

# **01MAS Matematická statistika**

Martin Kovanda  
(revize a korektury Václav Kůs)

25. srpna 2020

# Obsah

<b>1</b>	<b>Statistika - setup a základní postupy</b>	<b>1</b>
1.1	Statistické bodové odhady	2
1.2	Vlastnosti bodových odhadů	2
1.3	Výběrové charakteristiky a jejich vlastnosti	3
1.4	Výběrový kvantil (a jeho vlastnosti)	5
1.5	Neparametrické (empirické) odhady distribucí	5
<b>2</b>	<b>Metody pro hledání bodových odhadů</b>	<b>8</b>
2.1	Metoda momentů	8
2.2	Nestranné odhady s minimálním rozptylem (UMVUE)	9
2.3	Rao-Cramérova nerovnost	10
2.4	Metoda maximální věrohodnosti (MLE)	11
<b>3</b>	<b>Testování statistických hypotéz</b>	<b>13</b>
3.1	Základní strategie TSH	13
3.2	UMP testy pro parametr $\theta = \theta(F)$	14
3.3	Neyman-Pearsonovo lemma (N-PL)	14
3.4	Složené hypotézy a MLR systémy	15
3.5	Nestranné UMP testy (UMPU)	16
<b>4</b>	<b>Další metody testování hypotéz</b>	<b>18</b>
4.1	Test poměrem věrohodností (LRT)	18
4.2	Analýza variance (ANOVA)	19
4.3	Dvouvýběrové testy ( $2 \times \mathcal{N}_1$ )	20
4.4	Test koeficientu korelace ( $\mathcal{N}_2$ )	22
<b>5</b>	<b>Asymptotické testy hypotéz</b>	<b>24</b>
5.1	Asymptotické testy středních hodnot <i>iid</i> $\mathcal{L}_2$	24
5.2	Asymptotický LRT a Waldův test v $\mathbb{R}^k$	24
5.3	Testy dobré shody (GoF)	25
5.4	Modifikace $\chi^2$ -testů dobré shody	26
<b>6</b>	<b>Konfidenční množiny, Intervaly spolehlivosti</b>	<b>28</b>
6.1	Konstrukce $CM_{1-\alpha}$ pomocí pivotů (PQ)	28
6.2	Konstrukce $CM_{1-\alpha}$ pomocí TSH $\phi_\alpha$	29
6.3	Asymptotické konfidenční množiny	29

# 1 Statistika - setup a základní postupy

**Definice 1.1.**  $n$ -tici nezávislých náhodných veličin  $X_1, \dots, X_n$  stejně rozdělených s distribuční funkcí  $F$  nazýváme **náhodný výběr** z rozdělení  $F$ . Konkrétní realizací  $\mathbf{X}$  získáme vektor čísel  $\mathbf{x} = (x_1, \dots, x_n)$ , který nazveme **realizace** náhodného výběru  $X_1, \dots, X_n$ , neboli naměřená data.

**Definice 1.2.** Vybereme vektor individuí  $\omega^{(n)} = (\omega_1, \dots, \omega_n) \in \Omega^n$  a definujeme

$$X_j(\omega^{(n)}) := X(\omega_j), \quad \forall j \in \hat{n},$$

jako **pozorování** (v tomto tvaru máme skutečná data). Zavedeme dále  $\Omega^{(n)} := \Omega^n$ ,  $\mathcal{A}^{(n)} := \sigma(\mathcal{A}^n)$ ,  $\mathbb{P}^{(n)} := \bigotimes_1^n \mathbb{P}^X$  (součinnová míra), tedy

$$\mathbb{P}^{(n)}(A_1 \times \dots \times A_n) = \prod_{j=1}^n \mathbb{P}^X(A_j), \quad \forall A_j \in \mathcal{A}.$$

Definujeme **realizaci** náhodného výběru  $\mathbf{X} = (X_1, \dots, X_n)$  vztahem

$$\mathbf{x} = \mathbf{X}(\omega^{(n)}).$$

**Věta 1.3.**  $(X_j)_{j=1}^n \text{ iid } \mathbb{P}^X$ .

**Definice 1.4.** **Statistika** je libovolná funkce náhodného výběru  $X_1, \dots, X_n$ , jejíž funkční předpis nezávisí na parametrech příslušného rozdělení.

Příkladem statistiky může být **výběrový průměr** (sample mean)

$$\overline{X}_n = \frac{1}{n} \sum_{j=1}^n X_j,$$

případně **geometrický průměr**

$$\overline{X}_n^G = \sqrt[n]{X_1 X_2 \dots X_n},$$

**výběrový rozptyl**

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^2 \quad \text{nebo} \quad s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \overline{X}_n)^2,$$

kde  $\hat{\sigma}_n$  nebo případně  $s_n$  je **výběrová směrodatná odchylka**. Dále uveďme  **$r$ -tý výběrový obecný a centrální moment** (sample moments)

$$m'_r = \frac{1}{n} \sum_{j=1}^n X_j^r, \quad m_r = \frac{1}{n} \sum_{j=1}^n (X_j - \overline{X}_n)^r,$$

**medián**

$$\hat{X}_{\frac{1}{2}} = \begin{cases} X_{(\frac{n+1}{2})} & n \text{ liché,} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}) & n \text{ sudé,} \end{cases}$$

kde  $X_{(j)}$  značí  $j$ -tou zdola uspořádanou statistiku z náhodného výběru  $(X_{(1)} \leq \dots \leq X_{(n)})$ .

## 1.1 Statistické bodové odhady

Nechť  $\theta$  je parametr spojený s rozdělením  $\mathbb{P}^X$ , tzn.  $\theta = \theta(\mathbb{P}^X)$ . Máme dvě možnosti.

- a)  $\theta$  může být spojený s  $\mathbb{P}^X$ , např.  $\theta = \mathbb{E}X, DX, F_X(t), \dots$ , nebo
- b)  $\theta$  může být přímo parametr rozdělení  $\mathbb{P}_\theta^X$ .

V obou případech ale požadujeme tzv. identifikovatelnost rodiny  $\mathcal{P} = \{\mathbb{P}^X\}$ , resp.  $\mathcal{P} = \{\mathbb{P}_\theta^X : \theta \in \Theta\}$ , tzn., že každému parametru přísluší právě jedna pravděpodobnost  $X$ .

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_1^X \neq \mathbb{P}_2^X, \quad \text{resp.} \quad \theta_1 \neq \theta_2 \Rightarrow P_{\theta_1}^X \neq P_{\theta_2}^X.$$

Opačná implikace často neplatí, třeba pro danou střední hodnotu najdeme vícero různých rozdělení. Předpokládáme, že  $\theta \in \Theta \subset \mathbb{R}^k$ , kde  $\Theta$  se nazývá **parametrický prostor**. Dále máme  $\tau(\theta) : \Theta \rightarrow \mathbb{R}^s$ , kde  $\tau$  je tzv. **parametrická funkce** (většinou nás ale zajímá např. jedna vybraná složka, tedy  $s = 1$ ). Odhadnout celé rozdělení se nám většinou nepodaří (nebo to nepotřebujeme), proto hledáme odhad parametru  $\theta$ .

**Definice 1.5.** Libovolná (borelovsky) měřitelná funkce náhodného výběru  $X_1, \dots, X_n$  (tedy statistika)

$\hat{\theta}_n(\mathbf{X}) : \Omega^n \rightarrow \mathbb{R}^k$  se nazývá **odhadem** (estimator)  $\theta \in \Theta \subset \mathbb{R}^k$ .

$\hat{T}_n(\mathbf{X}) : \Omega^n \rightarrow \mathbb{R}^s$  se nazývá **odhadem** (estimator) parametrické funkce  $\tau(\theta) \in \tau(\Theta) \subset \mathbb{R}^s$ .

## 1.2 Vlastnosti bodových odhadů

**Definice 1.6.**  $T_n(\mathbf{X})$  nazýváme **nestranný odhad**  $\tau(\theta)$ , pokud

$$\mathbb{E}_\theta T_n(\mathbf{X}) = \tau(\theta), \quad \forall \theta \in \Theta, \quad \forall n \in \mathbb{N}.$$

$T_n(\mathbf{X})$  nazýváme **asymptoticky nestranný odhad**  $\tau$ , pokud

$$\mathbb{E}_\theta T_n(\mathbf{X}) \rightarrow \tau(\theta), \quad \forall \theta \in \Theta,$$

kde  $\mathbb{E}_\theta$  značí střední hodnotu vzhledem k rozdělení  $\mathbb{P}_\theta^X$ .

**Definice 1.7.**  $T_n(\mathbf{X})$  nazýváme **eficientní** (vydatný), pokud pro  $\forall \tilde{T}_n, \forall \theta \in \Theta$  platí

$$\begin{aligned} s = 1 : \mathbb{E} \left[ (T_n(\mathbf{X}) - \tau(\theta))^2 \right] &\leq \mathbb{E} \left[ (\tilde{T}_n(\mathbf{X}) - \tau(\theta))^2 \right], \\ s > 1 : \mathbb{E} \left[ \|T_n(\mathbf{X}) - \tau(\theta)\|_\varepsilon^2 \right] &\leq \mathbb{E} \left[ \|\tilde{T}_n(\mathbf{X}) - \tau(\theta)\|_\varepsilon^2 \right]. \end{aligned}$$

$\mathbb{E}(T_n - \tau)^2$  se nazývá střední kvadratická chyba - MSE (*mean squared error*) odhadu  $T_n(\mathbf{X})$ . Tedy efficientní odhad má nejnižší možnou MSE. Pro nestranné odhady ( $\tau(\theta) = \mathbb{E}T_n(\mathbf{X})$ ) tento vztah pro  $s = 1$  přechází na  $DT_n(\mathbf{X}) \leq D\tilde{T}_n(\mathbf{X})$ .

**Definice 1.8.** Máme-li  $T_n^1, T_n^2 : \Omega \rightarrow \mathbb{R}^1$  jako odhady  $\tau(\theta)$ , pak definujeme **relativní eficienti** vztahem

$$\text{RE}_{2,1} = \frac{\mathbb{E}(T_n^1 - \tau)^2}{\mathbb{E}(T_n^2 - \tau)^2} \xrightarrow{\text{nestranný odhad}} \frac{DT_n^1}{DT_n^2}.$$

**Definice 1.9.**  $T_n(\mathbf{X})$  se nazývá **konzistentní** odhad  $\tau$ , pokud

$$T_n(\mathbf{X}) \xrightarrow{\mathbb{P}, s.j.} \tau(\theta), \quad \forall \theta \in \Theta$$

(pro  $\xrightarrow{\mathbb{P}}$  slabě konzistentní, pro  $\xrightarrow{s.j.}$  silně konzistentní).

**Věta 1.10** (Kritérium konzistence). *Odhad  $T_n(\mathbf{X})$  je slabě konzistentní ( $T_n(\mathbf{X}) \xrightarrow{\mathbb{P}} \tau(\theta)$ ), pokud*

1.  $T_n(\mathbf{X})$  je asymptoticky nestranný ( $\mathbb{E}_\theta T_n(\mathbf{X}) \rightarrow \tau(\theta)$ , pro  $\forall \theta \in \Theta$ ) a
2. platí pro něj, že  $DT_n(\mathbf{X}) \rightarrow 0$ .

**Definice 1.11.**  $T_n(\mathbf{X})$  se nazývá **asymptoticky normálním** odhadem  $\tau(\theta)$  s asymptotickou kovarianční maticí  $\mathbb{C}(\theta)$  (matice tvaru  $s \times s$ ), pokud pro  $\forall \theta \in \Theta$

$$T_n(\mathbf{X}) \sim \mathcal{AN}_s\left(\tau(\theta), \frac{1}{n}\mathbb{C}(\theta)\right), \text{ tzn. } \sqrt{n}(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}_s(\mathbf{0}, \mathbb{C}(\theta)) \text{ (viz CLT).}$$

Pro  $s = 1$  definice přechází na  $\sqrt{n}(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(\theta))$ , kde  $\sigma^2(\theta)$  nazýváme asymptotický rozptyl odhadu  $T_n(\mathbf{X})$ .

**Definice 1.12.** Máme-li  $T_n^1, T_n^2$  jako odhady  $\tau(\theta)$ , které jsou oba  $\mathcal{AN}$  odhady  $\tau$  s asymptotickými rozptyly  $\sigma_1^2(\theta)$  a  $\sigma_2^2(\theta)$ , pak definujeme **asymptotickou relativní eficienti** vztahem  $\text{ARE}_{2,1} = \frac{\sigma_1^2(\theta)}{\sigma_2^2(\theta)}$ .

**Věta 1.13.**  $T_n(\mathbf{X})$  je asymptoticky normální  $\mathcal{AN}\left(\tau(\theta), \frac{1}{n}\sigma^2(\theta)\right)$ . Pak  $T_n(\mathbf{X}) \xrightarrow{\mathbb{P}} \tau(\theta)$ ,  $\forall \theta \in \Theta$ . Tato slabá konzistence je řádu  $o_p(n^{-\alpha})$ ,  $\alpha < \frac{1}{2}$ , tzn., že  $n^\alpha(T_n(\mathbf{X}) - \tau(\theta)) \xrightarrow{\mathbb{P}} 0$ ,  $\forall \alpha < \frac{1}{2}$ . To vyplývá z věty (MIP) (Mějme  $(X_n)_{n=1}^{+\infty}$ ,  $X_n \sim \mathcal{AN}(\mu, \sigma_n^2)$  tak, že  $\sigma_n \rightarrow 0$ . Pak  $X_n \xrightarrow{\mathbb{P}} \mu$ ).

