Hraniční efekty v jádrových odhadech distribuční funkce

Jan Koláček

Seminář „Vybrané partie z aplikované matematiky“

Ústav matematiky a statistiky
Janáčkovo nám. 2a
Brno

Obsah

- Základní pojmy
- Jádrové odhady hustoty a distribuční funkce
- Hraniční efekty
- Navrhovaný odhad
- Aplikace

Jádrová funkce

Nechť ν, k jsou celá nezáporná čísla taková, že platí $0 \leq \nu \leq k-2$, ν a k mají stejnou paritu. Funkci $K \in Lip[-1,1]$, nosič$(K) = [-1,1]$, splňující podmínky

$$(i) \qquad K(-1) = K(1) = 0$$

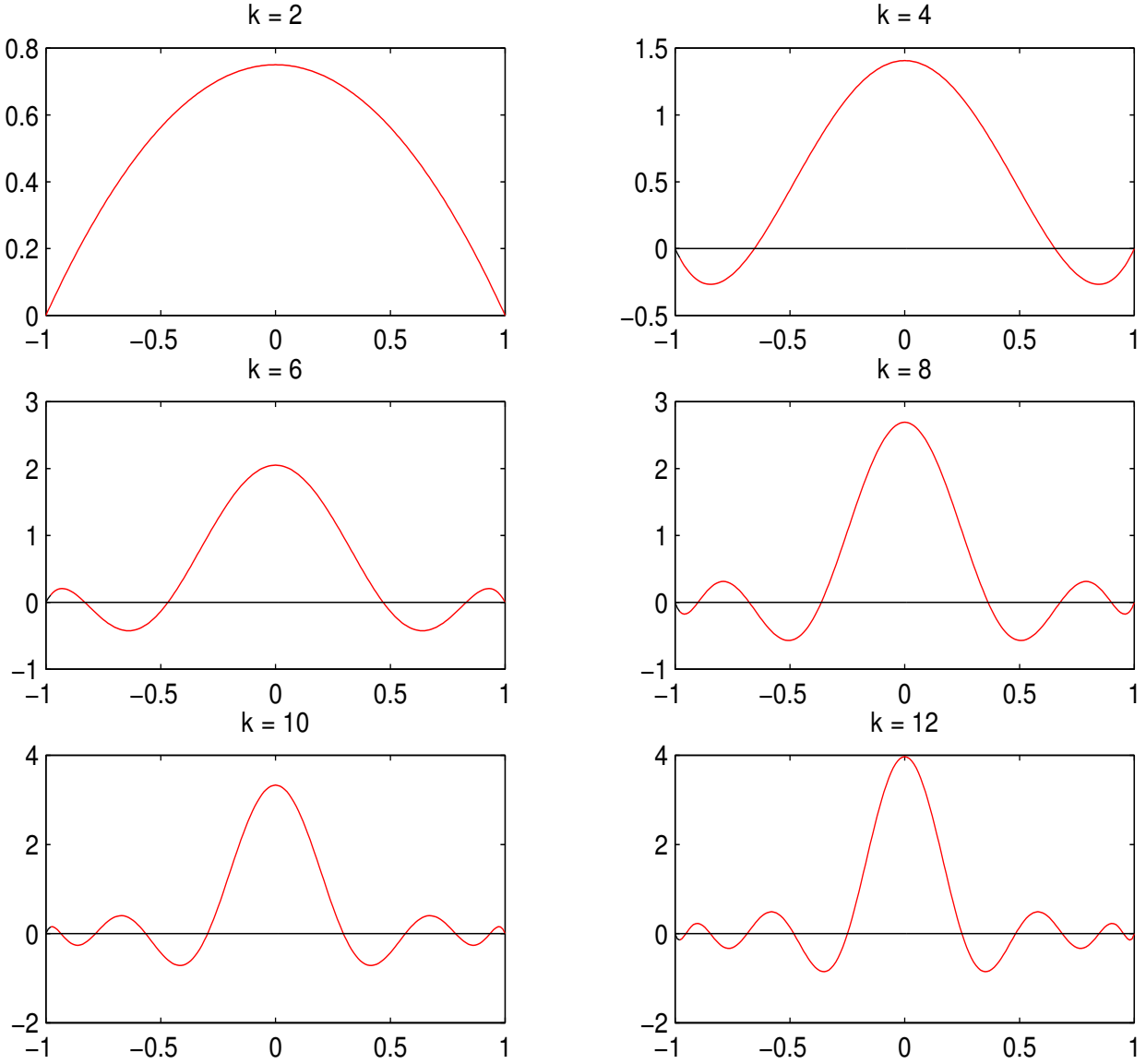
$$(ii) \qquad \int_{-1}^{1} x^j K(x)dx = \begin{cases} 0, & 0 \leq j < k, j \neq \nu \\ (-1)^\nu \nu!, & j = \nu \\ \beta_k \neq 0, & j = k, \end{cases}$$

nazýváme *jádrem* řádu (ν, k) a třídu všech takových jader značíme $S_{\nu k}$.

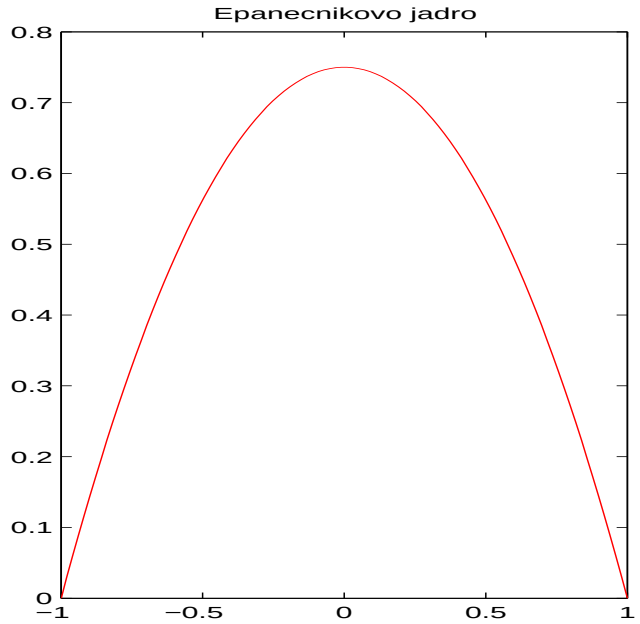
Tabulka jader

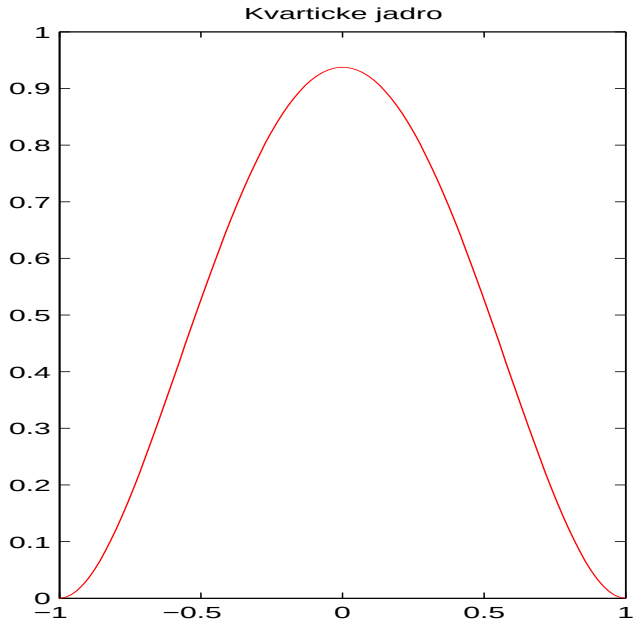
ν	k	Jádro (na $[-1,1]$)
0	2	$K_{0,2}(x) = \frac{3}{4}(1-x^2)$
0	2	$K_{0,2}(x) = \frac{15}{16}(1-x^2)^2$
0	2	$K_{0,2}(x) = \frac{35}{32}(1-x^2)^3$
0	4	$K_{0,4}(x) = \frac{15}{32}(x^2-1)(7x^2-3)$
2	4	$K_{2,4}(x) = \frac{105}{16}(1-x^2)(5x^2-1)$
1	3	$K_{1,3}(x) = \frac{15}{4}x(1-x^2)$

Jádra třídy S_{0k}, $k = 2, 4, 6, 8, 10, 12$.



Jádra třídy S_{02}





Jádrové odhady hustoty a distribuční funkce

Nechť $X_1,\ldots,X_n$ je náhodný výběr z rozdělení s příslušnou hustotou f, resp. distribuční funkcí F. Předpokládejme $f \in C^2$ pro $K \in S_{02}$.

- **Jádrový odhad hustoty**

$$\widehat{f}_{h,K}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

- **Jádrový odhad distribuční funkce**

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right), \qquad W(x) = \int_{-1}^{x} K(t)dt$$

kde $K \in S_{0,2}$, $K(x) \geq 0$ na $[-1,1]$ a parametr $h > 0$ je tzv. *šířka vyhlazovacího okna* $(h = h(n), \lim_{n\to\infty} h = 0, \lim_{n\to\infty} nh = \infty)$.

Optimální šířka okna pro $\widehat{F}_{h,K}$

- kritérium kvality odhadu je MISE (Mean Integrated Square Error)

$$\mathrm{MISE}(\widehat{F}_{h,K}) = \mathrm{E} \int (\widehat{F}_{h,K}(x) - F(x))^2 dx$$

- hlavní člen MISE (Bowman, A., Hall, P., Prvan, T. [2])

$$\overline{\mathrm{MISE}}(\widehat{F}_{h,K}) = \underbrace{\frac{1}{n} \int F(x)(1 - F(x))dx - q_1 \frac{h}{n}}_{\overline{var}(\widehat{F}_{h,K})} + \underbrace{q_2 h^4}_{\overline{bias}^2(\widehat{F}_{h,K})},$$

$$q_1 = \int_{-1}^{1} W(x)(1 - W(x))dx > 0, \qquad q_2 = \frac{\beta_2^2}{4} \int (F^{(2)}(x))^2 dx.$$

Odtud

$$h^F_{opt,0,2} = n^{-1/3} \left(\frac{q_1}{4q_2} \right)^{1/3}$$

Hraniční efekty

Předpoklady:

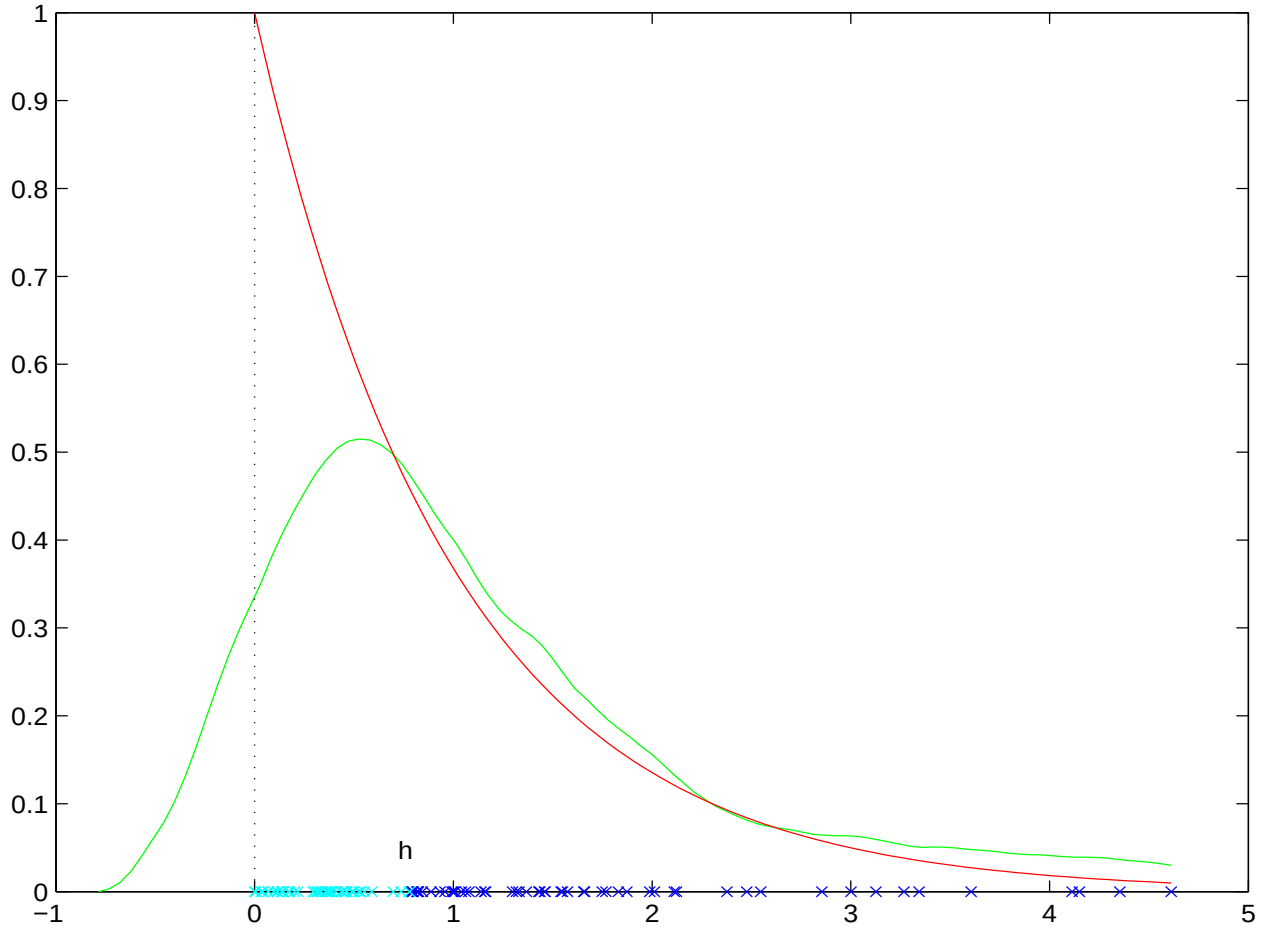
- hustota f má nosič $[0, \infty)$, tj. $X_i \geq 0,\ i = 1, \ldots, n$
- $f(0) \neq 0$
- distribuční funkce F má také nosič $[0, \infty)$

Hraniční efekty vznikají při odhadech v bodech „blízko“ levé hranice, tj. pro $x \in [0, h]$.

