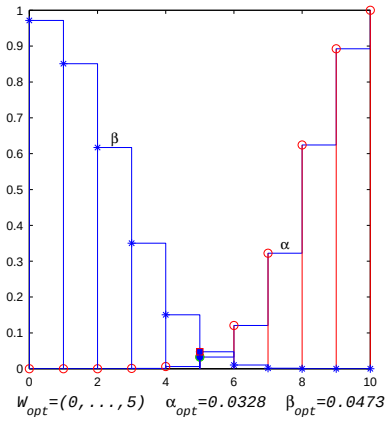


Hledání „spravedlivého“ testu pro

H_0: θ_0=0.8 proti H_1: θ_1=0.3

W_α=(0,..., 0)	α=0.0000	β=0.9718
W_α=(0,..., 1)	α=0.0000	β=0.8507
W_α=(0,..., 2)	α=0.0001	β=0.6172
W_α=(0,..., 3)	α=0.0009	β=0.3504
W_α=(0,..., 4)	α=0.0064	β=0.1503
W_α=(0,..., 5)	α=0.0328	β=0.0473
W_α=(0,..., 6)	α=0.1209	β=0.0106
W_α=(0,..., 7)	α=0.3222	β=0.0016
W_α=(0,..., 8)	α=0.6242	β=0.0001
W_α=(0,..., 9)	α=0.8926	β=0.0000
W_α=(0,...,10)	α=1.0000	β=0.0000

Chyby β $(*)$ a α $(\circ)$



Optimální test B: Pokud mezi deseti vybranými kuličkami bude méně než $k = 6$ bílých, pak zamítáme hypotézu, že pytlík s kuličkami patří Honzíkovi.

Teprve nyní je pravděpodobnost chyby prvního i druhého druhu vyvážená, srovnejme

$$\alpha = \int_{W_\alpha} p_0 \, d\nu = \sum_{i=1}^{k-1} 0.8^y(1-0.8)^{n-y} = \begin{cases} 0.3222 & \text{A} \\ 0.0328 & \text{B} \end{cases} \qquad 1-\alpha = \begin{cases} 0.6778 & \text{A} \\ 0.9672 & \text{B} \end{cases}$$

$$\beta = \int_{W_1} p_1 \, d\nu = \sum_{i=k}^{10} 0.3^y(1-0.3)^{n-y} = \begin{cases} 0.0016 & \text{A} \\ 0.0473 & \text{B} \end{cases} \qquad 1-\beta = \begin{cases} 0.9984 & \text{A} \\ 0.9527 & \text{B} \end{cases}$$

Tedy pravděpodobnost, že se v testu B vyvarujeme chyby 1. druhu je $\boxed{1-\alpha = 0.9672}$ a chyby 2. druhu je $\boxed{1-\beta = 0.9527}$.

V předchozím příkladě jsme se snažili najít takový test, aby obě dvě chyby vyhovovaly našim představám.

Nyní se opět vrátíme k příkladu 9.8 a ukážeme, že síla testu je pro pevně danou chybu prvého druhu ovlivněna rozsahem výběru.

Příklad 9.11 (Síla testu a rozsah výběru pro jednoduchou hypotézu a složenou alternativu v případě náhodného výběru z normálního rozdělení při známém rozptylu)**.** Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq N(\mu, \sigma^2)$ je normální náhodný výběr, ve kterém je μ je neznámý parametr a $\sigma^2 > 0$ je známá konstanta. Uvažujme test hypotéz

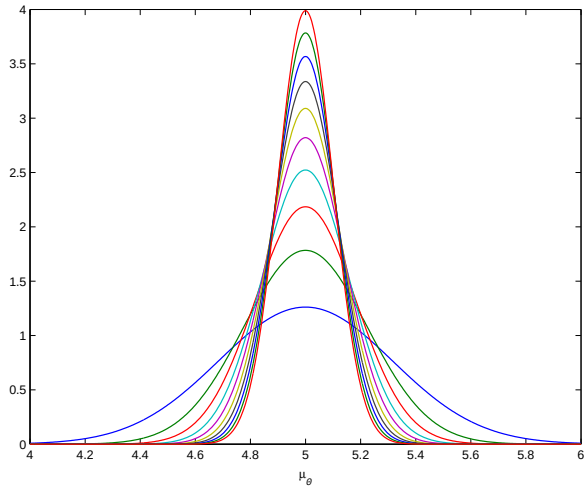
(a)	$H_0 : \mu = \mu_0$	proti	$H_1 : \mu \neq \mu_0$
(b)	$H_0 : \mu = \mu_0$	proti	$H_1 : \mu < \mu_0$
(c)	$H_0 : \mu = \mu_0$	proti	$H_1 : \mu > \mu_0$

V příkladu 9.8 jsme zkonstruovali **nezkreslený test** pro oboustrannou alternativu a v příkladu 9.4 **stejnoměrně nejsilnější testy** pro jednostranné alternativy.

Na následujících grafech ukážeme, jak při pevně dané chybě prvého druhu roste hodnota silofunkce při rostoucím rozsahu výběru. Toho se právě využívá, pokud si předepíšeme obě chyby a hledáme rozsah výběru, při kterém nepřekročíme stanovené chyby.

$\bar{X} \sim N(\mu_0, \sigma^2/\sqrt{n})$

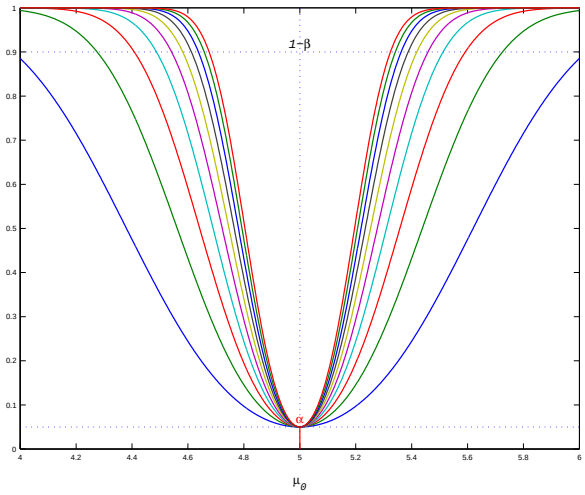
$\mu_0 = 5,\ \sigma = 1,\ n = 10, 20, \ldots, 90, 100$



Hustoty výběrových průměrů

(a) Silofunkce $\beta^*(\mu) = 1 - \beta(I\mu)$

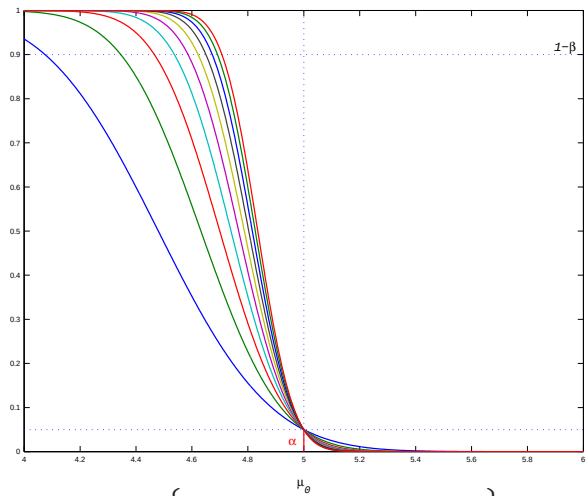
$\alpha = 0.05,\ \mu_0 = 5,\ \sigma = 1, n = 10, 20, \ldots, 90, 100$



$W_\alpha = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \notin \left(\mu_0 - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}, \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right)\right\}$

(b) Silofunkce $\beta^*(\mu) = 1 - \beta(I\mu)$

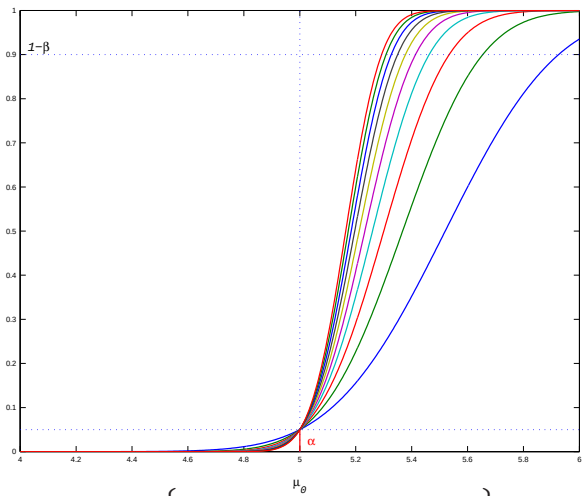
$\alpha = 0.05,\ \mu_0 = 5,\ \sigma = 1, n = 10, 20, \ldots, 90, 100$



$W_\alpha = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\}$

(c) Silofunkce $\beta^*(\mu) = 1 - \beta(I\mu)$

$\alpha = 0.05,\ \mu_0 = 5,\ \sigma = 1, n = 10, 20, \ldots, 90, 100$



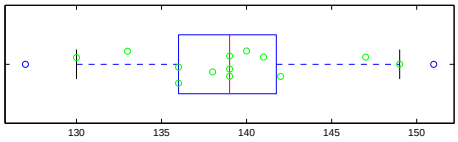
$W_\alpha = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\}$

Příklad 9.12 (Výška desetiletých chlapců)**.** V roce 1961 byla u 15 náhodně vybraných chlapců z populace všech desetiletých chlapců žijících v Československu zjištěna výška

Výšky 15 desetiletých chlapců

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
130	140	136	141	139	133	149	151	139	136	138	142	127	139	147

Je známo, že každá následující generace je v průměru o něco vyšší než generace předcházející. Můžeme se tedy ptát, zda průměr $\bar{x} = 139.133$ zjištěný v náhodném výběru rozsahu



$n = 15$ znamená, že na 5% hladině máme zamítnout nulovou hypotézu $H_0 : \mu = 136.1$ (zjištění z roku 1951) ve prospěch alternativní hypotézy $H_1 : \mu > 136.1$.

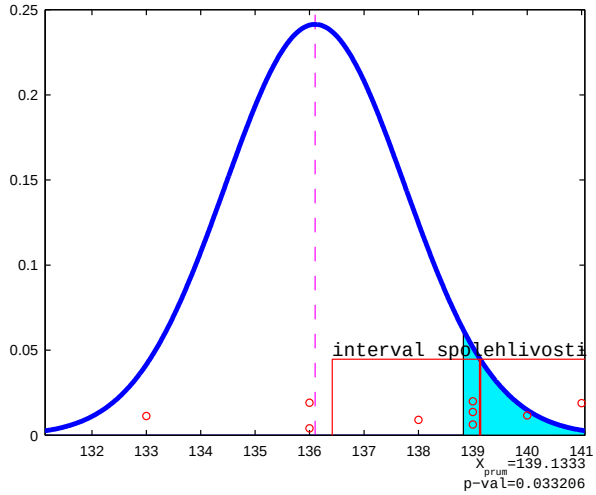
Rozptyl $\sigma^2 = 6.4^2$ cm^2, zjištěný v roce 1951 (kdy se provádělo rozsáhlé šetření), můžeme považovat za známý, neboť variabilita výšek zůstává (na rozdíl od střední výšky) téměř nezměněná.

(I) Testování nulové hypotézy pomocí pivotové statistiky $U_{\overline{X}}$ a kritické hodnoty. Protože kritický obor W_0 lze ekvivalentně vyjádřit i takto

$$W_0 = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \le k_2\} = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \le \mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\} = \left\{\mathbf{x} \in \mathbb{R}^n : u_{\bar{x}} = \tfrac{\bar{x}-\mu_0}{\sigma}\sqrt{n} \le u_{1-\alpha}\right\},$$

počítejme $u_{\bar{x}} = \frac{139.133-136.1}{6.4}\sqrt{15} = 1.835$. Protože $u_{\bar{x}} = 1.835$ překračuje kritickou hodnotu $u_{1-\alpha} = u_{0.95} = 1.645$ (získáme pomocí Matlabu, a to příkazem „`norminv(0.95)`“) nulovou hypotézu na 5% hladině **zamítneme ve prospěch alternativní hypotézy, že se střední výška desetiletých hochů zvětšila**.