**Věta 1.14** ( $\Delta$ -metoda). Nechť  $T_n(\mathbf{X}) \sim \mathcal{AN}\left(\tau(\theta), \frac{1}{n}\sigma^2(\theta)\right)$  a  $g : \mathbb{R}^1 \rightarrow \mathbb{R}^1$  spojitě diferencovatelnou v  $\tau(\theta)$ ,  $\forall \theta \in \Theta$ . Pak

$$g(T_n(\mathbf{X})) \sim \mathcal{AN}_1\left(g(\tau(\theta)), \frac{\sigma^2(\theta)}{n} [g'(\tau(\theta))]^2\right).$$

### 1.3 Výběrové charakteristiky a jejich vlastnosti

**Věta 1.15.** Mějme  $X \in \mathcal{L}_1$ , resp.  $\mathcal{L}_2$ , volme  $\theta = \theta(\mathbb{P}^X) = \mathbb{E}X = \mu$  a označme  $\sigma^2 = \text{DX} < +\infty$  ( $\mathcal{L}_2$ ). Pak **sample mean**

$$T_n(\mathbf{X}) = \overline{X_n} = \frac{1}{n} \sum_{j=1}^n X_j$$

je nestranným, konzistentním a  $\mathcal{AN}\left(\mathbb{E}X, \frac{1}{n}\sigma^2\right)$  odhadem  $\theta = \mathbb{E}X = \mu$ .

**Věta 1.16.**  $X \in \mathcal{L}_2$ , resp.  $\mathcal{L}_4$ , volme  $\theta = \theta(\mathbb{P}^X) = DX = \sigma^2$ . Pak **výběrový rozptyl**

$$T_n(\mathbf{X}) = \hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \quad i \quad T_n(\mathbf{X}) = s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2$$

jsou oba asymptoticky nestranné, konzistentní a  $\mathcal{AN}\left(\sigma^2, \frac{1}{n}(\mu_4 - \sigma^4)\right)$  odhady  $\sigma^2$ , kde  $\mu_4 = \mathbb{E}[(X - \mathbb{E}X)^4]$ . V případě  $T_n(\mathbf{X}) = s_n^2$  je  $T_n(\mathbf{X})$  navíc nestranný odhad  $\sigma^2$ ,  $\forall n \in \mathbb{N}$ .

**Důsledek 1.17.** Mějme  $X \in \mathcal{L}_2$ . Potom pro tzv. **t-statistiku** platí, že

$$t_n = t_n(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1).$$

**Věta 1.18** (Chinčín). Mějme  $X \in \mathcal{L}_r$ , resp.  $X \in \mathcal{L}_{2r}$ ,  $r \geq 1$ , volíme

$$\theta_1 = \theta_1(\mathbb{P}^X) = \mathbb{E}(X^r) = \mu'_r, \quad \theta_2 = \mathbb{E}(X - \mathbb{E}X)^r = \mu_r.$$

Pak  $r$ -tý výběrový moment

$$m'_r = m'_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n X_j^r$$

je nestranným, konzistentním a  $\mathcal{AN}$  odhadem  $\theta_1 = \mu'_1$  a

$$m_r = m_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^r$$

je konzistentním odhadem  $\mu_r$ .

**Věta 1.19.** Speciálně nyní mějme  $(X_j)_{j=1}^{n,+\infty}$  iid  $\mathcal{N}(\mu, \sigma^2)$ . Pak

a)  $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ ,  $\forall n \in \mathbb{N}$ . Dále potom  $\mathbb{E}\bar{X}_n = \mu$ ,  $D\bar{X}_n = \sigma^2$ ,  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$ ,

b)  $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \xrightarrow{\mathcal{N}} \frac{n\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1)$ ,  $\forall n \in \mathbb{N}$  (Pearsonovo rozdělení).

$s_n^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}_n)^2 \xrightarrow{\mathcal{N}} \frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2(n-1)$ ,  $\forall n \in \mathbb{N}$ . Dále platí

$$D(\hat{\sigma}_n^2) = \frac{\sigma^4}{n^2} 2(n-1) = \frac{n-1}{n^2} (2\sigma^4) \rightarrow 0,$$

$$D(s_n^2) = D\left(\frac{\sigma}{n-1} \chi^2(n-1)\right) = \frac{\sigma^4}{(n-1)^2} 2(n-1) = \frac{2\sigma^4}{n-1} \rightarrow 0,$$

$$D(s_n^2) > D(\hat{\sigma}_n^2).$$

c)  $\bar{X}_n$  a  $s_n^2$  jsou nezávislé a platí

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} = \frac{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}{\frac{s_n}{\sigma}} \sim \frac{\mathcal{N}(0, 1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1), \quad \forall n \in \mathbb{N} \quad (\text{Studentovo rozdělení}).$$

d) Pokud jsou  $X, Y$  na  $(\Omega, \mathcal{A}, \mathbb{P})$  nezávislé, pak pro

$X_1, \dots, X_n \text{ iid } \mathcal{N}(\mu_1, \sigma_1^2)$ , ze kterých známe  $\overline{X}_n, s_{1,n}^2$ ,

$Y_1, \dots, Y_m \text{ iid } \mathcal{N}(\mu_2, \sigma_2^2)$ , ze kterých známe  $\overline{Y}_m, s_{2,m}^2$ ,

platí, že

$$\frac{s_{1,n}^2(n-1)}{\sigma_1^2} \sim \chi^2(n-1), \quad \frac{s_{2,m}^2(m-1)}{\sigma_2^2} \sim \chi^2(m-1).$$

Navíc lze ukázat, že  $s_{1,n}^2$  a  $s_{2,m}^2$  jsou nezávislé.

**Důsledek 1.20.**

$$T_n(\mathbf{X}) = \frac{\frac{s_{1,n}^2}{\sigma_1^2}}{\frac{s_{2,m}^2}{\sigma_2^2}} = \frac{\frac{\chi^2(n-1)}{n-1}}{\frac{\chi^2(m-1)}{m-1}} \sim F(n-1, m-1) \quad (\text{Fisherovo rozdělení}).$$

## 1.4 Výběrový kvantil (a jeho vlastnosti)

**Definice 1.21.** Mějme náhodný výběr  $(X_1, \dots, X_n)$ . Pak  $(X_{(1)}, \dots, X_{(n)})$  nazveme **uspořádaný náhodný výběr** (vzestupně). **Výběrový  $\alpha$ -kvantil** definujeme jako  $\hat{X}_{\alpha,n} = X_{([\alpha n]+1)}$ , pro  $\alpha = \frac{1}{2}$  nazýváme  $\hat{X}_{\frac{1}{2}}$  **výběrovým mediánem**, který alternativně označujeme jako  $\hat{X}_{\text{med}}$ . **Výběrové rozpětí** pak definujeme jako  $d = X_{(n)} - X_{(1)}$  a **výběrové interkvartilové rozpětí** jako

$$d_{\frac{1}{4}} = X_{([\frac{3}{4}n]+1)} - X_{([\frac{1}{4}n]+1)}.$$

**Věta 1.22.** Mějme  $X_1, \dots, X_n \text{ iid } F$ ,  $\theta = x_\alpha$ ,  $\alpha \in (0, 1)$ . Nechť  $x_\alpha$  je jednoznačně určeno rovnicí  $F(x_\alpha) = \alpha$  a existuje  $F'(x_\alpha) > 0$ . Pak

$$\hat{X}_\alpha \sim \mathcal{AN}\left(x_\alpha, \frac{\alpha(1-\alpha)}{n[F'(x_\alpha)]^2}\right).$$

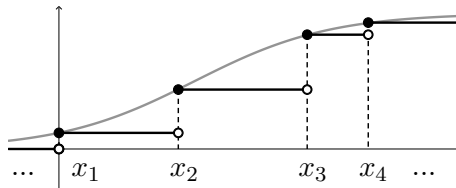
**Důsledek 1.23.** Z  $\mathcal{AN}$  plyne, že  $\hat{X}_\alpha \xrightarrow{\mathbb{P}} x_\alpha$  řádu  $o_p(n^{-\beta})$ ,  $\beta < \frac{1}{2}$ .

## 1.5 Neparametrické (empirické) odhady distribucí

**Definice 1.24.** Mějme  $X_1, \dots, X_n \text{ iid } F$ , označme pro dané  $t \in \mathbb{R}$  charakteristickou funkci na intervalu  $(-\infty, t]$  jako  $\mathbb{I}_{(-\infty, t]}$ . Pak **empirickou distribuční funkci** (EDF)  $F_n$  definujeme vztahem

$$F_n(t) = F_n(t, \mathbf{X}) = \frac{1}{n} \sum_{j=1}^n \mathbb{I}_{(-\infty, t]}(X_j), \quad \forall t \in \mathbb{R}.$$

Pro  $t$  fixní můžeme psát  $F_n(t, \mathbf{X}) = T_n(\mathbf{X})$ .



**Věta 1.25** (ZVMS). Mějme  $X_1, \dots, X_n$  iid  $F$ . Pak pro  $\forall t \in \mathbb{R}$  *fixní* platí, že  $F_n(t)$  je *ne-stranným, konzistentním a  $\mathcal{AN}$  odhadem  $\theta = F(t)$ , tzn.*

1.  $\mathbb{E}F_n(t) = F(t), \forall n,$
2.  $F_n(t) \xrightarrow{\mathbb{P}, s.j.} F(t), \forall t \in \mathbb{R},$
3.  $F_n(t) \sim \mathcal{AN}(F(t), \frac{1}{n}F(t)(1 - F(t))).$

Navíc dokonce platí, že

1.  $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{s.j.} 0, \text{ (Glivenko-Cantelliho lemma),}$
2.  $\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \varepsilon\right) \leq 8(n+1)e^{-\frac{n\varepsilon^2}{32}}, \forall n, \forall \varepsilon > 0, \text{ (Glivenko-Cantelli),}$
3.  $\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}, \forall n, \forall \varepsilon > 0 \text{ (Massart, 1990).}$

**Definice 1.26.** Mějme  $\theta = \theta(F)$  (funkcionál na prostoru distribučních funkcí). Pak

$$T_n(\mathbf{X}) = \theta\left(\underbrace{F_n}_{\rightarrow F}\right)$$

se nazývá **statistický funkcionál**.

**Definice 1.27** (Histogram). Mějme  $X \sim f$ ,  $\text{supp } f = [a, b]$ , resp. zde BÚNO  $[0, 1]$ . Zavedeme dělení intervalu  $[0, 1]$

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \dots, B_m = \left[\frac{m-1}{m}, 1\right]$$

a označíme dále  $h = \frac{1}{m} = \lambda(B_j)$ ,  $Y_j = \#\{i : X_i \in B_j\}_{i=1}^n$ ,  $\hat{p}_j = \frac{Y_j}{n}$  jako odhad  $p_j = \int_{B_j} f(x)dx$ .

Pak **histogramový odhad hustoty pravděpodobnosti** definujeme vztahem

$$\hat{f}_n^H(t) = \sum_{j=1}^m \frac{\hat{p}_j}{h} \mathbb{I}_{B_j}(t) = \frac{1}{nh} \sum_{j=1}^m Y_j \mathbb{I}_{B_j}(t) = \frac{1}{nh} \sum_{i=1}^n \mathbb{I}\{X_i \in B_j\}, \quad \forall t \in B_j.$$

**Věta 1.28** (IMSE). Pro  $\hat{f}_n^H(t)$  předpokládejme, že  $f'$  je absolutně spojitá a platí, že  $\int_{\mathbb{R}} (f'(u))^2 du < +\infty$ . Potom

$$R(\hat{f}_n^H, f) = \frac{h^2}{12} \int_0^1 (f'(x))^2 dx + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right),$$

což při volbě  $h_n = O(n^{-1/3})$  vede na řád integrované střední kvadratické chyby (Integrated Mean Square Error)

$$\text{IMSE} = R(\hat{f}_n^H, f) = \int_0^1 \mathbb{E} \left[ \hat{f}_n^H(t) - f(t) \right]^2 dt = O(n^{-2/3}).$$



## 1 Statistika - setup a základní postupy

**Definice 1.29.** Mějme jádro  $K(x)$  takové, že  $K(x) \geq 0$ ,  $\int_{\mathbb{R}} K(x) dx = 1$ ,  $\int_{\mathbb{R}} xK(x) dx = 0$ ,  $\sigma_K^2 = \int_{\mathbb{R}} x^2 K(x) dx > 0$ . Označíme  $h \in \mathbb{R}^+$  jako tzv. **šířku okna** (*bin width*), neboli vyhlazovací parametr (*smoothing parameter*). Pak definujeme **jádrový odhad hustoty** vztahem

$$\hat{f}_n^K(t) := \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{t - X_i}{h}\right), \quad \forall t \in \mathbb{R}.$$

## 2 Metody pro hledání bodových odhadů

### 2.1 Metoda momentů

Tato metoda je založená na užití výběrových momentů, ať už centrálních nebo necentrálních, případně i z momentů rozdělení. V praxi pak využijeme tu, která se dá vypočítat nejjednodušeji. V této metodě bereme v potaz všechny realizace.

Máme tedy  $(\Omega, \mathcal{A}, \mathbb{P})$ ,  $\theta \in \Theta \subset \mathbb{R}^k$ ,  $\tau(\theta)$  jako odhadovanou parametrickou funkci, vlastnost  $X \text{ iid } f_X(x, \theta)$ . Nechť  $X_1, \dots, X_n \text{ iid } \mathcal{L}_k$  (aby existovaly momenty do řádu  $k$ ) a označíme

$$\begin{aligned}\mu_r &= \mu_r(\theta) = \mathbb{E}X^r, \quad r \in \hat{k}, \\ \boldsymbol{\mu}(\theta) &= (\mu_1(\theta), \dots, \mu_k(\theta)) : \mathbb{R}^k \rightarrow \mathbb{R}^k.\end{aligned}$$

Dále požadujeme, aby  $\exists \boldsymbol{\mu}^{-1}$ , tedy například aby  $\boldsymbol{\mu}$  bylo regulární a prosté.

**Definice 2.1** (Momentový odhad). **Momentový odhad**  $\theta$  definujeme vztahem

$$\hat{\theta}_M := \hat{\theta}_M(\mathbf{X}) = \boldsymbol{\mu}^{-1}(m'_1(\mathbf{X}), \dots, m'_k(\mathbf{X})), \quad \text{kde } m'_r(\mathbf{X}) = \frac{1}{n} \sum_{j=1}^n X_j^r,$$

což znamená, že  $\hat{\theta}_M$  je řešením *soustavy momentových rovnic* (značíme  $ME_q$ ) ve tvaru

$$\mu_r(\theta) = m'_r(\mathbf{X}), \quad \forall r \in \hat{k}.$$

Definujeme dále **momentový odhad**  $\tau(\theta)$  vztahem  $T_M(\mathbf{X}) := \tau(\hat{\theta}_M(\mathbf{X}))$ , metoda momentů je tedy invariantní na transformace parametru  $\theta$  (je to jen jiné vyjádření pro vztah  $T_M(\mathbf{X}) = \tau(\widehat{\theta}_M) = \tau(\hat{\theta}_M)$ ).

