

POROVNÁNÍ NEPARAMETRICKÝCH BAYESOVSKÝCH ODHADŮ PŘI CENZOROVÁNÍ

Michal Friesl

*Katedra matematiky
Fakulta aplikovaných věd
Západočeská univerzita v Plzni*

Co nás čeká

Porovnat v Koziolově-Greenově modelu

- odhad funkce spolehlivosti (za K-G předpokladu)
- s obecnými odhady z cenzorovaných dat

Obsah

- model cenzorování (obecný $\times$ Koziolův-Greenův model)
- neparametrické bayesovské odhady
- odhady na datech
- bayesovská rizika

Náhodné cenzorování

Sledovaná veličina

- doba do poruchy X (doba života, ...)
- funkce spolehlivosti $S(t) = P(X > t) = \exp(-\Lambda(t))$, $t > 0$

Cenzorování zprava

- pozorování ukončeno nejpozději v okamžiku Y (X, Y nezáv.)
- pozorujeme jen $Z = \min(X, Y)$, $I = I_{[X \leq Y]}$

Koziolův-Greenův model

- předpokládáme navíc $S_Y(t) = (S(t))^\gamma$, $t > 0$,
pro nějaké $\gamma > 0$, tj. $\Lambda_Y(t) = \gamma \Lambda(t)$, $t > 0$
- informativní cenzorování, možnost zlepšení odhadů

Pozorování

K dispozici

- náhodný výběr $(Z_1, I_1), \dots, (Z_n, I_n)$

Vícenásobná pozorování

- $N \leq n$ navzájem různých časů $T_1 < \dots < T_N$
- dodefinujeme $T_0 = 0, T_{N+1} = \infty$
- „ i “-tý interval je $\langle T_{i-1}, T_i \rangle, i = 1, \dots, N+1$

Pro j -tý čas T_j

- $U_j = \#$ necenzorovaných pozorování se $Z_k = T_j$
- $C_j = \#$ cenzorovaných pozorování se $Z_k = T_j$
- $N_j = \#$ pozorování se $Z_k > T_j$

Apriorní rozdělení

Neparametrický bayesovský přístup

- neznámé parametry považujeme za náhodné veličiny
- zde parametr S nekonečněrozměrný $\rightarrow$ náhodný proces
- procesy zprava neutrální $\leftrightarrow \Lambda$ s nezávislými přírůsky

Gama proces $\mathcal{G}(n_0, \Lambda_0)$

- přírůsky $\Lambda(s, t) = \Lambda(t) - \Lambda(s) \sim G(n_0, n_0 \Lambda_0(s, t))$
- $n_0 > 0$, Λ_0 libovolná kumulativní intenzita (předp. spoj.)

Volba parametrů

- $E \Lambda(s, t) = E \Lambda_0(s, t)$, $\text{var } \Lambda(s, t) = \Lambda_0(s, t) / n_0$
- $E S(t) = E \exp(-\Lambda(t)) \neq \exp(-E \Lambda(t)) = \exp(-\Lambda_0(t))$

Obecný odhad

Aposteriorní rozdělení

- i při cenzorování neutrální zprava (Ferguson, Phadia)

$$\hat{S}_{\text{FP}}(t) = \prod_{j < i} \left[\underbrace{\left(\frac{M_{j-1}}{M_{j-1}^1} \right)^{n_0 \Lambda_0(T_{j-1}, T_j)}}_{\text{přes } j\text{-tý interval}} \underbrace{\frac{c_j^1}{c_j}}_{\text{skok v } T_j} \right] \cdot \left(\frac{M_{i-1}}{M_{i-1}^1} \right)^{n_0 \Lambda_0(T_{i-1}, t)}$$

kde $M_j^1 = n_0 + N_{j-1} + 1$, $c_j^1 = \sum_{k=1}^{U_j} (-1)^k \binom{U_j}{k} \ln \frac{M_j^1}{M_j^1 + k}$

