Některé postupy pro detekce změn ve statistických modelech

Funková, Garaj

1. května 2024

Statistický seminář

Formulace problému

- $X_1, ..., X_n$ jsou nezávislé náhodné veličny, $n \in \mathbb{N}$, které splňují

$$\begin{aligned} X_i &= \mu + \eta_1 \varepsilon_i, i = 1, ..., k_0 \\ &= \mu + \eta_2 \varepsilon_i, i = k_0 + 1, ..., n, \end{aligned}$$

  kde $\eta_1, \eta_2 > 0$ jsou neznámé parametry, $\mu \in \mathbb{R}$ je konstantní pro všechna i a $k_0 \in \{1, ..., n\}$ je neznámá nenáhodná poloha změny.

- $\varepsilon_1, ..., \varepsilon_n$ jsou iid se střední hodnotou 0 a rozptylem 1. Existuje $\delta > 0$ takové, že $E|\varepsilon_i|^{4+2\delta} < \infty$.

Hypotéza a alternativa

$H_0 : X_i = \mu + \eta\varepsilon_i, i = 1, ..., n$
H_1 : existuje $k_0 \in \{1, ..., n\}$ takové, že

$$X_i = \mu + \eta_1\varepsilon_i, i = 1, ..., k_0$$
$$X_i = \mu + \eta_2\varepsilon_i, i = k_0 + 1, ..., n$$

kde $\eta_1 \neq \eta_2$.

Metoda CUSUM

Pro $k \in \{1, ..., n\}$ definujeme kumulovaný součet

$$U_n(k) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{k} (X_i - \mu)^2 - \frac{k}{n} \sum_{j=1}^{n} (X_i - \mu)^2 \right).$$

Za H_0 platí

$$EU_n(k) = 0, \mathrm{var} U_n(k) = \frac{k}{n} \left(1 - \frac{k}{n} \right) \kappa^2, k \in \{1, ..., n\},$$

kde $\kappa^2 = \mathrm{var}(X_i - \mu)^2$ za H_0.

Testová statistika: $U_n = \max_{1 \leq k \leq n} |U_n(k)|$.

Metoda CUSUM: Rozdělení testové statistiky I

Věta

Za platnosti nulové hypotézy platí:

$$\frac{U_n}{\kappa^2} = \sup_{s\in(0,1)} \frac{|U_n(\lfloor ns \rfloor)|}{\kappa^2} \xrightarrow[n\to\infty]{\mathcal{D}} \sup_{u\in(0,1)} |B(u)|,$$

kde $\{B(t), t \in (0,1)\}$ *je standardní Brownův můstek.*

Metoda CUSUM: Rozdělení testové statistiky II

Věta

Nechť platí nulová hypotéza a nechť dále platí $E|\varepsilon_i|^4 \log\log(|\varepsilon_i|+1) < \infty$. *Potom pro každé* $x \in \mathbb{R}$ *platí*

$$P\left[A(\log n)\sup_{s\in(0,1)}\frac{|U_n(\lfloor ns\rfloor)|}{\sqrt{\kappa^2}(s(1-s))^{\frac{1}{2}}} - D(\log n) \le x\right] \to \exp(-2e^{-x})$$

$$kde\ A(\log n) = (2\log\log n)^{\frac{1}{2}}\ a$$

$$D(\log n) = 2\log\log n + \frac{1}{2}\log\log\log n - \frac{1}{2}\log\pi.$$

Metoda CUSUM: Kritické hodnoty

Podle metody I zamítáme H_0, když $U_n > \kappa^2 \omega_{1-\alpha}$.
Podle metody II zamítáme H_0, když

$$\sup_{s\in(0,1)} \frac{|U_n(\lfloor ns \rfloor)|}{\sqrt{\kappa^2}(s(1-s))^{\frac{1}{2}}} > -\frac{1}{A(\log n)} \left(\log\log \frac{1}{\sqrt{1-\alpha}} - D(\log n) \right).$$

Pozici změny k_0 odhadneme jako $\hat{k} = \underset{1\le k\le n}{\operatorname{argmax}} |U_n(k)|$. Pokud existuje takových indexů více, zvolíme ten nejmenší.

Metoda CUSUM simulace kritických hodnot

- Generujeme data délky n z nornálního a Laplaceova rozdělení, přičemž n bereme postupně z množiny $\{50, 100, 200\}$.
- Pro každé n vypočteme kritické hodnoty získané z 1000 simulací jako výběrový $(1-\alpha)$-kvantil testových statistik $\sup_{0<s<1}(\frac{|U_n(\lfloor ns \rfloor)|}{\sqrt{\kappa^2}})$ a $\sup_{0<s<1}(\frac{|U_n(\lfloor ns \rfloor)|}{\sqrt{\kappa^2}(s(1-s))^{\frac{1}{2}}})$.