**Věta 2.2.** Pokud je  $\boldsymbol{\mu}^{-1}$  spojitá funkce, pak  $\hat{\theta}_M(\mathbf{X})$  je **konzistentním** odhadem  $\theta$ . Pokud je navíc  $\tau$  spojitá, pak  $T_M(\mathbf{X})$  je konzistentní.

**Věta 2.3.** Nechť  $\hat{\theta}_M$  je odhad metodou momentů,  $(X_j)_{j=1}^{+\infty} \in \mathcal{L}_{2k}$ ,  $\boldsymbol{\mu}$  je difeomorfismus  $(\boldsymbol{\mu}, \boldsymbol{\mu}^{-1})$  spojitě diferencovatelné. Pak  $\forall \theta \in \Theta$  je

$$\hat{\theta}_M \sim \mathcal{N}\left(\theta, \frac{1}{n} \mathbb{C}_M(\theta)\right),$$

kde  $\mathbb{C}_M(\theta) = \mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta) \mathbb{C}(\theta) \mathbb{J}_{\boldsymbol{\mu}^{-1}}(\theta)^T$  a  $\mathbb{C} = \text{Cov}(X, X^2, \dots, X^k)$ . Pokud je navíc  $\tau(\theta)$  spojitě diferencovatelné a  $\nabla \tau(\theta) \neq 0$ , pak

$$T_M(\mathbf{X}) \sim \mathcal{N}\left(\tau(\theta), \frac{1}{n} \nabla \tau(\theta) \mathbb{C}_M(\theta) \nabla \tau(\theta)^T\right).$$

## 2.2 Nestranné odhady s minimálním rozptylem (UMVUE)

**Definice 2.4.** Mějme  $X_1, \dots, X_n$  iid  $F$ ,  $\theta \in \Theta \subset \mathbb{R}^k$ ,  $\tau(\theta) \in \mathbb{R}^1$ ,  $T(\mathbf{X})$  jako odhad  $\tau(\theta)$ . Definujeme

$$T_{\text{UMR}} = \underset{T}{\operatorname{argmin}} \mathbb{E}(T(\mathbf{X}) - \tau(\theta))^2, \quad \forall \theta \in \Theta.$$

Definujeme dále kvadratickou **ztrátovou funkci** (*loss function*) jako

$$\mathcal{L}_2(T, \theta) := (T(\mathbf{X}) - \tau(\theta))^2$$

a příslušnou **rizikovou funkci** (*risk function*) vztahem

$$R(T, \theta) := \mathbb{E}\mathcal{L}_2(T, \theta).$$

Potom  $T_{\text{UMR}} = \underset{T}{\operatorname{argmin}} R(T, \theta)$ , kde UMR je zkratka pro *uniformly minimum risk*.  $T_{\text{UMR}}$  je tedy odhad, který minimalizuje hodnotu rizikové funkce  $R$ .

**Definice 2.5.**  $S(\mathbf{X})$  se nazývá **postačující statistika** (*sufficiency*) pro  $\theta$ , pokud rozdělení  $\mathbf{X}$  podmíněné hodnotou  $S(\mathbf{X}) = s$  nezávisí na parametru  $\theta$ . Pro diskrétní případy tedy  $\mathbb{P}(\mathbf{X} = \mathbf{x} | S(\mathbf{X}) = s)$  nezávisí na  $\theta$ , případně  $f_{\mathbf{X}|S}(\mathbf{x}|s)$  nezávisí na  $\theta$ .

Postačující statistika je tedy taková funkce náhodného výběru (statistika), která umí sama o sobě nahradit původní výběr bez ztráty informace o parametru  $\theta$ .

**Definice 2.6.** Mějme  $X, Y$  náhodné veličiny na  $(\Omega, \mathcal{A}, \mathbb{P})$ . Pak

$$\mathbb{E}(X|Y = y) = \int_{\mathbb{R}} x d\mathbb{P}^{X|Y=y} = \left[ \text{pro ASR } f_{X|Y}(x|y) := \frac{f_{X,Y}(x, y)}{f_Y(y)} \right] = \dots$$

Z toho vyplývá, že  $\mathbb{E}(X|Y) : \Omega \rightarrow \mathbb{R}$  je náhodná veličina.

**Věta 2.7.** Pro náhodné veličiny  $(X, Y)$  s ASR  $f_{X,Y}$  platí vztah

$$\mathbb{E}[\mathbb{E}(X|Y)] \stackrel{\text{ASR}}{=} \int_{\mathbb{R}} \left( \int_{\mathbb{R}} x f_{X|Y} dx \right) f_Y dy \stackrel{\text{F.V.}}{=} \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f_{X|Y} f_Y dy \right) x dx = \int_{\mathbb{R}} \left( \int_{\mathbb{R}} f_{X,Y} dy \right) x dx = \int_{\mathbb{R}} x f_X dx = \mathbb{E}X.$$

Pro účely následující Rao-Blackwellovy věty označme

$$T_{\text{RB}}(\mathbf{X}) = T_{\text{RB}}(S(\mathbf{X})) := \mathbb{E}[T(\mathbf{X}) | S(\mathbf{X}) = s]$$

za předpokladu existence  $\mathbb{E}$  jako odhad zkonstruovaný v Rao-Blackwellově větě.  $T_{\text{RB}}(\mathbf{X})$  je tedy opět statistika, pro kterou platí, že

$$T_{\text{RB}}(\mathbf{X}) := \mathbb{E}[T(\mathbf{X}) | S(\mathbf{X}) = s] \stackrel{\text{ASR}}{=} \int T(\mathbf{x}) f_{T(\mathbf{X})|S(\mathbf{X})=s} d\mathbf{x}.$$

Na  $T_{\text{RB}}(\mathbf{X})$  pohlížíme jako na funkci  $\mathbf{X}$ , která vznikne ve dvou krocích:

1. spočítá se podmíněná střední hodnota  $\mathbb{E}(T(\mathbf{X}) | S(\mathbf{X}) = s) = T(s)$  při libovolně daném pevném  $s$ ,

## 2 Metody pro hledání bodových odhadů

2. za  $s$  se dosadí vazba  $s = S(\mathbf{X})$ , čímž vznikne  $T_{\text{RB}}(S(\mathbf{X}))$ .

**Věta 2.8** (Rao-Blackwell). *Mějme  $X_1, \dots, X_n$  iid  $F$ ,  $\theta \in \Theta \subset \mathbb{R}^k$ ,  $\tau(\theta) \in \mathbb{R}^1$ ,  $T(\mathbf{X})$  jako odhad  $\tau(\theta)$ , nechť  $S(\mathbf{X})$  je postačující statistika pro  $\theta$  a nechť  $\mathcal{L}(T, \theta)$  je konvexní funkcí v  $T$  pro  $\forall \theta \in \Theta$ . Pak*

$$R(T_{\text{RB}}(\mathbf{X}), \theta) \leq R(T(\mathbf{X}), \theta), \quad \forall \theta \in \Theta,$$

přičemž  $T_{\text{RB}}$  nezávisí na  $\theta$ .

**Definice 2.9.** Systém hustot  $\mathcal{F} = \mathcal{F}_X = \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$  se nazývá **úplný**, pokud pro  $\forall h : \mathbb{R} \rightarrow \mathbb{R}$ ,  $\mathbb{E}_\theta h(X) = 0$ ,  $\forall \theta \in \Theta$  platí, že

$$h(X) = 0 \text{ s.j. } \forall \theta \in \Theta, \quad \text{neboli} \quad \mathbb{P}_\theta(h(X) = 0) = 1.$$

**Definice 2.10.** Postačující statistika  $S$  se nazývá **úplná postačující statistika**, pokud systém rozdělení  $\mathcal{F}_S$  je úplný, tzn. pro libovolnou borelovskou funkci  $g : \mathbb{R} \rightarrow \mathbb{R}$  platí, že pokud

$$\mathbb{E}[g(S(X))] = 0, \quad \forall \theta \in \Theta, \quad \text{pak} \quad g(S(X)) = 0 \text{ s.j.}, \quad \forall \theta \in \Theta.$$

**Věta 2.11** (Lehmann-Scheffé). *Nechť jsou splněny předpoklady R.-B. věty a navíc  $T(\mathbf{X})$  je nestranný odhad  $\tau(\theta)$  (tzn.  $\mathbb{E}T = \tau$ ,  $\forall \theta \in \Theta$ ), a dále nechť  $S(\mathbf{X})$  je úplná postačující statistika. Pak  $T_{\text{RB}} = T_{\text{UMR}}$ , což pro volbu ztrátové funkce  $\mathcal{L}_2$  označíme jako  $T_{\text{UMVUE}}$ . (uniformly minimum variance unbiased estimator)*

## 2.3 Rao-Cramérova nerovnost

**Definice 2.12.** Mějme  $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$  a označíme

$$l(\theta) = \ln f(x, \theta), \quad l'_i(\theta) := \frac{\partial}{\partial \theta_i} \ln f(x, \theta), \quad \nabla_\theta l = \nabla l(\theta) = (l'_1(\theta), \dots, l'_k(\theta)).$$

Systém  $\mathcal{F}$  se nazývá **regulární systém hustot**, ozn.  $\mathcal{F}_{\text{reg}}$ , pokud

- 1)  $\text{supp } f$  nezávisí na  $\theta$  a  $\Theta$  je otevřená množina,
- 2) pro všechna  $\forall \theta \in \Theta$  existuje  $\nabla l(\theta)$  na  $\text{supp } f$ ,
- 3)  $\mathbb{E}[\nabla l(\theta)] = 0$ ,  $\forall \theta \in \Theta$ , což je zajištěno předpokladem následující záměny

$$\mathbb{E}[l'_i(\theta)] = \int_{\mathbb{R}} \frac{f'_i}{f} \cdot f dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta_i} f dx = \frac{\partial}{\partial \theta_i} \underbrace{\int_{\text{supp } f} f dx}_{=1} = 0,$$

- 4)  $\text{Cov}(\nabla l(\theta))$  je konečná a PD matice rozměru  $(k \times k)$ .

Systém  $\mathcal{F}$  označíme  $\mathcal{F}_{\text{reg}}^+$ , pokud navíc splňuje podmínku

$$5) \quad \mathbb{E}\left[\frac{f''_{i,j}(\mathbf{X}, \theta)}{f(\mathbf{X}, \theta)}\right] = 0, \quad \forall \theta \in \Theta, \quad \text{neboli} \quad \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\mathbf{x}, \theta) d\mathbf{x} = 0, \quad \forall \theta \in \Theta.$$

**Definice 2.13.**  $\mathbb{I}(\theta) = \text{Cov}(\nabla l(\theta))$  se nazývá **Fisherova informační matice** a platí, že

$$\mathbb{I}_{i,j}(\theta) := \text{Cov}(l'_i, l'_j) = \mathbb{E}(l'_i \cdot l'_j) - \underbrace{\mathbb{E}l'_i}_{=0} \underbrace{\mathbb{E}l'_j}_{=0} = \mathbb{E}\left(\frac{\partial \ln f}{\partial \theta_i} \cdot \frac{\partial \ln f}{\partial \theta_j}\right) \stackrel{5)}{=} -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta_i \partial \theta_j}\right].$$

**Lemma 2.14.** Mějme  $\mathbf{X} = (X_1, \dots, X_n)$  nezávislé,  $X_i \sim f_{X_i}(x_i, \theta)$ . Pak  $\mathbb{I}_{\mathbf{X}}(\theta) = \sum_{j=1}^n \mathbb{I}_{X_j}(\theta)$ . Pokud navíc  $X_1, \dots, X_n$  jsou iid  $f_X$ , pak  $\mathbb{I}_{\mathbf{X}}(\theta) = n\mathbb{I}_X(\theta)$ .

**Věta 2.15** (Rao-Cramérova nerovnost). Mějme  $T(\mathbf{X})$  jako nestranný odhad  $\tau(\theta)$ ,  $\mathcal{F}_{reg}$ , nechť dále  $(\forall \theta \in \Theta)(\exists \nabla \tau(\theta))$  a  $\mathbb{E}[T(\mathbf{X})]$  lze derivovat podle  $\theta_i$  pod znakem  $\mathbb{E}$  (tj. derivace pod integrálem) pro  $\forall i \in \hat{k}$ . Pak

$$D(T(\mathbf{X})) \geq \underbrace{\nabla \tau(\theta) \mathbb{I}_{\mathbf{X}}^{-1}(\theta) \nabla \tau(\theta)^T}_{\text{RCLB}_{\tau}(\theta)}, \quad \forall \theta \in \Theta. \quad (\text{RCLB} = \text{Rao-Cramérova spodní hranice}).$$

**Definice 2.16.** Mějme  $\mathcal{F}_{reg}$ . Pak definujeme **eficienci** nestranného odhadu  $T_n(\mathbf{X})$  funkce  $\tau(\theta)$  vztahem

$$e_n := \frac{\text{RCLB}_{\tau}(\theta)}{D(T_n(\mathbf{X}))}.$$

Pokud  $e_n = 1$ ,  $\forall n \in \mathbb{N}$ , případně  $\lim_{n \rightarrow +\infty} e_n = 1$ , pak  $T(\mathbf{X})$  nazýváme **(asymptoticky) eficientní**.

