Model selection (výběr modelu)

(i) Je náš model „vhodný"? Popisuje správně naše data?

(ii) Máme-li více modelů, který z nich je „nejlepší"? Který si vybrat?

Grafické metody pro posouzení vhodnosti modelu

- jsou založené na porovnání teoretického a empirického rozdělení

- vykreslení teoretické a empirické distr. funkce v jednom grafu

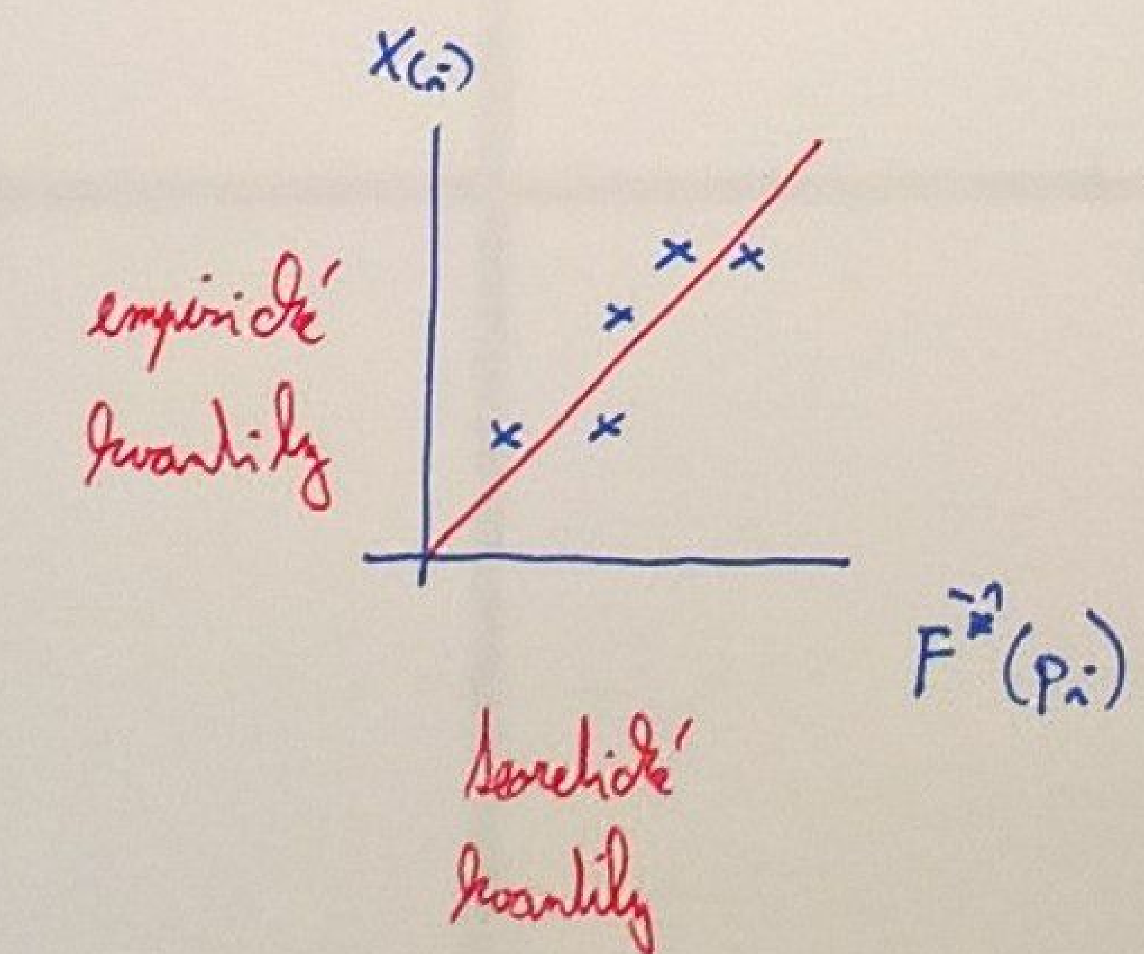
$D(x) = \hat{F}_n(x) - F^*(x)$, $x \in \mathbb{R}$, kde $\hat{F}_n(x)$ je empirická distr. fce a F^* je teoretická distr. fce (s odhadnutými parametry)

- vykreslení teoretické hustoty a jejího odhadu do jednoho grafu (histogram, jádrový odhad, ...)