Neparametrický bootstrap - princip

- Máme náhodný výběr $X_1, ..., X_n$.
- Provedeme B-krát náhodný výběr s opakováním, dostaneme výběr $Z^*_{1,b}, ..., Z^*_{n,b}, b = 1, ..., B$.
- Pro každé $b = 1, ..., B$ spočítáme hodnotu testové statistiky U_n, označíme $U^*_{n,b}$.
- Z vypočtených $U^*_{n,b}$ odhadneme rozdělení statistiky U_n a kritickou hodnotu dostaneme jako odhadnutý $(1 - \alpha)$ kvantil tohoto rozdělení.

Metoda CUSUM: Simulace kritických hodnot

n	metoda/hladina	0.01	0.05	0.1	0.01	0.05	0.1
50	simulace	4.54	3.64	3.15	5.41	3.94	3.29
50	aproximace	4.60	3.62	3.18	4.60	3.62	3.18
100	simulace	4.75	3.80	3.24	5.90	4.14	3.53
100	aproximace	4.57	3.64	3.23	4.57	3.64	3.23
200	simulace	5.32	3.90	3.28	6.30	4.41	3.39
200	aproximace	4.55	3.66	3.26	4.55	3.66	3.26

Tabulka: Kritické hodnoty z_α vlevo pro rozdělení $\mathcal{N}(0,1)$ a vpravo pro rozdělení $\mathcal{L}(0,\sqrt{2})$.

Metoda CUSUM: Simulace kritických hodnot

n	metoda/hladina	0.01	0.05	0.1	0.01	0.05	0.1
50	simulace	1.47	1.23	1.16	1.44	0.76	1.09
50	aproximace	1.63	1.36	1.22	1.63	1.36	1.22
100	simulace	1.49	1.30	1.20	1.48	0.78	1.12
100	aproximace	1.63	1.36	1.22	1.63	1.36	1.22
200	simulace	1.62	1.30	1.16	1.45	0.76	1.11
200	aproximace	1.63	1.36	1.22	1.63	1.36	1.22

Tabulka: Kritické hodnoty w_α vlevo pro rozdělení $\mathcal{N}(0,1)$ a vpravo pro rozdělení $\mathcal{L}(0, \sqrt{\frac{1}{2}})$.

Metoda CUSUM: Neparametrický bootstrap

n	data/hladina	0.01	0.05	0.1	0.01	0.05	0.1
50	beze změny	1.41	1.23	1.12	0.93	0.93	0.92
50	změna	1.39	1.22	1.12	9.45	9.45	9.45
100	beze změny	1.47	1.26	1.14	0.50	0.50	0.50
100	změna	1.44	1.26	1.15	19.88	19.88	19.88
200	beze změny	1.51	1.30	1.17	0.04	0.04	0.04
200	změna	1.50	1.27	1.15	38.85	38.85	38.85

Tabulka: Kritické hodnoty w_α vlevo pro normální rozdělení a vpravo pro Laplaceovo rozdělení. Jednou pro data z $\mathcal{N}(0,1)$ i pro data se změnou z $\mathcal{N}(0,1)$ a $\mathcal{N}(0,4)$. Taky pro data z $\mathcal{L}(0,\sqrt{\frac{1}{2}})$ a se změnou s rozdělením $\mathcal{L}(0,\sqrt{\frac{1}{2}})$ a $\mathcal{L}(0,\sqrt{2})$.

Metoda SIC

Schwarzovo informační kritérium:

$$SIC = -2l_n(\hat{\theta}) + p \log n,$$

kde $\theta \in \mathbb{R}^p$ je parametr, l_n je log-věrohodnost, n je rozsah výběru a $\hat{\theta}$ je odhad parametru θ metodou maximální věrohodnosti.

Testová statistika: $\min_{1 \leq k \leq n} SIC(k) - SIC(n)$,
kde $SIC(n)$ je SIC za platnosti nulové hypotézy a $SIC(k)$ je SIC za platnosti alternativy ($k_0 = k$).