**Věta 2.17** (Bhattacharya). Mějme  $\theta \in \Theta \subset \mathbb{R}^1$ ,  $T(\mathbf{X})$  jako odhad  $\tau(\theta)$ . Nechť dále existuje vektor derivací podle  $\theta$  do řádu  $m$ :  $\tilde{\tau}' = (\tau', \tau'', \dots, \tau^{(m)})$ . Pak za podobných dodatečných předpokladů jako v R.-C.N., platí, že

$$D(T(\mathbf{X})) \geq \tilde{\tau}'(\theta) \tilde{\mathbb{I}}_{\mathbf{X}}^{-1}(\theta) (\tilde{\tau}'(\theta))^T, \quad (\text{Bhattacharyova spodní hranice, BLB}_{\tau}(\theta)),$$

$$\text{kde } \tilde{\mathbb{I}}_{\mathbf{X}} = \text{Cov}\left(\left(\frac{\partial^i f}{\partial \theta^i} / f\right)_{i=1}^m\right).$$

## 2.4 Metoda maximální věrohodnosti (MLE)

**Definice 2.18** (Věrohodnostní funkce). Mějme  $X_1, \dots, X_n$  s odpovídajícím systémem hustot  $\mathcal{F} = \{f_{\mathbf{X}}(\mathbf{x}, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ . Pak definujeme **věrohodnostní funkci** vztahem

$$L(\theta) = f(\mathbf{x}, \theta), \quad \text{resp.} \quad L(\theta, \mathbf{x}) = h(\mathbf{x})f(\mathbf{x}, \theta), \quad \forall \theta \in \Theta, \quad \forall \mathbf{x} \in \mathbb{R}^n$$

a **logaritmickou věrohodnostní funkci** vztahem

$$l(\theta) = \ln L(\theta, \mathbf{x}).$$

Mějme nezávislý náhodný výběr  $X_1, \dots, X_n$  id. Pak věrohodnostní funkci můžeme zavést jako sdruženou hustotu tohoto výběru při dané realizaci  $\mathbf{x} = (x_1, \dots, x_n)$ , tedy

$$L(\theta) = \prod_{i=1}^n f_{X_i}(x_i, \theta), \quad \left( L(\theta) = \prod_{i=1}^n \mathbb{P}_{\theta}(X_i = x_i) - \text{pro diskrétní případ} \right).$$

## 2 Metody pro hledání bodových odhadů

**Definice 2.19** (Maximálně věrohodný odhad). Definujeme **maximálně věrohodný odhad** vztahem

$$\hat{\theta}_{\text{ML}}(\mathbf{X}) = \underset{\theta \in \Theta}{\operatorname{argsup}} L(\theta)$$

za předpokladu, že  $\hat{\theta}_{\text{ML}}$  je borelovsky měřitelná, jednoznačná a závisí na  $\mathbf{X}$ . Dále pak definujeme **maximálně věrohodný odhad** parametrické funkce  $\tau(\theta)$  jako

$$T_{\text{ML}}(\mathbf{X}) = \tau(\hat{\theta}_{\text{ML}}), \quad \text{tedy MLE je invariantní na transformace } \tau.$$

**Věta 2.20.** Mějme  $(X_j)_{j=1}^{+\infty}$  iid  $f(x, \theta_0)$ , kde  $\operatorname{supp} f_{\mathbf{X}}$  nezávisí na  $\theta$ ,  $\ln \frac{f(X, \theta)}{f(X, \theta_0)} \in \mathcal{L}_1$ . Dále předpokládáme identifikovatelnost rodiny hustot, tzn.

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2} \quad (\text{různé parametry definují různá rozdělení}).$$

Pak  $\mathbb{P}_{\theta_0}(L(\theta_0, \mathbf{x}) > L(\theta, \mathbf{x})) \xrightarrow{n \rightarrow +\infty} 1$ ,  $\forall \theta \neq \theta_0$ .

**Definice 2.21.** Systém hustot  $\mathcal{F} := \{f(x, \theta) : \theta \in \Theta \subset \mathbb{R}^k\}$  se nazývá **ML-regulární**, ozn.  $\mathcal{F}_{\text{reg}}^{\text{ML}}$ , pokud pro něj platí, že

1.  $\Theta$  je otevřená množina,  $\operatorname{supp} f_X$  nezávisí na  $\theta$ ,
2.  $f(x, \theta) \in \mathcal{C}^{(3)}$  vzhledem k  $\theta$  pro  $\forall x \in \mathbb{R}$ ,
3.  $\int_{\mathbb{R}} \frac{\partial f}{\partial \theta_r} dx = 0$  a  $\int_{\mathbb{R}} \frac{\partial^2 f}{\partial \theta_r \partial \theta_j} dx = 0$ ,  $\forall \theta \in \Theta$ , tzn. záměna derivací a integrálu je možná,
4. Fisherova informační matice  $\mathbb{I}_X(\theta)$  je PD a konečná,
5.  $(\forall \theta_0)(\exists H_{\theta_0})(\exists M(X) \in \mathcal{L}_1(\mathbb{P}_{\theta_0}))(\forall \theta \in H_{\theta_0})\left(\|\partial_{\theta}^3 \ln f\| \leq M(X), \text{ přičemž } \mathbb{E}_{\theta_0} M(X) < +\infty\right)$ .

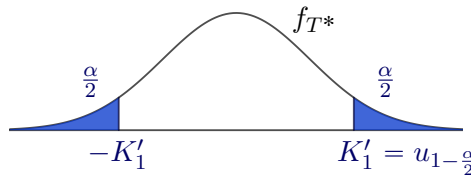
**Definice 2.22.** Mějme  $\hat{\theta}_n \sim \mathcal{AN}(\theta_0, \frac{1}{n}\mathbb{C}(\theta))$ .  $\hat{\theta}_n$  se nazývá **asymptoticky eficientní**, pokud  $\mathbb{C}(\theta) = \mathbb{I}_X^{-1}(\theta_0)$ .

**Věta 2.23.** Mějme  $X_1, \dots, X_n$  iid  $f(x, \theta_0) \in \mathcal{F}_{\text{reg}}^{\text{ML}}$ . Pak pro každé konzistentní řešení  $\hat{\theta}_n(\mathbf{X})$  soustavy věrohodnostních rovnic  $LE_q$  platí, že

$$\hat{\theta}_n \sim \mathcal{AN}_k\left(\theta_0, \frac{1}{n}\mathbb{I}_X^{-1}(\theta_0)\right), \quad \text{neboli} \quad \sqrt{n}(\hat{\theta}_n(\mathbf{X}) - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}_k\left(0, \frac{1}{\mathbb{I}_X(\theta_0)}\right).$$

Takové konzistentní řešení je tedy  $\mathcal{AN}$  a dle definice 2.22 i asymptoticky eficientním odhadem  $\theta_0$ , ozn.  $\hat{\theta}_{\text{ELE}}$  (ELE = efficient likelihood estimator).

**Věta 2.24.** Mějme  $\mathcal{F}_{\text{reg}}^{\text{ML}}$ ,  $\theta \in \mathbb{R}^1$ . Pak s pravděpodobností  $\xrightarrow{n \rightarrow +\infty} 1$  existuje konzistentní řešení  $LE_q$ .



## 3 Testování statistických hypotéz

### 3.1 Základní strategie TSH

**Definice 3.1.** Mějme populaci  $\Omega$  a na ní vlastnost  $X \sim F$ , kde  $F \in \mathcal{F}$ . Označme  $\theta = \theta(F)$  parametr modelu, který nás zajímá, kde  $\theta \in \Theta \subset \mathbb{R}^k$ . Označme  $H_0 : \theta \in \Theta_0$  jako **základní nulovou hypotézu** (*null hypothesis*) a  $H_1 : \theta \in \Theta_1$ , kde  $\Theta_1 = \Theta \setminus \Theta_0$ , jako **alternativní hypotézu**.

**Definice 3.2.** Definujeme  $R_{H_0}$  jako jev představující **zamítnutí**  $H_0$  (*rejection*) a  $\bar{R}_{H_0}$  jako **přijetí**  $H_0$  (*acceptation*). Pak **kritickou funkci testu** definujeme jako pravděpodobnost, že zamítneme  $H_0$  na základě naměřených dat  $\mathbf{x}$ , tzn.

$$\phi(\mathbf{x}) := \mathbb{P}(R_{H_0} | \mathbf{X} = \mathbf{x}) \in [0, 1], \quad \text{pro } \forall \mathbf{x} \in \mathcal{X},$$

kde  $\mathcal{X}$  je tzv. **výběrový prostor**,  $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \exists \omega \in \Omega, \mathbf{X}(\omega) = \mathbf{x}\}$ . Dále definujeme pro test  $H_0 : \theta \in \Theta_0$  vs.  $H_1 : \theta \in \Theta_1$ ,  $\Theta = \Theta_0 + \Theta_1$ , funkci

$$\begin{aligned} \beta_\phi(\theta) &:= \mathbb{E}_\theta[\phi(\mathbf{X})] = \mathbb{E}_\theta[\mathbb{P}(R_{H_0} | \mathbf{X} = \mathbf{x})] = \left| \begin{array}{l} \mathbb{P}(A) = \int_A 1 d\mathbb{P} = \int_\Omega \mathbb{I}_A d\mathbb{P} = \mathbb{E}[\mathbb{I}_A] \\ A = \{R_{H_0} | \mathbf{X} = \mathbf{x}\} \end{array} \right| = \\ &= \mathbb{E}_\theta[\mathbb{E}[\mathbb{I}_{R_{H_0}} | \mathbf{X} = \mathbf{x}]] \stackrel{\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}X}{=} \mathbb{E}_\theta[\mathbb{I}_{R_{H_0}}] = \mathbb{P}_\theta(R_{H_0}), \quad \text{pro } \forall \theta \in \Theta. \end{aligned}$$

Funkce  $\beta_\phi|_{\Theta_1}$ , tedy zúžení  $\beta_\phi$  z  $\Theta$  na obor  $\Theta_1$ , se nazývá **silofunkce testu**  $\phi$ .

Pokud testujeme hypotézu  $H_0$  oproti  $H_1$ , mohou nastat 4 navzájem se vylučující situace:

	<b>Zamítáme <math>H_0</math></b>	<b>Nezamítáme <math>H_0</math></b>
<b><math>H_0</math> platí</b>	Nastává chyba I. druhu	Správný výsledek
<b><math>H_0</math> neplatí</b>	Správný výsledek	Nastává chyba II. druhu

Chybu I. druhu považujeme za kritickou (tu horší) chybu. Pravděpodobnost chyby I. druhu vyjadřuje právě funkce  $\beta_\phi|_{\Theta_0}$ , což je zúžení  $\beta_\phi$  z  $\Theta$  na obor  $\Theta_0$ . Právě tuto pravděpodobnost budeme chtít mít pod kontrolou, tzn. pro vhodně malé zvolené číslo  $\alpha \in (0, 1)$  požadujeme, aby celé zúžení  $\beta_\phi|_{\Theta_0}$  bylo stejnoměrně pod zadanou hranicí  $\alpha$ . Číslo  $\alpha$  nazýváme **hladina významnosti** testu  $H_0$  versus  $H_1$  a požadujeme tedy, aby  $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \alpha$ .

Pravděpodobnost chyby II. druhu vyjadřuje funkce  $1 - \beta_\phi|_{\Theta_1}$  a budeme se ji snažit minimalizovat za vazební podmínky  $\beta_\phi|_{\Theta_0} \leq \alpha$  na chybu I. druhu.

Shrňme strategii testování  $H_0 : \theta \in \Theta_0$  vs.  $H_1 : \theta \in \Theta_1$  v následující sekci.

### 3.2 UMP testy pro parametr $\theta = \theta(F)$

**Definice 3.3** (UMP strategie testování  $H_0$  vs.  $H_1$ ). Hledáme takovou optimální kritickou funkci testu  $\phi^*$ , aby při zvolené hladině významnosti  $\alpha \in (0, 1)$  byla pravděpodobnost chyby II. druhu minimální, tzn. aby pro  $\forall \theta \in \Theta_1$  bylo  $\beta_{\phi^*}(\theta)$  stejnoměrně na  $\Theta_1$  **maximální silou testu**, za podmínky, že pravděpodobnost chyby I. druhu bude stále (stejnoměrně na  $\Theta_0$ ) pod hranicí  $\alpha$ , tzn.

$$\forall \theta \in \Theta_0, \beta_{\phi^*}(\theta) \leq \alpha.$$

Číslo  $\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta)$  se nazývá **hladina testu** (*size of test  $\phi^*$* ) a v praxi může být ostře pod nastavenou hranicí **hladiny významnosti** testu  $\alpha$  (*significance level  $\alpha$* ).

Konkrétní hodnotu  $\beta_{\phi^*}(\theta)$  pro  $\theta \in \Theta_1$  nazýváme **síla testu  $\phi^*$**  pro dané  $\theta \in \Theta_1$ , celé zúžení  $\beta_{\phi^*}|_{\Theta_1}$  pak nazýváme **silofunkce** testu  $\phi^*$ .

Pokud takový test  $\phi^*$  splňující uvedené podmínky existuje, nazýváme ho stejnoměrně nejsilnějším testem  $H_0$  oproti  $H_1$ , ozn. **UMP test** (*uniformly most powerful test*). Situaci UMP ilustruje obrázek ??.

**Definice 3.4.** Pokud  $\Theta_0$  je jednoprvková (1 stav), pak  $H_0$  nazveme **jednoduchá** hypotéza (*simple*), v opačném případě je  $H_0$  **složená** hypotéza (*composite*). Totéž platí pro  $\Theta_1$  a  $H_1$  alternativu.