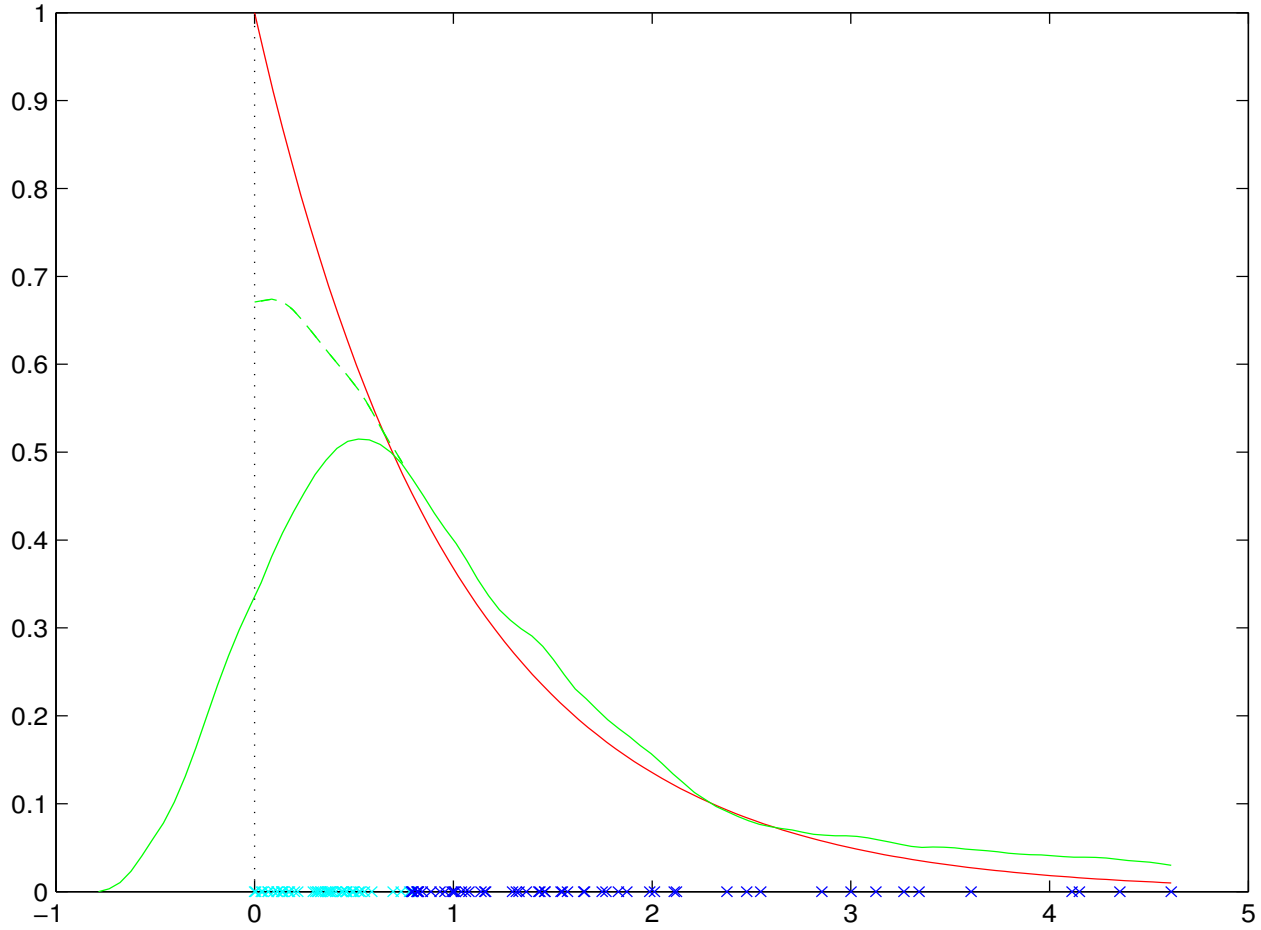
V dalším budeme psát

$$x = ch,\ 0 \leq c \leq 1.$$

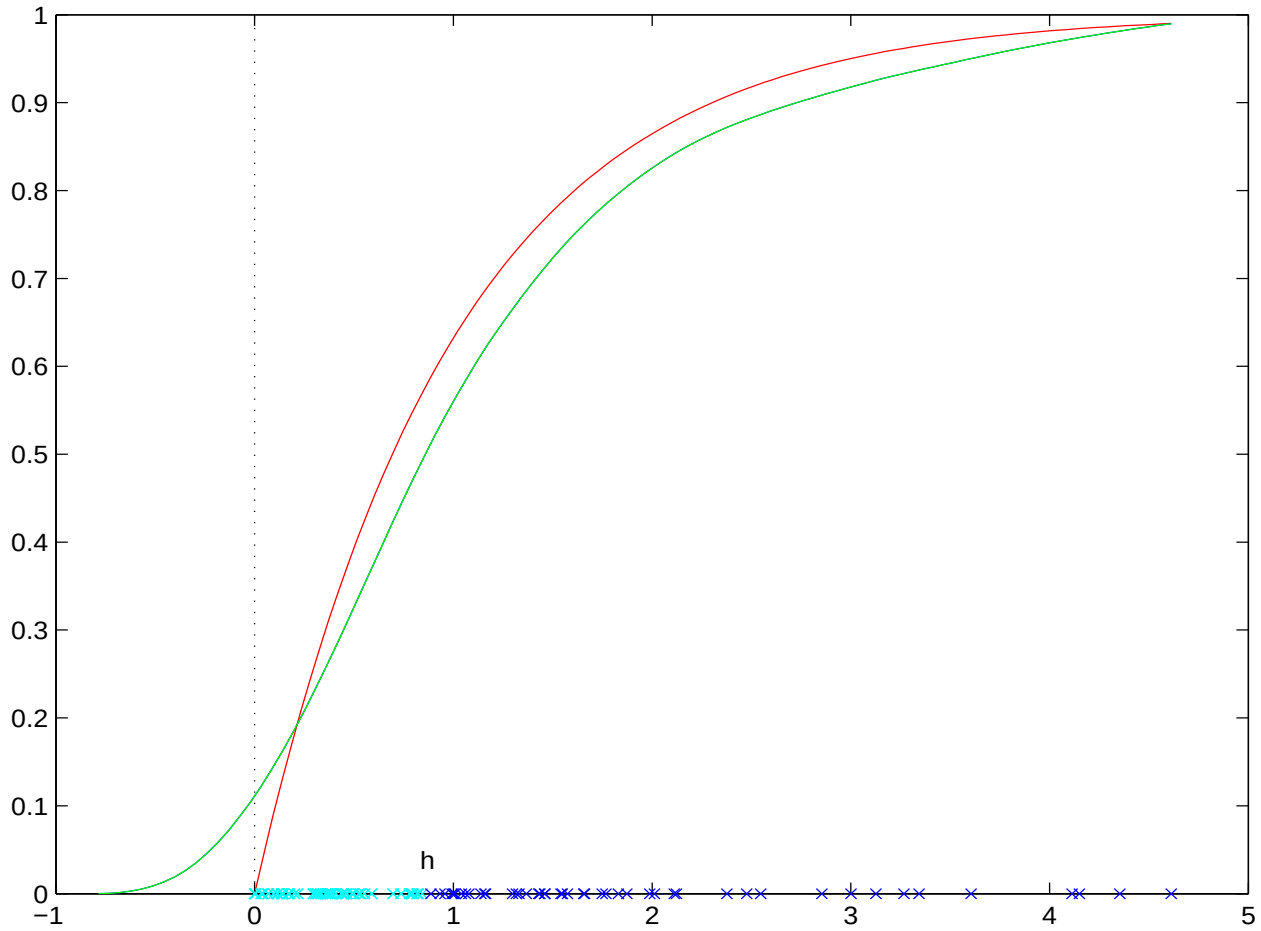
$X \sim Exp(1)$ – odhad hustoty f $(n = 100,\ h^f_{opt,0,2} = 0.786)$



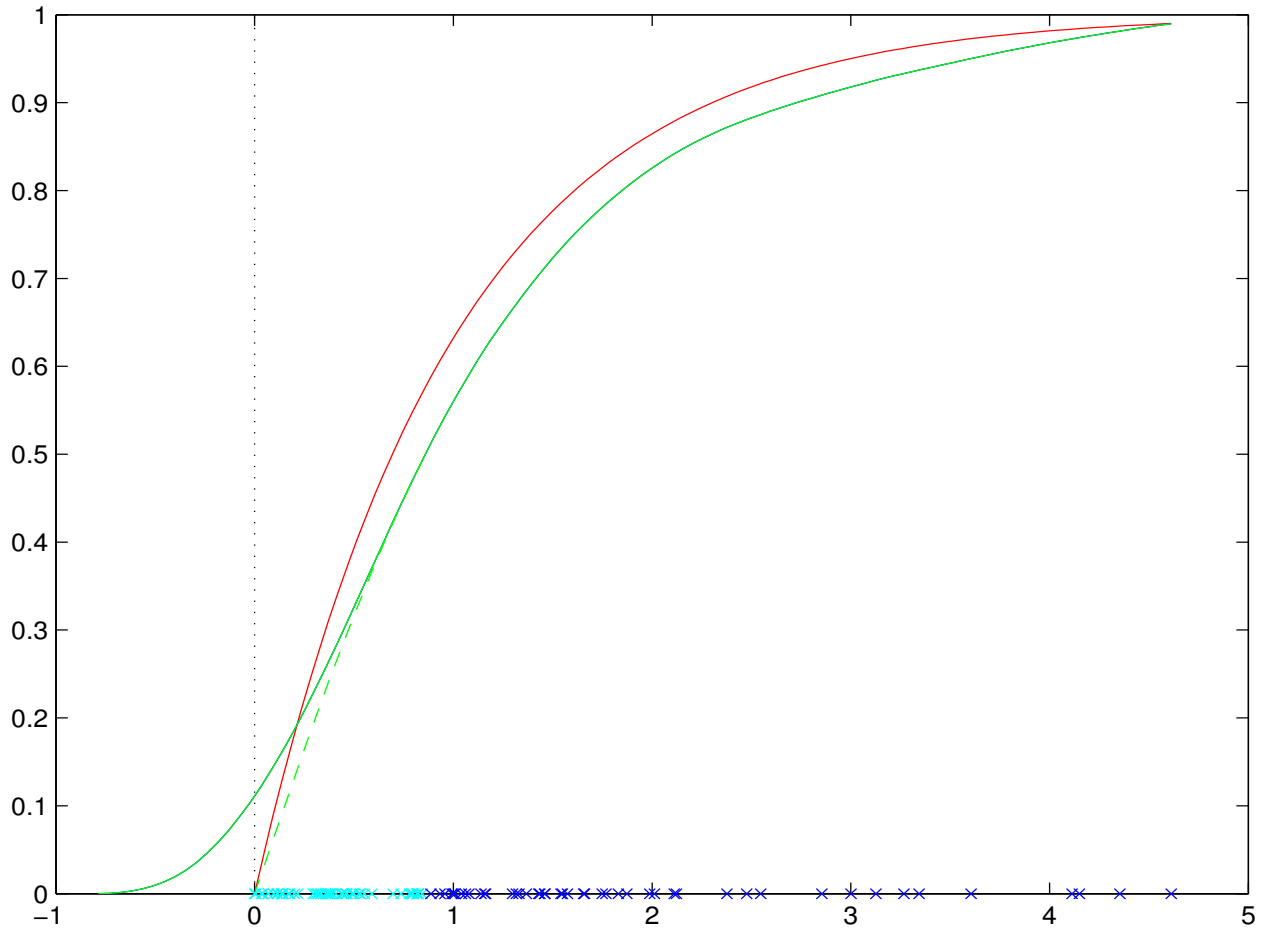
$X \sim Exp(1)$ – odhad hustoty f $(n = 100,\ h^f_{opt,0,2} = 0.786)$



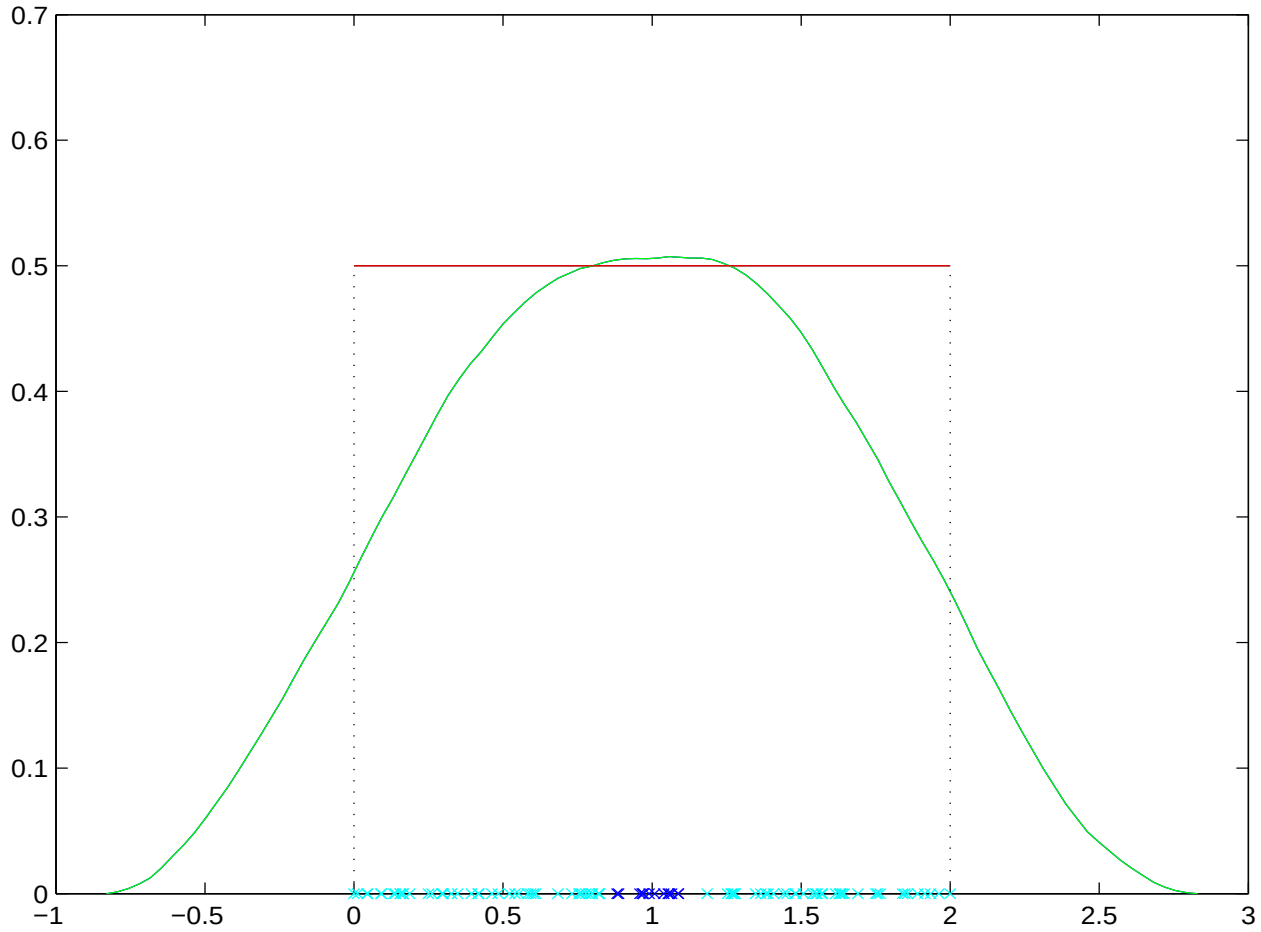
$X \sim Exp(1)$ – odhad distrib. funkce F $(n = 100,\ h^F_{opt,0,2} = 0.8479)$



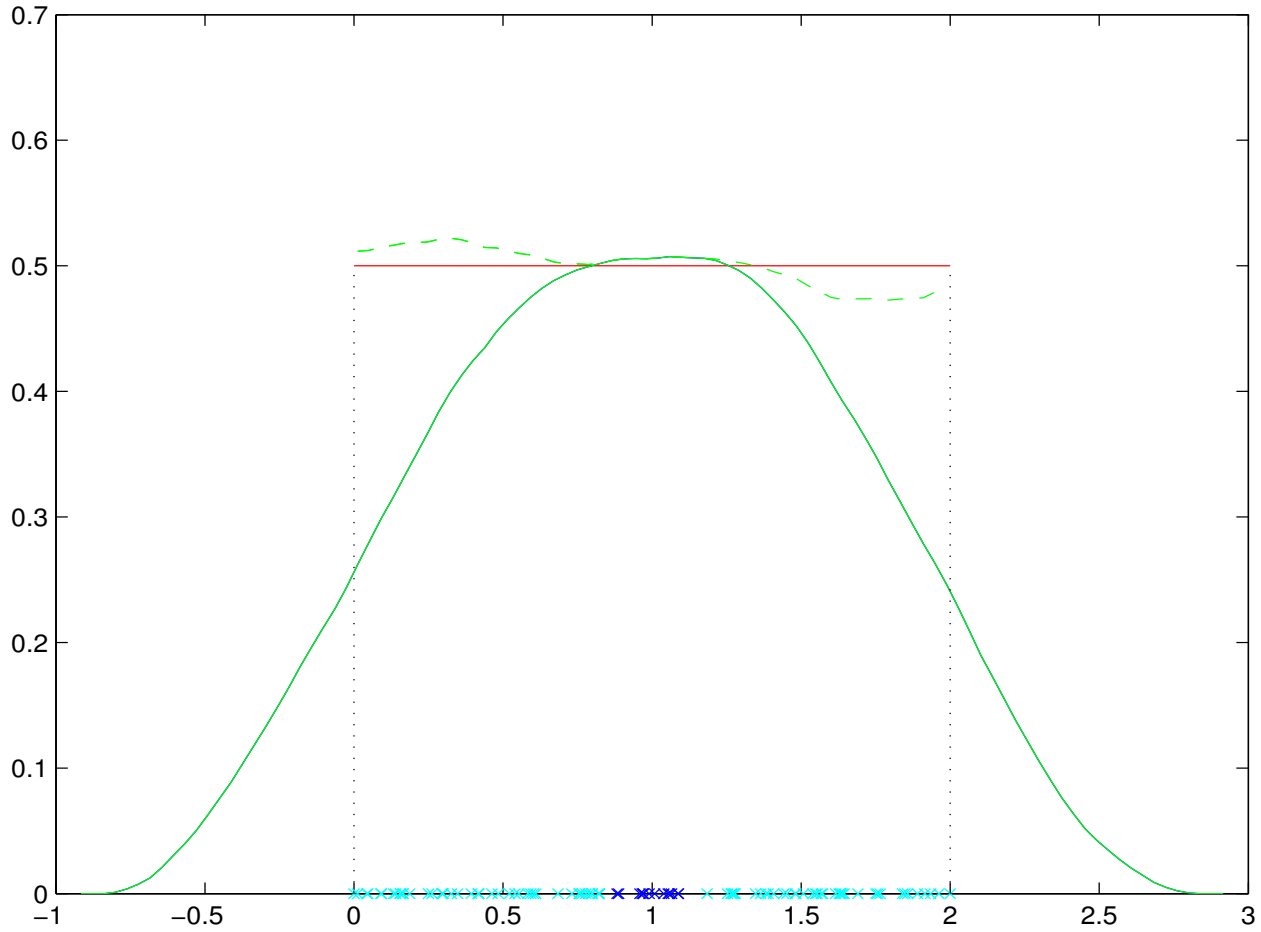
$X \sim Exp(1)$ – odhad distrib. funkce F $(n = 100,\ h^F_{opt,0,2} = 0.8479)$