(II) Testování nulové hypotézy pomocí p-hodnoty



Dosažená hladina odpovídající testové statistice (tj. tzv. p-hodnota, anglicky *P-value*, *significance value*), což je nejmenší hladina testu, při které bychom ještě hypotézu H_0 zamítli, je rovna 0.033 (opět získáme pomocí Matlabu příkazem „`1 - normcdf(mean(x),136.1,6.4/sqrt(n))`“), takže například při $\alpha = 2.5\%$ by již dosažený výsledek nebyl statisticky významný.

Protože p-hodnota je menší než zvolená hladina významnosti $\alpha = 0.05$, hypotézu **zamítáme**.

(III) Testování nulové hypotézy pomocí intervalu spolehlivosti $\langle D, +\infty)$

Protože jde o jednostranný test, použijeme **dolní odhad** střední hodnoty μ

$$d = \bar{x} - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha} = 139.133 - \tfrac{6.4}{\sqrt{15}} 1.645 = 136.415$$

Protože interval spolehlivosti $\langle 136.415, +\infty)$ nepokrývá hodnotu 136.1, proto nulovou hypotézuna na hladině významnosti $\alpha = 0.05$ **zamítáme**.

Příklad 9.13 (Počet pozorování při dané chybě prvního a druhého druhu)**.**

Mějme $\perp\!\!\!\perp\{X_1,\dots,X_n\} \simeq N(\mu,\sigma^2)$, kde $\sigma^2 = 25$ je známé. Chceme testovat hypotézu

$$\boxed{H_0}: \mu = \mu_0 = 5 \qquad \text{proti} \qquad \boxed{H_1}: \mu = \mu_1 = 4.$$

Naším úkolem je zjistit rozsah výběru tak, aby chyba 1. druhu byla rovna 0.05 a druhého druhu 0.01.

V příkladě 9.4 jsme, ukázali, že kritický obor pro rovnoměrně nejsilnější test pro alternativu typu $\mu_0 > \mu_1$ je tvaru

$$W_0 = \{\mathbf{x} \in \mathbb{R}^n : \bar{x} \le k_2\} = \left\{\mathbf{x} \in \mathbb{R}^n : \bar{x} \le \mu_0 - \tfrac{\sigma}{\sqrt{n}} u_{1-\alpha}\right\}.$$

Jeli $\alpha = 0.05$, pak $u_{1-\alpha} = 1.645$. Při této volbě máme zajištěnu chybu prvního druhu rovnou 0.05, tj.

$$P_{\mu_0}(\bar{X} \le k_2) = \Phi\left(\tfrac{k_2-\mu_0}{\sigma/\sqrt{n}}\right) = \alpha = 0.05.$$

Nyní musíme zvolit n tak, aby pro chybu druhého druhu platilo

$$P_{\mu_1}(\bar{X} > k_2) = 1 - \Phi\left(\tfrac{k_2-\mu_1}{\sigma/\sqrt{n}}\right) \le \beta = 0.01,$$

takže

$$u_{1-\beta} = \frac{k_2-\mu_1}{\sigma/\sqrt{n}} = \frac{\mu_0 - \frac{\sigma}{\sqrt{n}}u_{1-\alpha} - \mu_1}{\sigma/\sqrt{n}} = \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - u_{1-\alpha}$$

a odtud již dostaneme, že

$$u_{1-\beta} + u_{1-\alpha} = \tfrac{\mu_0-\mu_1}{\sigma/\sqrt{n}},$$

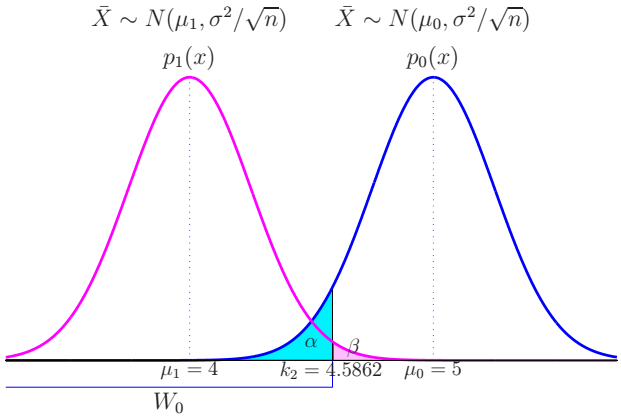
takže

$$\sqrt{n} = \tfrac{u_{1-\beta}+u_{1-\alpha}}{\mu_0-\mu_1}\sigma = 19.8560$$

tj.

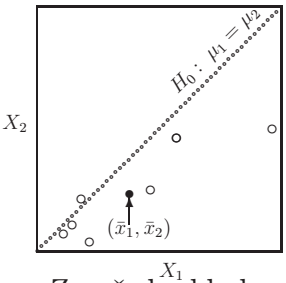
$$n = \left\lceil \tfrac{(u_{1-\beta}+u_{1-\alpha})^2}{(\mu_0-\mu_1)^2}\sigma^2 \right\rceil = \lceil 394.2610 \rceil = 395,$$

kde symbol $\lceil c \rceil$ značí zaokrouhlení na celé číslo nahoru.



Pokud ovšem bychom σ neznali, **pak by úloha nešla vyřešit**.

Příklad 9.14. PÁROVÝ TEST



Na sedmi rostlinách byl posuzován vliv fungicidního přípravku podle počtu skrvn na listech před a týden po použití přípravku. Otestujte, zdali má přípravek vliv na počet skrvn na listech. Data udávající počet skrvn na listech před a po použití přípravku:

POČET SKRVN NA LISTECH

před použitím přípravku	X_1	9	17	31	7	8	20	10
po použití přípravku	X_2	10	11	18	6	7	17	5

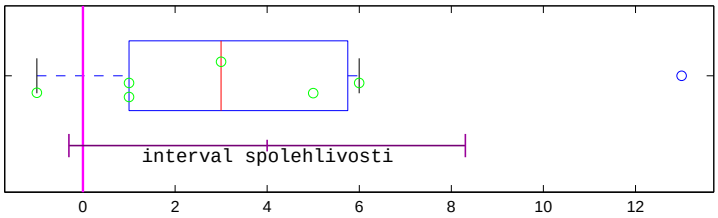
Za předpokladu, že náhodný výběr pochází z normálního rozdělení, tj.

$$\perp\!\!\!\perp \left\{ \begin{pmatrix} X_{1,1} \\ X_{2,1} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1,n} \\ X_{2,n} \end{pmatrix} \right\} \sim N_2 \left(\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \text{ kde } \rho \in (0,1)$$

pak $\begin{matrix} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{matrix}$, $Z = X_1 - X_2 \sim N(\mu_z = \mu_1 - \mu_2, \sigma_z^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2)$

a statistika $T = \frac{\bar{Z}}{S_Z/\sqrt{n}} = \frac{\bar{X}_1 - \bar{X}_2}{S_Z/\sqrt{n}}$ má za platnosti nulové hypotézy $\boxed{H_0 : \mu_1 - \mu_2 = 0}$ Studentovo rozdělení o $n-1$ stupních volnosti.

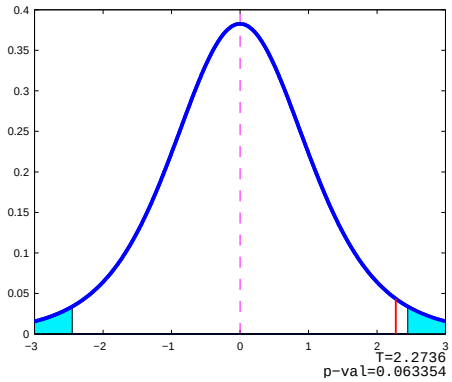
(I) TESTOVÁNÍ NULOVÉ HYPOTÉZY POMOCÍ INTERVALU SPOLEHLIVOSTI



$$[\bar{X}_1 - \bar{X}_2 - t_{1-\alpha/2}(n-1) \cdot S/\sqrt{n};$$
$$\bar{X}_1 - \bar{X}_2 + t_{1-\alpha/2}(n-1) \cdot S/\sqrt{n}] =$$
$$[4 \pm 2.4469 \cdot 4.6547/2.6458] =$$
$$[-0.30492; 8.3049]$$

Protože interval spolehlivosti pokrývá hodnotu $Z = 0$, na dané hladině významnosti **hypotézu nemůžeme zamítnout**.

(II) TESTOVÁNÍ NULOVÉ HYPOTÉZY POMOCÍ STATISTIKY T A KRITICKÉ HODNOTY



Vypočítáme-li hodnotu statistiky

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}}$$

a porovnáme s kvantilem Studentova rozdělení, tj.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s/\sqrt{n}} = 2.2736 \not> t_{1-\alpha/2}(n-1) = 2.4469,$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

nezamítáme.

(III) TESTOVÁNÍ NULOVÉ HYPOTÉZY POMOCÍ p-HODNOTY

Vypočítáme-li p-hodnotu a porovnáme se zvolenou hladinou významnosti $\alpha = 0.05$

$$p = P(|T| > t) = 2(1 - P(|T| \leq t)) = 0.06335 > \alpha$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

nezamítáme.

Shrneme-li předchozí výsledky slovně, pak nulovou hypotézu o tom, že

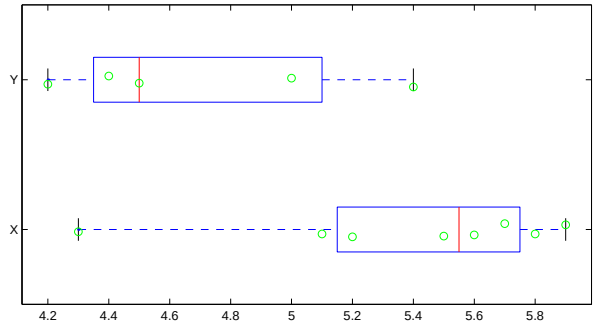
PŘÍPRAVEK NEMÁ VLIV NA POČET SKVRN

na hladině významnosti $\alpha = 0.05$ **nemůžeme zamítnout** oproti alternativě o jeho vlivu.

Příklad 9.15 (Dva nezávislé náhodné výběry z normálního rozdělení při neznámých ale stejných rozptylech)**.** Bylo vybráno 13 polí stejné kvality. Na 8 z nich se zkoušel nový způsob hnojení, zbývajících 5 bylo ošetřeno běžným způsobem. Výnosy pšenice uvedené v tunách na hektar jsou označeny X_i u nového a Y_i u běžného způsobu hnojení. (převzato z knihy Anděl, J.: *Statistické metody*, str. 82, př. 8.2).

Je třeba zjistit, zda způsob hnojení má vliv na výnos pšenice.

X_i	5.7	5.5	4.3	5.9	5.2	5.6	5.8	5.1
Y_i	5.0	4.5	4.2	5.4	4.4			



Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_{n_X}\} \sim N(\mu_X, \sigma_X^2)$ je náhodný výběr rozsahu n_X z normálního rozdělení $N(\mu_X, \sigma_X^2)$, $\bar{X}$ je jeho výběrový průměr a S_X^2 jeho výběrový rozptyl.

Dále nechť $\perp\!\!\!\perp\{Y_1, \ldots, Y_{n_Y}\} \sim N(\mu_Y, \sigma_Y^2)$ je náhodný výběr rozsahu n_Y z normálního rozdělení $N(\mu_Y, \sigma_Y^2)$, $\bar{Y}$ je jeho výběrový průměr a S_Y^2 jeho výběrový rozptyl.