Rozumí se

- za předpokladu nezávislosti mezi Y a S, X
- což v Koziolově-Greenově modelu neplatí

V Koziolově-Greenově modelu

Apriorní rozdělení

- ke gama procesu $\Lambda \sim \mathcal{G}(n_0, n_0 \Lambda_0)$ nezávisle
- γ s hustotou $\pi(\gamma)$, $\gamma > 0$, vzhledem k nějaké μ na $(0, \infty)$

Aposteriorní rozdělení Λ — podmíněně nezávislé přírůstky

- mezi časy pozorování

$$(\Lambda(s, t) \mid \text{data}, \gamma) \sim G(M_{j-1}(\gamma), n_0 \Lambda_0(s, t))$$

kde $M_j^m(\gamma) = n_0 + N_j(1 + \gamma) + m$, $j = 0, \dots, N$,

- v časech pozorování skoky s hustotou (při daném γ)

$$x^{-1} e^{-(M_j(\gamma) + C_j)x} (1 - e^{-x})^{U_j} (1 - e^{-\gamma x})^{C_j} / c_j(\gamma)$$

kde $c_j(\gamma)$ je normovací konstanta

Odhad

Aposteriorní střední hodnota

- při daném γ je $E(e^{-\Lambda(t)} | \text{data}, \gamma) = \prod E(e^{-\Lambda(\cdot, \cdot)} | \text{data}, \gamma)$
- vážíme aposteriorní hustotou γ

$$\pi(\gamma | \text{data}) \propto \left(\prod_{j=1}^N q_j(\gamma) \right) \pi(\gamma) \quad \text{vzhledem k } \mu$$

$$\text{kde } q_j^m(\gamma) = (M_{j-1}^m(\gamma))^{-n_0 \Lambda_0(T_{j-1}, T_j)} c_j^m(\gamma), \quad j = 1, \dots, N$$

Dostáváme

- $A(s) = \int \left(\prod_{j < i} q_j^1(\gamma) \right) \left(\prod_{j \geq i} q_j(\gamma) \right) \left(\frac{M_{i-1}(\gamma)}{M_{i-1}^1(\gamma)} \right)^{n_0 \Lambda_0(T_{i-1}, s)} \times \pi(\gamma) d\mu(\gamma)$
- tj. odhad $\widehat{S}(t) = A(t)/A(0)$

Počítáme

Explicitní vzorec

- až na integraci vzhledem ke γ
- žádná integrace, když γ diskrétní

Problematická normovací konstanta

- $c_j^m(\gamma) = \sum_{k=1}^{U_j} \sum_{\ell=1}^{C_j} (-1)^{k+\ell} \binom{U_j}{k} \binom{C_j}{\ell} \ln \frac{M_j^m(\gamma) + C_j}{M_j^m(\gamma) + C_j + k + \ell \gamma}$
- vícenásobná pozorování: sčítanci $\approx 1/M_j$, $c_j^m(\gamma) \approx (1/M_j)^{??}$

Neparametrické odhady

- Kaplanův-Meierův odhad
- Abdushukurovův-Chengův-Linův (ACL) odhad

V.A.C.U.R.G.

Zdroj

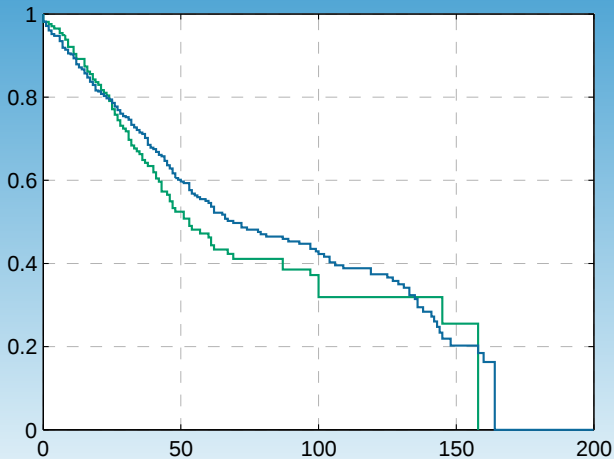
- doby života po ošetření u pacientů s rakovinou prostaty
- Hollander, Proschan (1979)