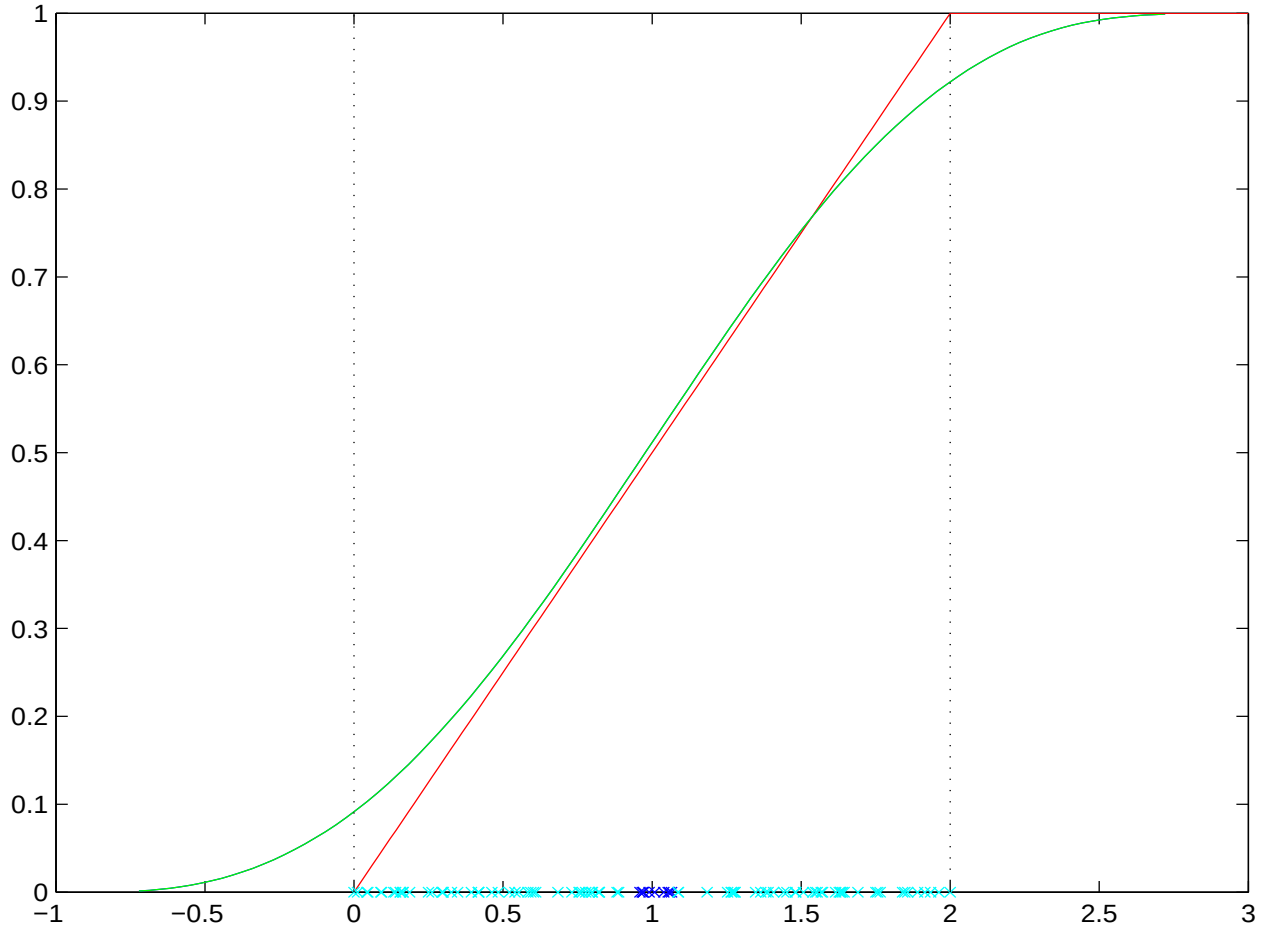
$X \sim Rs(0,2)$ – odhad hustoty f $(n = 100,\ \hat{h}^f_{opt,0,2} = 0.8304)$



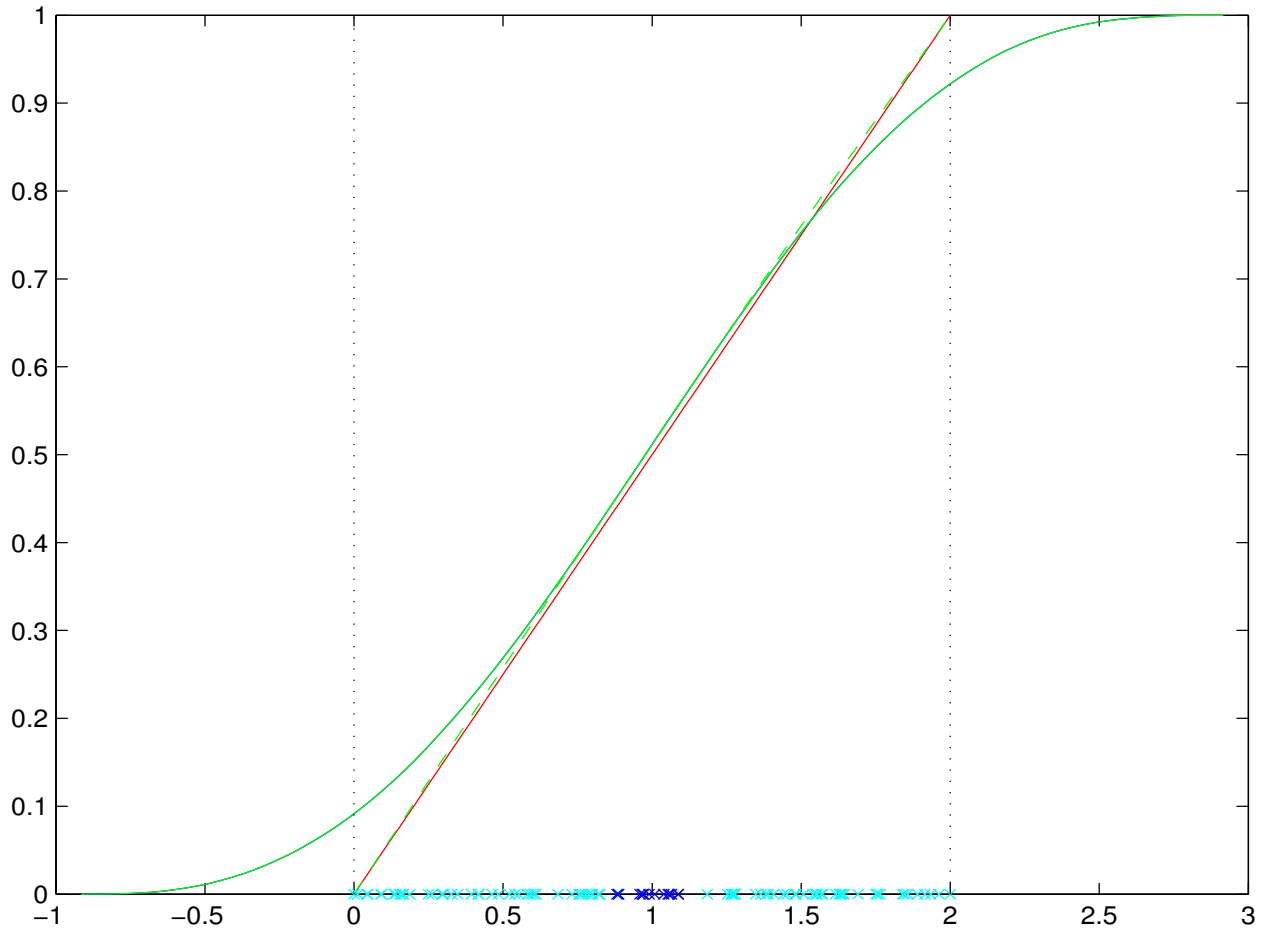
$X \sim Rs(0,2)$ – odhad hustoty f $(n = 100,\ \hat{h}^f_{opt,0,2} = 0.8304)$



$X \sim Rs(0,2)$ – odhad distrib. funkce F $(n = 100,\ \hat{h}^F_{opt,0,2} = 0.9163)$



$X \sim Rs(0,2)$ – odhad distrib. funkce F $(n = 100,\ \hat{h}^F_{opt,0,2} = 0.9163)$



Vychýlení odhadu $\widehat{F}_{h,K}(x)$ v bodě $x = ch$,

- „blízko“ hranice $(0 \leq c < 1)$:

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = hf(0) \int_{-1}^{-c} W(t)dt + h^2 f^{(1)}(0) \left\{ \frac{c^2}{2} + c \int_{-1}^{-c} W(t)dt - \int_{-1}^{c} tW(t)dt \right\} + o(h^2)$$

- „uvnitř“ $(c \geq 1)$:

$$\mathrm{E}(\widehat{F}_{h,K}(x)) - F(x) = \frac{h^2}{2} f^{(1)}(0) \int_{-1}^{1} tW(t)dt + o(h^2)$$

Řešení problému

- hraniční jádra
- pseudo-data
- transformace dat
- zrcadlení

$$\widehat{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x-X_i}{h}\right) - W\left(-\frac{x+X_i}{h}\right)\right\}$$

- kombinace výše uvedených

Navrhovaný odhad

„Zobecněná“ metoda zrcadlení (pro hustoty – viz [5])

$$\widetilde{F}_{h,K}(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{W\left(\frac{x - g_1(X_i)}{h}\right) - W\left(-\frac{x + g_2(X_i)}{h}\right)\right\},$$

$$g_1 = g_2 \Rightarrow \widetilde{F}_{h,K}(0) = 0$$

Položme $g := g_1 = g_2$

Předpoklady pro funkci g:

- g je spojitá, nezáporná rostoucí funkce na $[0, \infty)$
- g^{-1} existuje
- $g(0) = 0$
- $g^{(1)}(0) = 1$
- $g^{(i)},\ i = 2, 3$ existují a jsou spojité na $[0, \infty)$.

Vychýlení odhadu $\widetilde{F}_{h,K}(x)$ v bodě $x = ch$, $\boxed{0 \le c < 1}$

$$\begin{aligned}\mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = h^2 \Big\{ & f^{(1)}(0)[c^2/2 + 2cI_1 - I_2] \\ & - f(0)g^{(2)}(0)[c^2 + 2cI_1 - I_2]\Big\} \\ & + \frac{1}{2}ch^3 \Big\{ f^{(2)}(0)\beta_2 - g^{(2)}(0)[f^{(1)}(0) - f(0)g^{(2)}(0)]\times \\ & \times(3\beta_2 + c^2) - f(0)g^{(3)}(0)(\beta_2 + c^2/3)\Big\} \\ & + O(h^4),\end{aligned}$$

kde $I_1 = \int_{-1}^{-c} W(t)dt,\ I_2 = \int_{-c}^{c} tW(t)dt$

Vychýlení odhadu $\widetilde{F}_{h,K}(x)$ v bodě $x = ch$, $\boxed{c > 1}$

$$\begin{aligned}\mathrm{E}(\widetilde{F}_{h,K}(x)) - F(x) = {} & \frac{1}{2}h^2 \left\{ f^{(1)}(0)\beta_2 - f(0)g^{(2)}(0)[c^2 + \beta_2] \right\} \\ & + \frac{1}{2}ch^3 \left\{ f^{(2)}(0)\beta_2 - g^{(2)}(0)[f^{(1)}(0) - f(0)g^{(2)}(0)] \times \right. \\ & \left. \times (3\beta_2 + c^2) - f(0)g^{(3)}(0)(\beta_2 + c^2/3) \right\} \\ & + O(h^4)\end{aligned}$$

Z předchozího volíme

$$g^{(2)}(0) = \begin{cases} d_1 \frac{\frac{c^2}{2}+2cI_1-I_2}{c^2+2cI_1-I_2}, & \text{pro } 0 \le c < 1 \\ d_1 \frac{\beta_2}{c^2+\beta_2}, & \text{pro } c > 1 \end{cases} \qquad (= A_c)$$

$$g^{(3)}(0) = \begin{cases} d_2 \frac{3\beta_2}{c^2+3\beta_2} - d_1^2 \frac{3c^2(c^2+4cI_1-2I_2)}{4(c^2+2cI_1-I_2)^2}, & \text{pro } 0 \le c < 1 \\ d_2 \frac{3\beta_2}{c^2+3\beta_2} - d_1^2 \frac{3c^2\beta_2}{(c^2+\beta_2)^2}, & \text{pro } c > 1 \end{cases} \qquad (= B_c)$$

kde

$$d_1 = \frac{f^{(1)}(0)}{f(0)}, \; d_2 = \frac{f^{(2)}(0)}{f(0)}.$$

Konstrukce funkce $g(y)$

Odhad d_1, d_2

$$d_1 = \frac{f^{(1)}(0)}{f(0)} = (\ln f(x))^{(1)}_{x=0} \approx \hat{d}_1 = \frac{\ln f(h_1) - \ln f(0)}{h_1}, \; h_1 \approx n^{-\frac{1}{6}}$$

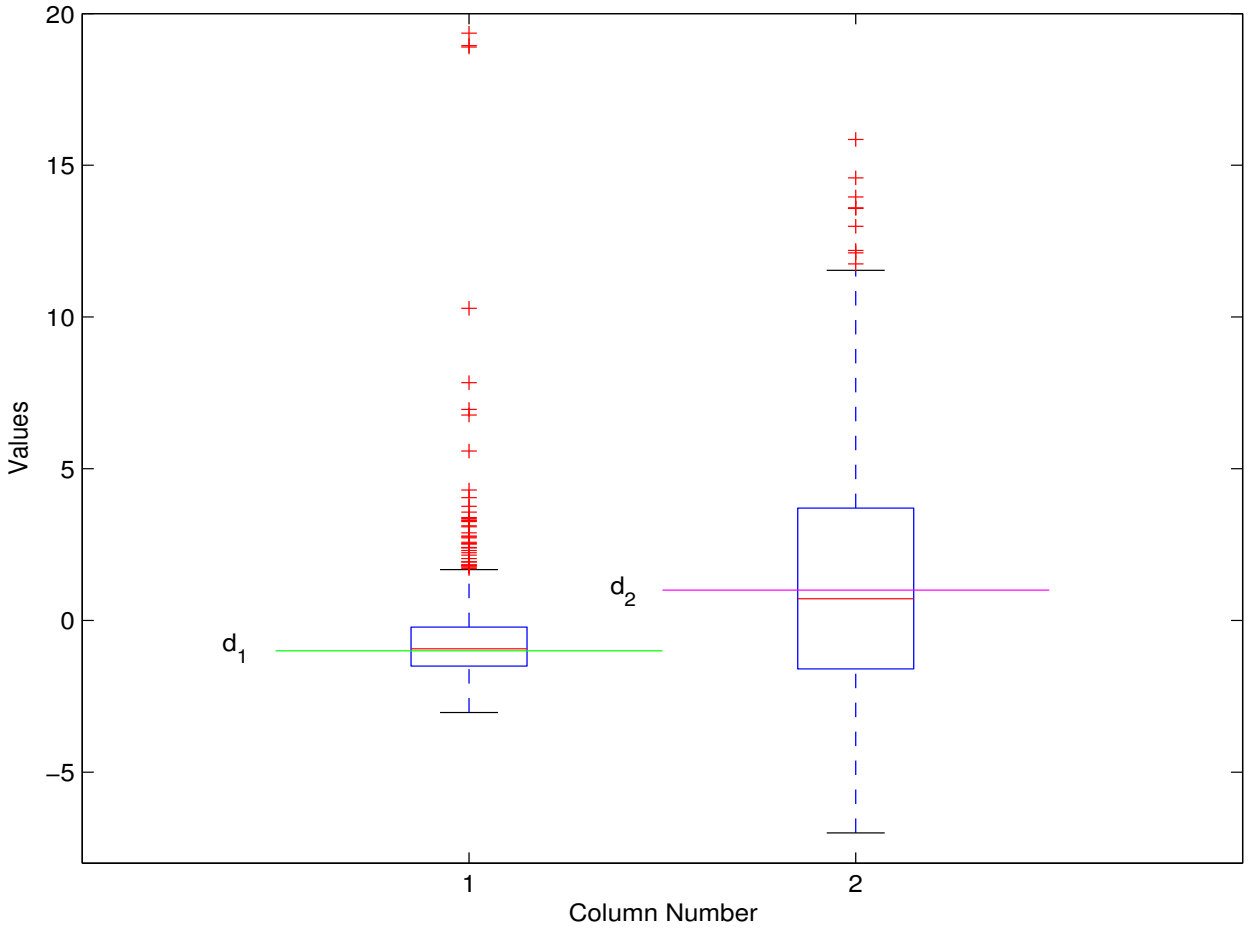
$$d_2 = (\ln f(x))^{(2)}_{x=0} + d_1^2,$$

$\hat{d}_2$ podobně, viz Karunamuni R.J., Alberts T. [5].