Testová statistika pro normální a Laplaceovo rozdělení

Pro normální rozdělení máme

$$SIC(n) = n \log 2\pi + n \log \hat{\sigma}^2_{1,n} + n + \log n$$
$$SIC(k) = n \log 2\pi + k \log \hat{\sigma}^2_{1,k} + (n-k) \log \hat{\sigma}^2_{k+1,n} + n + 2 \log n.$$

Pro Laplaceovo rozdělení máme

$$SIC(n) = n \log 2 + n \log \hat{\sigma}^2_{1,n} + 2n + \log n$$
$$SIC(k) = n \log 2 + k \log \hat{\sigma}^2_{1,k} + (n-k)k \log \hat{\sigma}^2_{k+1,n} + 2n + 2 \log n,$$

kde $\hat{\sigma}^2_{1,n}, \hat{\sigma}^2_{1,k}, \hat{\sigma}^2_{k+1,n}$ jsou maximálně věrohodné odhady rozptylů za nulové hypotézy a za alternativy.

Metoda SIC: Rozdělení testové statistiky

Věta

Označme

$$\Delta_n = \min_{1 \le k \le n-1} SIC(k) - SIC(n)$$

$$\lambda_n = \left\{ \max_{1 \le k \le n-1} [n \log \hat{\sigma}^2_{1,n} - k \log \hat{\sigma}^2_{1,k} - (n-k) \log \hat{\sigma}^2_{k+1,n}] \right\}^{\frac{1}{2}}$$

Potom za nulové hypotézy pro každé $x \in \mathbb{R}$ *platí*

$$P\left[A(\log n)\lambda_n - D(\log n) \le x\right] \to exp(-2e^{-x}).$$

Metoda SIC: Kritické hodnoty

Zamítáme H_0, když existuje $k \in \{1, ..., n\}$ takové, že

$$SIC(n) > SIC(k) + \left(-\frac{1}{A(\log n)} \log \log (1-\alpha)^{-\frac{1}{2}} + \frac{D(\log n)}{A(\log n)} \right)^2 - \log n.$$

Pozici změny k_0 odhadneme jako $\hat{k}$, které splňuje $SIC(\hat{k}) = \min_{1 \leq k \leq n-1} SIC(k)$. Pokud existuje takových indexů více, zvolíme ten nejmenší.

Metoda SIC simulace kritických hodnot

- Generujeme data délky n z nornálního a Laplaceova rozdělení, přičemž n bereme postupně z množiny $\{50, 100, 200\}$.
- Pro každé n vypočteme kritické hodnoty získané z 1000 simulací jako výběrový $(1 - \alpha)$-kvantil testovej statistiky $SIC(n) - \min_{1 \leq k \leq n-1} SIC(k)$.

Metoda SIC: Neparametrický bootstrap

n	data/hladina	0.01	0.05	0.1	0.01	0.05	0.1
50	beze změny	1.38	1.20	1.10	7.66	4.96	3.54
50	změna	1.36	1.19	1.10	8.99	5.64	4.08
100	beze změny	1.41	1.24	1.12	7.57	4.60	3.24
100	změna	1.40	1.22	1.11	9.01	5.73	4.21
200	beze změny	1.45	1.26	1.14	7.27	4.50	3.04
200	změna	1.46	1.26	1,14	8.99	6.03	4.08

Tabulka: Kritické hodnoty spočtené vlevo pro normální rozdělení a vpravo pro Laplaceovo rozdělení. Jednou pro data z $\mathcal{N}(0,1)$ i pro data se změnou z $\mathcal{N}(0,1)$ a $\mathcal{N}(0,4)$. Taky pro data z $\mathcal{L}(0,\sqrt{\frac{1}{2}})$ a se změnou z $\mathcal{L}(0,\sqrt{\frac{1}{2}})$ a $\mathcal{L}(0,\sqrt{2})$.

Simulační studie - porovnání CUSUM a SIC metody

Generujeme data $X_1, ..., X_n$, kde n volíme postupně z množiny $\{50, 100, 200, 1000\}$ následujícím způsobem:

- $X_i \sim \mathcal{N}(0, 1) \qquad i = 1, ..., \lfloor ns_0 \rfloor$
- $X_i \sim \mathcal{N}(0, \delta) \qquad i = \lfloor ns_0 \rfloor + 1, ...n,$

kde s_0 volíme postupně z množiny $\{0.25, 0.5, 0.75\}$ a
δ z $\{0.5, 2, 3\}$.
Pro každou kombinaci generujeme 1000 posloupností dat.
Pro každou posloupnost je testována H_0 na hladině
$\alpha \in \{0.01, 0.5, 0.1\}$.