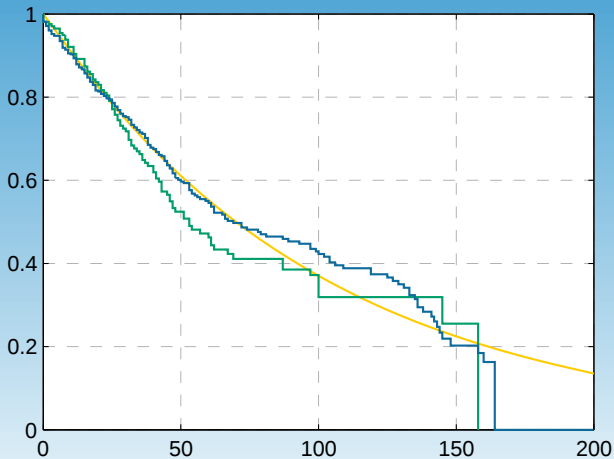
Data

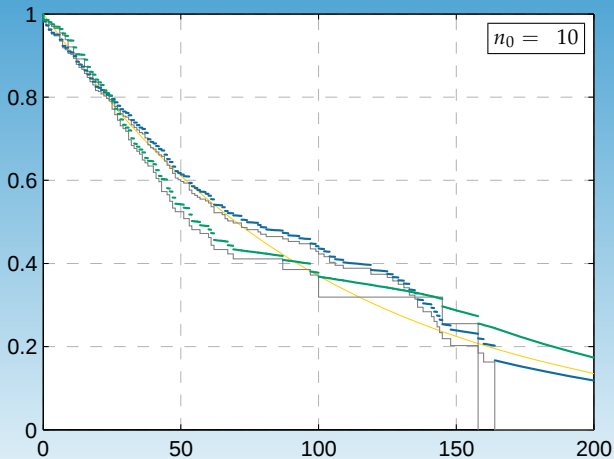
- 211 pozorování, z nich 90 necenzorováno
- minimum 0, maximum 164 měsíců
- 97 různých hodnot, z nich 53 vícenásobných

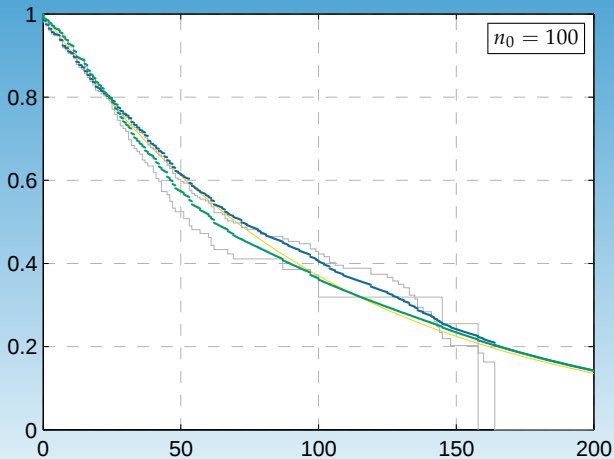
Koziol, Green (1976)

- nezamítají $\text{Exp}(100)$ za předpokladu K-G modelu
- data z K-G ale asi nepocházejí









Channig House Data

Délka života (domov důchodců)