Předpokládejme, že oba výběry jsou stochasticky nezávislé, tj. $\mathbf{X} \perp \mathbf{Y}$.

Chceme-li testovat hypotézu, že rozdíl středních hodnot je nulový (při neznámém rozptylu $\sigma^2 = \sigma_X^2 = \sigma_Y^2$), za pivotovou statistiku zvolíme statistiku

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_{XY}} \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \sim t(n_X + n_Y - 2),$$

kde

$$S_{XY}^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}.$$

Chceme-li použít $T_{\bar{X}-\bar{Y}}$, měli bychom být přesvědčeni o tom, že rozptyly obou výběrů se významně neliší. Budeme tedy nejprve testovat hypotézu $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$, že podíl obou rozptylů je roven jedné proti alternativě, že se nerovná $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$. Za pivotovou statistiku zvolíme statistiku

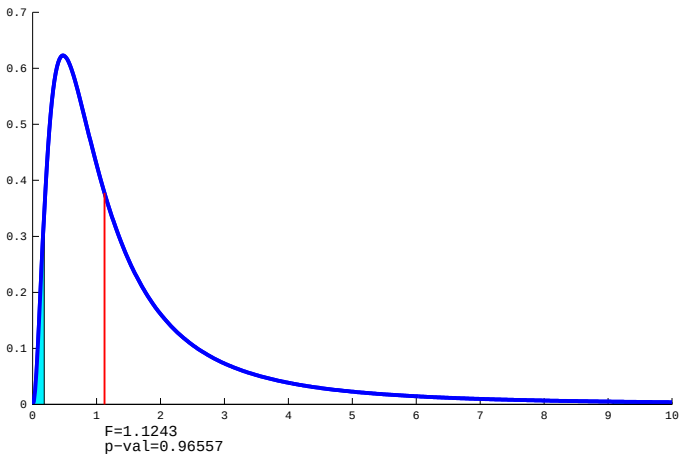
$$F = \frac{S_X^2}{S_Y^2} \frac{\sigma_Y^2}{\sigma_X^2} \sim F(n_X - 1, n_Y - 1).$$

(a) Můžeme například vypočítat statistiku F za platnosti nulové hypotézy a porovnat ji s příslušnými oboustrannými kvantily.

Protože

$f = 1.1243$
$F_{\frac{\alpha}{2}}(n_X - 1, n_Y - 1) = 0.1811$
$F_{1-\frac{\alpha}{2}}(n_X - 1, n_Y - 1) = 9.0741$

vidíme, že f není ani větší než horní kritický bod, ani menší než dolní kritický bod, takže hypotézu o rovnosti rozptylů proti alternativě nerovnosti **nezamítáme** a můžeme konstatovat, že data nejsou v rozporu s testovanou hypotézou.



(b) Další možností je spočítat dosaženou hladinu významnosti, tj. p-hodnotu (pomocí Matlabu: `2*min(1-fcdf(var(x)/var(y),n1-1,n2-1),fcdf(var(x)/var(y),n1-1,n2-1)`) a srovnat se zvolenou hladinou testu α:

$$p - value = 0.9656 \gg 0.05$$

Protože p-hodnota je výrazně větší než zvolená hladina testu, hypotézu o rovnosti rozptylů proti alternativě nerovnosti **nezamítáme**. Můžeme také říci, že data **nejsou v rozporu s testovanou hypotézou**.

(c) A naposledy můžeme ještě zkostruovat $100(1-\alpha)\%$ interval spolehlivosti pro podíl rozptylů $\frac{\sigma_X^2}{\sigma_Y^2}$

$$\left\langle \frac{S_X^2}{S_Y^2}\frac{1}{F_{1-\frac{\alpha}{2}}(n_X-1,n_Y-1)}, \frac{S_X^2}{S_Y^2}\frac{1}{F_{\frac{\alpha}{2}}(n_X-1,n_Y-1)} \right\rangle.$$

a zjistit, zda pokrývá hodnotu 1. Protože dostáváme interval $\langle 0.1239, 6.2088\rangle$, který pokrývá jedničku, hypotézu nezamítáme.

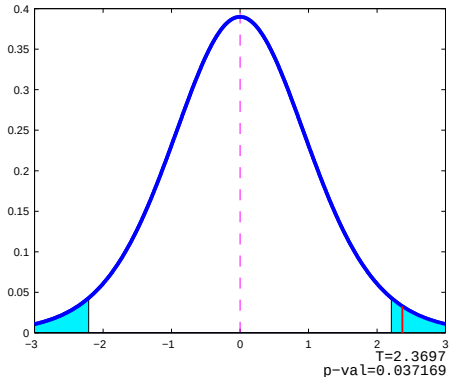
Díky předchozímu zjištění již můžeme bez obav testovat hypotézu $H_0 : \mu_x - \mu_Y = 0$ proti alternativě $H_1 : \mu_x - \mu_Y \neq 0$ a provedeme to opět třemi způsoby:

(I) Testování nulové hypotézy pomocí intervalu spolehlivosti

$$\left\langle \bar{X}-\bar{Y}-t_{1-\frac{\alpha}{2}}(\nu)\ S\sqrt{\tfrac{n_X+n_Y}{n_Xn_Y}}; \bar{X}-\bar{Y}+t_{1-\frac{\alpha}{2}}(\nu)\ S\sqrt{\tfrac{n_X+n_Y}{n_Xn_Y}} \right\rangle = \langle 0.6875 \pm 2.201 \cdot 0.5089/1.7541\rangle$$
$$= \langle 0.048958; 1.326\rangle$$

Protože interval spolehlivosti nepokrývá nulu, na dané hladině významnosti **hypotézu zamítáme** ve prospěch alternativy.

(II) Testování nulové hypotézy pomocí statistiky T a kritické hodnoty



Vypočítáme-li hodnotu statistiky

$$T_{\bar{X}-\bar{Y}} = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_{XY}}\sqrt{\frac{n_Xn_Y}{n_X+n_Y}}$$

a porovnáme s kvantilem Studentova rozdělení, tj.

$$t_{\bar{x}-\bar{y}} = 2.3697 > t_{1-\alpha/2}(11) = 2.201,$$

takže **hypotézu**

$$H_0 : \mu_X - \mu_Y = 0$$

zamítáme.

(III) Testování nulové hypotézy pomocí p-hodnoty

Vypočítáme-li p-hodnotu a porovnáme se zvolenou hladinou významnosti $\alpha = 0.05$

$$p = P(|T_{\bar{X}-\bar{Y}}| > t) = 2(1 - P(|T_{\bar{X}-\bar{Y}}| \leq t)) = 0.037169 < \alpha$$

takže **hypotézu**

$$H_0 : \mu_1 - \mu_2 = 0$$

zamítáme.

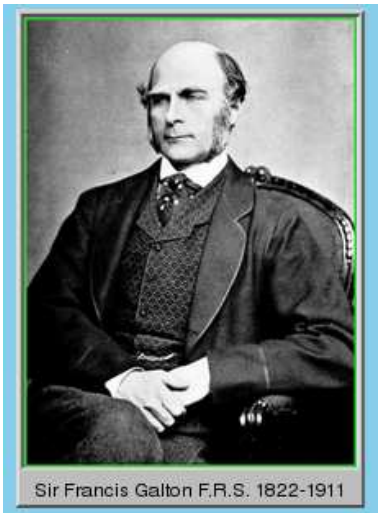
Shrneme-li předchozí výsledky slovně, pak nulovou hypotézu o tom, že

hnojení je stejně účinné

na hladině významnosti $\alpha = 0.05$ **zamítáme** ve prospěch alternativy, že má rozdílné účinky.

10. Regresní analýza

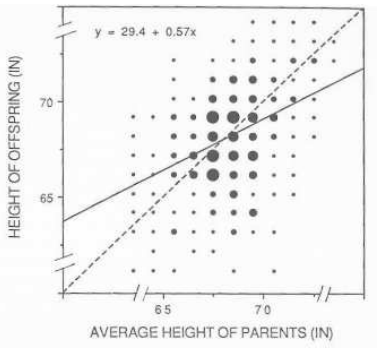
10.1. Pojem regrese. Název *regrese* pochází z prací antropologa a meteorologa *Francise Galtona*, které předložil veřejnosti v letech 1877 až 1885. Galton se zabýval obecnými otázkami dědičnosti a mimo jiné také o vztah mezi výškou otců a jejich prvorozených synů. Pozorováním a analýzou údajů došel k rovnici, ze které vyplývá, že



- ◇ vysocí otcové sice mají i vysoké syny, ale v průměru jsou větší než jejich synové,
- ◇ a podobně i malí otcové mají i malé syny, ale v průměru jsou menší než jejich synové.

Směrnice regresní přímky má hodnotu menší než 1 (přibližně kolem 0.5). To znamená, že otcové, kteří jsou například o 10 cm vyšší, než je průměrná výška mužů jejich generace, mají syny v průměru jen o 5 cm vyšší, než je průměrná výška muže v generaci synů (jde samozřejmě o výšku v dospělosti).

Směrnice regresní přímky, která číselně charakterizuje velikost této tendence, dostala proto název **regresní koeficient**.



Tuto tendenci návratu následující generace směrem k průměru nazval Galton **regresí** (původně tomuto jevu říkal *reversion*, než později změnil na *regression* = krok zpět).

Současné pojetí regresní analýzy má sice jen málo společného s původním záměrem Galtona, nicméně myšlenka přístupu k empirickým datům zůstala zachována a pojem regrese se natolik vžil, že se používá dodnes.

10.2. Definice modelu. Regresní analýza je velmi široké téma, proto se v této úvodní přednášce omezíme jen na studium modelu s regresní přímkou, který definujeme takto:

DEFINICE 10.1. Nechť

$(M1)$
$Y_1, \dots, Y_n$ (1) jsou **nezávislé náhodné veličiny** se středními hodnotami $\boxed{EY_i = \beta_0 + \beta_1 x_i}$ $i = 1, \dots, n$

(2) jsou **homoskedastické** náhodné veličiny tj. mají všechny stejný rozptyl $\boxed{DY_i = \sigma^2}$ $i = 1, \dots, n$

kde

$x_1, \dots, x_n$ jsou **známé** konstanty, z nichž alespoň dvě jsou různé,
$\beta_0, \beta_1 \in R$ jsou **neznámé parametry**

Uvedený model $(M1)$ nazveme MODELEM LINEÁRNÍ REGRESE (s regresní přímkou).

Tento model se často vyskytuje v praxi, kdy mezi (nenáhodnými) veličinami x a y existuje lineární závislost $\boxed{y = \beta_0 + \beta_1 x}$,

- jejíž parametry však neznáme
- a informaci o nich získáváme jen **experimentálně**, tj. tak, že pro zvolené hodnoty $\boxed{x_i}$ naměříme odpovídající hodnoty $\boxed{y_i}$ zatížené **chybou měření** $\boxed{\varepsilon_i}$

Naměřené veličiny jsou tedy rovny $\boxed{Y_i = y_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \dots, n}$.

Jsou-li chyby $\boxed{\varepsilon_i}$ **nezávislé náhodné bez systematické složky**, což vyjádříme požadavkem $E\varepsilon_i = 0$ **měřené stejně přesně** $D\varepsilon_i = \sigma^2$

pak dospějeme k uvedenému modelu.

10.3. Odhady neznámých parametrů pomocí metody nejmenších čtverců.

Metodou, která se nejčastěji používá k získání **bodových odhadů neznámých parametrů**, je tzv. METODA NEJMENŠÍCH ČTVERCŮ, která spočívá v proložení dat (x_i, Y_i) křivkou tak, aby součet čtverců odchylek byl minimální. Pokud body prokládáme přímku, nazveme ji REGRESNÍ PŘÍMKOU.