**Definice 3.5.** Pokud uvažujeme kritickou funkci ve tvaru

$$\phi(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in W \subset \mathbb{R}^n, \\ 0 & \text{jinak,} \end{cases}$$

pak  $W$  nazveme **kritický obor testu** (*critical region*). Je to tedy obor naměřených hodnot, při kterém zamítáme  $H_0$ . Tomuto tvaru testu se říká **neznáhodněný test** a o přijetí  $H_0$  rozhodujeme následovně:

$$\begin{aligned} \text{nastává jev } \{\mathbf{X} \in W\} &\Rightarrow \text{zamítáme } H_0, \\ \text{nastává opačný jev } \{\mathbf{X} \notin W\} &\Rightarrow \text{nezamítáme } H_0. \end{aligned}$$

### 3.3 Neyman-Pearsonovo lemma (N-PL)

**Věta 3.6** (Neyman-Pearsonovo lemma). Mějme dvě hypotézy  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta = \theta_1$  a číslo  $\alpha \in (0, 1)$  jako hladinu významnosti testu. Označme nyní hustotu pravděpodobnosti  $f_0 := f(\mathbf{x}, \theta_0)$  a  $f_1 := f(\mathbf{x}, \theta_1)$  obě vzhledem ke vhodné dominující míře  $\lambda$ . Pak existuje  $K > 0$  a UMP test  $\phi^*$  ve tvaru

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & f_1 > K f_0, \\ \gamma & f_1 = K f_0, \\ 0 & f_1 < K f_0, \end{cases} \quad \text{tak, že } \beta_{\phi^*}(\theta_0) = \alpha.$$

Pokud  $\phi$  je nějaký jiný UMP test na hladině  $\alpha$ , pak  $\phi$  je nutně stejného tvaru jako  $\phi^*$  na množině  $\{f_1 \neq K f_0\}$ . Výjimkou je situace, kdy existuje test  $\phi$  s  $\beta_{\phi}(\theta_1) = 1$ , přičemž  $\beta_{\phi}(\theta_0) < \alpha$ , což znamená, že test  $\phi$  nemůže dosáhnout zadané hranice  $\alpha$  pro svou pravděpodobnost chyby I. druhu tak, jako ji dosahuje test  $\phi^*$ .



### 3 Testování statistických hypotéz

**Důsledek 3.7.** Pokud platí, že  $\mathbb{P}_{\theta_0}(f_1 = Kf_0) = 0$ , pak můžeme psát

$$\phi^* = \begin{cases} 1 & x \in W^* = \{f_1 \geq Kf_0\}, \\ 0 & x \in (W^*)^c = \{f_1 < Kf_0\}. \end{cases}$$

Tedy v případě, že hranice  $\{f_1 = Kf_0\}$  je nulové  $\mathbb{P}_{\theta_0}$  míry, potom existuje neznáhodněný test s UMP kritickou oblastí  $W^*$  pro testování  $H_0$  versus  $H_1$ , přičemž pravděpodobnost chyby I. druhu je přímo rovna požadované signifikanci  $\alpha$ ,  $\beta_{W^*}(\theta_0) = \alpha$ , zatímco síla testu  $\beta_{W^*}(\theta_1)$  je maximální možná.

## 3.4 Složené hypotézy a MLR systémy

**Postup použití N-PL v praxi pro test z důsledku 3.7.**

Hledáme takový test  $\phi^*$  tvaru

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in \{f_1 \geq Kf_0\} = W^* \dots \text{UMP CR}, \\ 0 & \mathbf{x} \in \{f_1 < Kf_0\} = (W^*)^c, \end{cases}$$

pro který je dosažena hladina testu  $\beta_{\phi^*}(\theta_0) = \alpha$ , přičemž síla (sílofunkce) testu  $\beta$  je optimální.

- 1) Nejdříve najdeme **tvar**  $W^*$  jako řešení nerovnice  $f_1 \geq Kf_0$ . Získáme ho v nějakém tvaru  $W^* = \{T(\mathbf{x}) \geq K'\}$ , resp.  $W^* = \{T(\mathbf{x}) \leq K'\}$ , s blíže nespecifikovanou volnou konstantou  $K'$ , tedy

$$\{f_1 \geq Kf_0\} \sim \{T(\mathbf{X}) \geq K'\},$$

resp.

$$\{f_1 \geq Kf_0\} \sim \{T(\mathbf{X}) \leq K'\},$$

kde  $T(\mathbf{X})$  se nazývá **testovací statistika**.

- 2) Konkrétní hodnotu  $K'$  pak určíme z rovnice  $\mathbb{P}_{\theta_0}(T(\mathbf{X}) \geq K') = \alpha$ , resp.  $\mathbb{P}_{\theta_0}(T(\mathbf{X}) \leq K') = \alpha$ . K vyřešení této nerovnosti však nutně potřebujeme umět vyjádřit  $\mathbb{P}_{\theta_0}(T \geq K')$ , resp.  $\mathbb{P}_{\theta_0}(T \leq K')$  za předpokladu platnosti hypotézy  $H_0$ , tzn. při  $\theta_0$ . Odvození rozdělení  $T(\mathbf{X})$  při  $\theta_0$  se říká "distributional problem" testování hypotéz a jde o stěžejní část úspěšné aplikace.
- 3) V případě, že  $H_1$  je složená, tzn.  $H_1 : \theta \in \Theta_1$ , postupujeme takto: volíme  $\theta_1 \in \Theta_1$  libovolně pevně a testujeme hypotézu  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta = \theta_1$  na hladině  $\alpha$ . Z Neyman-Pearsonova lemmatu existuje  $\text{UMP}_\alpha$  test  $\phi^*$ , případně  $\text{UMPCR}$   $W^*$ . Pokud tento  $\phi^*$ , případně  $W^*$ , nezávisí na volbě  $\theta_1$ , máme finální  $\text{UMP}_\alpha$  test při složené alternativě  $H_1$ .
- 4) Pokud i  $H_0$  je složená, tzn.  $H_0 : \theta \in \Theta_0$ , pak, pokud to lze, ještě navíc ukážeme, že  $\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta) \leq \alpha$ , tzn., že  $\forall \theta'_0 \in \Theta$ , platí, že  $\beta_{\phi^*}(\theta'_0) \leq \alpha$ . Průchodnost bodů 3) a 4) zajišťuje například následující koncept MLR.

**Definice 3.8.** Systém hustot  $\mathcal{F}$  se nazývá **MLR** (*Monotone likelihood ratio*), pokud  $\exists T(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^1$  tak, že pro  $\forall \theta_0 < \theta_1$  platí, že  $\frac{f_1}{f_0}$  je monotónní funkcí statistiky  $T(\mathbf{x})$ , tzn.  $\frac{f_1}{f_0} = g(T(\mathbf{x}))$ , kde  $g$  je monotónní. Podíl  $\frac{f_1}{f_0} = \frac{L(\theta_1)}{L(\theta_0)}$  se nazývá **věrohodnostním poměrem** (*likelihood ratio*), ozn.  $\text{LR}(\mathbf{x})$ ,  $\mathbf{x} \in \mathcal{X}$ .

### 3 Testování statistických hypotéz

**Věta 3.9.** Mějme rostoucí MLR systém hustot  $\mathcal{F}$  se statistikou  $T(\mathbf{X})$ . Testujeme hypotézu  $H_0 : \theta \leq \theta_0$  vs.  $H_1 : \theta > \theta_0$ ,  $\theta \in \mathbb{R}^1$ , tzv. jednostrannou hypotézu oproti jednostranné alternativě, na zadané hladině  $\alpha \in (0, 1)$ . Pak existuje  $\text{UMP}_\alpha$  test  $\phi^*$  ve tvaru

$$\phi^*(\mathbf{x}) = \begin{cases} 1 & T(\mathbf{x}) > K, \\ \gamma & T(\mathbf{x}) = K, \\ 0 & T(\mathbf{x}) < K, \end{cases}$$

přičemž  $K$  a  $\gamma$  jsou určeny podmínkou  $\beta_{\phi^*}(\theta_0) = \alpha$ , tzn.  $\mathbb{E}_{\theta_0}[\phi^*(\mathbf{X})] = \alpha$ , tedy

$$\mathbb{P}_{\theta_0}(T(\mathbf{X}) > K) + \gamma \mathbb{P}_{\theta_0}(T(\mathbf{X}) = K) + 0 \cdot \mathbb{P}_{\theta_0}(T(\mathbf{X}) < K) = \alpha.$$

Pro případ klesajícího MLR systému  $\mathcal{F}$  stačí v tvrzení zaměnit nerovnosti za opačné.

### 3.5 Nestranné UMP testy (UMPU)

V aplikacích TSH v praxi vyvstává nutnost testovat další složitější hypotézy, jako například

$$\begin{aligned} H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta \neq \theta_0, \quad \text{nebo} \\ H_0 : \theta_1 \leq \theta \leq \theta_2 \text{ vs. } H_1 : \theta \notin [\theta_1, \theta_2]. \end{aligned}$$

Pro takovéto případy, kdy alternativní hypotéza  $H_1$  je tzv. oboustranná ( $\theta < \theta_1$  nebo  $\theta > \theta_2$ ), zpravidla neexistuje stejnoměrně nejsilnější  $\text{UMP}_\alpha$  test  $\phi^*$  na hladině  $\alpha$  a jsme nuceni z optimality UMP testu slevit. Zavedeme proto UMPU testy.

**Definice 3.10.** Testujme  $H_0 : \theta \in \Theta_0$  vs.  $H_1 : \theta \in \Theta_1$ . Test  $\phi$  se nazývá nestranný, pokud  $\sup_{\theta \in \Theta_0} \beta_\phi(\theta) \leq \inf_{\theta \in \Theta_1} \beta_\phi(\theta)$ .

**Věta 3.11.** Každý UMP test  $\phi^*$  je nestranný, tzn.  $\beta_{\phi^*}|_{\Theta_0} \leq \beta_{\phi^*}|_{\Theta_1}$ , tj.

$$\sup_{\theta \in \Theta_0} \beta_{\phi^*}(\theta) \leq \inf_{\theta \in \Theta_1} \beta_{\phi^*}(\theta).$$

**Definice 3.12.** Stejnoměrně nejsilnější test mezi všemi nestrannými testy se nazývá UMPU. (UMP Unbiased).

**Věta 3.13.** Testujeme hypotézu  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$ , kde  $\theta \in \Theta \subset \mathbb{R}^1$  a  $\theta_0$  je vnitřním bodem  $\Theta$ , tedy,  $\theta_0 \in \Theta^\circ$ . Nechť  $\mathcal{F}$  je exponenciální třída hustot z příkladu ?? s  $Q(\theta)$  ryze rostoucí a diferencovatelnou, tedy  $\mathcal{F}_n = \left\{ f(\mathbf{x}, \theta) \stackrel{iid}{=} \prod_{i=1}^n f(x_i, \theta) \right\}$  je MLR systém se statistikou  $T(\mathbf{x}) = \sum_{i=1}^n T(x_i)$ . Pak existuje  $\text{UMPU}_\alpha$  test tvaru

$$\phi_u^*(\mathbf{x}) = \begin{cases} 1 & T(\mathbf{x}) < K_1 \vee T(\mathbf{x}) > K_2, \\ \gamma_1 & T(\mathbf{x}) = K_1, \\ \gamma_2 & T(\mathbf{x}) = K_2, \\ 0 & T(\mathbf{x}) \in (K_1, K_2), \end{cases}$$

takový, že  $\beta_{\phi_u^*}(\theta_0) = \alpha$ . Konstanty  $K_1, K_2, \gamma_1, \gamma_2$  určíme tak, aby byla splněna podmínka

$$\mathbb{P}(T(\mathbf{X}) < K_1) + \mathbb{P}(T(\mathbf{X}) > K_2) + \gamma_1 \mathbb{P}(T(\mathbf{X}) = K_1) + \gamma_2 \mathbb{P}(T(\mathbf{X}) = K_2) = \alpha.$$

### 3 Testování statistických hypotéz

Na základě věty 3.13 umíme nalézt alespoň  $\text{UMP}_{\alpha}$  optimální testy mezi všemi nestrannými testy, např. pro hypotézy ve tvaru

$$\begin{aligned} H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu \neq \mu_0, & \quad \text{při Gaussovském modelu } X \sim \mathcal{N}(\mu, \sigma^2 \text{ známé}), \\ H_0 : \sigma^2 = \sigma_0^2 \text{ vs. } H_1 : \sigma^2 \neq \sigma_0^2, & \quad \text{při Gaussovském modelu } X \sim \mathcal{N}(\mu \text{ známé}, \sigma^2). \end{aligned}$$

## 4 Další metody testování hypotéz

### 4.1 Test poměrem věrohodností (LRT)

**Definice 4.1.** Mějme rodinu  $\mathcal{F} = \{f(x, \theta) : \theta \in \Theta\}$  a testujme obecnou hypotézu  $H_0 : \theta \in \Theta_0$  vs.  $H_1 : \theta \in \Theta_1$  na zadané hladině významnosti  $\alpha \in (0, 1)$ . Zaved'me funkci

$$\Lambda(\mathbf{x}) := \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}, \quad \text{kde } L(\theta) = f(\mathbf{x}, \theta)$$

je věrohodnostní funkcí testovaného modelu, založenou na vzorku  $\mathbf{x} \in \mathcal{X}$  z náhodného výběru  $\mathbf{X} = (X_j)_{j=1}^n$  iid  $f(x, \theta)$ . Definujme test tvaru

$$\phi_\Lambda(\mathbf{x}) = \begin{cases} 1 & \mathbf{x} \in W_\Lambda \subset \mathbb{R}^n, \\ 0 & \mathbf{x} \notin W_\Lambda, \end{cases}$$

kde  $W_\Lambda = \{\mathbf{x} \in \mathbb{R}^n : \Lambda(\mathbf{x}) \leq K\}$  je taková, že pro nějakou konstantu  $K \in [0, 1]$  platí  $\beta_{\phi_\Lambda}|_{\Theta_0} \leq \alpha$ , tzn.

$$\beta_{W_\Lambda}(\theta) := \beta_{\phi_\Lambda}(\theta) = \mathbb{E}_\theta \phi_\Lambda(\mathbf{X}) = \mathbb{P}_\theta(\phi_\Lambda(\mathbf{X}) = 1) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

Takové  $\phi_\Lambda$ , pokud existuje, se nazývá **LRT test** pro testování  $H_0 \times H_1$  na hladině významnosti  $\alpha$ , ozn.  $\text{LRT}_\alpha$ .  $W_\Lambda$  je odpovídající LRT kritická oblast tohoto testu ( $\text{LRT}_\alpha$  CR). Pokud nastal jev  $\{\mathbf{X} \in W_\Lambda\}$ , zamítáme  $H_0$ , pokud nastává opačný jev, pak  $H_0$  nezamítneme.