Metody hodnotíme podle dvou kritérií:

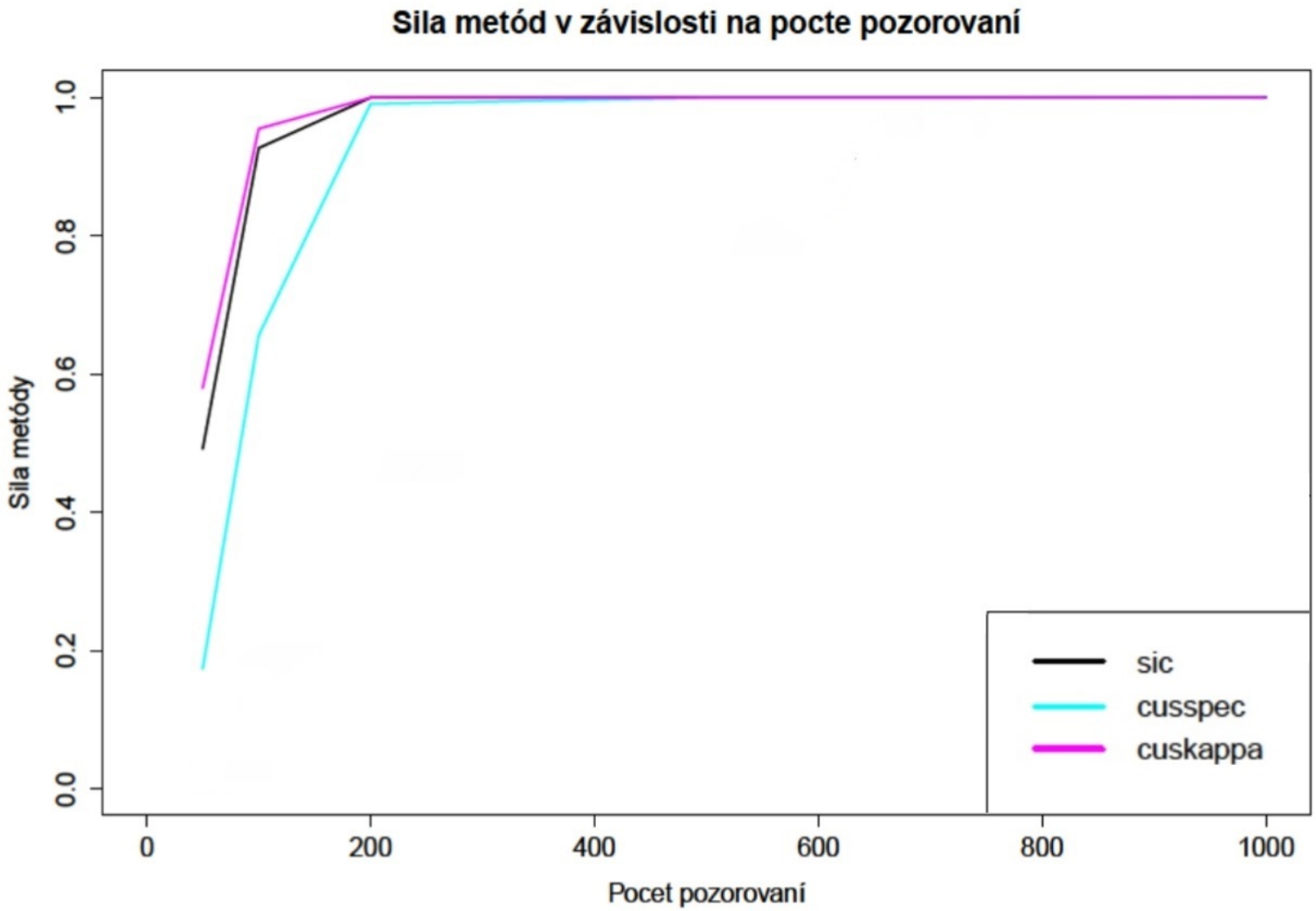
- empirické síly testu - podíl počtu zamítnutých hypotéz k celkovému počtu testů,
- průměrné odchylky - průměrná vzdálenost odhadnutého bodu změny $\hat{k}$ od skutečného bodu změny k_0.

δ		CUSUM	SIC
0.5	emp. síla	0.954	0.916
0.5	prům. odchylka	4.690	3.990
2	emp. síla	1.000	0.927
2	prům. odchylka	4.880	4.170
3	emp. síla	0.996	1.000
3	prům. odchylka	3.240	1.580