DEFINICE 10.2. Náhodné veličiny $\widehat{\beta_0}$ a $\widehat{\beta_1}$, které pro daná $Y_1, \ldots, Y_n$ **minimalizují** součet čtverců

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2,$$

nazýváme odhady parametrů β_0, β_1 **metodou nejmenších čtverců**.

V následujících dvou větách ještě nebudeme činit žádný předpoklad o typu rozdělení náhodných veličin $Y_i - EY_i$, nemusejí být ani stejně rozdělené.

Ještě dříve než vyslovíme první větu, zaveďme následující značení

$$\boxed{\bar{Y} = \tfrac{1}{n} \sum_{i=1}^{n} Y_i} \qquad \text{a} \qquad \boxed{\bar{x} = \tfrac{1}{n} \sum_{i=1}^{n} x_i}$$

a dále

$$\boxed{S_{XX}} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 > 0 \quad \text{(neboť alespoň dvě } x_i \text{ jsou různá)}$$

$$\boxed{S_{XY}} = \sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}$$

$$\boxed{S_{YY}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$$

VĚTA 10.3. *V modelu (M_1 mají odhady neznámých parametrů β_0 a β_1 pomocí metody nejmenších čtverců následující tvar*

$$\boxed{\widehat{\beta_1}} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \boxed{\frac{S_{XY}}{S_{XX}}} \quad a \quad \boxed{\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{x}},$$

přičemž ***reziduální součet čtverců*** *nabývá hodnoty*

$$S_e^2 = S(\widehat{\beta_0}, \widehat{\beta_1}) = S_{YY} - \frac{S_{XY}^2}{S_{XX}}.$$

Důkaz. Odhady $\widehat{\beta_0}$ a $\widehat{\beta_1}$ musí nutně vyhovovat soustavě rovnic

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \qquad \text{a} \qquad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Provedeme-li uvedené derivace, dostaneme

$$\left.\begin{aligned} -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) &= 0 \\ -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i) x_i &= 0 \end{aligned}\right\} \Rightarrow \boxed{\begin{aligned} \sum_{i=1}^{n} Y_i &= n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} Y_i x_i &= \beta_0 + \beta_1 \sum_{i=1}^{n} x_i^2 \end{aligned}} \quad \text{tzv. SYSTÉM NORMÁLNÍCH ROVNIC}$$

Vzhledem k předpokladu, že alespoň dvě hodnoty x_i jsou od sebe různé, pak determinant soustavy rovnic

$$\begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} = n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 = n\sum_{i=1}^n (x_i-\bar{x})^2 = nS_{XX} > 0,$$

takže tato soustava má právě jedno řešení, které označíme $\widehat{\beta}_0$ a $\widehat{\beta}_1$. S využitím notace pomocí $\bar{x}$ a $\bar{Y}$ lze **systém normálních rovnic** napsat jako

$$n\widehat{\beta}_0 + n\widehat{\beta}_1\bar{x} = n\bar{Y}$$
$$n\widehat{\beta}_0\bar{x} + \widehat{\beta}_1\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_iY_i$$

Z první rovnice okamžitě dostaneme, že $\boxed{\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1\bar{x}}$.

Pokud první rovnici vynásobíme výrazem $-\bar{x}$ a obě rovnice sečteme, máme

$$\widehat{\beta}_1 \underbrace{\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right]}_{S_{XX}} = \underbrace{\sum_{i=1}^n x_iY_i - n\bar{x}\bar{Y}}_{S_{XY}} \quad \Rightarrow \quad \boxed{\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}}.$$

Nutnou podmínku k existenci minima jsme již splnili. Nyní bude třeba dokázat, že jde skutečně o minimum, tj. že pro libovolné $\beta_0, \beta_1 \in \mathbb{R}$ platí $S(\widehat{\beta}_0, \widehat{\beta}_1) \le S(\beta_0, \beta_1)$.

Připomeňme, že

$$S_{XX} = \sum_{i=1}^n (x_i-\bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \quad \Rightarrow \quad \boxed{\sum_{i=1}^n x_i^2 = S_{XX} + n\bar{x}^2}$$

a upravujme

$$\begin{aligned}
S(\beta_0,\beta_1) &= \sum_{i=1}^n (Y_i-\beta_0-\beta_1x_i)^2 = \sum_{i=1}^n \left[(Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i) - (\beta_0-\widehat{\beta}_0) - (\beta_1-\widehat{\beta}_1)x_i\right]^2 \\
&= \sum_{i=1}^n (Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)^2 + n(\beta_0-\widehat{\beta}_0)^2 + (\beta_0-\widehat{\beta}_0)^2\sum_{i=1}^n x_i^2 \\
&\quad - 2(\beta_0-\widehat{\beta}_0)\underbrace{\sum_{i=1}^n (Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)}_{-\frac{1}{2}\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_0}=0} - 2(\beta_0-\widehat{\beta}_0)\underbrace{\sum_{i=1}^n x_i(Y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i)}_{-\frac{1}{2}\frac{\partial S(\beta_0,\beta_1)}{\partial\beta_1}=0} \\
&\quad + 2(\beta_0-\widehat{\beta}_0)(\beta_1-\widehat{\beta}_1)\sum_{i=1}^n x_i \\
&= S(\widehat{\beta}_0,\widehat{\beta}_1) + n(\beta_0-\widehat{\beta}_0)^2 + (\beta_1-\widehat{\beta}_1)^2\left[S_{XX}+n\bar{x}^2\right] + 2(\beta_0-\widehat{\beta}_0)(\beta_1-\widehat{\beta}_1)n\bar{x} \\
&= S(\widehat{\beta}_0,\widehat{\beta}_1) + \underbrace{n(\beta_0-\widehat{\beta}_0)^2}_{*} + (\beta_1-\widehat{\beta}_1)^2S_{XX} + \underbrace{n(\beta_1-\widehat{\beta}_1)^2\bar{x}^2}_{*} + \underbrace{2(\beta_0-\widehat{\beta}_0)(\beta_1-\widehat{\beta}_1)n\bar{x}}_{*} \\
&= \underbrace{S(\widehat{\beta}_0,\widehat{\beta}_1)}_{=S_e^2} + \underbrace{(\beta_1-\widehat{\beta}_1)^2S_{XX}}_{=S_1^2\ge 0} + \underbrace{n\left[(\beta_0-\widehat{\beta}_0)+(\beta_1-\widehat{\beta}_1)\bar{x}\right]^2}_{=S_0^2\ge 0}. \qquad (18)
\end{aligned}$$

Takže pro libovolné $\beta_0, \beta_1 \in \mathbb{R}$ skutečně dostáváme, že

$$S(\beta_0, \beta_1) \geq S(\widehat{\beta}_0, \widehat{\beta}_1)$$

což znamená, že $\widehat{\beta}_0, \widehat{\beta}_1$ jsou odhady parametrů β_0, β_1 metodou nejmenších čtverců.

Ještě než dopočítáme reziduální součet čtverců, označme

$$\boxed{\widehat{Y}_i} = \widehat{\beta}_0 - \widehat{\beta}_1 x_i = \bar{Y} - \tfrac{S_{XY}}{S_{XX}}\bar{x} + \tfrac{S_{XY}}{S_{XX}} x_i = \boxed{\bar{Y} + \tfrac{S_{XY}}{S_{XX}}(x_i - \bar{x})}$$

a počítejme

$$\begin{aligned} S(\widehat{\beta}_0, \widehat{\beta}_1) &= \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} \left[Y_i - \bar{Y} - \tfrac{S_{XY}}{S_{XX}}(x_i - \bar{x}) \right]^2 \\ &= \sum_{i=1}^{n} \left[(Y_i - \bar{Y})^2 - 2\tfrac{S_{XY}}{S_{XX}}(x_i - \bar{x})(Y_i - \bar{Y}) + \tfrac{S_{XY}^2}{S_{XX}^2}(x_i - \bar{x})^2 \right] \\ &= S_{YY} - 2\frac{S_{XY}}{S_{XX}} S_{XY} + \frac{S_{XY}^2}{S_{XX}^2} S_{XX} = \boxed{S_{YY} - \frac{S_{XY}^2}{S_{XX}}} = \frac{S_{YY} S_{XX} - S_{XY}^2}{S_{XX}} \end{aligned}$$

□

Naším dalším úkolem bude

- popsat vlastnosti odhadů $\widehat{\beta}_0$ a $\widehat{\beta}_0$ získaných pomocí metody nejmenších čtverců
- a najít odhad neznámého parametru σ^2.

Pro tyto účely budou velmi výhodné následující **transformace**:

(I) **Centrování**: $\mathbf{V} = \mathbf{Y} - \boldsymbol{\mu}$ pomocí $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, kde $EY_i = \mu_i = \beta_0 + \beta_1 x_i$ pro $i = 1, \ldots, n$, takže platí

(a) $EV_i = 0 \Rightarrow E\mathbf{V} = \mathbf{0}$

(b) $DV_i = D(Y_i - \beta_0 - \beta_1 x_i) = DY_i = \sigma^2$

(c) $C(V_i, V_j) = C(Y_i, Y_j) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$, což plyne z nezávislosti $Y_1, \ldots, Y_n$.

(II) **Ortogonalizace**: $\mathbf{Z} = \mathbf{B}\mathbf{V} = \mathbf{B}(\mathbf{Y} - \boldsymbol{\mu})$ přičemž $\mathbf{B}$ je **ortonormální matice** tvaru

$$\mathbf{B} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{x_1 - \bar{x}}{\sqrt{S_{XX}}} & \frac{x_2 - \bar{x}}{\sqrt{S_{XX}}} & \frac{x_3 - \bar{x}}{\sqrt{S_{XX}}} & \cdots & \frac{x_n - \bar{x}}{\sqrt{S_{XX}}} \\ b_{31} & b_{32} & b_{33} & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \cdots & b_{nn} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1' \\ \mathbf{b}_2' \\ \mathbf{b}_3' \\ \vdots \\ \mathbf{b}_n' \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_n \end{pmatrix},$$

přičemž $\mathbf{b}_j'\mathbf{b}_k = \begin{cases} 1 & j = k, \\ 0 & j \neq k \end{cases}$ $\mathbf{s}_j'\mathbf{s}_k = \begin{cases} 1 & j = k, \\ 0 & j \neq k. \end{cases}$ takže celkově platí $\boxed{\mathbf{B}\mathbf{B}' = \mathbf{B}'\mathbf{B} = \mathbf{I}_n}$.