- Q-Q plot

uspořádáme pozorování $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$

$X_{(i)}$ je $p_i = \frac{i-\beta}{n+1-2\beta}$ $(0 \le \beta < 1)$ - tý kvantil



v praxi: $\beta = 0{,}5$
$\beta = 0{,}3175$

Q-Q plot je graf $[F^{*-1}(p_i), X_{(i)}]$ pro $i = 1, \dots, n$

F^{*-1} je teoretická kvantilová funkce (s odhadnutými parametry)

Poznámka

N-P plot je Q-Q plot pro „ověřování normality dat"

$\beta = 0{,}3175$ pro $n \leq 10$

$\beta = 0{,}5$ pro $n > 10$

F^{*-1} je Φ^{-1} kvantilová funkce $N(0,1)$.

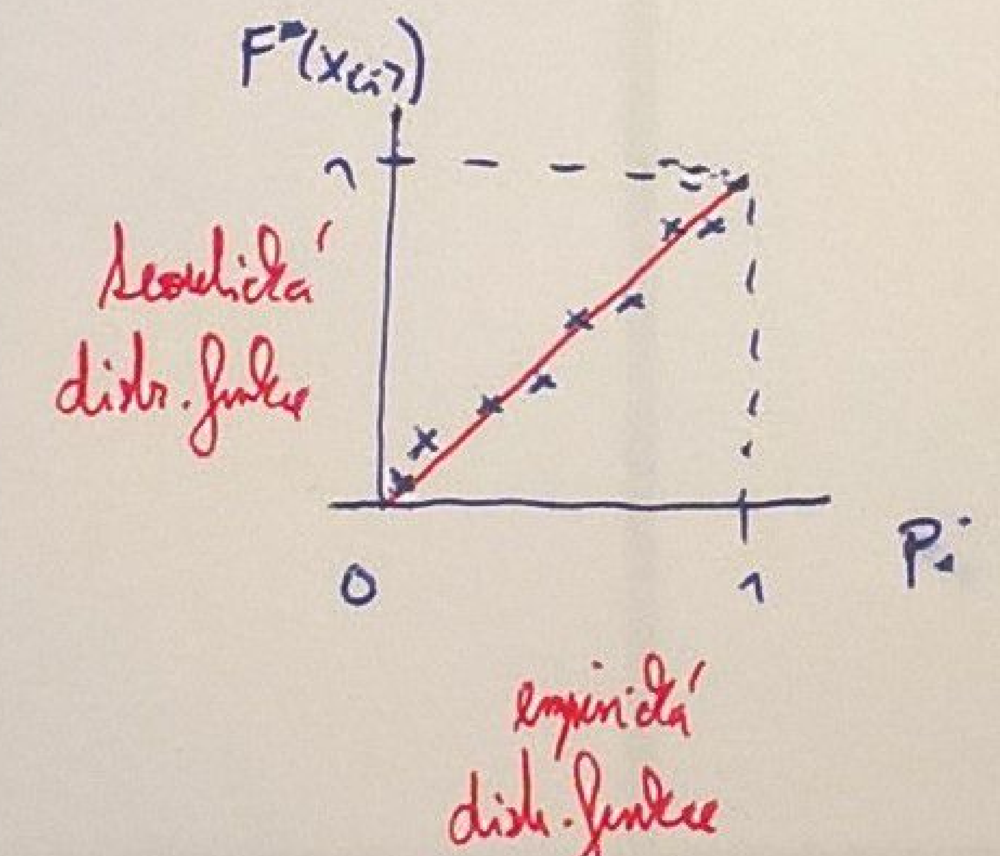
- P-P plot

- porovnává empirickou a teoretickou distr. funkci

$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$ ... drobná modifikace $\tilde{F}_n(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$ ... nabývá hodnot $p_i = \frac{i}{n+1}$ pro $i = 1, 2, \dots, n$.

P-P plot je graf $[p_i, F^*(X_{(i)})]$ pro $i = 1, \dots, n$

F^* je teoretická distribuční funkce (s odhadnutými parametry)