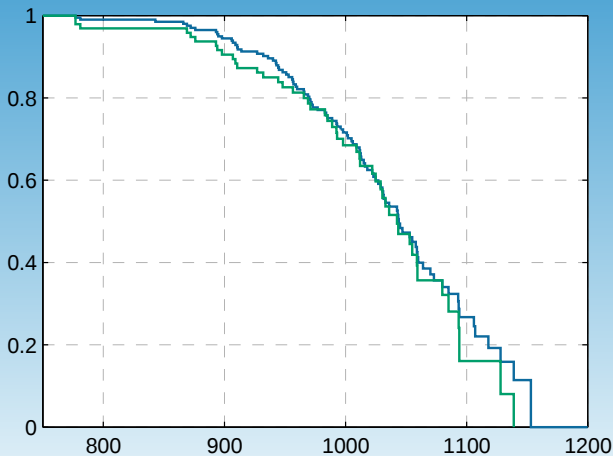
- Hyde (1977), muži, useknutí zleva ignorujeme
- Csörgő (1988) — splňují Koziolův-Greenův model

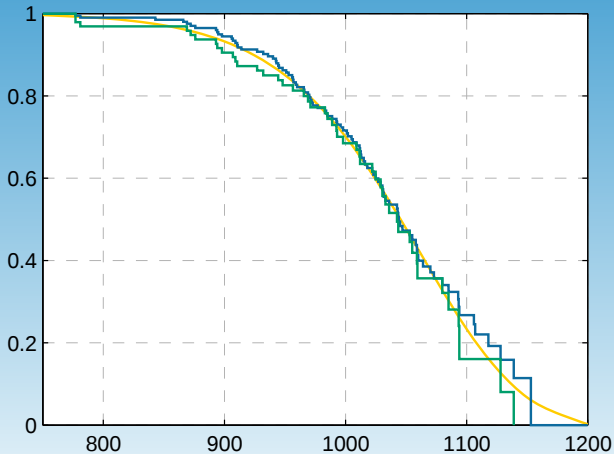
Data

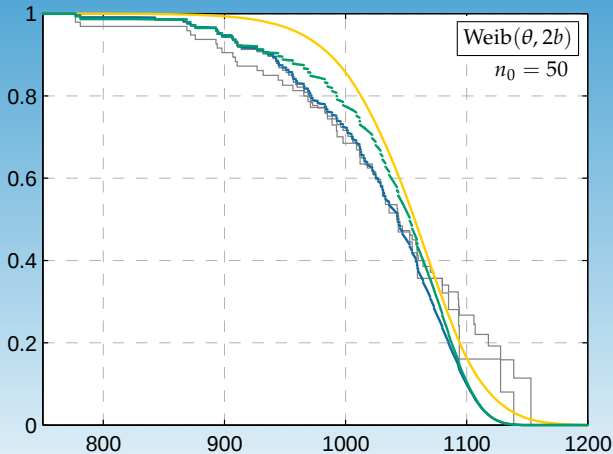
- 97 pozorování, z nich 46 necenzorovaných
- minimum 775, maximum 1153 měsíců
- 83 různých časů, z nich 10 je dvakrát a 2 třikrát