Zkoumejme vlastnosti této transformace:

(1) $\sum_{i=1}^{n} Z_i^2 = \mathbf{Z}'\mathbf{Z} = (\mathbf{Y}-\boldsymbol{\mu})'\mathbf{B}'\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu}) = \sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 x_i)^2 = S(\beta_0,\beta_1)$

$= S(\widehat{\beta}_0,\widehat{\beta}_1) + n\left[(\beta_0-\widehat{\beta}_0)+(\beta_1-\widehat{\beta}_1)\bar{x}\right]^2 + (\beta_1-\widehat{\beta}_1)^2 S_{XX}$

$= S(\widehat{\beta}_0,\widehat{\beta}_1) + S_0^2 + S_1^2$

(2) $Z_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n'(\mathbf{Y}-\boldsymbol{\mu}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 x_i) = \frac{1}{\sqrt{n}}(n\bar{Y}-n\beta_0-n\beta_1\bar{x})$

$= \sqrt{n}(\underbrace{\bar{Y}-\widehat{\beta}_1\bar{x}}_{=\widehat{\beta}_0} - \beta_0 + \widehat{\beta}_1\bar{x} - \beta_1\bar{x}) = \sqrt{n}\left[(\widehat{\beta}_0-\beta_0)+(\widehat{\beta}_1-\beta_1)\bar{x}\right]$

$\Rightarrow \quad Z_1^2 = S_0^2$

(3) $Z_2 = \mathbf{b}_2'(\mathbf{Y}-\boldsymbol{\mu}) = \frac{1}{\sqrt{S_{XX}}}\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\beta_0-\beta_1 x_i)$

$= \frac{1}{\sqrt{S_{XX}}}\sum_{i=1}^{n}[Y_i(x_i-\bar{x}) - \beta_0(x_i-\bar{x}) - \beta_1(x_i-\bar{x})x_i]$

$= \frac{1}{\sqrt{S_{XX}}}\underbrace{\left(\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}\right)}_{=S_{XY}} - \frac{\beta_0}{\sqrt{S_{XX}}}\underbrace{\sum_{i=1}^{n}(x_i-\bar{x})}_{=0} - \frac{\beta_1}{\sqrt{S_{XX}}}\underbrace{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}_{=S_{XX}}$

$= \underbrace{\frac{S_{XY}}{S_{XX}}}_{=\widehat{\beta}_1}\sqrt{S_{XX}} - \beta_1\sqrt{S_{XX}} = (\widehat{\beta}_1-\beta_1)\sqrt{S_{XX}}$

$\Rightarrow \quad Z_2^2 = S_1^2$

(4) $\sum_{i=3}^{n} Z_i^2 = S(\widehat{\beta}_0,\widehat{\beta}_1)$ neboť $S(\beta_0,\beta_1) = \sum_{i=1}^{n} Z_i^2 = S(\widehat{\beta}_0,\widehat{\beta}_1) + S_0^2 + S_1^2$

(5) $EZ_j = E\sum_{i=1}^{n} b_{ji}(Y_i-\mu_i) = E\sum_{i=1}^{n} b_{ji}V_i = \sum_{i=1}^{n} b_{ji}\underbrace{EV_i}_{=0} = 0$

$DZ_j = EZ_j^2 = D\sum_{i=1}^{n} b_{ji}V_i \overset{nez.}{=} \sum_{i=1}^{n} b_{ji}^2 DV_i = \sigma^2\underbrace{\sum_{i=1}^{n} b_{ji}^2}_{=1} = \sigma^2$

pro $l \neq k$

$C(Z_l,Z_k) = C\left(\sum_{i=1}^{n} b_{li}V_i, \sum_{j=1}^{n} b_{kj}V_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{li}b_{kj}C(V_i,V_j)$

$= \sum_{i=1}^{n} b_{li}b_{ki}\underbrace{C(V_i,V_i)}_{=\sigma^2} = \sigma^2 \underbrace{\mathbf{b}_l'\mathbf{b}_k}_{=0 \text{ pro } l\neq k} = 0$

Předchozích poznatků nyní využijeme ve větě:

VĚTA 10.4. *V modelu* $(M1)$ *platí*
(1) Odhady $\boxed{\widehat{\beta}_0}$ *a* $\boxed{\widehat{\beta}_1}$ *jsou* ***nestrannými*** *odhady parametrů* β_0 *a* β_1.

(2) Statistika $\boxed{S^2_{M_1} = \dfrac{S_e^2}{n-2}}$ *je* ***nestranným*** *odhadem parametru* $\boxed{\sigma^2}$.

(3) Veličina $\boxed{\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x}$ *je* ***nestranným*** *odhadem veličiny* $\boxed{y = \beta_0 + \beta_1 x}$ *pro* $\forall x \in \mathbb{R}$.

Důkaz.

(1) Počítejme postupně

$$\boxed{E\bar{Y}} = E\left(\tfrac{1}{n}\sum_{i=1}^n Y_i\right) = \tfrac{1}{n}\sum_{i=1}^n EY_i = \tfrac{1}{n}\sum_{i=1}^n E(\beta_0+\beta_1 x_i) = \beta_0 + \beta_1 \tfrac{1}{n}\sum_{i=1}^n x_i = \boxed{\beta_0 + \beta_1 \bar{x}}$$

$$\begin{aligned}\boxed{E\widehat{\beta}_1} &= E\left(\tfrac{S_{XY}}{S_{XX}}\right) = \tfrac{1}{S_{XX}} E\left(\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})\right) = \tfrac{1}{S_{XX}}\sum_{i=1}^n (x_i-\bar{x}) E(Y_i - \bar{Y}) \\ &= \tfrac{1}{S_{XX}}\sum_{i=1}^n (x_i-\bar{x})(EY_i - E\bar{Y}) = \tfrac{1}{S_{XX}}\sum_{i=1}^n (x_i-\bar{x})(\beta_0+\beta_1 x_i - \beta_0 - \beta_1\bar{x}) \\ &= \tfrac{1}{S_{XX}}\beta_1 \underbrace{\sum_{i=1}^n (x_i-\bar{x})^2}_{=S_{XX}} = \boxed{\beta_1}\end{aligned}$$

$$\boxed{E\widehat{\beta}_0} = E(\bar{Y} - \widehat{\beta}_1\bar{x}) = E(\beta_0 + \underbrace{\beta_1}_{=E\widehat{\beta}_1}\bar{x} - \widehat{\beta}_1\bar{x}) = \beta_0 + E\widehat{\beta}_1\bar{x} - E\widehat{\beta}_1\bar{x} = \boxed{\beta_0}$$

(2) Dále počítejme

$$\boxed{ES^2_{M_1}} = E\left(\frac{S_e^2}{n-2}\right) = \frac{1}{n-2}ES_e^2 = \frac{1}{n-2}\sum_{i=3}^n \underbrace{EY_i^2}_{=\sigma^2} = \frac{1}{n-2}(n-2)\sigma^2 = \boxed{\sigma^2}$$

(3) Z nestrannosti $\widehat{\beta}_0$ a $\widehat{\beta}_1$ plyne

$$\boxed{E\widehat{Y}} = E(\widehat{\beta}_0 + \widehat{\beta}_1 x) = \beta_0 + \beta_1 x = \boxed{y}.$$

□

VĚTA 10.5. *Nechť v modelu* $(M1)$ *pro* $i = 1, \dots, n$ *platí, že náhodné veličiny* $\boxed{Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)}$. *Pak*

(1) Odhad parametru $\boxed{\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)}$.

(2) Odhad parametru $\boxed{\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right)}$.

(3) Odhad pro $y = \beta_0 + \beta_1 x$ $\boxed{\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x \sim N\left(\beta_0 + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)\right)}$.

(4) Náhodný vektor $\begin{pmatrix}\widehat{\beta}_0 \\ \widehat{\beta}_1\end{pmatrix}$ *a statistika* $\boxed{K = \frac{(n-2)S_{M1}^2}{\sigma^2}}$ *jsou* ***nezávislé****.*

(5) Statistika $\boxed{K \sim \chi^2(n-2)}$.

Důkaz. Pokud předpokládáme, že pro $i = 1, \dots, n$ mají náhodné veličiny Y_i normální rozdělení

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2),$$

pak

$$V_i = Y_i - \beta_0 - \beta_1 x_i \sim N(0, \sigma^2)$$

a také

$$Z_i = \mathbf{b}_i' \mathbf{V} = \sum_{k=1}^{n} b_{ik} V_i \sim N(0, \sigma^2 \underbrace{\mathbf{b}_i' \mathbf{b}_i}_{=1}).$$

Navíc vzhledem k tomu, že Z_i jsou normální náhodné veličiny, pak z nekorelovanosti plyne také nezávislost.

(1) Protože $\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ a statistika $Z_2 = (\widehat{\beta}_1 - \beta_1)\sqrt{S_{XX}}$, pak odhad $\widehat{\beta}_1$ lze vyjádřit pomocí Z_2 takto

$$\widehat{\beta}_1 = \frac{Z_2}{\sqrt{S_{XX}}} + \beta_1 \sim N(\beta_1, \sigma^2 S_{XX}^{-1}).$$

(2) Protože

$$Z_1 = \sqrt{n}\left[(\widehat{\beta}_0 - \beta_0) + (\widehat{\beta}_1 - \beta_1)\bar{x}\right]$$

a

$$\widehat{\beta}_1 - \beta_1 = \frac{Z_2}{\sqrt{S_{XX}}},$$

pak

$$\widehat{\beta}_0 = \tfrac{Z_1}{\sqrt{n}} - \tfrac{Z_2}{\sqrt{S_{XX}}}\,\bar{x} + \beta_0 \sim N\left(\beta_0, \sigma^2\left(\tfrac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right)$$

(3) Počítejme postupně

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = \beta_0 + \frac{Z_1}{\sqrt{n}} - \frac{Z_2}{\sqrt{S_{XX}}}\,\bar{x} + \left(\frac{Z_2}{\sqrt{S_{XX}}} + \beta_1\right) x$$

$$= \beta_0 + \beta_1 x + \frac{Z_1}{\sqrt{n}} + \frac{Z_2}{\sqrt{S_{XX}}}(x - \bar{x}) \sim N\left(\beta_0 + \beta_1 x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}\right)\right)$$

(4) Protože $\widehat{\beta}_0$ a $\widehat{\beta}_1$ závisí pouze na Z_1 a Z_2, kdežto $S_e^2 = \sum_{i=3}^{n} Z_i^2$ a $Z_1, \ldots, Z_n$ jsou nezávislé, pak také statistika

$$K = \frac{(n-2)S_{M1}^2}{\sigma^2} = \frac{S_e^2}{\sigma^2} \quad \text{a náhodný vektor} \quad \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}$$

jsou nezávislé.

(5) Protože

$$\frac{Z_i}{\sigma} \quad \sim \quad N(0,1),$$

pak

$$K = \frac{(n-2)S_{M1}^2}{\sigma^2} = \frac{S_e^2}{\sigma^2} = \sum_{i=3}^{n} \left(\frac{Z_i}{\sigma}\right)^2 \quad \sim \quad \chi^2(n-2).$$

□

DŮSLEDEK 10.6. *Nechť v modelu* $(M1)$ *pro* $i = 1, \ldots, n$ *platí, že náhodné veličiny* $\boxed{Y_i \quad \sim \quad N(\beta_0 + \beta_1 x_i, \sigma^2)}$. *Pak platí*

(1) Statistika

$$T_1 = \frac{\widehat{\beta}_0 - \beta_0}{S_{M1}\sqrt{\frac{1}{n} + \frac{\bar{x}}{S_{XX}}}} \quad \sim \quad t(n-2).$$

(2) Statistika

$$T_2 = \frac{\widehat{\beta}_1 - \beta_1}{S_{M1}}\sqrt{S_{XX}} \quad \sim \quad t(n-2).$$

(3) Statistika

$$T_3 = \frac{\widehat{Y} - (\beta_0 + \beta_1 x)}{S_{M1}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}}} \quad \sim \quad t(n-2).$$

Důkaz. Postupně dokazujme jednotlivá tvrzení:

(1) Víme, že v modelu $(M1)$ má LS-odhad parametru β_0 normální rozdělení

$$\widehat{\beta}_0 \quad \sim \quad N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)\right).$$

Po provedení standardizace dostaneme

$$U_{\widehat{\beta}_0} = \frac{\widehat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \quad \sim \quad N(0,1).$$

Se statistikou $U_{\widehat{\beta}_0}$ je nezávislá statistika

$$K = \frac{(n-2)S_{M1}^2}{\sigma^2} \quad \sim \quad \chi^2(n-2).$$

Protože platí, že

$$\frac{U_{\widehat{\beta}_0}}{\sqrt{\frac{K}{n-2}}} \quad \sim \quad t(n-2),$$

pak po dosazení a úpravách dostaneme

$$\frac{U_{\widehat{\beta}_0}}{\sqrt{\frac{K}{n-2}}} = \frac{\frac{\widehat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}}}}{\sqrt{\frac{\frac{(n-2)S^2_{M1}}{\sigma^2}}{n-2}}} = \frac{\widehat{\beta}_0-\beta_0}{S_{M1}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}}} = T_1 \quad \sim \quad t(n-2).$$