Statistické testy pro posouzení vhodnosti modelu

$X_1, \dots, X_n$ je náh. výběr s distribuční funkcí F

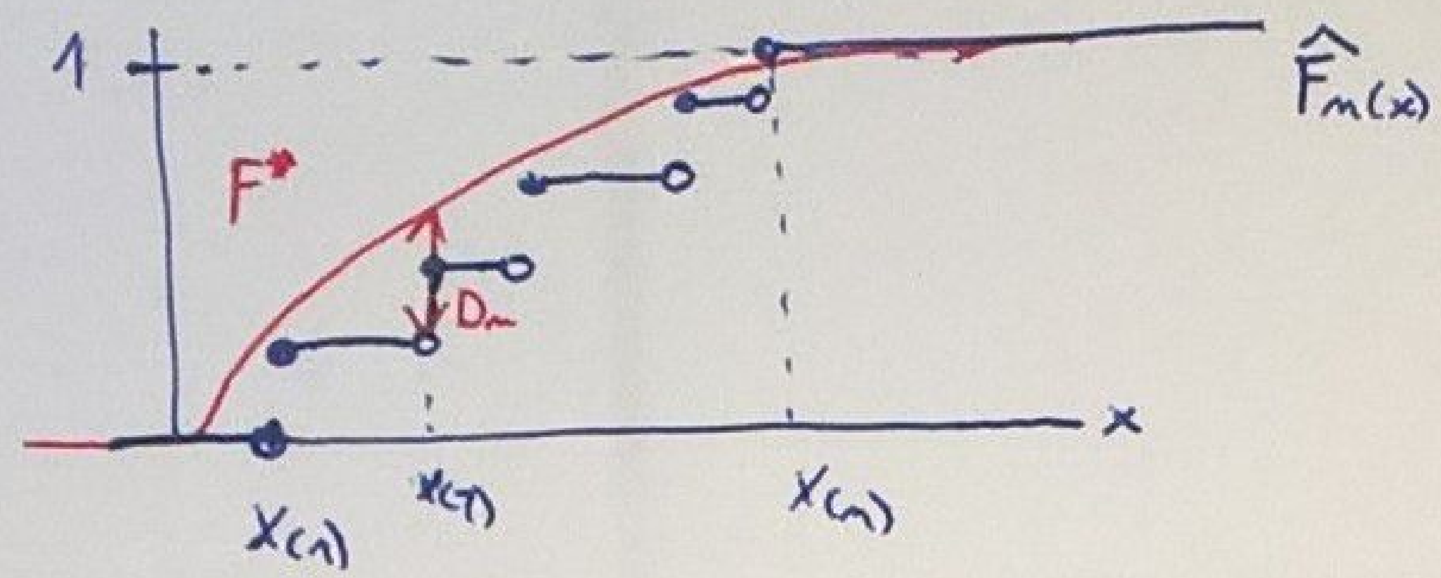
$H_0: F = F^*$

$H_1: F \neq F^*$

F^* je nějaká známá distribuční funkce

• Kolmogorovův - Smirnovův test

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}$$ je empirická distribuční funkce



$$D_n = \max_{x \in \mathbb{R}} |\hat{F}_n(x) - F^*(x)| = \max_{i=1,\dots,n} \left|\frac{i}{n} - F^*(X_{(i)})\right|$$

<u>Tvrzení</u>

Je-li F^* spojitá, pak za platnosti H_0 $\sqrt{n}\, D_n$ má asymptoticky při $n \to \infty$ rozdělení $\sup_{t \in [0,1]} |B(t)|$, kde $B(t)$ je Brownův most v $C(0,1)$.

<u>Poznámka</u>

Rozdělení náh. veličiny $Y = \sup_{t \in [0,1]} |B(t)|$ je známé, ale nedá se vyjádřit v uzavřené podobě. Její distribuční funkce je

$$F_Y(y) = 1 - 2 \cdot \sum_{j=1}^{\infty} (-1)^{j+1} \cdot e^{-2j^2y^2}, \; y > 0$$

$\to$ přibližná kvantilová funkce $F_Y^{-1}(\alpha) = \sqrt{\frac{1}{2} \log \frac{2}{1-\alpha}}$

$$\doteq 1 - 2 e^{-2y^2}, \; y > 0$$

Použití:

spočítáme $\sqrt{n}\, D_n$. Je-li $\sqrt{n}\, D_n > F_Y^{-1}(1-\alpha)$ ... zamítáme H_0 na hladině významnosti α

nezávisí na F^*

Poznámky

- test lze použít jen v případě, že H_0 je plně specifikována modelem. Pokud jsou data použita nejprve použita k odhadu parametrů, skutečná krit. hodnota je menší (test je příliš konzervativní)
- kritická hodnota v tomto případě závisí na testované distribuci
- pro normální rozdělení je spočítaná – Lillieforsův test (95% kvantil se z 1,36 sníží na 0,886)
- pro ostatní rozdělení se p-hodnota testu určí pomocí MC simulací:

$X_1, \dots, X_n$ je náh. výběr s distribuční funkcí $F(x, \theta)$

$H_0: F = F(x, \theta)$ pro nějaké θ

(i) odhadneme parametr θ z dat $\rightarrow \hat{\theta}$

(ii) určíme hodnotu testové statistiky pro hypotetickou distribuční funkci $F(x, \hat{\theta})$, označme ji T_0

(iii) nagenerujeme mnoho náhodných výběrů o rozsahu n z rozdělení s distr. fcí $F(x, \hat{\theta})$

(iv) odhadneme parametr θ z těchto dat $\rightarrow \tilde{\theta}$

(v) určíme hodnotu testové statistiky pro hypotetickou distr. fci $F(x, \tilde{\theta})$, označme ji T

(vi) kroky (iii) $\rightarrow$ (v) několikrát opakujeme, p-hodnotu odhadneme jako podíl případů, kdy $T > T_0$.

- Andersonův-Darlingův test

patří do třídy testů dané statistikou: $n \cdot \int_{-\infty}^{\infty} \left(\hat{F}_n(x) - F^*(x)\right)^2 \cdot w(F^*(x))\, f^*(x)\, dx$ pro nějakou váhovou funkci w.

$w(y) = \frac{1}{y(1-y)}$ $0<y<1$ ... dává větší váhu pozorováním na chvostech

$$A^2 = n \cdot \int_{-\infty}^{\infty} \frac{(\hat{F}_n(x) - F^*(x))^2}{F^*(x)(1-F^*(x))} f^*(x)\, dx = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)\left[\log F^*(X_{(i)}) + \log\left(1 - F^*(X_{(n+1-i)})\right)\right].$$