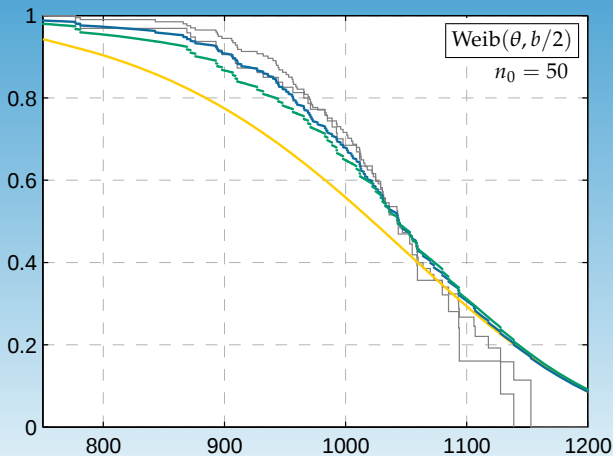
Apriorní rozdělení

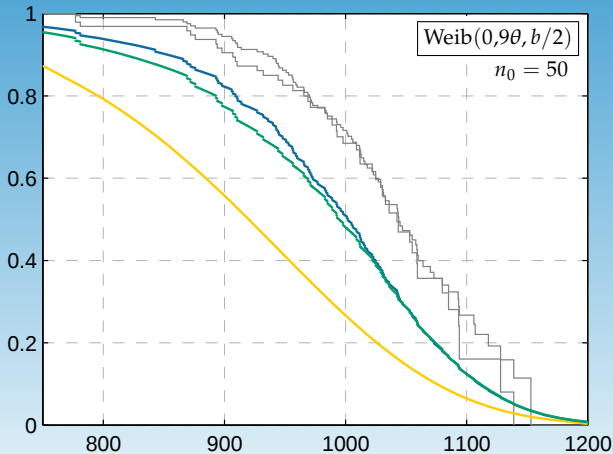
- Weibullovo $\Lambda_0(t) = (t/\theta)^b$, $t > 0$,
- s $\theta = 1071$ a $b = 15,9$, a dále s $b/2$ a $2b$

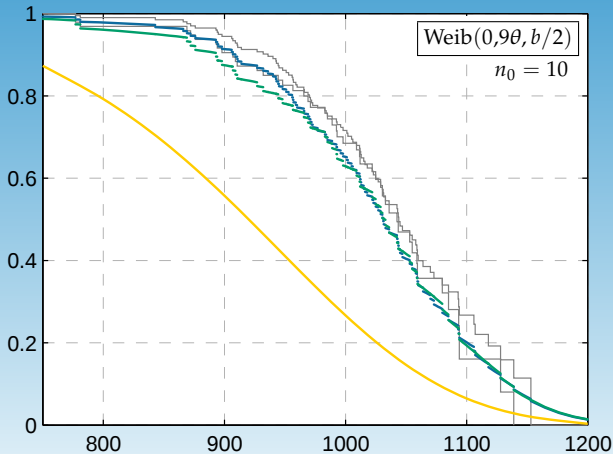












Zpátky k teorii

Porovnání odhadů

- pomocí bayesovského rizika
- $BR \hat{\theta} = E(\hat{\theta} - \theta)^2$

Oč je $\hat{S}$ lepší než $\hat{S}_{FP}$

- za předpokladu Koziolova-Greenova modelu
- pro bayesovský odhad $BR \hat{S}(t) = E \text{var}(S(t) | \text{data})$
- zatímco $BR \hat{S}_{FP}(t) = BR \hat{S}(t) + E(\hat{S}(t) - \hat{S}_{FP}(t))^2$

Teorie → simulace

- aposteriorní rozptyl umíme spočítat ($E(S^2(t) | \text{data})$): 1 → 2)
- E přes data — ze simulovaných dat

Simulace

Chtěli bychom

- generovat parametr γ a proces S dle apriorního rozdělení
- a pak výběr $(X_1, Y_1), \dots, (X_n, Y_n)$ z S (resp. S^γ)

Proces S

$$\hat{S}, \hat{S}_{FP}$$

- realizace jsou funkcemi spolehlivosti diskrétních rozdělení (NTR)
- má ale nekonečně skoků v každém intervalu ($\mathcal{G}$)

Aproximace

- diskretizovat čas a po přírůstcích
- největší skoky $\wedge$ (inverzní Lévyova míra)
- speciální metody pro gama proces, ...