(2) Při důkazu druhého tvrzení budeme postupovat zcela analogicky jako v předchozím případě:

$$\widehat{\beta}_1 \quad \sim \quad N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \quad \Rightarrow \quad U_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1-\beta_1}{\sigma\sqrt{\frac{1}{S_{XX}}}} \quad \sim \quad N(0,1).$$

Dále

$$U_{\widehat{\beta}_1} \perp K \quad \Rightarrow \quad \boxed{T_2} = \frac{U_{\widehat{\beta}_1}}{\sqrt{\frac{K}{n-2}}} = \frac{\frac{\widehat{\beta}_1-\beta_1}{\sigma\sqrt{\frac{1}{S_{XX}}}}}{\sqrt{\frac{\frac{(n-2)S^2_{M1}}{\sigma^2}}{n-2}}} = \boxed{\frac{\widehat{\beta}_1-\beta_1}{S_{M1}}\sqrt{S_{XX}}} \quad \sim \quad t(n-2).$$

(3) Postupujme opět analogicky jako v předchozích dvou případech

$$\widehat{Y} = \widehat{\beta}_0+\widehat{\beta}_1 x \sim N\left(\beta_0+\beta_1 x, \sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)\right) \quad \Rightarrow \quad U_{\widehat{Y}} = \frac{\widehat{Y}-(\beta_0+\beta_1 x)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}}} \sim N(0,1).$$

Dále

$$U_{\widehat{Y}} \perp K \quad \Rightarrow \quad \boxed{T_3} = \frac{U_{\widehat{Y}}}{\sqrt{\frac{K}{n-2}}} = \frac{\frac{\widehat{Y}-(\beta_0+\beta_1 x)}{\sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}}}}{\sqrt{\frac{\frac{(n-2)S^2_{M1}}{\sigma^2}}{n-2}}} = \boxed{\frac{\widehat{Y}-(\beta_0+\beta_1 x)}{S_{M1}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}}}} \quad \sim \quad t(n-2).$$

□

10.4. Intervalové odhady a testy hypotéz v regresním modelu. V předchozím odstavci jsme nečinili žádný předpoklad o typu rozdělení náhodných veličin Y_i (resp. ε_i) pro $i = 1, \ldots, n$.

Abychom mohli konstruovat intervalové odhady a provádět testy hypotéz, musíme připojit předpoklad o typu rozdělení, a to předpoklad normálního rozdělení.

Důsledek 10.7. *Nechť v modelu $(M1)$ pro $i = 1, \ldots, n$ platí, že náhodné veličiny* $\boxed{Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)}$. *Pak intervalový odhad (se spolehlivostí $1-\alpha$)*

(1) pro β_0 je tvaru

$$\left(\widehat{\beta}_0 - S_{M1}\sqrt{\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta}_0 + S_{M1}\sqrt{\tfrac{1}{n} + \tfrac{\bar{x}^2}{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2)\right).$$

(2) pro β_1 je tvaru

$$\left(\widehat{\beta}_1 - \tfrac{S_{M1}}{\sqrt{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta}_1 + \tfrac{S_{M1}}{\sqrt{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2)\right).$$

(3) pro $y = \beta_0 + \beta_1 x$ je tvaru

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 x - S_{M1}\sqrt{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta}_0 + \widehat{\beta}_1 x + S_{M1}\sqrt{\tfrac{1}{n} + \tfrac{(x-\bar{x})^2}{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2)\right).$$

(4) pro σ^2 je tvaru

$$\left(\frac{(n-2)S_{M1}^2}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}, \frac{(n-2)S_{M1}^2}{\chi^2_{\frac{\alpha}{2}}(n-2)}\right).$$

Důkaz. Při dokazování prvních tří tvrzení použijeme pivotové statistiky T_j $(j = 1, 2, 3)$ uvedené v předchozím důsledku, tj. vyjdeme ze vztahu

$$1 - \alpha = P\left(-t_{1-\frac{\alpha}{2}}(n-2) \leq T_j \leq t_{1-\frac{\alpha}{2}}(n-2)\right)$$

a pomocí jednoduchých úprav dostaneme první tři tvrzení.

Pro důkaz čtvrtého tvrzení využijeme pivotovou statistiku $K = \frac{(n-2)S_{M1}^2}{\sigma^2} \sim \chi^2(n-2)$, tj.

$$1 - \alpha = P\left(\chi^2_{\frac{\alpha}{2}}(n-2) \leq K \leq \chi^2_{1-\frac{\alpha}{2}}(n-2)\right)$$

a po jednoduchých úpravách dojdeme k poslednímu tvrzení. □

Všimněme si nyní testování hypotéz v regresním modelu $(M1)$. Testy lze obecně sestavit např. metodou podílu věrohodností. V následující tabulce je popíšeme pomocí kritických oblastí W_α.

H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{Y} \in W_\alpha$, tj.
$\beta_0 = 0$	$\beta_0 \neq 0$	$\lvert\widehat{\beta}_0\rvert / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}} \geq S_{M1}t_{1-\frac{\alpha}{2}}(n-2)$
$\beta_0 = 0$	$\beta_0 > 0$	$\widehat{\beta}_0 / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}} \geq S_{M1}t_{1-\alpha}(n-2)$
$\beta_0 = 0$	$\beta_0 < 0$	$\widehat{\beta}_0 / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}} \leq -S_{M1}t_{1-\alpha}(n-2)$
$\beta_1 = 0$	$\beta_1 \neq 0$	$\lvert\widehat{\beta}_1\rvert\sqrt{S_{XX}} \geq S_{M1}t_{1-\frac{\alpha}{2}}(n-2)$
$\beta_1 = 0$	$\beta_1 > 0$	$\widehat{\beta}_1\sqrt{S_{XX}} \geq S_{M1}t_{1-\alpha}(n-2)$
$\beta_1 = 0$	$\beta_1 < 0$	$\widehat{\beta}_1\sqrt{S_{XX}} \leq -S_{M1}t_{1-\alpha}(n-2)$

10.5. Některé speciální případy regresních modelů.

10.5.1. *Regresní přímka procházející počátkem.* Pokud vztah mezi veličinami x a y je vztahem přímé úměrnosti, pak v regresním modelu $(M1)$ klademe

$$\beta_0 = 0$$

a body (x_i, Y_i) prokládáme regresní přímkou procházející počátkem. Označme nejprve

$$S^*_{XX} = \sum_{i=1}^n x_i^2 \qquad S^*_{XY} = \sum_{i=1}^n x_i Y_i \qquad S^*_{YY} = \sum_{i=1}^n Y_i^2.$$

Odhad parametru β_1 pomocí metody nejmenších čtverců vypočteme, když nejprve položíme první derivaci funkce

$$S(\beta_1) = \sum_{i=1}^n (Y_i - \beta_1 x_i)^2$$

rovnu nule, tj.

$$-2\sum_{i=1}^n (Y_i - \beta_1 x_i) x_i = 0$$

a odtud pak

$$\boxed{\widehat{\beta}_1} = \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2} = \boxed{\frac{S^*_{XY}}{S^*_{XX}}}.$$

Přesvědčíme se, že jde o minimum:

$$\begin{aligned} S(\beta_1) &= \sum_{i=1}^n (Y_i - \beta_1 x_i)^2 = \sum_{i=1}^n \Big[(Y_i - \widehat{\beta}_1 x_i) - (\beta_1 - \widehat{\beta}_1) x_i \Big]^2 \\ &= \underbrace{\sum_{i=1}^n (Y_i - \widehat{\beta}_1 x_i)^2}_{S(\widehat{\beta}_1)} - 2(\beta_1 - \widehat{\beta}_1) \underbrace{\sum_{i=1}^n (Y_i - \widehat{\beta}_1 x_i) x_i}_{-\frac{1}{2}\frac{dS(\beta_1)}{d\beta_1} = 0} + (\beta_1 - \widehat{\beta}_1)^2 \sum_{i=1}^n x_i^2 \\ &= S(\widehat{\beta}_1) + \underbrace{(\beta_1 - \widehat{\beta}_1)^2 \sum_{i=1}^n x_i^2}_{\geq 0} = S(\widehat{\beta}_1) + \underbrace{(\beta_1 - \widehat{\beta}_1)^2 S^*_{XX}}_{S_1^2} \end{aligned}$$

takže pro libovolné $\beta_1 \in \mathbb{R}$ platí $\boxed{S(\widehat{\beta}_1) \leq S(\beta_1)}$. Nyní explicitně vyjádřeme $S(\widehat{\beta}_1)$:

$$\begin{aligned} \boxed{S(\widehat{\beta}_1)} &= \sum_{i=1}^n (Y_i - \widehat{\beta}_1 x_i)^2 = \sum_{i=1}^n \left(Y_i - \frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2} x_i \right)^2 \\ &= \sum_{i=1}^n Y_i^2 - 2\frac{\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n x_i^2} \sum_{i=1}^n Y_i x_i + \frac{(\sum_{i=1}^n Y_i x_i)^2}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \\ &= \sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i x_i)^2}{\sum_{i=1}^n x_i^2} = \boxed{S^*_{YY} - \frac{S^{*2}_{XY}}{S^*_{XX}}} \end{aligned}$$

Abychom mohli odvodit vlastnosti odhadů opět použijeme transformaci vektoru $\mathbf{Y}$, a to ortogonalizaci $\mathbf{Z} = \mathbf{BV} = \mathbf{B}(\mathbf{Y} - \boldsymbol{\mu})$ přičemž $\mathbf{B}$ je **ortonormální matice** tvaru

$$\mathbf{B} = \begin{pmatrix} \frac{x_1}{\sqrt{S^*_{XX}}} & \frac{x_2}{\sqrt{S^*_{XX}}} & \frac{x_3}{\sqrt{S^*_{XX}}} & \cdots & \frac{x_n}{\sqrt{S^*_{XX}}} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & b_{n3} & \cdots & b_{nn} \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1' \\ \mathbf{b}_2' \\ \vdots \\ \mathbf{b}_n' \end{pmatrix},$$

pričemž $\mathbf{b}_j'\mathbf{b}_k = \begin{cases} 1 & j=k, \\ 0 & j \neq k \end{cases}$ takže celkově platí $\boxed{\mathbf{B}\mathbf{B}' = \mathbf{B}'\mathbf{B} = \mathbf{I}_n}$

a $\mathbf{V} = \mathbf{Y} - \boldsymbol{\mu}$ pomocí $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, kde $EY_i = \mu_i = \beta_1 x_i$ pro $i = 1, \ldots, n$.

Postupně spočítejme

(a) $EV_i = 0 \Rightarrow E\mathbf{V} = \mathbf{0}$

(b) $DV_i = D(Y_i - \beta_1 x_i) = DY_i = \sigma^2$

(c) $C(V_i, V_j) = C(Y_i, Y_j) = \begin{cases} \sigma^2 & i=j \\ 0 & i \neq j \end{cases}$, což plyne z nezávislosti $Y_1, \ldots, Y_n$.