Jde o test založený na limitních vlastnostech statistického modelu, přičemž smysluplnost zavedení tohoto LRT testů vyplývá z lemmatu, dokázaného v sekci 2 MLE odhadů, které říká, že pro  $(X_j)_{j=1}^{+\infty}$  iid  $f(x, \theta_0)$ , kde  $\text{supp } f$  je nezávislý na  $\theta$ , platí, že

$$\mathbb{P}_{\theta_0}(L(\theta_0) > L(\theta)) \xrightarrow{n \rightarrow +\infty} 1, \quad \forall \theta \neq \theta_0.$$

a) Pokud  $H_0$  platí, a tedy skutečná hodnota parametru  $\theta_0$  leží jak v  $\Theta_0$ , tak v  $\Theta_0 \cup \Theta_1$ , pak

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)} = 1 \text{ s pravděpodobností } \mathbb{P}_{\theta_0} \text{ jdoucí k } 1 \text{ při } n \rightarrow +\infty.$$

b) Pokud  $H_0$  neplatí, a tedy skutečná hodnota parametru  $\theta_0$  neleží v  $\Theta_0$ , ale stále je obsažena ve  $\Theta_0 \cup \Theta_1$ , pak  $\Lambda(\mathbf{x}) \leq K < 1$  je ostře odraženo od 1 s pravděpodobností  $\mathbb{P}_{\theta_0}$  jdoucí k 1 při  $n \rightarrow +\infty$ . Právě tak jsme nastavili v definici  $\phi_\Lambda$  kritický obor  $W_\Lambda$  pro přijetí/zamítnutí  $H_0$ .

## 4.2 Analýza variance (ANOVA)

Analýza rozptylu (*analysis of variance*, ANOVA) je metoda, která umožňuje zjistit, jestli má na Gaussovskou náhodnou veličinu vliv některý ze znaků u jednotlivých jedinců, např. zda na plat zaměstnanců má vliv dosažené vzdělání, pohlaví, věk apod.

Mějme nezávislé náhodné výběry  $X_{i1}, \dots, X_{in_i} \sim \mathcal{N}(\mu_i, \sigma^2)$ ,  $i \in I$ ,  $N = \sum_{i=1}^I n_i$ . Potom sdružená hustota z  $\mathcal{F}_N$  je tvaru

$$f(\mathbf{x}|\mu_1, \dots, \mu_I, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^I \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2 \right\}.$$

Testujeme hypotézu

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I (= \mu) \quad \text{vs.} \quad H_1 : \text{alespoň jedna nerovnost}$$

na hladině  $\alpha \in (0, 1)$  za dodatečného předpokladu  $\sigma_1^2 = \dots = \sigma_I^2 = \sigma^2$  neznámé, tzn. předpokládáme homogenitu rozptylů jednotlivých testovaných podskupin  $i \in I$ .

### Odvození ANOVA LRT $_{\alpha}$ testu

Mějme

$$\Lambda(\mathbf{x}) = \frac{\sup\{f(\mathbf{x}|\mu, \mu, \dots, \mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}}{\sup\{f(\mathbf{x}|\mu_1, \mu_2, \dots, \mu_I, \sigma^2) : \mu_i \in \mathbb{R}, \sigma^2 > 0\}}.$$

Při řešení extrémů prostřednictvím diferenciálního počtu  $\partial_r f = 0$  v čitateli získáme 2 rovnice a ve jmenovateli  $I + 1$  rovnic, které vyřešíme a příslušné hodnoty maximálně věrohodných odhadů  $\hat{\mu}, \hat{\sigma}^2$ , resp.  $\hat{\mu}_i, \hat{\sigma}^2$ ,  $i \in I$ , zpětně dosadíme do  $\Lambda(\mathbf{x})$ . Dále potom nalezneme tvar LRT kritické oblasti  $W_{\Lambda} = \{\mathbf{x} : \Lambda(\mathbf{x}) \leq K\}$ , kdy platí

$$\Lambda(\mathbf{x}) \leq K \quad \Leftrightarrow \quad \mathcal{F}_{\Lambda}(\mathbf{x}) = \frac{(N - I)S_A}{(I - 1)S_e} \geq C,$$

kde

$$S_A = \sum_{i=1}^I n_i (\bar{x}_i - \bar{\bar{x}})^2, \quad S_e = \sum_{i=1}^I \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \text{a} \quad \bar{\bar{x}} = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} x_{ij}.$$

**PŘÍKLAD 4.2.** Distribuční problém tohoto LRT testu ANOVA spočívá v odvození rozdělení testovací statistiky  $\mathcal{F}_{\Lambda}(\mathbf{x})$  za předpokladu platnosti  $H_0 : \mu_1 = \dots = \mu_I = \mu$ . Postupujeme následovně:

$$\begin{aligned} \sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij}^2 &= \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i + \bar{X}_i)^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + \underbrace{2 \sum_{i=1}^I \sum_{j=1}^{n_i} \bar{X}_i (X_{ij} - \bar{X}_i)}_0 + \sum_{i=1}^I n_i \bar{X}_i^2 = \\ &= \underbrace{\sum_{i=1}^I \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}_{Q_1 = S_e} + \sum_{i=1}^I n_i (\bar{X}_i - \bar{\bar{X}} + \bar{\bar{X}})^2 = \\ &= S_e + \underbrace{\sum_{i=1}^I n_i (\bar{X}_i - \bar{\bar{X}})^2}_{Q_2 = S_A} + \underbrace{2 \sum_{i=1}^I n_i \bar{\bar{X}} (\bar{X}_i - \bar{\bar{X}})}_0 + \underbrace{\sum_{i=1}^I \bar{\bar{X}}^2 n_i}_{N \cdot \bar{\bar{X}}^2 =: Q_3} = S_e + S_A + Q_3 = \sum_{i=1}^3 Q_i, \end{aligned}$$

#### 4 Další metody testování hypotéz

což je součet tří kvadratických forem. Dá se ukázat (viz lineární algebra), že součet hodnotí těchto tří kvadratických forem dává plnou dimenzi úlohy  $N$ ,

$$\sum_{i=1}^3 h(Q_i) = h(S_e) + h(S_A) + h(Q_3) = (N - I) + (I - 1) + 1 = N.$$

Z Cochranovy věty pak vyplývá, že  $Q_i$  jsou **nezávislé** a  $Q_i(\mathbf{X}) \sim \chi^2(h(Q_i))$ , důsledkem čehož

$$\mathcal{F}_\Lambda(\mathbf{x})|_{H_0} = \frac{S_A/(I-1)}{S_e/(N-I)} \sim \frac{\chi^2(I-1)/(I-1)}{\chi^2(N-I)/(N-I)} \sim F(I-1, N-I),$$

tedy  $\mathcal{F}_\Lambda(\mathbf{x})$  má za platnosti  $H_0$  Fisherovo rozdělení s  $(I-1)$  a  $(N-I)$  stupni volnosti.

Nyní hledáme konstantu  $C$  tak, aby platilo  $\mathbb{P}_{H_0}(\mathcal{F}_\Lambda(\mathbf{X}) \geq C) = \alpha$ , což vede na LRT kritický obor  $W_\Lambda = \{\mathbf{x} : \mathcal{F}_\Lambda(\mathbf{x}) \geq F_{1-\alpha}(I-1, N-I)\}$ , kde  $F_{1-\alpha}$  značí  $(1-\alpha)$ -kvantil příslušného Fisherova rozdělení  $F$ . Skončí-li experiment v tomto kritickém oboru, pak zamítáme  $H_0$ .

### 4.3 Dvouvýběrové testy ( $2 \times \mathcal{N}_1$ )

**Dvouvýběrový nepárový t-test** porovnává střední hodnoty dvou Gaussovských výběrů. Příkladem toho může být třeba střední hodnota tlaku krve u kuřáků a nekuřáků, atp.

**PŘÍKLAD 4.3** (Dvouvýběrový t-test). Uvažujme dva náhodné výběry ze dvou různých Gaussovských distribucí

$$\begin{aligned} X_1, \dots, X_{n_1} \text{ iid } \mathcal{N}(\mu_1, \sigma_1^2) &\Rightarrow \bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \\ Y_1, \dots, Y_{n_2} \text{ iid } \mathcal{N}(\mu_2, \sigma_2^2) &\Rightarrow \bar{Y}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right), \quad \frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}. \end{aligned}$$

Budeme testovat hypotézu shodnosti středních hodnot obou souborů  $H_0 : \mu_1 = \mu_2$  vs.  $H_1 : \mu_1 \neq \mu_2$  na hladině  $\alpha$ . Rozlišujeme tři případy:

a) Známe-li  $\sigma_1^2, \sigma_2^2$ , potom

$$\frac{\bar{X}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1),$$

protože z reprodukční vlastnosti  $\mathcal{N}$  víme, že  $(\bar{X}_1 - \bar{Y}_2) \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$ .

Při  $H_0 : \mu_1 = \mu_2$  pak nalezneme rozdělení testovací statistiky

$$U = U(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1).$$

Vyřešením rovnice  $\mathbb{P}(|U| \geq K_\alpha) = \alpha$  dostaneme  $K_\alpha = u_{1-\frac{\alpha}{2}}$  s následným RT kritickým oborem  $W_\alpha = \{(\mathbf{x}, \mathbf{y}) : |U(\mathbf{x}, \mathbf{y})| \geq u_{1-\frac{\alpha}{2}}\}$ , kde  $u_{1-\frac{\alpha}{2}}$  značí příslušný kvantil  $\mathcal{N}(0, 1)$  rozdělení.

#### 4 Další metody testování hypotéz

- b) Pokud neznáme  $\sigma_1^2, \sigma_2^2$ , ale víme, že  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  (analogie ANOVA pro  $I = 2$ , kde  $\sigma^2$  neznáme), pak volíme testovací statistiku jako

$$T = T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \text{ při platnosti } H_0,$$

kde  $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$  se nazývá *pooled sample variance*. Studentovo  $t(n_1 + n_2 - 2)$  rozdělení plyne z faktu, že  $T = \frac{U}{s/\sigma}$ , přičemž

$$(n_1 + n_2 - 2) \frac{s^2}{\sigma^2} = \left[ \frac{(n_1 - 1)s_1^2}{\sigma_1^2} + \frac{(n_2 - 1)s_2^2}{\sigma_2^2} \right] \sim \chi^2(n_1 - 1) + \chi^2(n_2 - 1) \stackrel{id}{\sim} \chi^2(n_1 + n_2 - 2),$$

což plyne z reprodukční vlastnosti  $\chi^2$  rozdělení. Podobně jako v a) dostáváme RT kritický obor  $W_\alpha = \left\{ |T(\mathbf{x}, \mathbf{y})| \geq t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) \right\}$ , kde  $t_{1-\frac{\alpha}{2}}$  značí kvantil příslušného  $t(n_1 + n_2 - 2)$  Studentova rozdělení.

- c) Pokud  $\sigma_1^2, \sigma_2^2$  neznáme, ale víme, že  $\sigma_1^2 \neq \sigma_2^2$ , pak užíváme testovací statistiku

$$T_\nu = T_\nu(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu) \quad (\text{Welchova aproximace}),$$

kde

$$\nu = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{1}{n_1-1} \left( \frac{s_1^2}{n_1} \right)^2 + \frac{1}{n_2-1} \left( \frac{s_2^2}{n_2} \right)^2}.$$

Následný kritický obor  $W_\alpha = \left\{ |T_\nu| \geq t_{1-\frac{\alpha}{2}}(\nu) \right\}$  definuje dvouvýběrový **t-test**. Pro ne-celočíselné  $\nu$  interpolujeme  $t(\nu)$  z hodnot sousedních  $t([\nu])$  a  $t([\nu] + 1)$ .

**PŘÍKLAD 4.4** (Test homogenity rozptylů = F-test). Za stejných předpokladů jako u dvouvýběrového t-testu z příkladu 4.3 testujeme hypotézu homogenity rozptylů dvou Gaussovských výběrů

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2 \quad \text{na hladině } \alpha \in (0, 1).$$

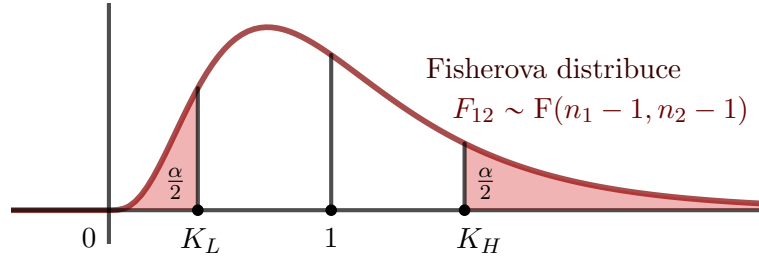
Testovací statistiku volíme

$$F_{12} = F(\mathbf{X}, \mathbf{Y}) = \frac{s_1^2}{s_2^2} = \left| H_0 : \sigma_1^2 = \sigma_2^2 \right| \xrightarrow{H_0} \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}}{\frac{(n_2-1)s_2^2}{\sigma_2^2}} \cdot \frac{n_2-1}{n_1-1} \sim \frac{\frac{\chi^2(n_1-1)}{n_1-1}}{\frac{\chi^2(n_2-1)}{n_2-1}} \stackrel{id}{\sim} F(n_1-1, n_2-1),$$

za platnosti  $H_0$ . Pak RT kritický obor F-testu je při symetrické volbě kvantilů Fisherova F rozdělení

$$W_\alpha = \left\{ (\mathbf{x}, \mathbf{y}) : F(\mathbf{x}, \mathbf{y}) \geq F_{1-\frac{\alpha}{2}}(n_1-1, n_2-1) \text{ nebo } F(\mathbf{x}, \mathbf{y}) \leq F_{\frac{\alpha}{2}}(n_1-1, n_2-1) \right\},$$

viz obrázek 4.1. Tuto volbu odůvodňuje fakt, že  $s_{1,2}^2 \xrightarrow{s.j.} \sigma_{1,2}^2$  a  $\mathbb{E}s_{1,2}^2 = \sigma_{1,2}^2$ .



Obrázek 4.1: Kritický obor F-testu homogenity rozptylů.

#### 4.4 Test koeficientu korelace ( $\mathcal{N}_2$ )

Předpokládejme  $(X_j, Y_j)_{j=1}^n \text{ iid } \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$  z dvourozměrného nede degenerovaného Gaussova rozdělení při  $\sigma_1 > 0$ ,  $\sigma_2 > 0$ ,  $|\varrho| < 1$ . Testujeme nekorelovanost  $X, Y$ , tzn. nulovou hodnotu korelačního koeficientu  $\varrho = \varrho(X, Y)$ :

$$H_0 : \varrho = 0 \text{ vs. } H_1 : \varrho \neq 0 \quad (\text{tzn. test nezávislosti } X \text{ a } Y \text{ v } \mathcal{N}_2 \text{ modelu})$$

na hladině významnosti  $\alpha \in (0, 1)$ .