Odtud $\hat{d}_1$, $\hat{d}_2 \Rightarrow \widehat{A}_c$, $\widehat{B}_c$

$$\widehat{g}(y) = \frac{\widehat{B}_c}{6} y^3 + \frac{\widehat{A}_c}{2} y^2 + y$$

Odhady d_1, d_2 pro $X \sim Exp(1)$ (1 000 simulací, $n = 100$)



Příklady

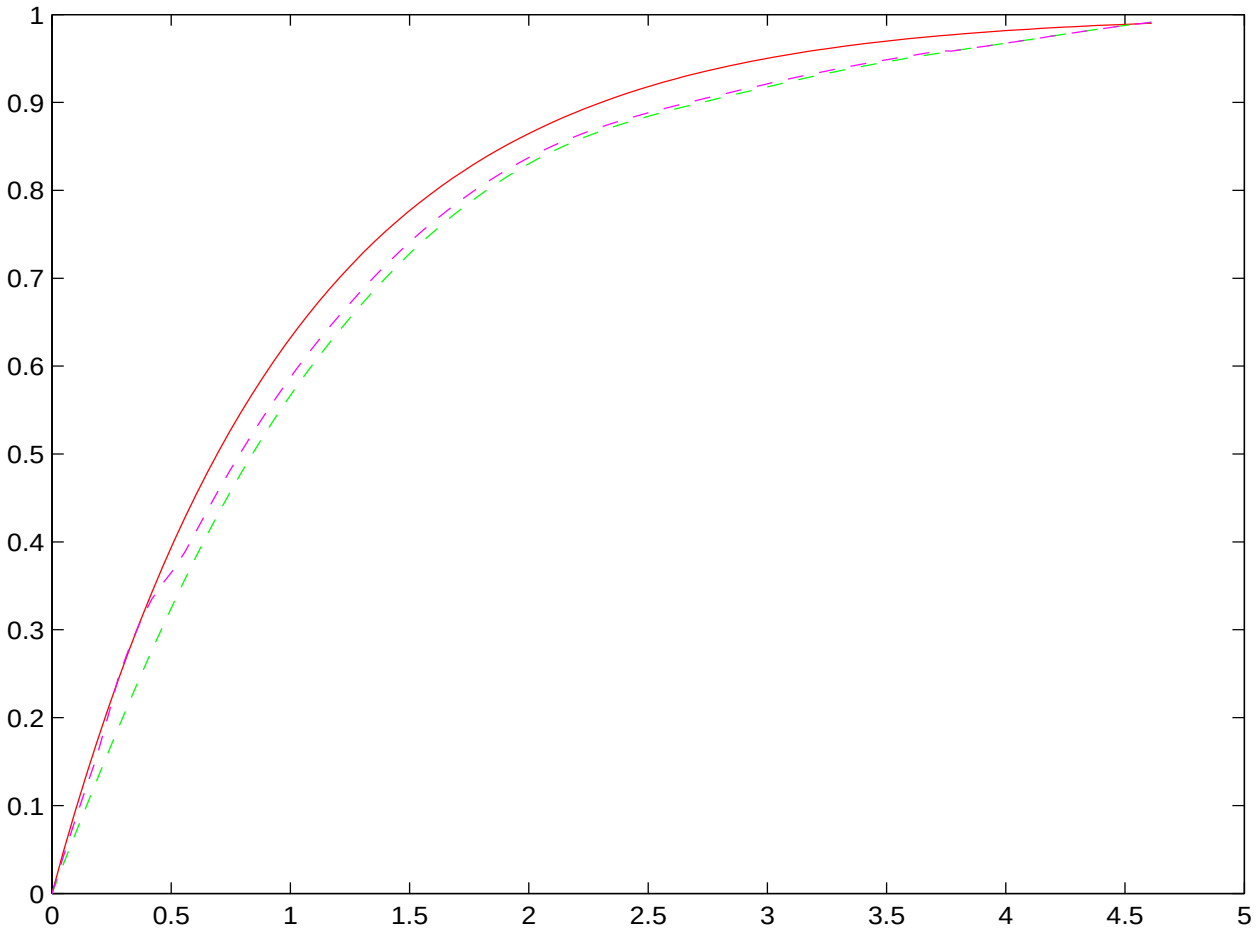
- Pro výsledný jádrový odhad používáme Epanečnikovo jádro

$$K_{0,2}(x) = \frac{3}{4}(1 - x^2)I_{[-1,1]},$$

  kde I_A je indikátor množiny A.

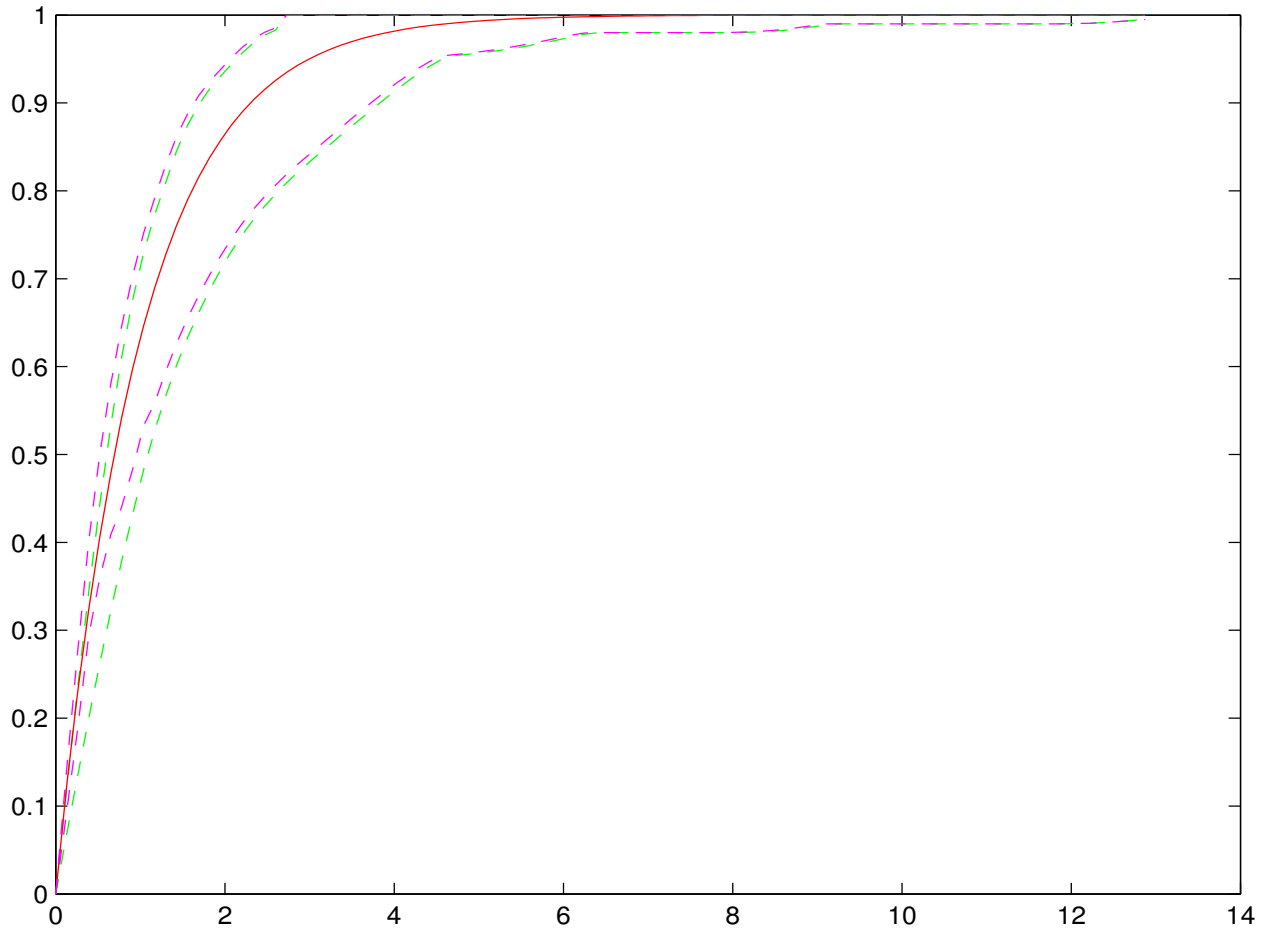
- Pro odhad optimální šířky okna používáme iterační metodu popsanou v Horová I., Zelinka J. [4]

$X \sim Exp(1)$ – odhad distrib. funkce F $(n = 100,\ h^F_{opt,0,2} = 0.8479)$



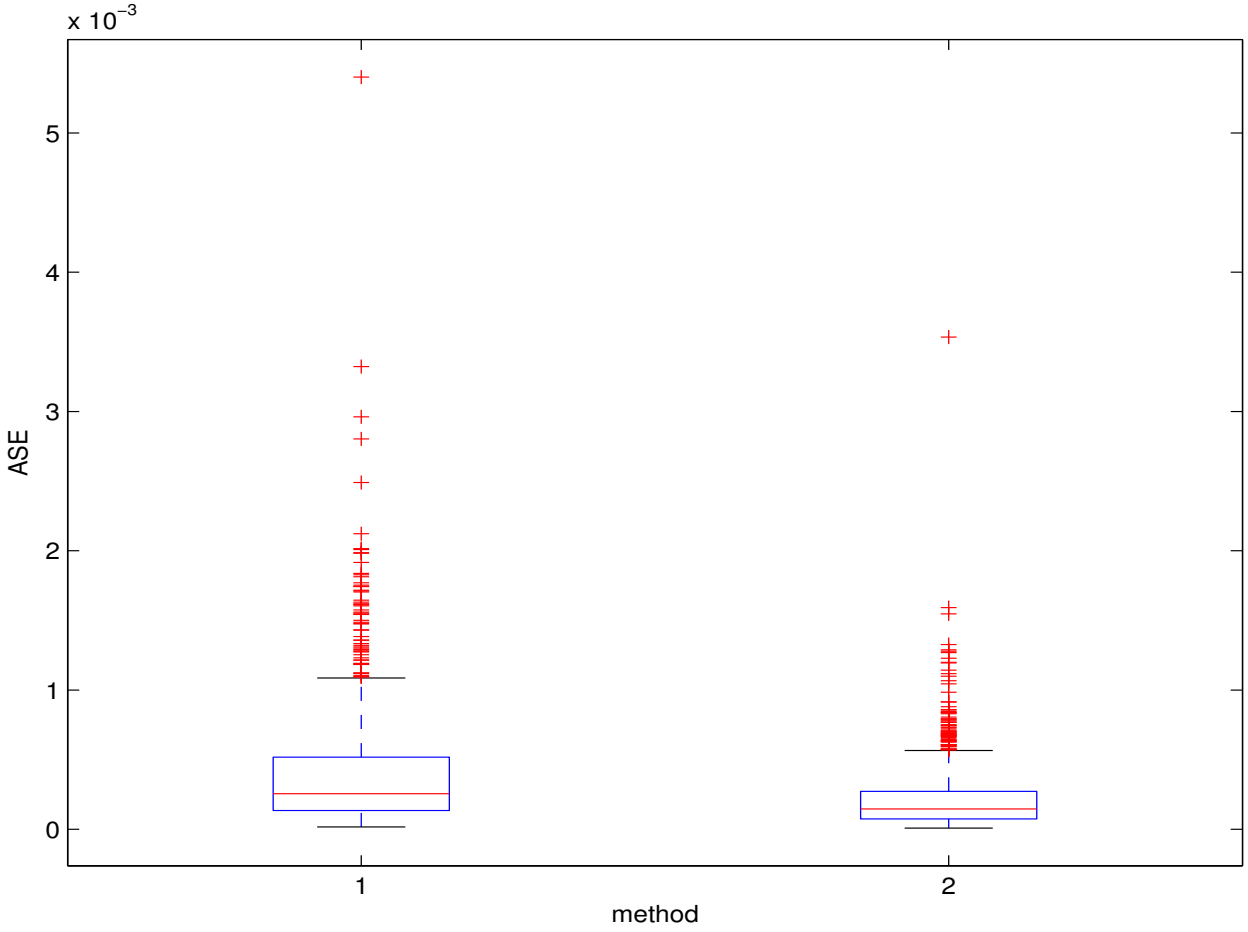
– – pro $\widehat{F}$, – – pro $\widetilde{F}$

$X \sim Exp(1)$ – odhad distrib. funkce F (1 000 simulací, $n = 100$)



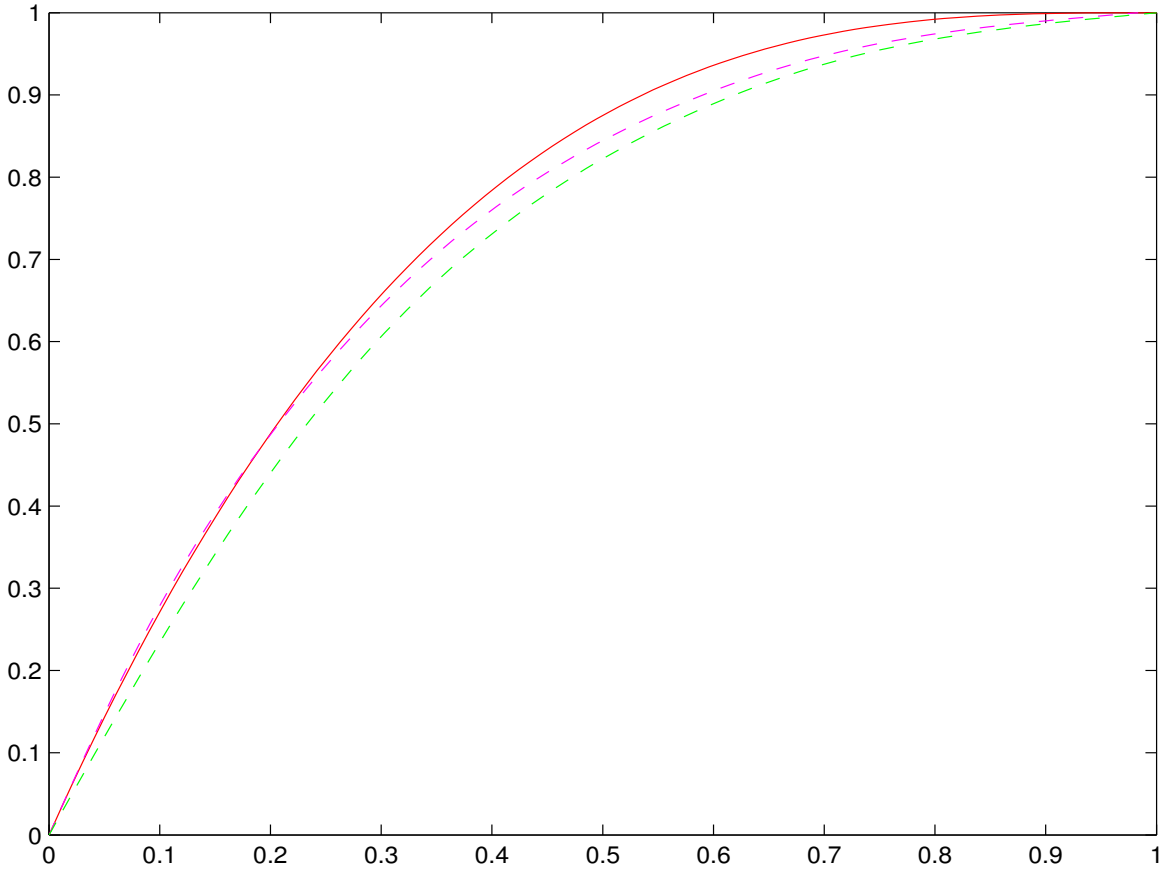
– – pro $\widehat{F}$, – – pro $\widetilde{F}$

$X \sim Exp(1)$ – boxplot pro ASE (1 000 simulací, $n = 100$)