(d) $EZ_j = E\sum_{i=1}^n b_{ji}(Y_i - \mu_i) = E\sum_{i=1}^n b_{ji}V_i = \sum_{i=1}^n b_{ji}\underbrace{EV_i}_{=0} = 0$

$$DZ_j = EZ_j^2 = D\sum_{i=1}^n b_{ji}V_i \overset{nez.}{=} \sum_{i=1}^n b_{ji}^2 DV_i = \sigma^2 \underbrace{\sum_{i=1}^n b_{ji}^2}_{=1} = \sigma^2$$

pro $l \neq k$

$$\begin{aligned} C(Z_l, Z_k) &= C\left(\sum_{i=1}^n b_{li}V_i, \sum_{j=1}^n b_{kj}V_j\right) = \sum_{i=1}^n\sum_{j=1}^n b_{li}b_{kj}C(V_i, V_j) \\ &= \sum_{i=1}^n b_{li}b_{ki}\underbrace{C(V_i, V_i)}_{=\sigma^2} = \sigma^2 \underbrace{\mathbf{b}_l'\mathbf{b}_k}_{=0 \text{ pro } l\neq k} = 0 \end{aligned}$$

(e) Všimněme si, že

$$\begin{aligned} \sum_{i=1}^n Z_i^2 &= \mathbf{Z}'\mathbf{Z} = (\mathbf{Y}-\boldsymbol{\mu})'\mathbf{B}'\mathbf{B}(\mathbf{Y}-\boldsymbol{\mu}) = \sum_{i=1}^n (Y_i - \beta_1 x_i)^2 = S(\beta_1) \\ &= S(\widehat{\beta}_1) + (\beta_1 - \widehat{\beta}_1)^2 S_{XX}^* = S(\widehat{\beta}_0, \widehat{\beta}_1) + S_1^2 \end{aligned}$$

(f) A dále

$$\begin{aligned} Z_1 &= \mathbf{b}_1'(\mathbf{Y}-\boldsymbol{\mu}) = \tfrac{1}{\sqrt{S_{XX}^*}}\sum_{i=1}^n x_i(Y_i - \beta_1 x_i) = \tfrac{1}{\sqrt{S_{XX}^*}}\sum_{i=1}^n x_i Y_i - \tfrac{\beta_1}{\sqrt{S_{XX}^*}}\sum_{i=1}^n x_i^2 \\ &= \underbrace{\frac{S_{XY}^*}{S_{XX}^*}}_{=\widehat{\beta}_1}\sqrt{S_{XX}^*} - \beta_1\sqrt{S_{XX}^*} = (\widehat{\beta}_1 - \beta_1)\sqrt{S_{XX}^*} \qquad \Rightarrow \qquad Z_1^2 = S_1^2 \end{aligned}$$

(g) Nakonec

$$\sum_{i=2}^n Z_i^2 = S(\widehat{\beta}_1) \quad \text{neboť} \quad S(\beta_1) = \sum_{i=1}^n Z_i^2 = \underbrace{S(\widehat{\beta}_1)}_{=S_e^2} + S_1^2$$

Pomocí předchozí transformace snadno spočítáme vlastnosti odhadů, když si uvědomíme, že platí

$$Z_1 = (\widehat{\beta}_1 - \beta_1)\sqrt{S_{XX}^*} \quad \sim \quad \mathcal{L}(0, \sigma^2) \qquad \Rightarrow \qquad \widehat{\beta}_1 = \beta_1 + \frac{Z_1}{\sqrt{S_{XX}^*}} \quad \sim \quad \mathcal{L}\left(\beta_1, \tfrac{\sigma^2}{S_{XX}^*}\right),$$

tj. $\widehat{\beta}_1$ je **nestranným** odhadem parametru $\boxed{\beta_1}$.

Opět ukážeme, že statistika $\boxed{S^2_{M_1} = \frac{S_e^2}{n-1}}$ je **nestranným** odhadem parametru $\boxed{\sigma^2}$.

$$ES^2_{M_1} = E\left(\frac{S_e^2}{n-1}\right) = \frac{1}{n-1}\sum_{i=2}^{n}\underbrace{EZ_i^2}_{=\sigma^2} = \boxed{\sigma^2}$$

Přidáme-li podmínku normality, tj. $Y_i \sim N(\beta_1 x_i, \sigma^2)$ pro $i = 1, \ldots, n$, pak LS-odhad parametru β_1 má normální rozdělení

$$\widehat{\beta}_1 \sim N\left(\beta_1, \tfrac{\sigma^2}{S^*_{XX}}\right) \quad \Rightarrow \quad U_{\widehat{\beta}_1} = \tfrac{\widehat{\beta}_1-\beta_1}{\sigma}\sqrt{S^*_{XX}} \sim N(0,1)$$

a je nezávislý se statistikou

$$K = \tfrac{(n-1)S^2_{M1}}{\sigma^2} \sim \chi^2(n-1)\,.$$

Díky těmto vlastnostem můžeme získat statistiku

$$T = \frac{U_{\widehat{\beta}_1}}{\sqrt{\frac{K}{n-1}}} = \tfrac{\widehat{\beta}_1-\beta_1}{S_{M1}}\sqrt{S^*_{XX}} \sim t(n-1)\,.$$

Na závěr si ještě všimněme TESTOVÁNÍ HYPOTÉZ v regresním modelu s regresní přímkou procházející počátkem. Testy lze obecně opět sestavit např. metodou podílu věrohodností. V následující tabulce je popíšeme pomocí kritických oblastí W_α.

H_0	H_1	Hypotézu H_0 zamítáme, pokud $\mathbf{Y} \in W_\alpha$, tj.
$\beta_1 = 0$	$\beta_1 \neq 0$	$\lvert\widehat{\beta}_1\rvert\sqrt{S_{XX}} \geq S_{M1}t_{1-\frac{\alpha}{2}}(n-1)$
$\beta_1 = 0$	$\beta_1 > 0$	$\widehat{\beta}_1\sqrt{S_{XX}} \geq S_{M1}t_{1-\alpha}(n-1)$
$\beta_1 = 0$	$\beta_1 < 0$	$\widehat{\beta}_1\sqrt{S_{XX}} \leq -S_{M1}t_{1-\alpha}(n-1)$

10.5.2. *Dva nezávislé náhodné výběry.* Nechť $\perp\!\!\!\perp\{X_1, \ldots, X_{n_x}\} \sim N(\mu_X, \sigma_X^2)$ je náhodný výběr rozsahu n_x z normálního rozdělení $N(\mu_X, \sigma_X^2)$, $\bar{X}_{n_x}$ je jeho výběrový průměr a S_X^2 jeho výběrový rozptyl.

Dále nechť $\perp\!\!\!\perp\{Y_1, \ldots, Y_{n_y}\} \sim N(\mu_Y, \sigma_Y^2)$ je náhodný výběr rozsahu n_Y z normálního rozdělení $N(\mu_Y, \sigma_Y^2)$, $\bar{Y}_{n_y}$ je jeho výběrový průměr a S_Y^2 jeho výběrový rozptyl.

Položíme-li $\boxed{n = n_x + n_y}$ a zavedeme-li následující značení

$$\begin{array}{lcllcl} Y_1 & = & X_1 & x_1 & = & 1 \\ & \vdots & & & \vdots & \\ Y_{n_x} & = & X_{n_x} & x_{n_x} & = & 1 \\ Y_{n_x+1} & = & Y_1 & x_{n_x+1} & = & 0 \\ & \vdots & & & \vdots & \\ Y_n & = & Y_{n_y} & x_n & = & 0 \end{array}$$

dostáváme regresní model $(M1)$, ve kterém

$$\boxed{\bar{x}} = \tfrac{1}{n}\sum_{i=1}^{n} x_i = \boxed{\tfrac{n_x}{n_x+n_y}}$$

$$\boxed{\bar{Y}} = \tfrac{1}{n}\sum_{i=1}^{n} Y_i = \boxed{\tfrac{n_x}{n_x+n_y}\bar{X}_{n_x} + \tfrac{n_y}{n_x+n_y}\bar{Y}_{n_y}}$$

$$\begin{aligned} \boxed{S_{XX}} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = n_x - (n_x+n_y)\left(\tfrac{n_x}{n_x+n_y}\right)^2 \\ &= \tfrac{n_x(n_x+n_y)-n_x^2}{n_x+n_y} = \tfrac{n_x(n_x+n_y-n_x)}{n_x+n_y} = \boxed{\tfrac{n_x n_y}{n_x+n_y}} \end{aligned}$$

$$\begin{aligned} \boxed{S_{XY}} &= \sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \\ &= n_x\bar{X}_{n_x} - (n_x+n_y)\tfrac{n_x}{(n_x+n_y)}\left[\tfrac{n_x}{n_x+n_y}\bar{X}_{n_x} + \tfrac{n_y}{n_x+n_y}\bar{Y}_{n_y}\right] \\ &= \frac{n_x\left[n_y\bar{X}_{n_x} - n_y\bar{Y}_{n_y}\right]}{n_x+n_y} = \boxed{\tfrac{n_x n_y}{n_x+n_y}\left(\bar{X}_{n_x} - \bar{Y}_{n_y}\right)} \end{aligned}$$

$$\begin{aligned} \boxed{S_{YY}} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \\ &= \sum_{i=1}^{n} Y_i^2 - (n_x+n_y)\left[\tfrac{n_x}{n_x+n_y}\bar{X}_{n_x} + \tfrac{n_y}{n_x+n_y}\bar{Y}_{n_y}\right]^2 \\ &= \boxed{\sum_{i=1}^{n} Y_i^2 - \tfrac{(n_x\bar{X}_{n_x}+n_y\bar{Y}_{n_y})^2}{n_x+n_y}} \end{aligned}$$

$$\boxed{\widehat{\beta_1}} = \frac{S_{XY}}{S_{XX}} = \frac{\frac{n_x n_y}{n_x+n_y}\left(\bar{X}_{n_x} - \bar{Y}_{n_y}\right)}{\frac{n_x n_y}{n_x+n_y}} = \boxed{\bar{X}_{n_x} - \bar{Y}_{n_y}}$$

$$\boxed{\widehat{\beta_0}} = \bar{Y} - \widehat{\beta}_1 \bar{x} = \tfrac{n_x}{n_x+n_y}\bar{X}_{n_x} + \tfrac{n_y}{n_x+n_y}\bar{Y}_{n_y} - \left(\bar{X}_{n_x} - \bar{Y}_{n_y}\right)\tfrac{n_x n_y}{n_x+n_y} = \boxed{\bar{Y}_{n_y}}$$

$$\begin{aligned}
\boxed{S_e^2} &= S_{YY} - \tfrac{S_{XY}^2}{S_{XX}} = S_{YY} - \widehat{\beta}_1 S_{XY} \\
&= \sum_{i=1}^{n} Y_i^2 - \tfrac{(n_x\bar{X}_{n_x}+n_y\bar{Y}_{n_y})^2}{n_x+n_y} - \tfrac{n_x n_y}{n_x+n_y}\left(\bar{X}_{n_x} - \bar{Y}_{n_y}\right)^2 \\
&= \sum_{i=1}^{n} Y_i^2 - \tfrac{1}{n_x+n_y}\left[n_x^2\bar{X}_{n_x}^2 + 2n_x n_y \bar{X}_{n_x}\bar{Y}_{n_y} + n_y\bar{Y}_{n_y}^2 + n_x n_y \bar{X}_{n_x}^2 - n_x n_y \bar{X}_{n_x}\bar{Y}_{n_y} + n_x n_y \bar{Y}_{n_y}^2\right] \\
&= \sum_{i=1}^{n} Y_i^2 - \tfrac{1}{n_x+n_y}\left[n_x(n_x+n_y)\bar{X}_{n_x}^2 + n_y(n_x+n_y)\bar{Y}_{n_y}^2\right] \\
&= \underbrace{\sum_{i=1}^{n_x} X_i^2 - n_x\bar{X}_{n_x}^2}_{(n_x-1)S_X^2} + \underbrace{\sum_{i=n_x+1}^{n_x+n_y} Y_i^2 - n_y\bar{Y}_{n_y}^2}_{(n_y-1)S_Y^2} = \boxed{(n_x-1)S_X^2 + (n_y-1)S_Y^2}
\end{aligned}$$

$$\boxed{S_{M1}^2} = \frac{S_e^2}{n-2} = \boxed{\frac{(n_x-1)S_X^2 + (n_y-1)S_Y^2}{n_x+n_y-2}}$$