**Odvodíme LRT test  $H_0$ :**

Logaritmická věrohodnostní funkce modelu  $\mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$  při označení  $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$  je

$$\begin{aligned} l(\theta) = \ln L(\theta) = & \frac{-1}{2(1-\varrho^2)} \left[ \sum_{j=1}^n \frac{(X_j - \mu_1)^2}{\sigma_1^2} + \sum_{j=1}^n \frac{(Y_j - \mu_2)^2}{\sigma_2^2} - \right. \\ & \left. - 2\varrho \sum_{j=1}^n \frac{(X_j - \mu_1)(Y_j - \mu_2)}{\sigma_1 \sigma_2} \right] - n \ln(2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}). \end{aligned}$$

Následně vyhodnotíme výraz

$$\Lambda(\mathbf{x}, \mathbf{y}) = \frac{\sup\{L(\theta) : \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho = 0\}}{\sup\{L(\theta) : \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, |\varrho| < 1\}} \sim \frac{4 \text{ rovnice typu } \partial_r \ln L = 0}{5 \text{ rovnic typu } \partial_l \ln L = 0}.$$

Řešením obou extrémů jsou MLE odhady:  $\hat{\mu}_1 = \bar{X}_n$ ,  $\hat{\mu}_2 = \bar{Y}_n$ ,  $\hat{\sigma}_1^2 = \hat{\sigma}_{n,X}^2$  a  $\hat{\sigma}_2^2 = \hat{\sigma}_{n,Y}^2$  a dále

$$\hat{\varrho}_{XY} = \hat{\varrho}_n(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{j=1}^n (X_j - \bar{X}_n)(Y_j - \bar{Y}_n)}{\sqrt{\sum_{j=1}^n (X_j - \bar{X}_n)^2 \sum_{j=1}^n (Y_j - \bar{Y}_n)^2}} = \frac{\widehat{\text{Cov}}_{XY}}{\hat{\sigma}_1 \cdot \hat{\sigma}_2},$$

který se nazývá **Pearsonův výběrový koeficient korelace**. Dosazením těchto odhadů do  $\Lambda(\mathbf{x}, \mathbf{y})$  získáme

$$\begin{aligned} \ln \Lambda(\mathbf{x}, \mathbf{y}) &= \ln L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \varrho = 0) - \ln L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2, \hat{\varrho}_{XY}) = \\ &= -n \ln(2\pi\hat{\sigma}_1\hat{\sigma}_2) + n \ln(2\pi\hat{\sigma}_1\hat{\sigma}_2\sqrt{1-\hat{\varrho}_{XY}^2}) = \\ &= \frac{n}{2} \ln(1-\hat{\varrho}_{XY}^2) \leq K \Leftrightarrow |\hat{\varrho}_{XY}| \geq K' \Leftrightarrow \frac{|\hat{\varrho}_{XY}|\sqrt{n-2}}{\sqrt{1-\hat{\varrho}_{XY}^2}} \geq K'', \end{aligned}$$



#### 4 Další metody testování hypotéz

kde  $K, K', K''$  jsou vhodné konstanty nezávislé na  $(\mathbf{x}, \mathbf{y})$ . Lze ukázat, že při platnosti  $H_0 : \varrho = 0$  má testovací LRT statistika rozdělení

$$T = T(\mathbf{X}, \mathbf{Y}) = \frac{\hat{\varrho}_{XY} \sqrt{n-2}}{\sqrt{1 - \hat{\varrho}_{XY}^2}} \sim t(n-2),$$

tedy Studentovo rozdělení s  $(n-2)$  stupni volnosti. To vede na kritický obor testu  $H_0 : \varrho = 0$  ve tvaru

$$W_\alpha = \left\{ (\mathbf{x}, \mathbf{y}) : |T(\mathbf{x}, \mathbf{y})| \geq t_{1-\frac{\alpha}{2}}(n-2) \right\},$$

opět s použitím  $\left(1 - \frac{\alpha}{2}\right)$ -kvantilu příslušného Studentova rozdělení.

## 5 Asymptotické testy hypotéz

### 5.1 Asymptotické testy středních hodnot iid $\mathcal{L}_2$

**Věta 5.1** (Jednovýběrový asymptotický test  $\mu = \mu_0$ ). Mějme náhodný výběr  $X_1, \dots, X_n$  iid  $\mathcal{L}_2$  pocházející z libovolného rozdělení s  $\mathbb{E}X_j = \mu$  a s konečným rozptylem  $\mathrm{D}X_j = \sigma^2 > 0$ , který je neznámý. Testujeme hypotézu  $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$  (resp.  $\mu \geq \mu_0$  apod.) na hladině  $\alpha \in (0, 1)$ . Pak testovací statistika

$$T_n = T_n(\mathbf{X}) = \sqrt{n} \frac{\bar{X}_n - \mu_0}{s_n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

za platnosti  $H_0$ . Následně test  $H_0 : \mu = \mu_0$ , založený na kritické oblasti  $W_\alpha = \{|T_n(\mathbf{x})| \geq u_{1-\frac{\alpha}{2}}\}$ , kde  $u_{1-\frac{\alpha}{2}}$  značí kvantil  $\mathcal{N}(0, 1)$ , zamítá  $H_0$  na **asymptotické hladině**  $\alpha$ .

**Věta 5.2.** Nechť  $X$  a  $Y$  jsou nezávislé z  $\mathcal{L}_2$  a mějme dva náhodné výběry (např. testovací a kontrolní)  $(X_i)_{i=1}^{n_1}$  iid  $\mathcal{L}_2(\mu_1, \sigma_1^2 > 0)$  a  $(Y_j)_{j=1}^{n_2}$  iid  $\mathcal{L}_2(\mu_2, \sigma_2^2 > 0)$ . Pak

$$T_{12} = T_{12}(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1), \quad \text{při } n_1, n_2 \rightarrow +\infty.$$

**Důsledek 5.3** (Dvouvýběrový asymptotický test  $\mu_1 = \mu_2$ ). Díky větě 5.2 lze zkonstruovat asymptotický test pro testování hypotézy  $H_0 : \mu_1 = \mu_2$  v obecném  $\mathcal{L}_2$  modelu, kdy máme k dispozici dva nezávislé náhodné výběry ze dvou potenciálně zcela typově odlišných libovolných distribucí  $F_X$  a  $F_Y$ , o kterých víme pouze to, že obě distribuce mají konečné neznámé rozptyly  $\sigma_1^2 > 0$ ,  $\sigma_2^2 > 0$ . Test  $H_0 : \mu_1 = \mu_2$ , založený na kritické oblasti

$$W_\alpha = \left\{ |T_{12}(\mathbf{x}, \mathbf{y})| \geq u_{1-\frac{\alpha}{2}} \right\}, \quad \text{kde } T_{12}(\mathbf{x}, \mathbf{y}) = \frac{\bar{x}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

zamítá  $H_0$  na asymptotické hladině  $\alpha$ . Hranice zamítnutí  $u_{1-\frac{\alpha}{2}}$  zde opět označuje příslušný kvantil Gaussova rozdělení  $\mathcal{N}(0, 1)$ .

### 5.2 Asymptotický LRT a Waldův test v $\mathbb{R}^k$

**Věta 5.4.** Mějme  $\theta \in \Theta \subset \mathbb{R}^k$  a testujeme hypotézu  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$  na základě náhodného výběru  $X_1, \dots, X_n$  iid  $f \in \mathcal{F}$ . Nechť jsou dále splněny předpoklady z věty 2.24 o asymptotické normalitě MLE odhadů, tedy  $\mathcal{F} = \mathcal{F}_{\text{reg}}^{\text{ML}}$ , a nechť  $\mathbb{I}(\theta)$  je spojitá ( $k \times k$  Fisherova informační matice) v bodě  $\theta_0$ . Pak za platnosti  $H_0$  platí

$$\lambda_n(\mathbf{X}) = -2 \ln \Lambda(\mathbf{X}) \xrightarrow{\mathcal{D}} \chi^2(k).$$

## 5 Asymptotické testy hypotéz

**Důsledek 5.5** (Asymptotický LRT). *Za předpokladů věty 5.4 je test hypotézy  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$ , založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{x} : \lambda_n(\mathbf{x}) \geq \chi_{1-\alpha}^2(k) \right\},$$

kde  $\chi_{1-\alpha}^2(k)$  značí  $(1 - \alpha)$ -kvantil příslušného  $\chi^2(k)$  rozdělení, je tzv. **asymptotickým LRT testem**, který zamítá  $H_0$  na asymptotické hladině  $\alpha$ .

### 5.3 Testy dobré shody (GoF)

**Definice 5.6.** Mějme náhodnou veličinu (vlastnost)  $X$  na  $(\Omega, \mathcal{A}, \mathbb{P})$  a příslušný náhodný výběr  $X_1, \dots, X_n$  pocházející z nějaké neznámé distribuce  $F$ . Volme jednu konkrétní distribuci  $F_0 \in \mathcal{F}$ . Pak test hypotézy

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0 \quad (\text{resp. } F = F_1)$$

se nazývá **testem dobré shody** (*GoF* - *Goodness-of-Fit*) modelu  $F_0$ . Opět volíme hladinu významnosti  $\alpha \in (0, 1)$  a testujeme  $H_0$  na této signifikantní hranici pro chybu I.druhu. Tedy, za kritickou chybu považujeme rozhodnutí o zamítnutí modelu  $F_0$ , přestože ten je správný.

#### $\chi^2$ -testy GoF

Za jeden z nejznámějších GoF testů lze považovat následující  $\chi^2$ -test, který převádí celou úlohu na specifický *asymptotický* test v Multinomickém parametrickém modelu za cenu jistého *binování* (diskretizace) dostupného náhodného výběru  $\mathbf{X}$ . Označme  $H_X$  obor hodnot náhodné veličiny  $X \sim F$  a vytvořme dělení  $\{A_1, \dots, A_k\}$  oboru hodnot  $H_X$  na  $k$  disjunktních boxů či tříd (binů). Dále zavedeme

$$p_j = \mathbb{P}_F(X \in A_j), \quad p_{0j} = \mathbb{P}_{F_0}(X \in A_j), \quad \forall j \in \hat{k}.$$

Mějme nyní k dispozici náhodný výběr  $\mathbf{X} = (X_i)_{i=1}^n$  iid  $F$  a nechť

$$Y_j = \#\{i : X_i \in A_j\} = \sum_{i=1}^n \mathbb{I}\{X_i \in A_j\} = \sum_{i=1}^n \mathbb{I}_{A_j}(X_i)$$

je počet těch pozorování  $X_i$  z  $\{X_1, \dots, X_n\}$ , která se vyskytují v  $j$ -tém binu  $A_j$ ,  $j \in \hat{k}$ . Vzhledem k iid předpokladu pro jednotlivá  $X_i$  pak plyne, že každé  $Y_j \sim \text{Bi}(n, p_j)$ , a proto i celý vektor  $\mathbf{Y} = (Y_1, \dots, Y_k)$  má Multinomické rozdělení rozdělení  $\mathbf{Y} \sim \text{Mult}(n, \mathbf{p})$ , při označení  $\mathbf{p} = (p_1, \dots, p_k)$ . Namísto testu  $H_0 : F = F_0$  vs.  $H_1 : F \neq F_0$  pak testujeme parametrickou hypotézu na hladině  $\alpha > 0$

$$H_0 : \mathbf{p} = \mathbf{p}_0 \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \mathbf{p}_0 \quad \text{kde } \mathbf{p}_0 = (p_{01}, \dots, p_{0k})$$

v Multinomickém modelu  $\mathbf{Y} \sim \text{Mult}(n, \mathbf{p})$ , přičemž  $\sum_{j=1}^k p_j = \sum_{j=1}^k p_{0j} = 1$ .

## 5 Asymptotické testy hypotéz

**Věta 5.7.** *Nechť  $\hat{p}_j = Y_j/n$  značí MLE odhady parametrů  $p_j$  v binomickém modelu  $\text{Bi}(n, p_j)$ ,  $\forall j \in \hat{k}$ . Následující tři testovací statistiky v  $\text{Mult}(n, \mathbf{p})$  modelu*

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{n(\hat{p}_j - p_{0j})^2}{p_{0j}} = \sum_{j=1}^k \frac{(Y_j - np_{0j})^2}{np_{0j}} \quad (\text{Pearsonova}),$$

$$\tilde{\chi}^2(\mathbf{Y}) = \sum_{j=1}^k \frac{n(\hat{p}_j - p_{0j})^2}{\hat{p}_j} = \sum_{j=1}^k \frac{(Y_j - np_{0j})^2}{Y_j} \quad (\text{Neymanova}),$$

$$\lambda_n(\mathbf{Y}) = -2 \ln \Lambda(\mathbf{Y}) = -2 \ln \prod_{j=1}^k \left( \frac{p_{0j}}{\hat{p}_j} \right)^{n\hat{p}_j} = -2 \sum_{j=1}^k Y_j \ln \left( \frac{np_{0j}}{Y_j} \right) \quad (\text{LRT stat.}),$$

jsou za platnosti  $H_0 : \mathbf{p} = \mathbf{p}_0$  asymptoticky ekvivalentní a všechny konvergují v distribuci k limitnímu  $\chi^2(k-1)$  rozdělení.

**Důsledek 5.8** (Pearson  $\chi^2$  GoF). *Test  $H_0 : \mathbf{p} = \mathbf{p}_0$  založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{y} : \chi^2(\mathbf{y}) \geq \chi_{1-\alpha}^2(k-1) \right\} \quad (\chi_{1-\alpha}^2 \text{ značí kvantil } \chi^2 \text{ rozdělení})$$

je asymptotickým Pearsonovým  $\chi^2$ -testem dobré shody dosahujícím asymptotické signifikance  $\alpha$ . Podobně pro  $\tilde{\chi}^2(\mathbf{Y})$  a  $\lambda_n(\mathbf{Y})$  testovací statistiky.