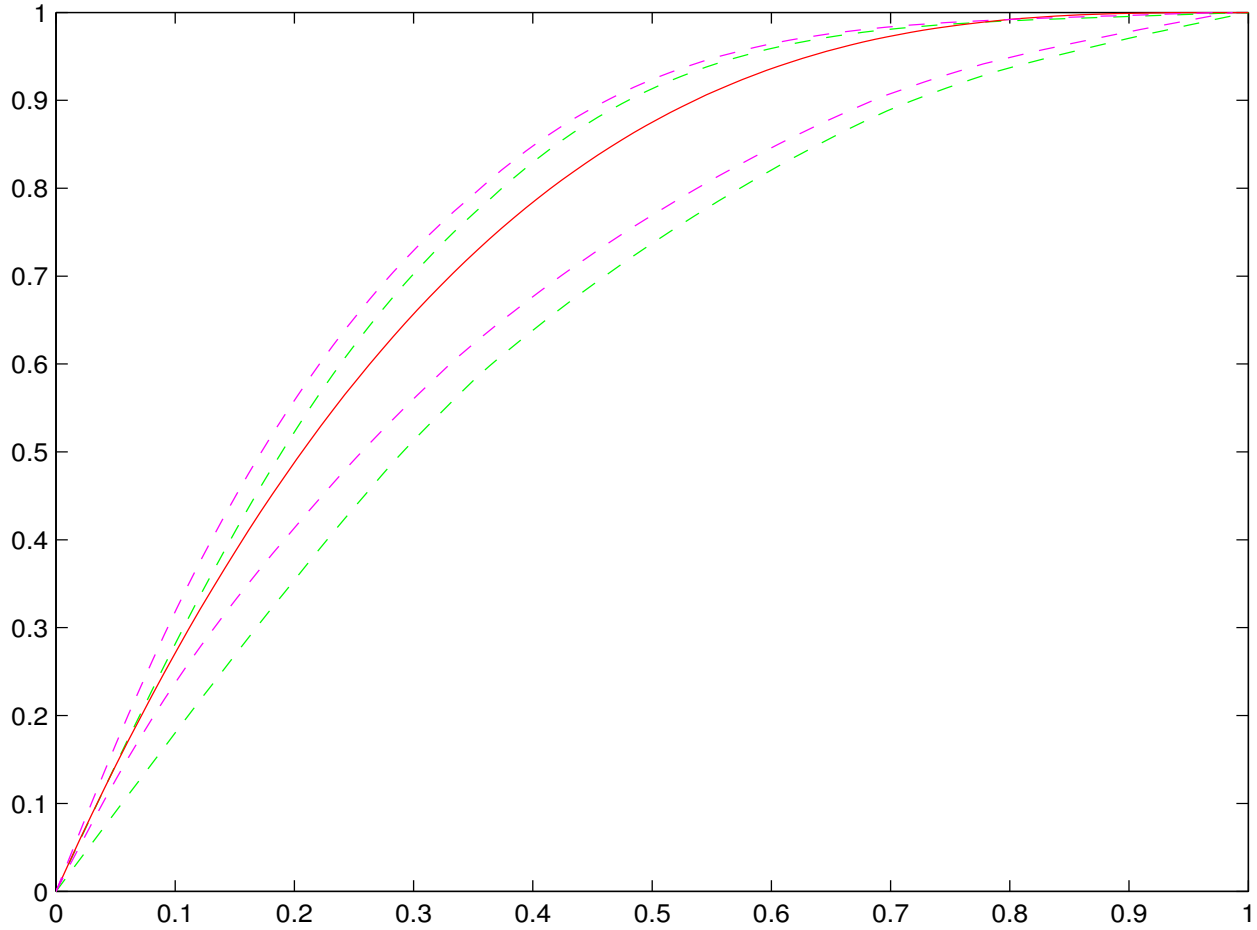
1 pro $\widehat{F}$, 2 pro $\widetilde{F}$

$$F(x) = (x-1)^3 + 1; \qquad n = 100,\ h^F_{opt,0,2} = 0.401$$



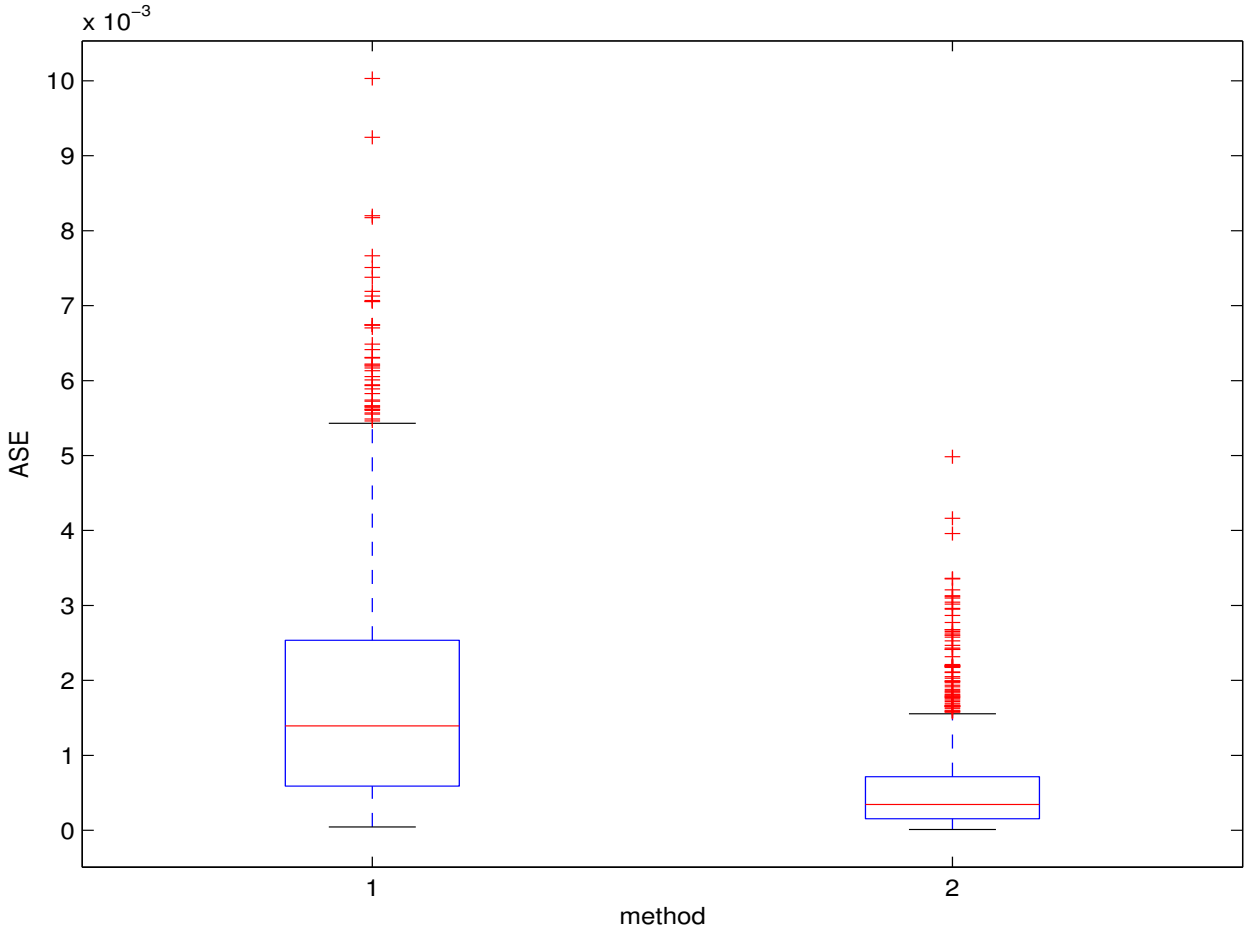
- - pro $\widehat{F}$, - - pro $\widetilde{F}$

$F(x) = (x-1)^3 + 1;$ 1 000 simulací, $n = 100$



- - pro $\widehat{F}$, - - pro $\widetilde{F}$

$F(x) = (x-1)^3 + 1;$ boxplot pro ASE (1 000 simulací, $n = 100$)

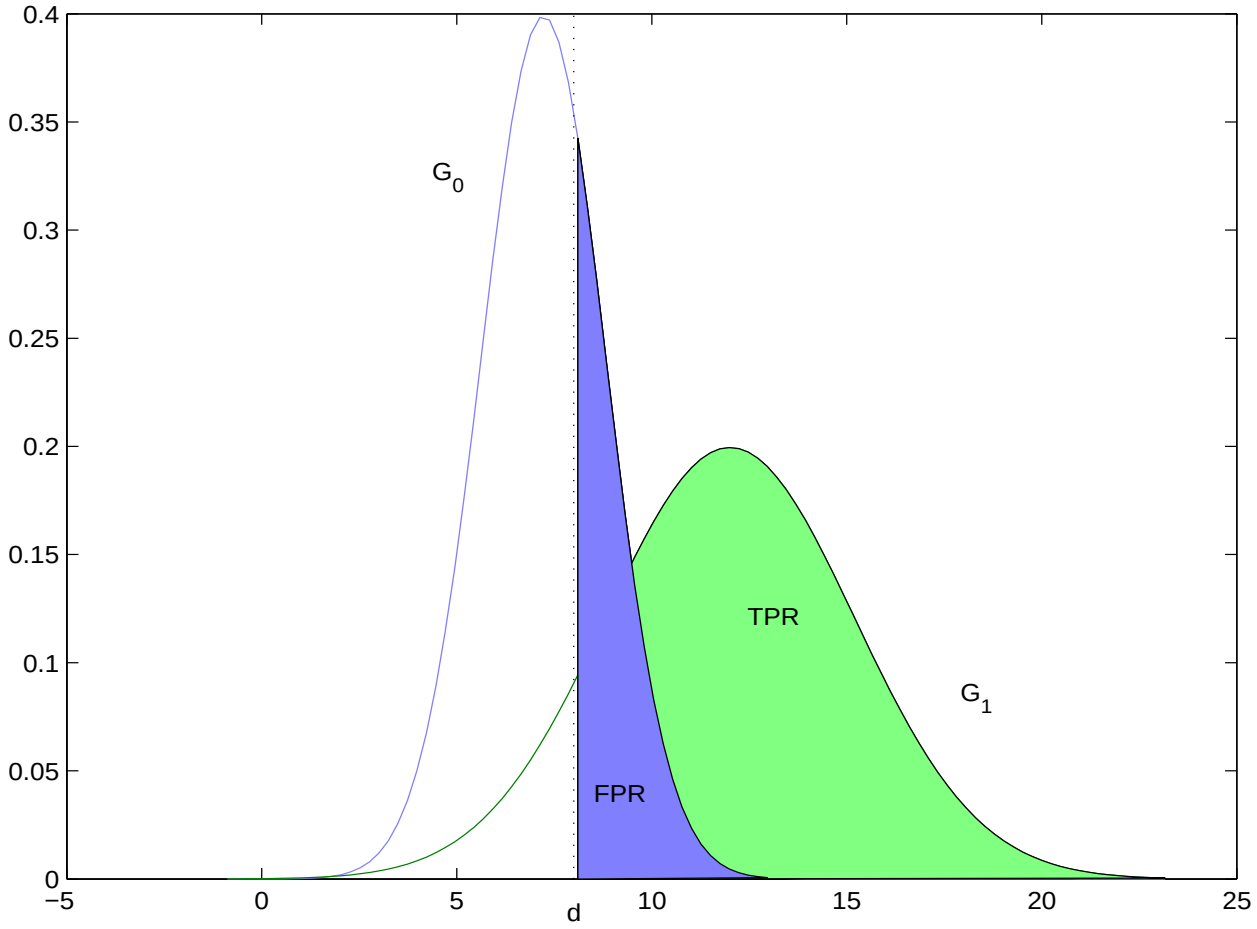


1 pro $\widehat{F}$, 2 pro $\widetilde{F}$

ROC křivky

- Uvažujeme 2 skupiny objektů $\mathcal{G}_0$ (negativní) a $\mathcal{G}_1$ (pozitivní).
- ROC křivka (Receiver Operating Characteristic) popisuje kvalitu diagnostického testu, který rozděluje dané subjekty do skupin $\mathcal{G}_0$ a $\mathcal{G}_1$ na základě hodnot spojité n.v. X (prediktoru), tj. subjekt je klasifikován jako prvek $\mathcal{G}_1$ jestliže $X \geq d$, jinak jako prvek $\mathcal{G}_0$ pro danou hodnotu $d \in \mathbb{R}$.
- ROC křivka je definována jako pravděpodobnost nesprávně klasifikovaných objektů z $\mathcal{G}_0$ (*FPR*) proti pravděpodobnosti správně klasifikovaných objektů z $\mathcal{G}_1$ (*TPR*) pro všechny možné hodnoty $d \in \mathbb{R}$, tj. ROC je dána [FPR, TPR].

ROC

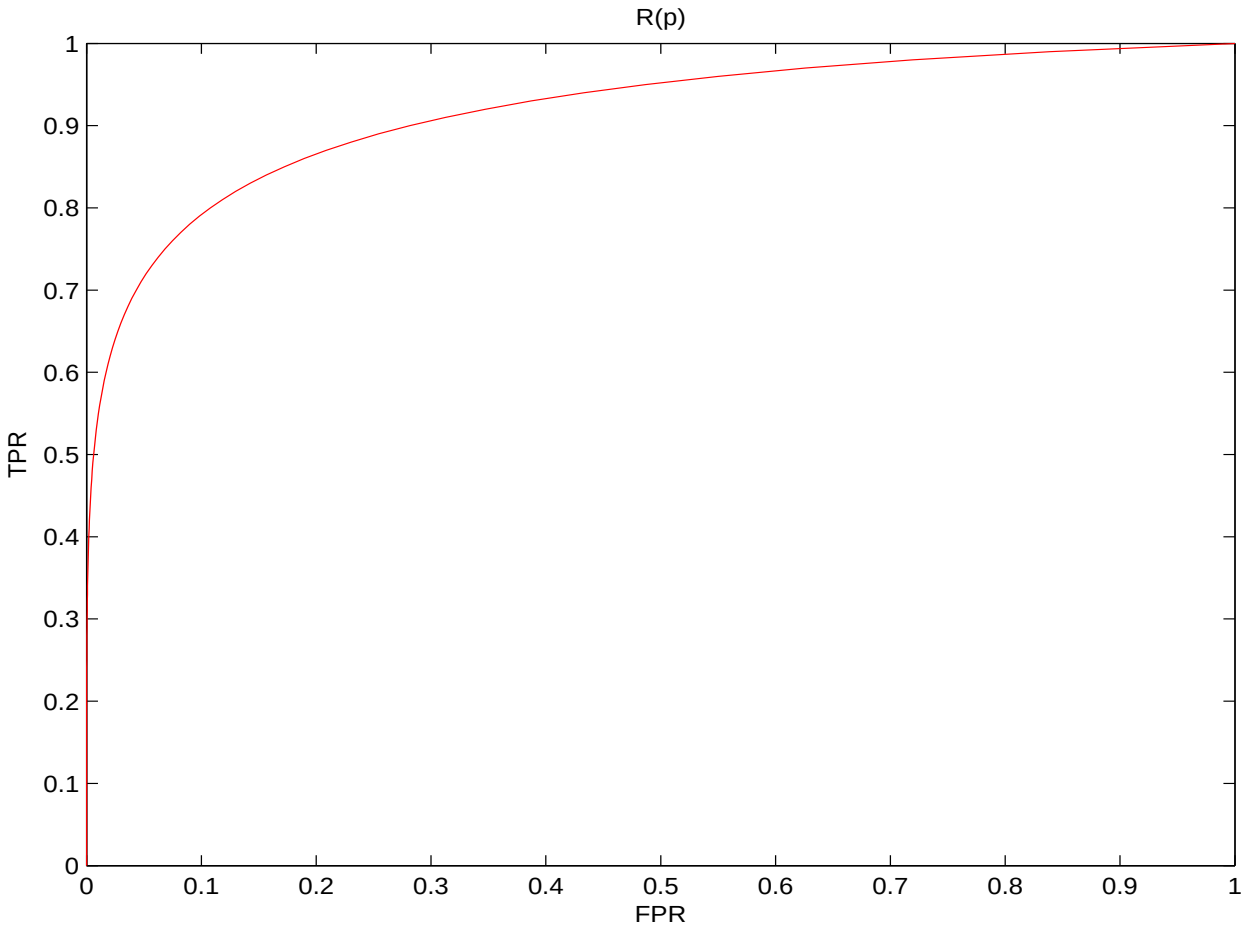


- Nechť F_0 a F_1 jsou distribuční funkce rozdělení X v $\mathcal{G}_0$ resp. $\mathcal{G}_1$. Pak