Vzhledem k tomu, že výběrové průměry jsou nestrannými odhady středních hodnot, pak neznámé parametry β_0 a β_1 lze interpretovat takto

$$\begin{aligned}
\beta_0 &= \mu_Y \\
\beta_1 &= \mu_X - \mu_Y
\end{aligned}$$

Na závěr si ještě všimněme, že (oboustranný) interval spolehlivosti, který jsme odvodili pro neznámý parametr β_1

$$\left(\widehat{\beta}_1 - \tfrac{S_{M1}}{\sqrt{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta}_1 + \tfrac{S_{M1}}{\sqrt{S_{XX}}}t_{1-\frac{\alpha}{2}}(n-2)\right)$$

po dosazení má tvar pro β_1 je tvaru

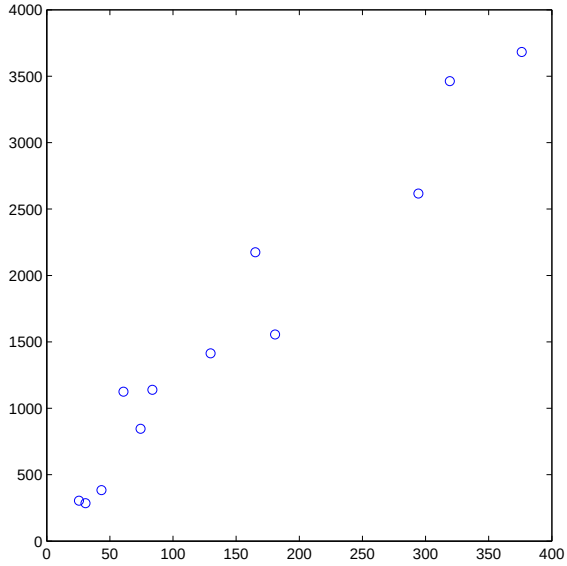
$$\left(\widehat{\beta}_1 - \tfrac{S_{M1}}{\sqrt{\frac{n_x n_y}{n_x+n_y}}}t_{1-\frac{\alpha}{2}}(n-2), \widehat{\beta}_1 + \tfrac{S_{M1}}{\sqrt{\frac{n_x n_y}{n_x+n_y}}}t_{1-\frac{\alpha}{2}}(n-2)\right).$$

a je naprosto shodný s intervalem, který jsme odvodili pro rozdíl středních hodnot dvou nezávislých náhodných výběrů z normálního rozdělení.

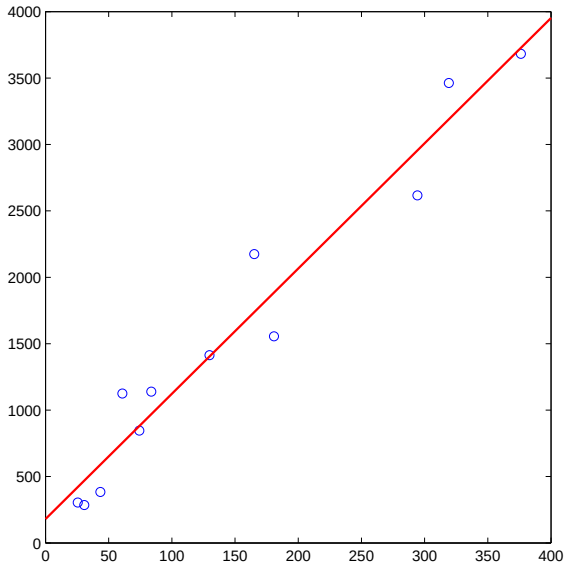
Příklad 10.8. Máme analyzovat data o počtu pracovních hodin za měsíc spojených s provozováním anesteziologické služby v závislosti na velikosti spádové populace nemocnice (viz následující tabulka). Údaje byly získány ve 12 nemocnicích ve Spojených státech.

Poř. číslo	Počet pracovních hodin	Velikost populace spádové oblasti (osoby v tisích)
1	304,37	25,5
2	2616,32	294,3
3	1139,12	83,7
4	285,43	30,7
5	1413,77	129,8
6	1555,68	180,8
7	383,78	43,4
8	2174,27	165,2
9	845,30	74,3
10	1125,28	60,8
11	3462,60	319,2
12	3682,33	376,2

Závislost počtu pracovních hodin na velikosti populace



Graf naznačuje lineární vztah mezi pracovní dobou a velikostí populace, a tak budeme pokračovat kvantifikací tohoto vztahu pomocí přímky $y = \beta_0 + \beta_1 x$.



Používáme-li model regresní analýzy pro statistické zpracování našich dat, je dobré ověřit předpoklady, ze kterých model vychází. Shrňme je v následujících třech bodech.

(1) Závisle proměnná Y (pracovní doba) má normální rozdělení pro každou hodnotu nezávisle proměnné x (velikost populace).
(2) Rozptyl závisle proměnné Y je stejný pro každou hodnotu nezávisle proměnné x.
(3) Závislost veličiny Y na x je lineární.

Pro tuto chvíli předpokládejme, že pro náš příklad jsou tyto předpoklady splněny.

Odhad absolutního členu β_0 a směrnice β_1 regresní přímky a jejich statistické charakteristiky jsou uvedeny v další tabulce. Směrodatná chyba koeficientu je výběrová směrodatná odchylka odhadovaného parametru, tj. $\widehat{s}_{\beta_0} = S_{M1}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$ a $\widehat{s}_{\beta_1} = \frac{S_{M1}}{\sqrt{S_{XX}}}$ (Ve statistických programech je obvykle označována anglicky jako Standard Error.)

Statistické charakteristiky lineární regrese

Parametr	Koeficient	Směrodatná chyba koef.	t-statistika	p-hodnota
Absolutní člen β_0	180,658	128,381	1,407	0,1896823
Směrnice β_1	9,429	0,681	13,847	7.520972e-08

Z tabulky tedy dostáváme:

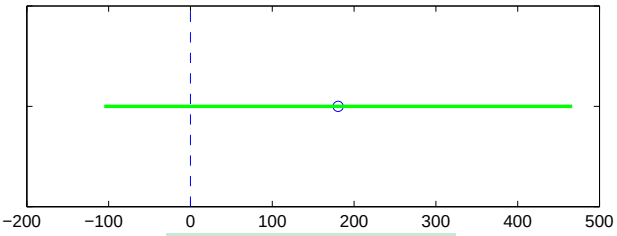
$$\text{pracovní doba} = 180{,}658 + 9{,}429 \cdot \text{velikost populace}.$$

To je třeba interpretovat jako odhad průměrné hodnoty počtu pracovních hodin pro populaci s danou velikostí. Očekáváme, že na každých dalších 1 000 lidí stoupne za měsíc počet pracovních hodin o 9,429, což je směrnice regresní přímky. Uvědomte si, že absolutní člen $(180,658)$ značí průměrný počet pracovních hodin, když je populace rovna nule. To zřejmě nedává smysl a mělo by nám to připomenout, že model by se měl používat pouze v tom rozmezí obou veličin, v němž se pohybovaly pozorované hodnoty. V tomto případě to znamená x od 26 do 370. Je ovšem pravda, že dosažená hladina významnosti pro absolutní člen je přibližně $0,19$, a nelze tedy říci, že by se absolutní člen β_0 významně lišil od nuly.

Připomeňme, že tyto výsledky jsme spočítali pro náhodný výběr 12 nemocnic. Kdybychom teď zvolili jiný náhodný výběr 12 nemocnic, dostali bychom odlišný odhad směrnice a absolutního členu. Určeme proto intervaly spolehlivosti neznámých parametrů β_0 a β_1.

Oboustranný interval spolehlivosti pro β_0

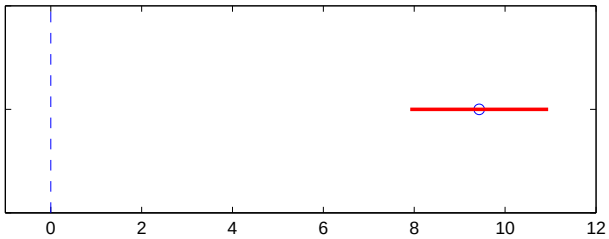
$180{,}6575 \pm 2{,}228 \cdot 128{,}3812 = 180{,}6575 \pm 286{,}051$



$(-105{,}394;\ 466{,}709)$

Oboustranný interval spolehlivosti pro β_1

$9{,}429 \pm 2{,}228 \cdot 0{,}681 = 9{,}429 \pm 1{,}517$



$(7{,}912;\ 10{,}946)$

Na základě výběru 12 nemocnic můžeme říci, že neznámý parametr β_0 leží mezi $-105,394$ a $466,709$ a neznámý parametr β_1, tj. parametr změny průměrného počtu pracovních hodin v závislosti na změně velikosti populace (v tisících), leží mezi $7,912$ a $10,946$ pracovními hodinami za měsíc.

Protože interval spolehlivosti pro β_0 pokrývá nulu, nelze potvrdit, že se významně liší od nuly. Naproti tomu interval spolehlivosti pro β_1 nulu nepokrývá, tedy se významně liší od nuly, jinak řečeno počet pracovních hodin skutečně lineárně závisí na rozsahu spádové populace.

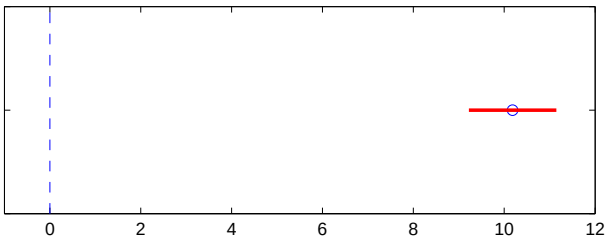
Pokud bychom uvažovali **regresi procházející počátkem** (plná čára) a výsledek srovnali s obecnou regresní přímkou (čárkovaná čára), dostaneme následující odhady

$\widehat{\beta}_1^* = 10,185$ $\widehat{s}_{\beta_1^*} = 0,4371,$

$t^* = 3,30157,\ p^* - hodnota = 1.0318e - 10$

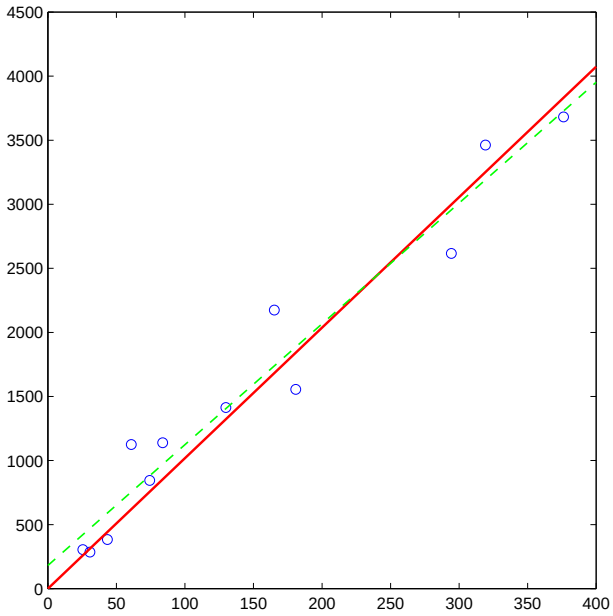
Oboustranný interval spolehlivosti pro β_1^*

$10{,}185 \pm 2{,}2 \cdot 0{,}4371 = 10{,}185 \pm 0{,}962$



$(9{,}223;\ 11{,}147)$

Protože interval spolehlivosti pro β_1^* nulu nepokrývá, opět jsme prokázali, že se významně liší od nuly, tj. počet pracovních hodin skutečně lineárně závisí na rozsahu spádové populace.



pracovní doba $= 10{,}185 \cdot$ velikost populace.