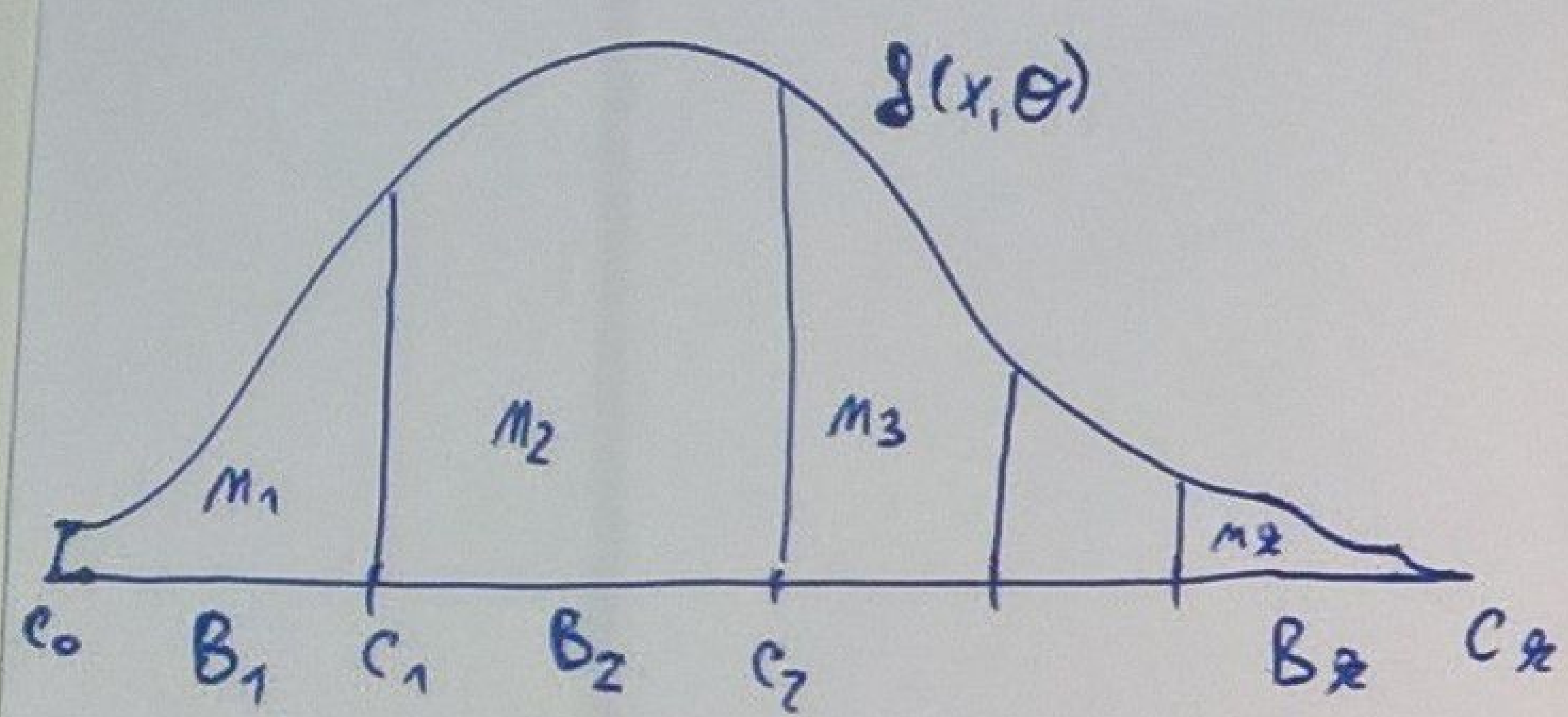
Poznámka

I pro F^* známé hodnota A^2 závisí na testové distribuci F^* $\Rightarrow$ neexistují „univerzální" kvantily.

p-hodnota se určí pomocí simulací Monte Carlo.

Cramérův-von Misesův test – používá váhovou funkci $w(y) = 1$.

- Pearsonův χ^2-test dobré shody



○ obor hodnot n.v. X_i rozdělíme na k intervalů $B_1, \dots, B_k$

○ označme n_j počet pozorování, které padnou do intervalu B_j

○ označme $p_j(\theta) = P(X_i \in B_j) = \int_{c_{j-1}}^{c_j} f(x,\theta)\,dx = F(c_j,\theta) - F(c_{j-1},\theta)$

○ celkem máme n pozorování, tedy v intervalu B_j by mělo být $n \cdot p_j(\theta)$ pozorování (očekávané četnosti).

○ ty porovnáme s pozorovanými četnostmi n_j:

$$\chi^2 = \sum_{j=1}^{k} \frac{[n_j - n \cdot p_j(\theta)]^2}{n \cdot p_j(\theta)}.$$

Tvrzení

Za platnosti H_0 má testová statistika χ^2 při $n \to \infty$ asymptoticky χ^2 rozdělení s $(k-1)$ stupni volnosti

Je-li $\chi^2 > \chi^2_{1-\alpha}(k-1)$ … zamítáme H_0.

Poznámky

Pokud jsou data nejprve použita k odhadu neznámého p-rozměrného parametru, pak $X^2 \overset{H_0}{\approx} \chi^2(k-1-p)$. (metoda minimálního χ^2)

Volba k? – heuristická pravidla: $k \approx 2 \cdot n^{\frac{2}{5}}$ či $k \approx 15 \cdot \left(\frac{n}{100}\right)^{\frac{2}{5}}$

Intervaly $B_1, \dots, B_k$ se pak volí „stejně pravděpodobné", tedy $c_1, c_2, \dots, c_{k-1}$ jsou $\frac{1}{k}, \frac{2}{k}, \dots, \frac{k-1}{k}$ – kvantily rozdělení s hustotou $f(x, \hat{\theta})$.

Výběr vhodného modelu (z několika kandidátů)

– princip Occamovy břitvy (princip logické úspornosti) – vybereme co nejjednodušší model (vyhneme se přeučení, lepší interpretovatelnost)

• judgement-based přístup (založený na subjektivním úsudku analytika)

– rozhodnutí založené na různých grafech a tabulkách (tail vs. mod fit)

– rozhodnutí založené na dřívější zkušenosti (Paretovo rozdělení pro výši příjmů, Benfordovo pro rozdělení četností prvních číslic)

– model je plně určen situací, kterou má popisovat (alternativní rozdělení pro házení mincí)

• score-based přístup (založený na číselných charakteristikách; objektivní)

– nejnižší hodnota testové statistiky vybraného testu

– nejvyšší p-hodnota vybraného testu

– nejvyšší hodnota věrohodnosti (logaritmické věrohodnosti)

} nebere do úvahy složitost modelu (počet parametrů)

– nejnižší hodnota nějakého penalizačního kritéria, např.

$$AIC = -2 \log \underbrace{L(\hat{\theta})}_{\text{věrohodnost}} + 2 \cdot \underbrace{p}_{\text{počet odhadovaných parametrů}}$$

$$BIC = -2 \log L(\hat{\theta}) + p \cdot \log n$$