$$ROC(d) = [1 - F_0(d),\, 1 - F_1(d)], \qquad d \in \mathbb{R}$$

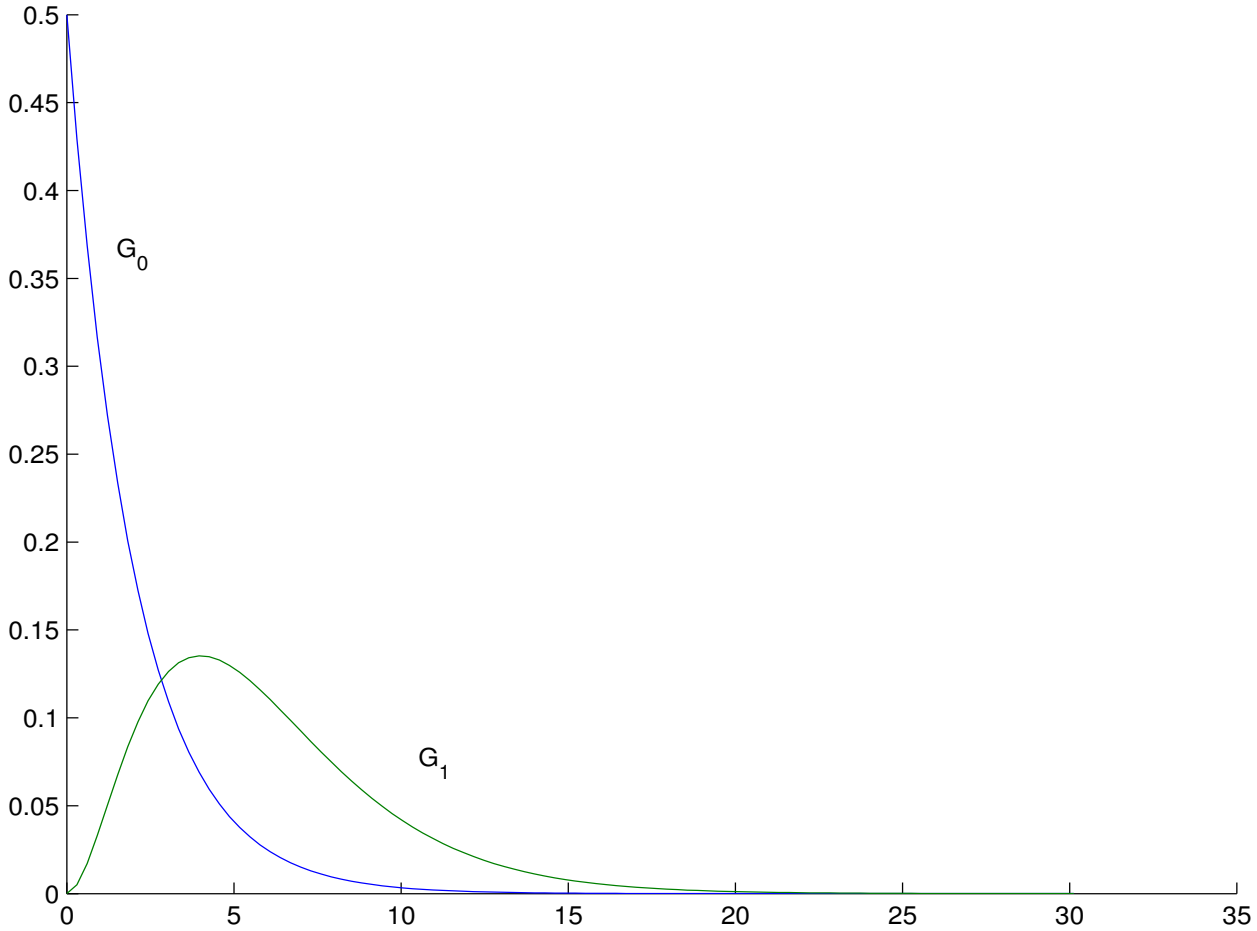
- Jiný zápis ROC

$$R(p) = 1 - F_1(F_0^{-1}(1 - p)), \quad 0 < p < 1.$$

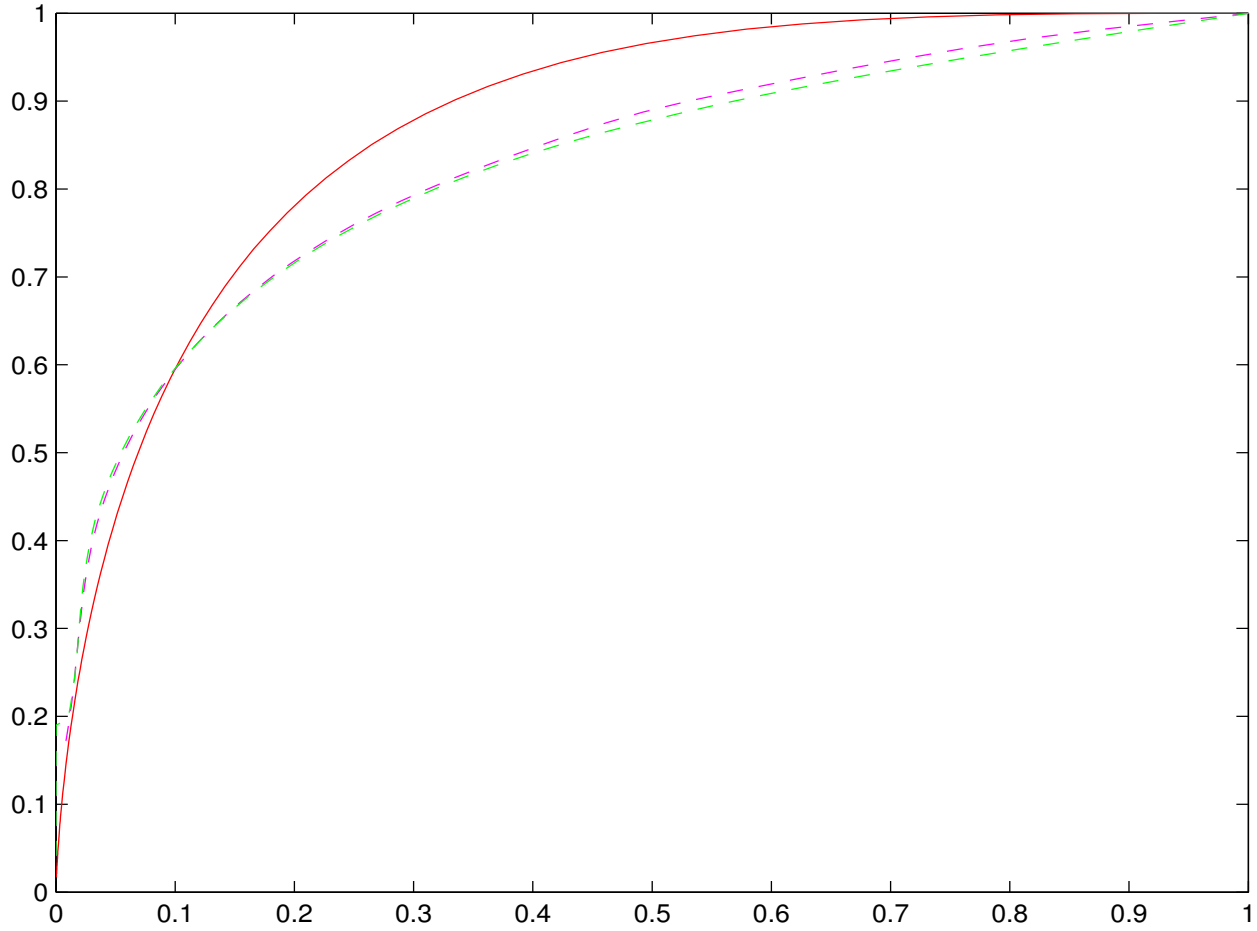

R(p)
TPR
FPR
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

ROC simulace

$X_0 \sim Exp(2),\ X_1 \sim Gamma(3,2),\ n_0 = n_1 = 50$

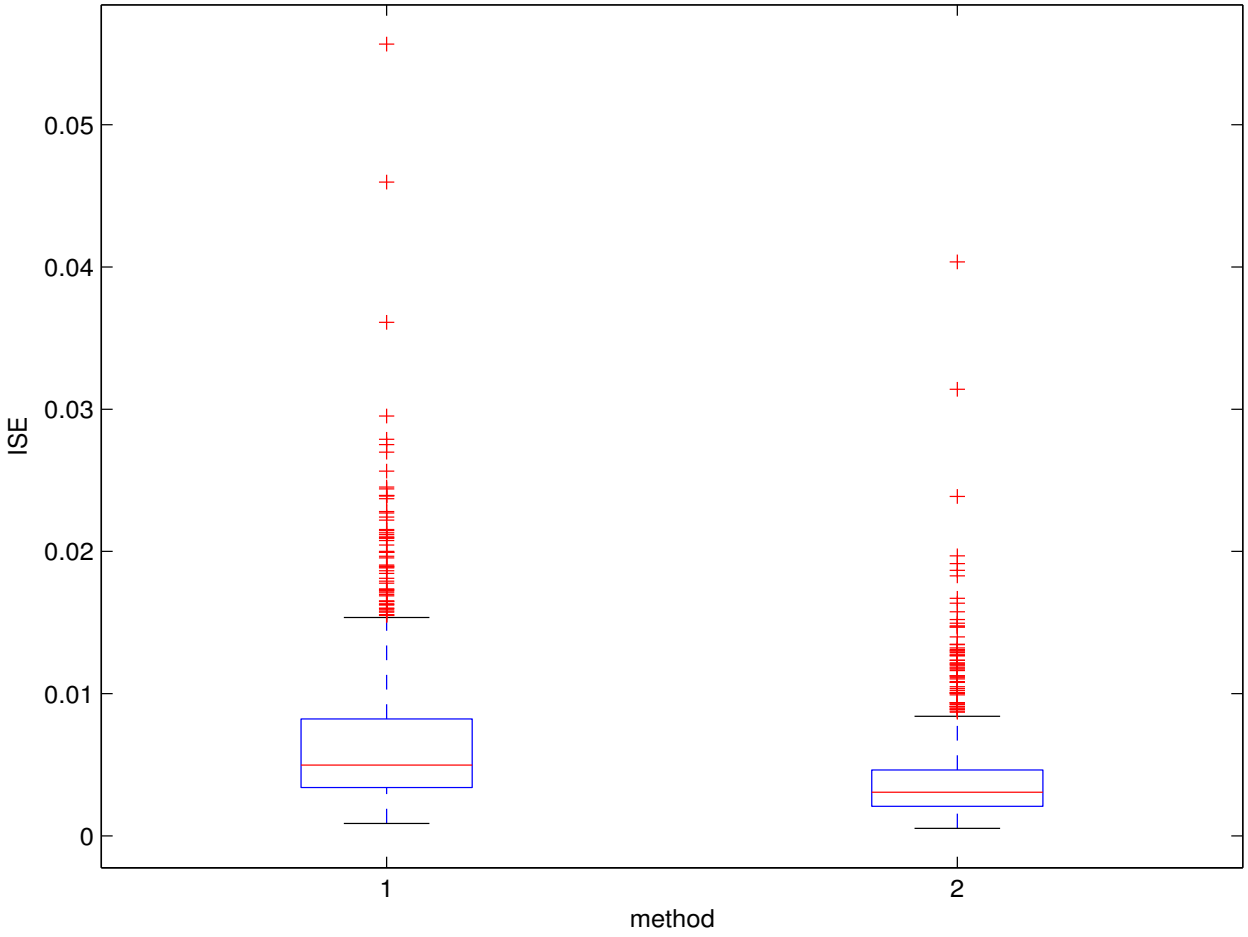


ROC



- - pro $\widehat{F}$, - - pro $\widetilde{F}$

boxplot pro ISE (1 000 simulací)



1 pro $\widehat{F}$, 2 pro $\widetilde{F}$

Reálná data

Půjčky zákazníkům

Použití (blíže nespecifikované) scoringové funkce pro ohodnocení zákazníka.

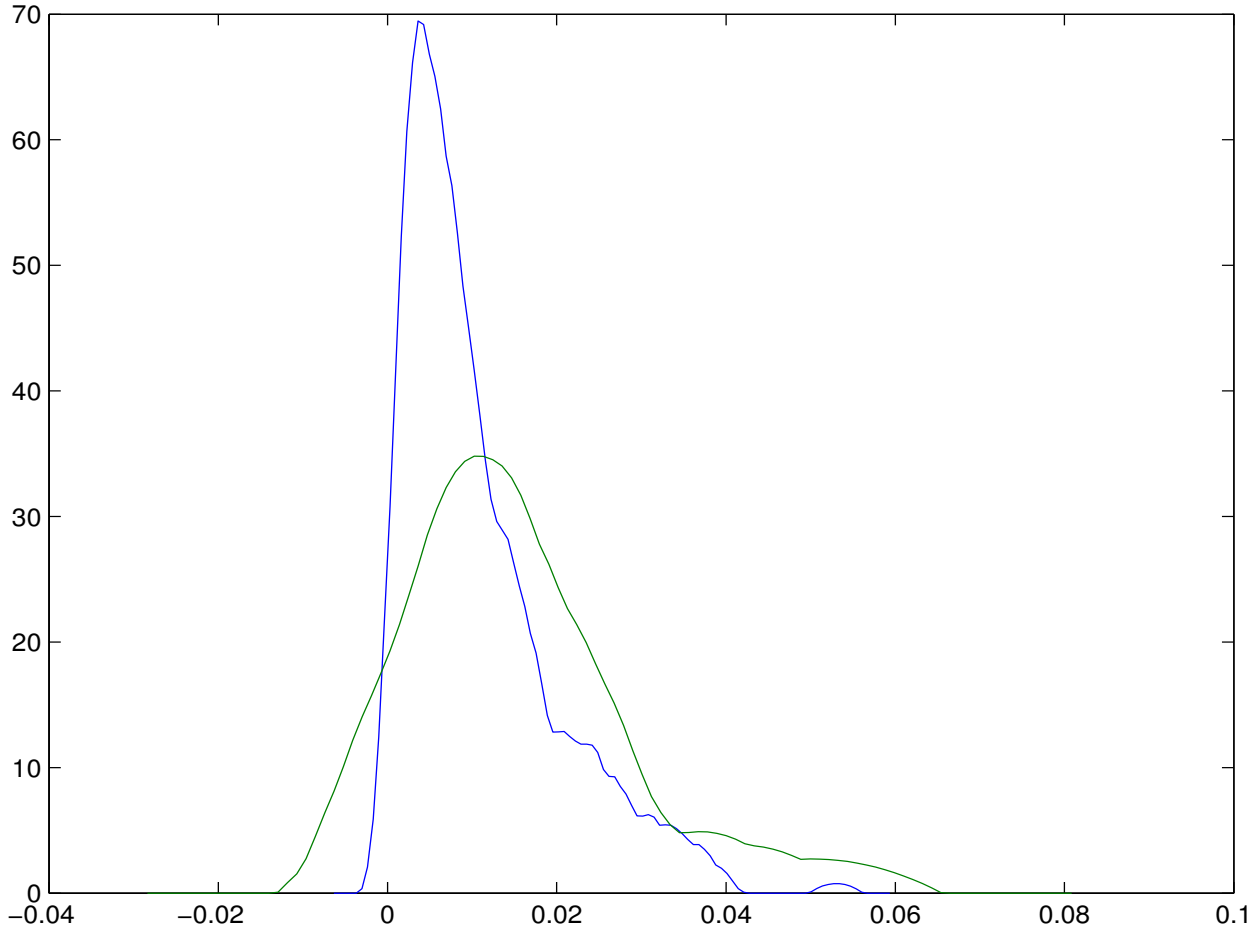
Zajímá nás, kteří zákazníci budou schopni splácet svoji půjčku.

Testovací množina: 327 zákazníků – 309 splatilo úvěr (skupina $\mathcal{G}_0$), 18 neplatilo (skupina $\mathcal{G}_1$).

Použijeme ROC křivku, abychom zhodnotili rozdělení mezi zákazníky s dobrou a se špatnou schopností splácet.

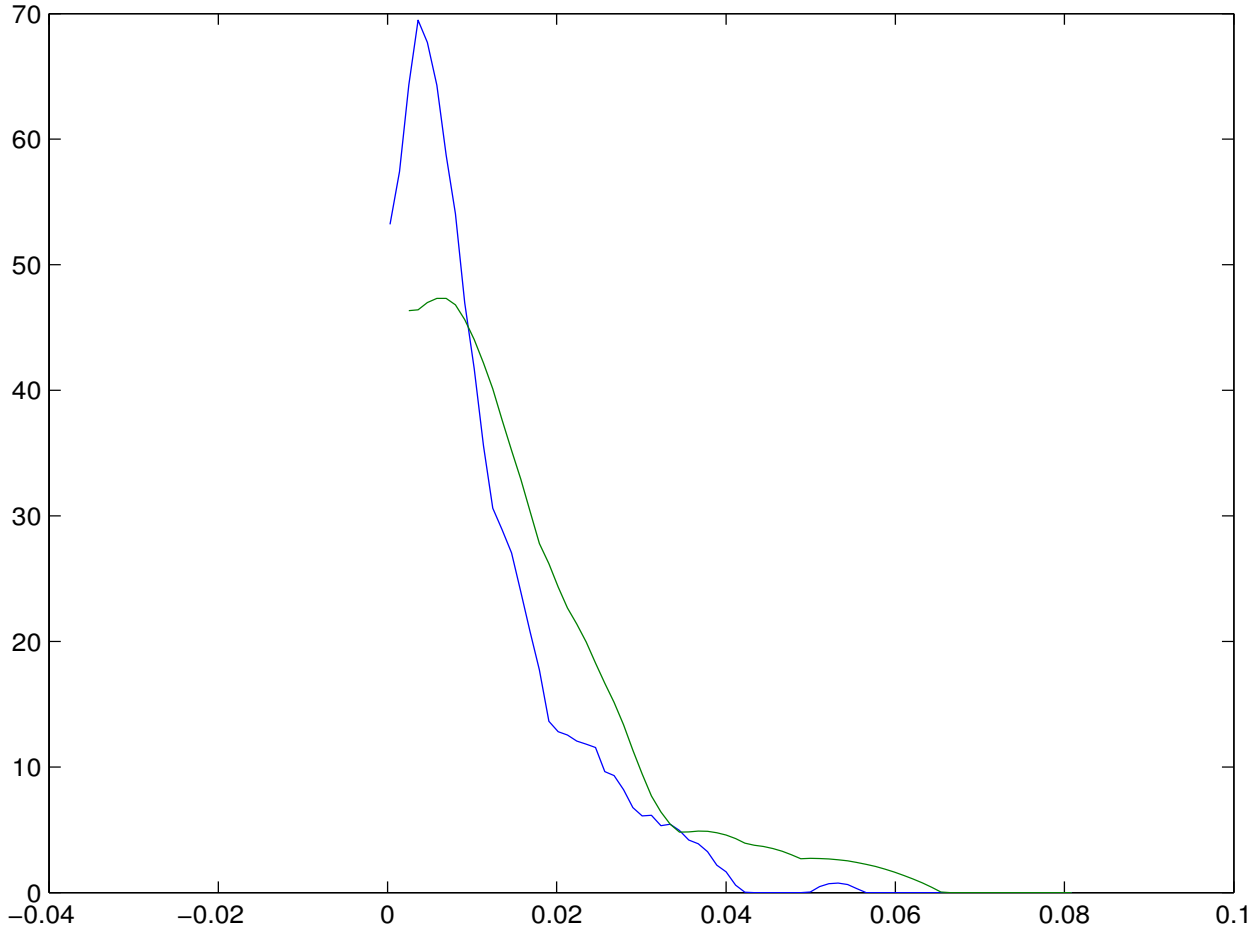
Zajímá nás, jestli naše scoringová funkce je dobrý „prediktor“.