Přesná simulace

Pěkné prediktivní rozdělení

- např. $P(X_{i+1} > x \mid \gamma, *) = E \left[\underbrace{P(X_{i+1} > x \mid \gamma, *, \Lambda)}_{\exp(-\Lambda(x))} \mid \gamma, * \right]$
- jako tvar $\hat{S}$

Nepotřebujeme-li S samotný

- místo náhodného výběru (X, Y) z nepřesného S
- generujeme postupně γ , $(X_1 \mid \gamma)$, $(Y_1 \mid \gamma, X_1)$, $(X_2 \mid \gamma, X_1, Y_1)$, $(Y_2 \mid \gamma, X_1, Y_1, X_2), \dots$

Případně vybraná $(S(t) \mid \gamma, X_1, Y_1, \dots, X_n, Y_n)$

- přírůstky mezi časy pozorování — gama rozdělení
- skoky v T_i — zamítací metodou z $xf(x)/(1 - e^{-x})$

Příklad

Apriorní rozdělení

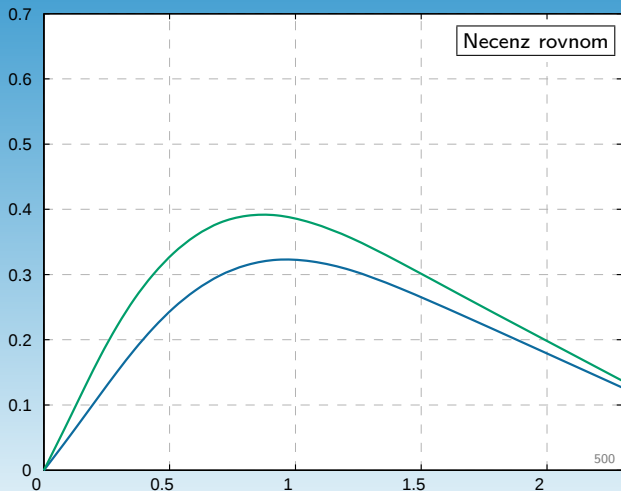
- pro Λ : gama proces s $n_0 = 10$ a $\Lambda_0(t) = t$ (exponenciální rozdělení)
- pro γ : diskrétní rozdělení na $1/p - 1$, $p = 1/10, 2/10, \dots, 9/10$

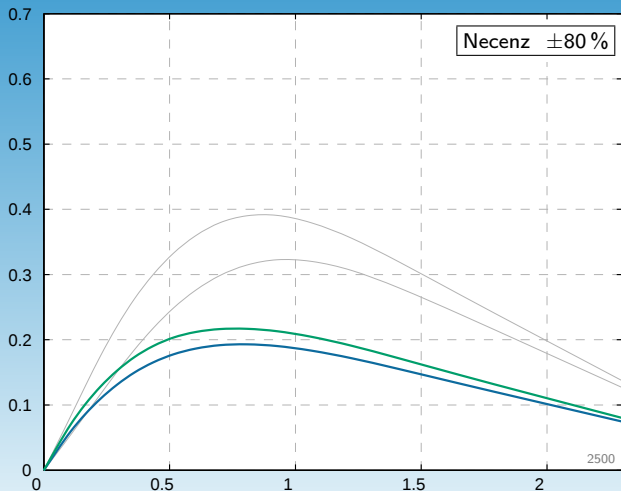
Počítáno pro

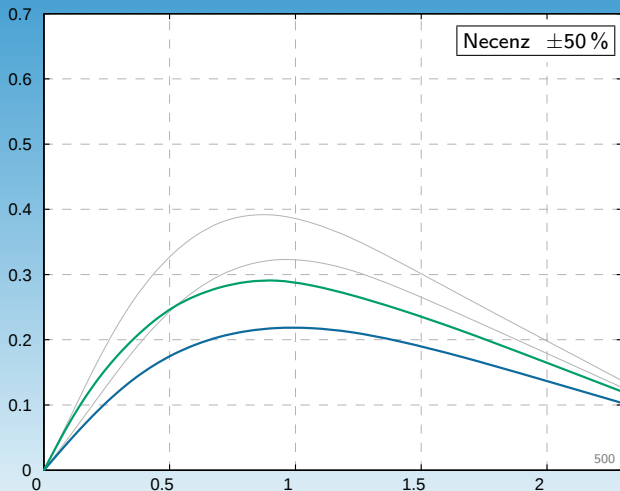
- rozsahy výběrů $n = 10, 20, \dots, 100$,
- časy $t =$ kvantily 10 %–90 % rozdělení $\text{Exp}(1)$