Jde o asymptotický test, tedy je vyžadován jednak dostatečný počet  $n$  pozorování  $(x_i)_1^n$ , obvykle alespoň  $n \geq 50$ , a současně by měla být splněna podmínka  $np_{0j} \geq 5$ ,  $\forall j \in \hat{k}$ . Hodnoty testovacích statistik  $\chi^2(\mathbf{y})$  a  $\tilde{\chi}^2(\mathbf{y})$  představují součet vážených kvadratických odchylek tzv. pozorovaných četností  $y_j$  od teoretických četností  $np_{0j}$  přes všechny biny  $A_j$ ,  $j \in \hat{k}$ .

### 5.4 Modifikace $\chi^2$ -testů dobré shody

Pokud potřebujeme testovat shodu dat s modelem ve složené podobě, tzn.

$$H_0 : \mathbf{F} = \mathbf{F}_\theta \quad \text{vs.} \quad H_1 : \mathbf{F} \neq \mathbf{F}_\theta \quad \text{na hladině } \alpha \in (0, 1),$$

kde  $\theta \in \Theta \subset \mathbb{R}^s$  je neznámý parametr takový, že  $\dim(\Theta) = s$ , pak postup testování upravíme následovně. Označíme opět

$$p_j = \mathbb{P}_{\mathbf{F}}(X \in A_j), \quad p_{0j} = p_{0j}(\theta) = \mathbb{P}_{\mathbf{F}_\theta}(X \in A_j), \quad \forall j \in \hat{k},$$

kde nyní  $\mathbf{p} = (p_1, \dots, p_k)$  a  $\mathbf{p}_0 = \mathbf{p}_0(\theta) = (p_{01}(\theta), \dots, p_{0k}(\theta))$ . Na základě binovaných náhodných veličin  $(Y_j)_{j=1}^k$  testujeme v  $\text{Mult}(n, \mathbf{p})$  modelu parametrickou hypotézu

$$H_0 : \mathbf{p} = \mathbf{p}_0(\theta) \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \mathbf{p}_0(\theta) \quad \text{na hladině } \alpha > 0.$$

Protože však nyní vektor  $\mathbf{p}_0$  je funkcí neznámého parametru  $\theta$ , musíme tento parametr odhadnout za platnosti  $H_0$ . Takový vhodný odhad  $\hat{\theta}_n = \hat{\theta}_n(\mathbf{Y})$  by měl disponovat dostatečně dobrými asymptotickými vlastnostmi při  $n \rightarrow +\infty$ . Dostupnou realizaci odhadu  $\hat{\theta}_n(\mathbf{y})$  pak dosadíme do funkce  $\mathbf{p}_0(\theta)$  a testujeme již jednoduchou bodovou hypotézu v Multinomickém modelu na hladině  $\alpha$

$$H_0 : \mathbf{p} = \hat{\mathbf{p}}_0 \quad \text{vs.} \quad H_1 : \mathbf{p} \neq \hat{\mathbf{p}}_0, \quad \text{kde } \hat{\mathbf{p}}_0 = \hat{\mathbf{p}}_0(\hat{\theta}_n(\mathbf{y})).$$

Toto zanesení odhadu  $\hat{\theta}_n$  do testované hypotézy  $H_0$  však modifikuje asymptotické vlastnosti použitých testovacích statistik.

## 5 Asymptotické testy hypotéz

**Věta 5.9.** *Mějme test  $H_0 : \mathbf{p} = \hat{\mathbf{p}}_0$  vs.  $H_1 : \mathbf{p} \neq \hat{\mathbf{p}}_0$  za výše uvedených podmínek, kdy  $\dim(\Theta) = s$ . Nechť  $\hat{\theta}_n$  značí maximálně věrohodný odhad  $\hat{\theta}_{\text{ML}}(\mathbf{Y})$  za platnosti  $H_0$ . Pak testovací statistika*

$$\chi^2(\mathbf{Y}) = \sum_{j=1}^k \frac{(Y_j - n\hat{p}_{0j})^2}{n\hat{p}_{0j}} \xrightarrow{\mathcal{D}} \chi^2(k - s - 1)$$

*a test založený na kritické oblasti*

$$W_\alpha = \left\{ \mathbf{y} : \chi^2(\mathbf{y}) \geq \chi_{1-\alpha}^2(k - s - 1) \right\} \quad (\chi_{1-\alpha}^2 \text{ značí kvantil } \chi^2 \text{ rozdělení})$$

*je Pearsonovým  $\chi^2$ -testem dobré shody dosahujícím asymptotické signifikance  $\alpha$ . Podobně pro testovací statistiky  $\tilde{\chi}^2(\mathbf{Y})$  a  $\lambda_n(\mathbf{Y})$ .*

## 6 Konfidenční množiny, Intervaly spolehlivosti

Bodové odhady  $\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n$ , resp.  $T_n(\mathbf{X}) = T_n$ , sice poskytují odhad hodnoty daného parametru, nicméně například ve spojitém statistickém modelu ( $\text{ASR}_\lambda$ ) vždy platí, že  $\mathbb{P}(\hat{\theta}_n(\mathbf{X}) = \theta_0) = 0$ ,  $\forall \theta_0$ , resp.  $\mathbb{P}(T_n(\mathbf{X}) = \tau(\theta)) = 0$ ,  $\forall \theta \in \Theta$ . Z tohoto důvodu se zavádí tzv. interval spolehlivosti, což je interval, ve kterém se neznámý parametr  $\theta$  nachází s námi zadanou pravděpodobností (například 95%). Přesnost takového odhadu je pak dán šířkou tohoto intervalu.

**Definice 6.1.** Nechť  $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P} \in \mathcal{P}$  na  $(\Omega, \mathcal{A})$ , označme  $\theta = \theta(\mathbb{P}) \in \Theta \subset \mathbb{R}^1$  funkcionál na  $\mathcal{P}$ , resp.  $\theta \in \Theta \subset \mathbb{R}^k$ . Označíme  $\mathcal{B}_\Theta$  systém borelovských podmnožin  $\Theta$  a zvolíme číslo  $\alpha \in (0, 1)$ . Pak  $C(\mathbf{X}) \in \mathcal{B}_\Theta$  se nazývá **konfidenční množina** ( $\text{CM}_{1-\alpha}$ ) pro  $\theta$  na hladině  $(1 - \alpha)$ , pokud

$$\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\theta \in C(\mathbf{X})) \geq 1 - \alpha.$$

Číslo  $\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\theta \in C(\mathbf{X}))$  se pak nazývá **konfidenční koeficient** (koeficient spolehlivosti). Pokud speciálně  $C(\mathbf{X}) = [\underline{\theta}(\mathbf{X}), \bar{\theta}(\mathbf{X})] \subset \Theta$ , nazveme ji **konfidenční interval** ( $\text{CI}_{1-\alpha}$ ) (interval spolehlivosti) na hladině  $1 - \alpha$ .

### 6.1 Konstrukce $\text{CM}_{1-\alpha}$ pomocí pivotů (PQ)

**Definice 6.2.** Borelovsky měřitelná funkce  $\mathcal{R}(\mathbf{X}, \theta)$  se nazývá pivotální veličina (*pivot*) (PQ) pro parametr  $\theta$ , pokud **rozdělení**  $\mathcal{R}(\mathbf{X}, \theta)$  nezávisí na volbě  $\mathbb{P} \in \mathcal{P}$ .

**Metoda konstrukce  $\text{CM}_{1-\alpha}$  pro  $\theta$ :**

- 1) Nalezneme vhodnou **pivotální veličinu**  $\mathcal{R}(\mathbf{X}, \theta)$ , pokud taková existuje.
- 2) Volíme pevně  $\mathbb{P} \in \mathcal{P}$  a najdeme konstanty  $c_1, c_2$  takové, že platí

$$\mathbb{P}(c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2) \geq 1 - \alpha, \quad \text{případně} \quad \mathbb{P}(c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2) = 1 - \alpha.$$

- 3) Pak  $C(\mathbf{X}) = \left\{ \theta \in \Theta : c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2 \right\}$  je  $\text{CM}_{1-\alpha}$ .

Problémy konstrukce  $C(\mathbf{X})$ :

- a) existence alespoň jednoho pivotu  $\mathcal{R}(\mathbf{X}, \theta)$ ,
- b) existence mnoha různých  $\mathcal{R}(\mathbf{X}, \theta)$ : tedy nevíme, kterou z nich v konstrukci  $C(\mathbf{X})$  použít, tzn. která poskytuje nejmenší střední objem/délku  $C(\mathbf{X})$ ,
- c) volba  $c_1, c_2$ : kritériem může být například  $\alpha/2$ -symetrie, tzn. volba  $c_1, c_2$  tak, aby  $\mathbb{P}(\mathcal{R}(\mathbf{X}, \theta) > c_2) = \frac{\alpha}{2}$  a  $\mathbb{P}(\mathcal{R}(\mathbf{X}, \theta) < c_1) = \frac{\alpha}{2}$ .

- d) výpočet  $C(\mathbf{X})$  ze soustavy nerovnic  $c_1 \leq \mathcal{R}(\mathbf{X}, \theta) \leq c_2$ : pokud je  $\mathcal{R}(\mathbf{X}, \theta)$  např. ryze rostoucí v proměnné  $\theta \in \mathbb{R}^1$ , pak  $C(\mathbf{X}) = \left\{ \theta : \mathcal{R}^{-1}(c_1, \mathbf{X}) \leq \theta \leq \mathcal{R}^{-1}(c_2, \mathbf{X}) \right\}$ . Podobně pro funkci  $\mathcal{R}(\mathbf{X}, \theta)$  regulární a prostou vzhledem k  $\theta \in \mathbb{R}^k$ .

## 6.2 Konstrukce $\text{CM}_{1-\alpha}$ pomocí TSH $\phi_\alpha$

**Definice 6.3.** Nechť  $\phi(\mathbf{x})$  je test hypotézy  $H_0$  vs.  $H_1$ . Pak množinu  $A(H_0) = \{\mathbf{x} : \phi(\mathbf{x}) < 1\}$  nazýváme **přípustná oblast testu  $\phi$**  (ozn. AR - acceptance region). Speciálně pro test  $H_0 : \theta = \theta_0$  označíme přípustnou oblast AR jako  $A(\theta_0)$ .

Mějme nyní náš statistický model  $\mathbb{P} \in \mathcal{P}$ , parametr zájmu  $\theta = \theta(\mathbb{P})$  a náhodný výběr  $X_1, \dots, X_n$  iid  $\mathbb{P}_\theta$ .

### Metoda konstrukce $\text{CM}_{1-\alpha}$ pro $\theta$ :

1. Volíme  $\theta_0 \in \Theta$  libovolně pevně.
2. Testujeme  $H_0 : \theta = \theta_0$  na hladině významnosti  $\alpha$ . Příslušný test označíme  $\phi_\alpha$ , resp.  $W_\alpha$  pro případ neznáhodněného testu.
3. Vyjádříme  $A(\theta_0)$  pro test  $\phi_\alpha$ , resp.  $W_\alpha$ , při libovolném  $\theta_0 \in \Theta$ . Tím získáme  $A(\theta)$  pro  $\forall \theta \in \Theta$ .
4. Pak  $C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A(\theta)\}$  je  $\text{CM}_{1-\alpha}$  pro  $\theta$ . Navíc, pokud je test  $\phi_\alpha$  neznáhodněný a je na hladině  $\alpha$ ,  $\beta_{\phi_\alpha}(\theta_0) = \alpha$  pro  $\forall \theta_0 \in \Theta$ , pak uvedená  $C(\mathbf{X})$  je  $\text{CM}_{1-\alpha}$  s **konfidenčním koeficientem** rovným  $(1 - \alpha)$ .

## 6.3 Asymptotické konfidenční množiny

**Definice 6.4.** Mějme  $\mathbb{P} \in \mathcal{P}$  na  $(\Omega, \mathcal{A})$ ,  $\theta = \theta(\mathbb{P}) \in \Theta \subset \mathbb{R}^1$ , resp.  $\Theta \subset \mathbb{R}^k$  a  $\mathcal{B}_\Theta$  Borelovské. Mějme  $X_1, \dots, X_n \sim \mathbb{P} \in \mathcal{P}$  a  $\alpha \in (0, 1)$ . Pak  $C(\mathbf{X}) \in \mathcal{B}_\Theta$  se nazývá **asymptotická  $\text{CM}_{1-\alpha}$** , ozn  $\text{ACM}_{1-\alpha}$ , pokud pro

$$\forall \mathbb{P} \in \mathcal{P}, \quad \lim_{n \rightarrow +\infty} \mathbb{P}(\theta \in C(\mathbf{X})) \geq 1 - \alpha.$$

### Metody konstrukce:

- I) Najdeme takovou vhodnou náhodnou veličinu  $\mathcal{R}_n(\mathbf{X}, \theta)$ , která je **asymptoticky pivotální** veličinou (ozn. APQ), tzn. její *limitní* rozdělení nezávisí na  $\mathbb{P} \in \mathcal{P}$ :

$$\mathcal{R}_n(\mathbf{X}, \theta) \xrightarrow{\mathcal{D}} G, \quad \text{kde } G \text{ nezávisí na } \mathbb{P} \in \mathcal{P}.$$

Toto limitní  $G$  použijeme pro konstrukci  $\text{ACM}_{1-\alpha}$  stejně jako v případě neasymptotických  $\text{CM}_{1-\alpha}$  (viz sekce 6.1).

- II) Stejně jako v sekci 6.2, pro konstrukci  $\text{ACM}_{1-\alpha}$  použijeme přípustnou oblast  $A(\theta_0)$  založenou na asymptotickém testu  $\phi_\alpha$  dosahujícím asymptotické hladiny  $\alpha$  pro testování  $H_0 : \theta = \theta_0$ . Tedy  $C(\mathbf{X}) = \{\theta \in \Theta : \mathbf{X} \in A(\theta)\}$ , kde  $A(\theta)$  je AR asymptotického testu  $\phi_\alpha$ , je  $\text{ACM}_{1-\alpha}$  pro parametr  $\theta$ .