Odhad hustot $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 0.0032$) a $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 0.0153$) s hraničními efekty.



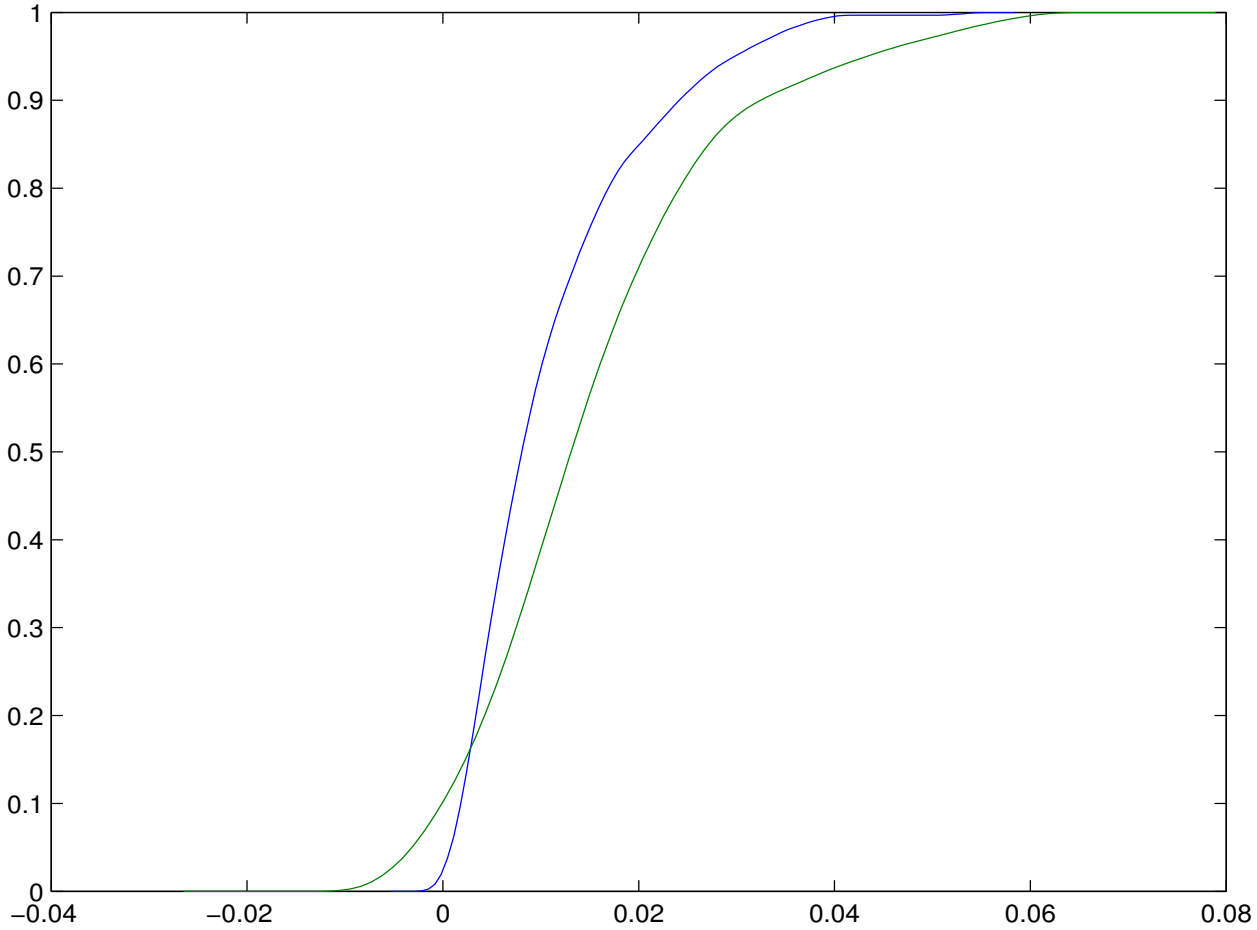
– pro $\widehat{f_0}$, – pro $\widehat{f_1}$

Odhad hustot $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 0.0032$) a $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 0.0153$) BEZ hraničních efektů.



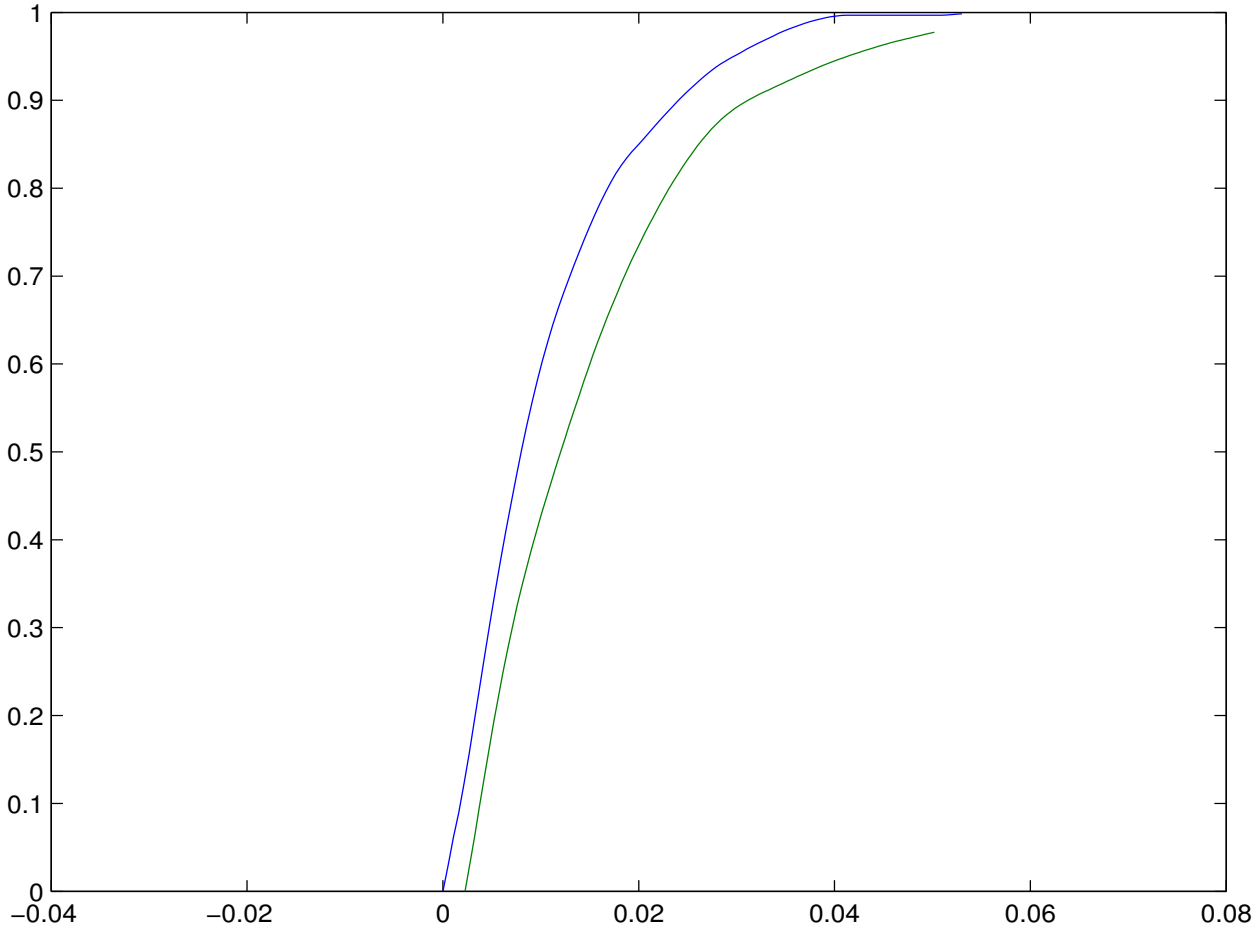
– pro $\widehat{f_0}$, – pro $\widehat{f_1}$

Odhad distribučních funkcí $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 0.0068$) a $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 0.0286$) s hraničními efekty.



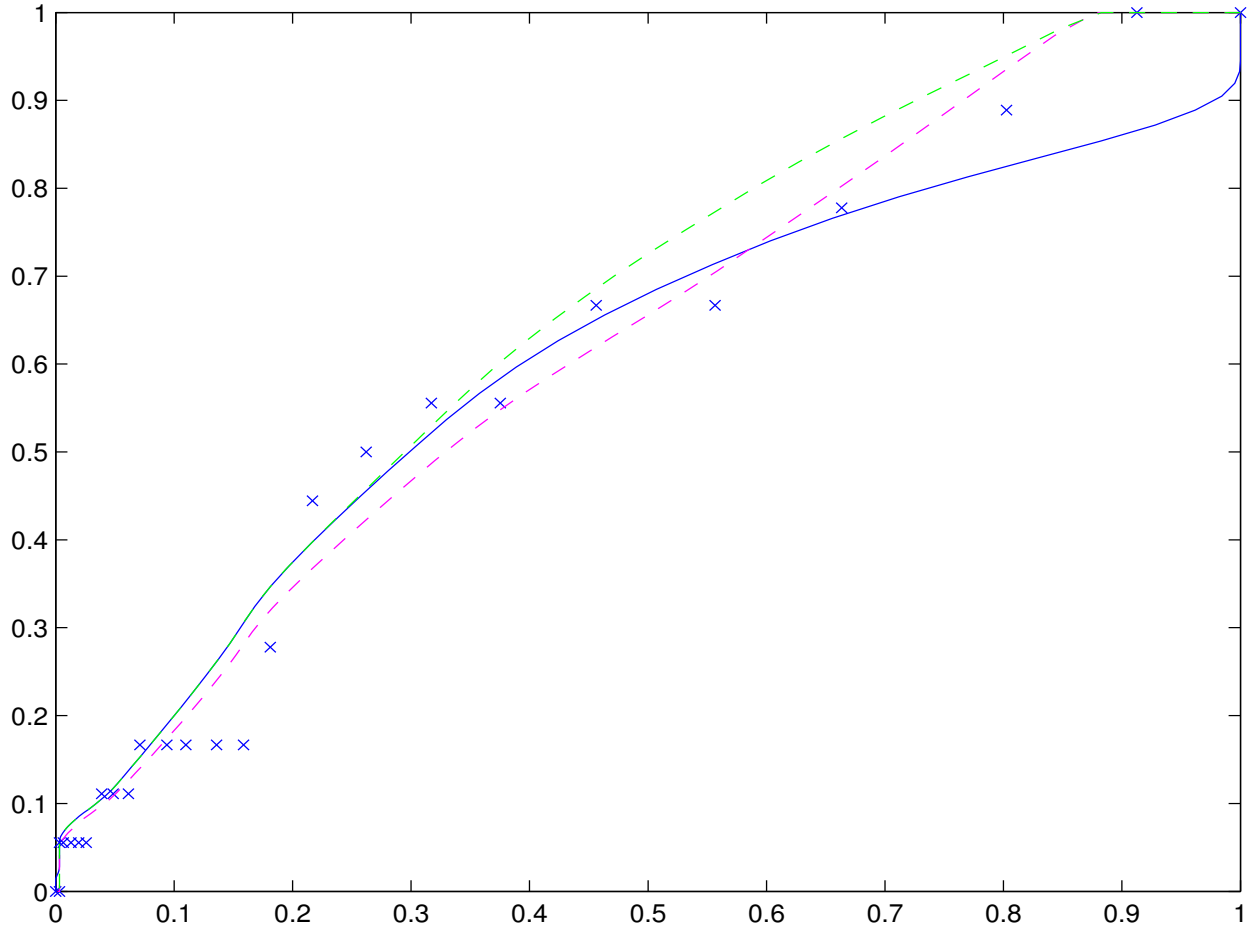
– pro $\widehat{F}_0$, – pro $\widehat{F}_1$

Odhad distribučních funkcí $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 0.0068$) a $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 0.0286$) BEZ hraničních efektů.



– pro $\widehat{F}_0$, – pro $\widehat{F}_1$

Odhad ROC křivky



– pro hraniční efekty, - - pro $\widehat{F}$, - - pro $\widetilde{F}$

Reálná data II

Zranění hlavy

Použití množství isoenzymu CK–BB (creative kinase – BB) naměřeného během 24 hodin od poranění hlavy pro předpověď následků tohoto poranění.

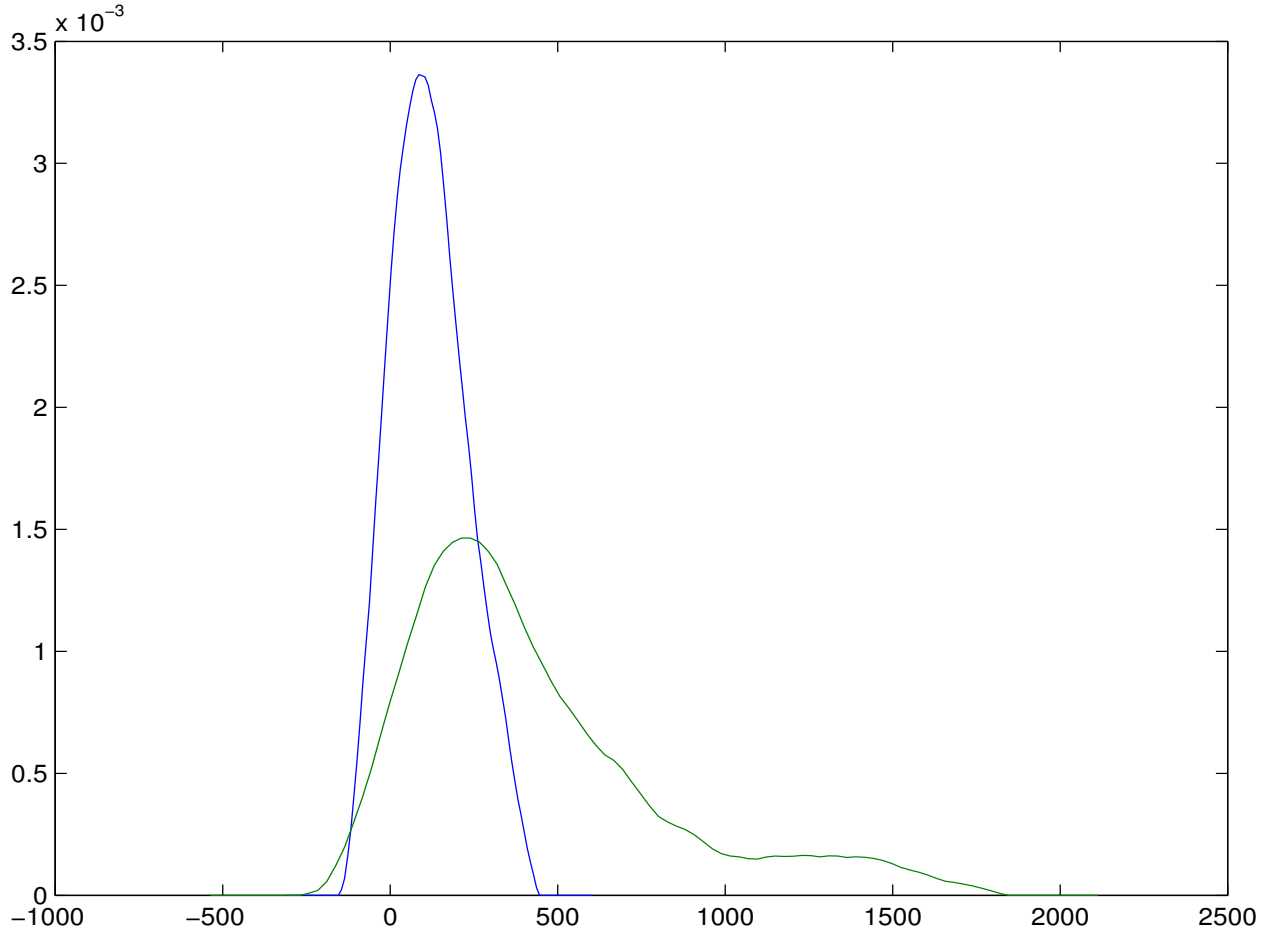
Zajímá nás, kteří pacienti budou mít trvalé následky (resp. smrt) po zranění hlavy.

60 pacientů: 19 – dobré nebo úplné uzdravení, 41 – trvalé následky nebo smrt.

Použijeme ROC křivku, abychom popsali kvalitu testu rozdělení mezi pacienty s trvalými následky a bez nich.