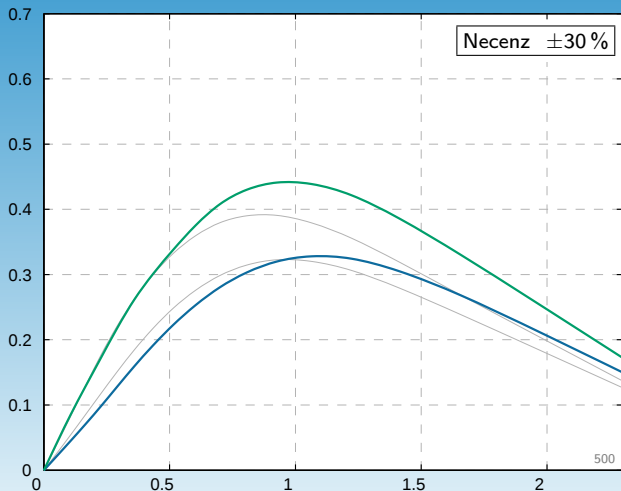
Grafy $n \text{ BR } \hat{S}(t)$ a $n \text{ BR } \hat{S}_{\text{FP}}(t)$

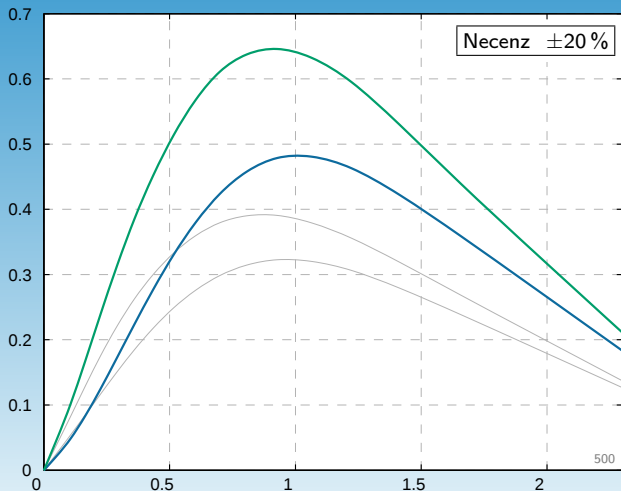
- jako funkce t pro $n_0 = 100$
- jako funkce n pro $t = 20\%$ kvantil