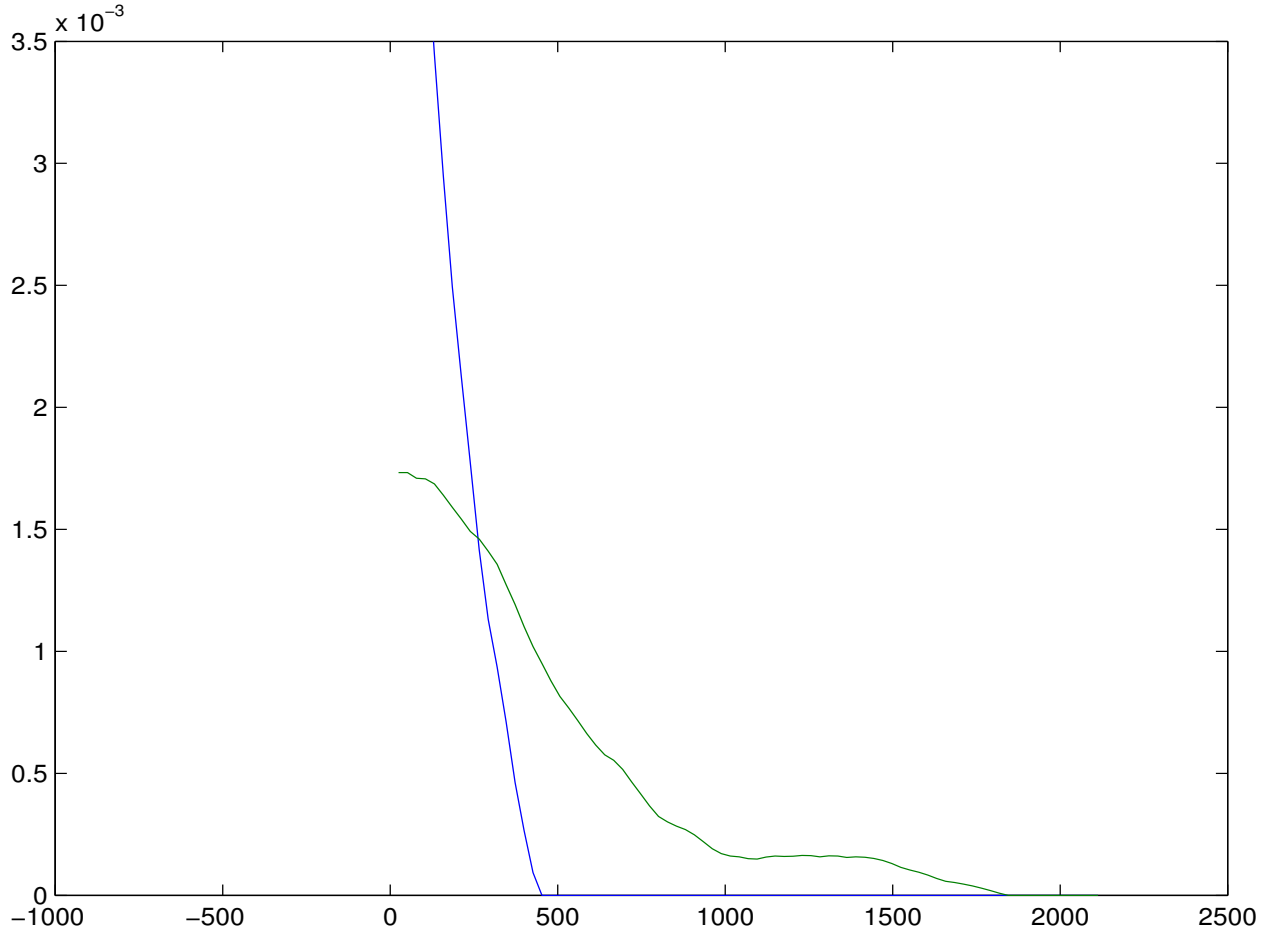
Zajímá nás, jestli CK–BB isoenzym je dobrý „prediktor“.

Odhad hustot $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 145.7135$) a $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 253.6472$) s hraničními efekty.



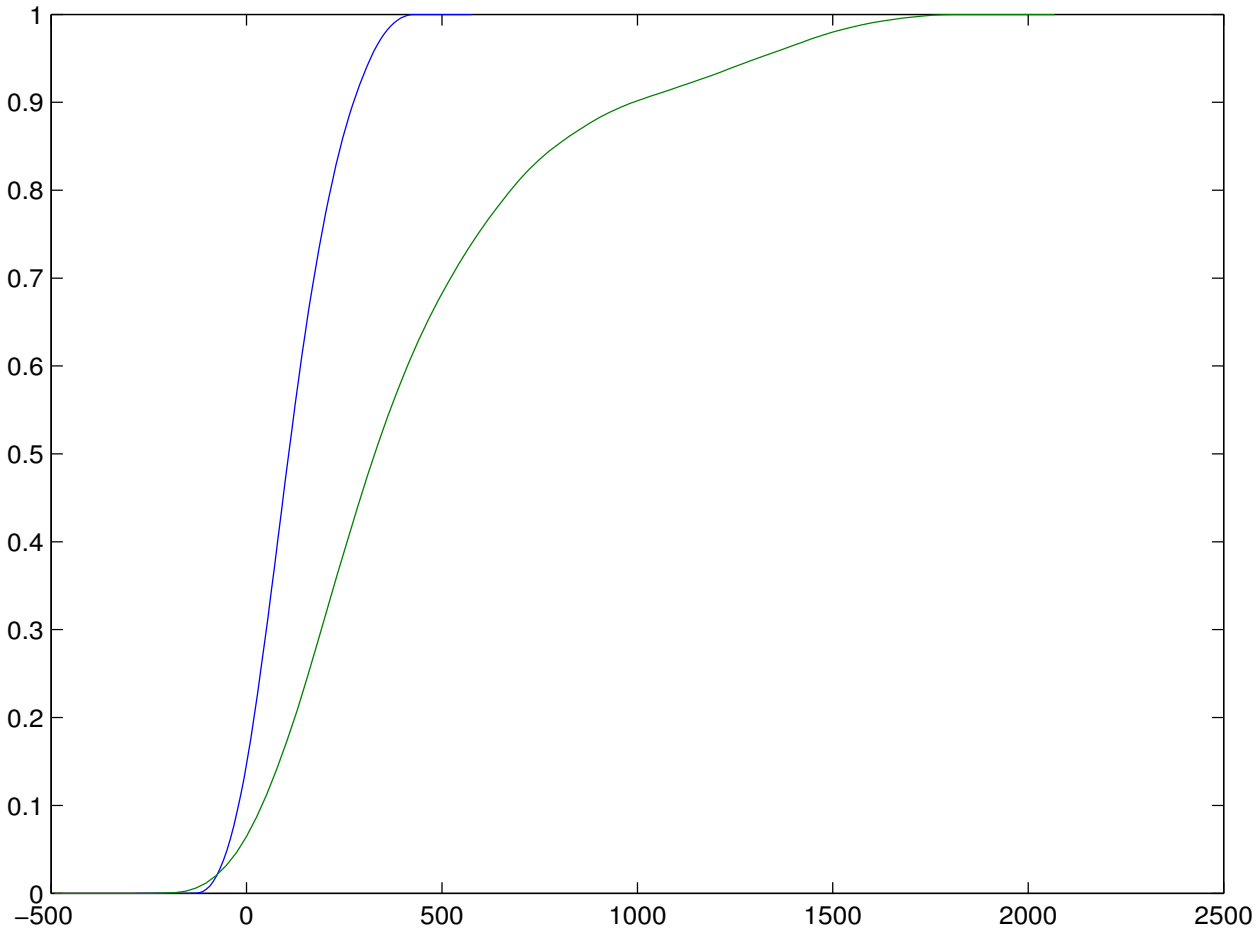
– pro $\widehat{f_0}$, – pro $\widehat{f_1}$

Odhad hustot $f_0(x)$ ($\hat{h}^{f0}_{opt,0,2} = 145.7135$) a $f_1(x)$ ($\hat{h}^{f1}_{opt,0,2} = 253.6472$) BEZ hraničních efektů.



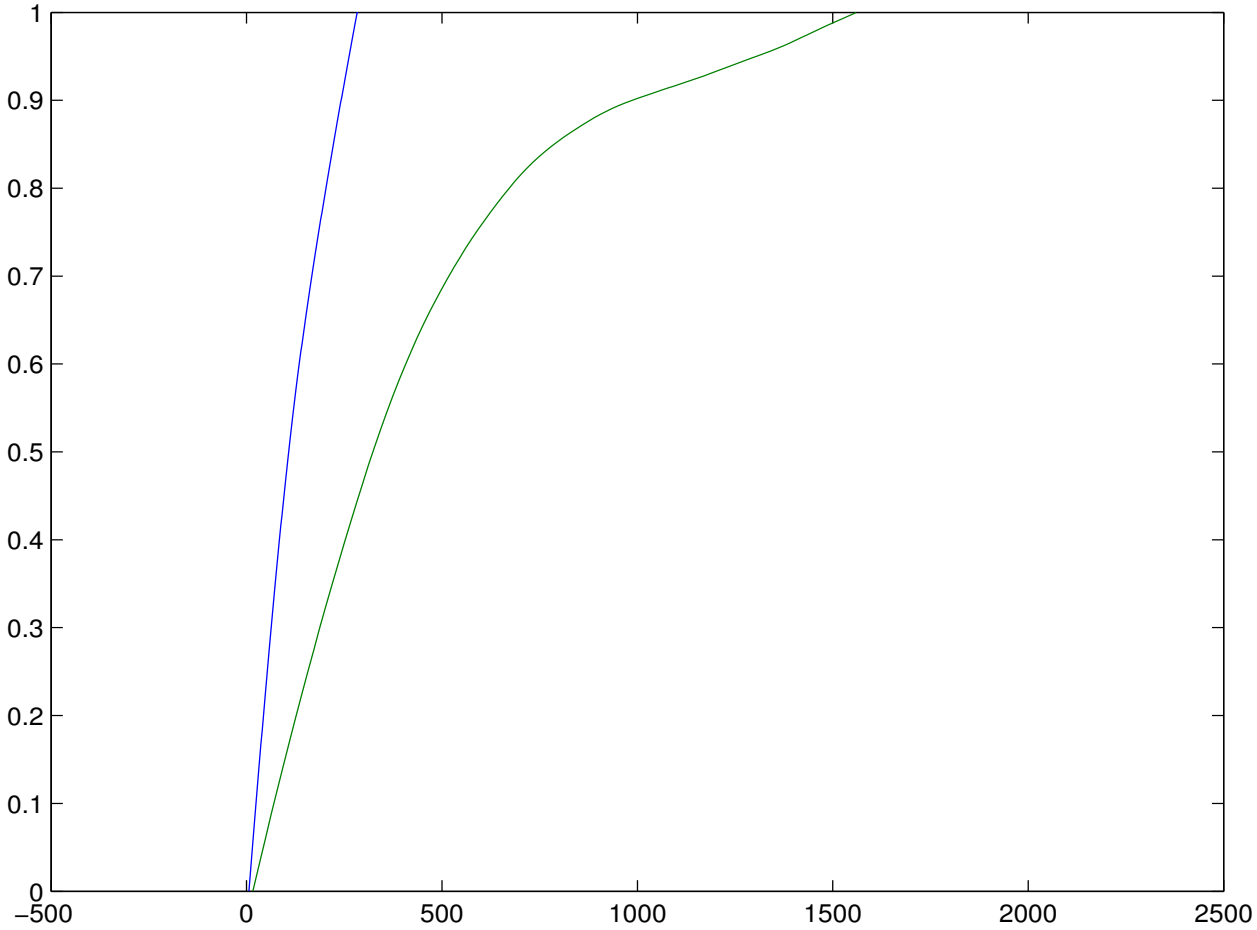
– pro $\widehat{f_0}$, – pro $\widehat{f_1}$

Odhad distribučních funkcí $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 158.6975$) a $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 276.5697$) s hraničními efekty.



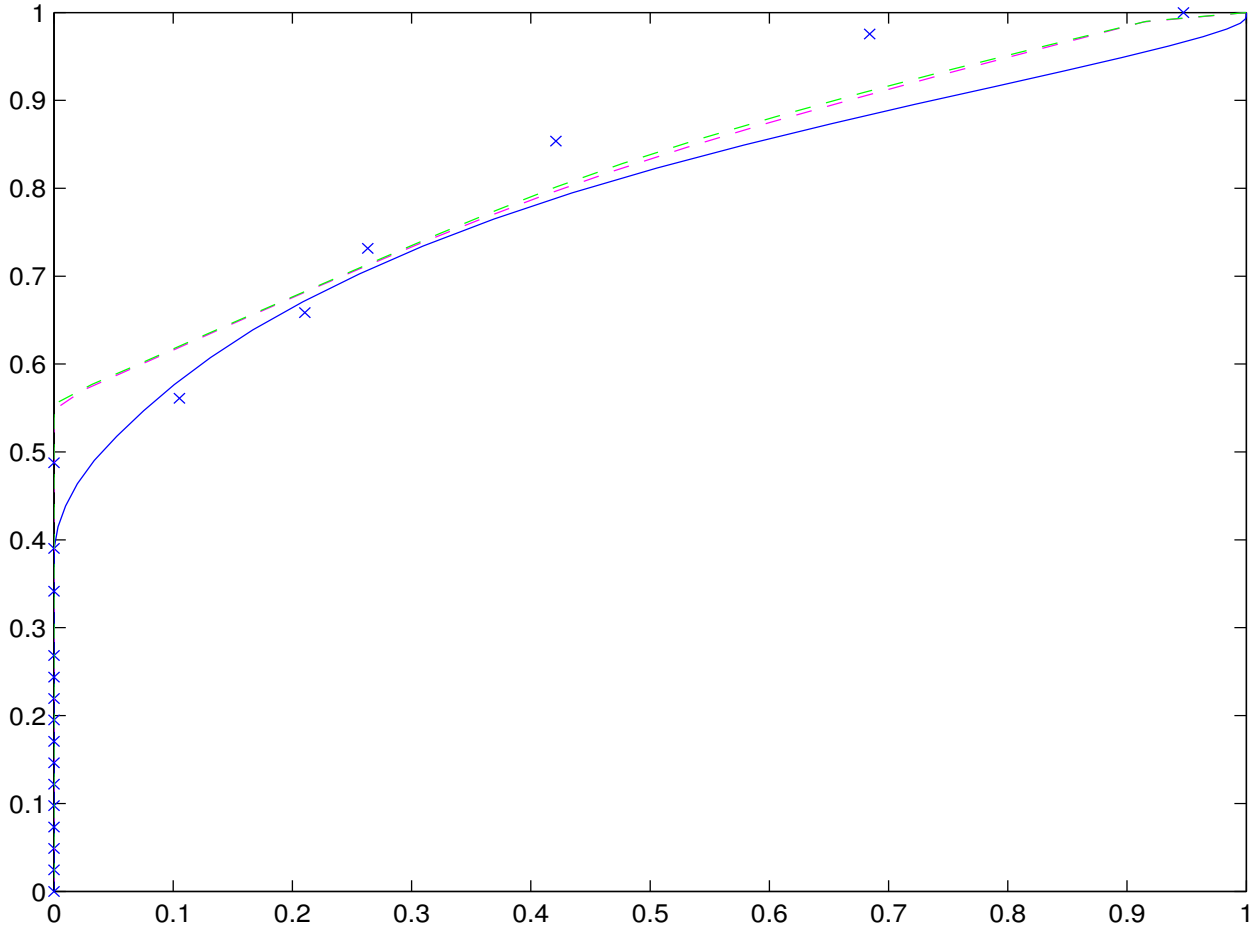
– pro $\widehat{F}_0$, – pro $\widehat{F}_1$

Odhad distribučních funkcí $F_0(x)$ ($\hat{h}^{F0}_{opt,0,2} = 158.6975$) a $F_1(x)$ ($\hat{h}^{F1}_{opt,0,2} = 276.5697$) BEZ hraničních efektů.



– pro $\widehat{F}_0$, – pro $\widehat{F}_1$

Odhad ROC křivky



– pro hraniční efekty, - - pro $\widehat{F}$, - - pro $\widetilde{F}$

Literatura

[1] Azzalini, A.: *A note on the estimation of a distribution function and quantiles by a kernel method*. Biometrika, 68, No 1, pp. 326–328, 1981.

[2] Bowman, A., Hall, P., Prvan, T.: *Bandwidth selection for the smoothing of distribution functions*. Biometrika, 85, No 4, pp. 799–808, 1998.

[3] Härdle, W.: *Applied nonparametirc regression*. Cambridge University Press, 1991.

[4] Horová, I., Zelinka, J.: *Different approaches to ROC curve fitting for a continuous diagnostic test*. CSDA, submitted, 2007.

[5] Karunamuni, R.J., Alberts T.: *On boundary correction in kernel density estimation*. Statistical Methodology 2, pp. 191–212, 2005.

[6] Lloyd, C.J., Zhou Yong: *Kernel estimators of the ROC curve are better than empirical*. Statistics and Prob. Letters 44, pp. 221–228, 1999.

[7] Silverman, B.W.: *Density estimation for statistics and Data Analysis*. Chapman and Hall, New York, 1986.

[8] Terrell, G. R.: *The maximal smoothing principle in density estimation*. Journal of the American Statistical Association. Vol. 85, No. 410, pp. 440-447, 1990.

[9] Wand, I.P. and Jones, I.C.: *Kernel smoothing*. Chapman & Hall, London, 1995.