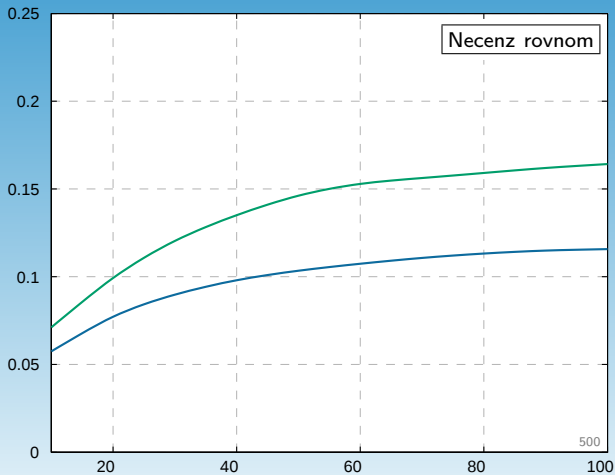


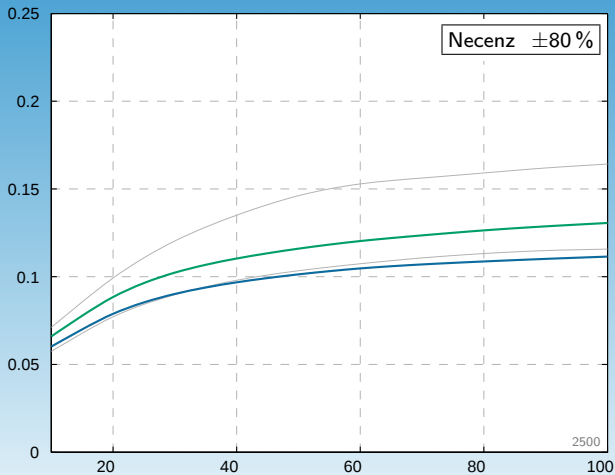


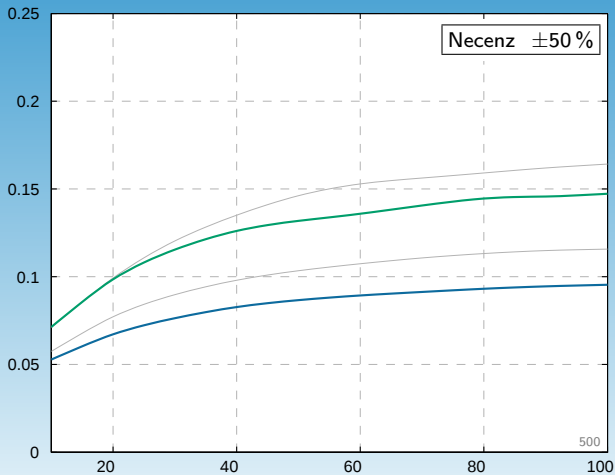


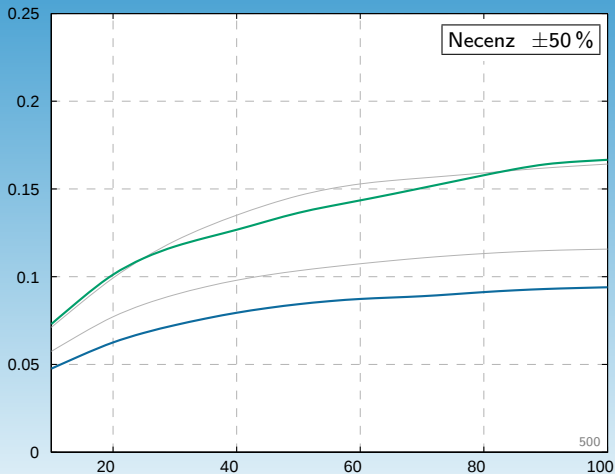


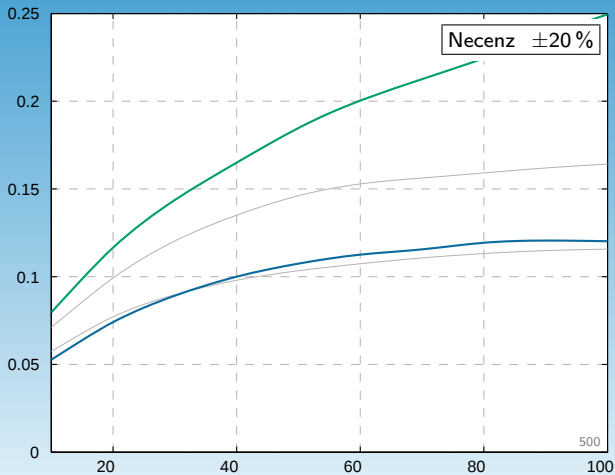












POROVNÁNÍ NEPARAMETRICKÝCH BAYESOVSKÝCH ODHADŮ PŘI CENZOROVÁNÍ

Michal Friesl

*Katedra matematiky
Fakulta aplikovaných věd
Západočeská univerzita v Plzni*

- <http://home.zcu.cz/~friesl/Archiv/Rob06.html>