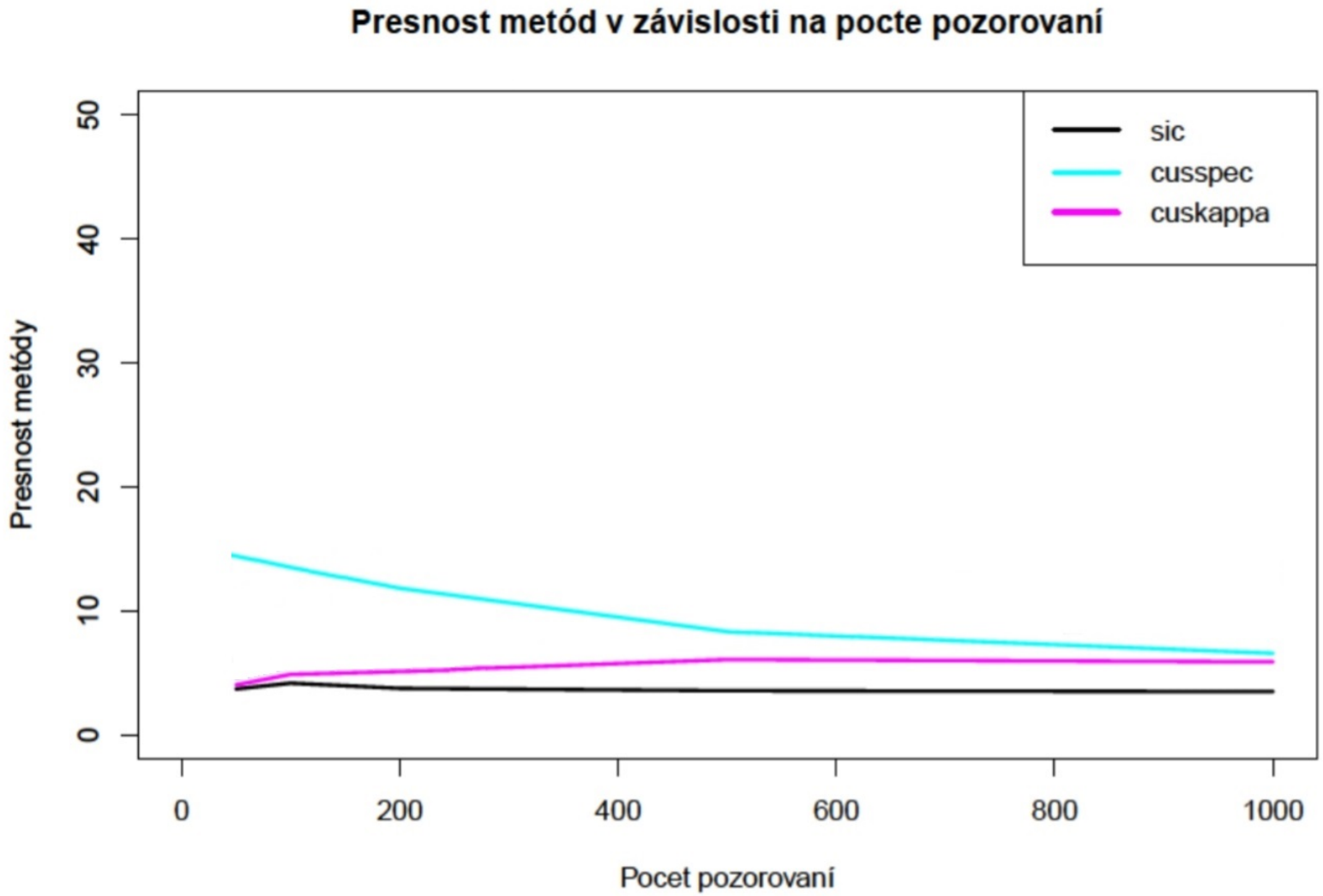
Tabulka: Vliv velikosti změny v rozptylu na vlastnosti metody SIC a CUSUM s kvantily spočtenými dle Věty 1 pro $n = 100$, $s_0 = 0.05$ a $\alpha = 0.05$.

s_0		CUSUM	SIC
0.25	emp. síla	0.962	0.996
0.25	prům. odchylka	16.480	3.780
0.50	emp. síla	1.000	1.000
0.50	prům. odchylka	5.120	3.750
0.75	emp. síla	0.995	1.24
0.75	prům. odchylka	3.100	3.750

Tabulka: Vliv polohy změny v rozptylu na vlastnosti metody SIC a CUSUM s kvantily spočtenými dle Věty 1 pro $n = 200$, $\delta = 2$ a $\alpha = 0.05$.



Obrázek: Empirické síly jednotlivých metod pro $s_0 = 0.5$, $\delta = 2$ a $\alpha = 0.05$ (metoda SIC - černá, metoda CUSUM Věta1 - fialová, metoda CUSUM Věta2 - modrá).



Obrázek: Průměrné odchylky jednotlivých metod pro $s_0 = 0.5$, $\delta = 2$ a $\alpha = 0.05$ (metoda SIC - černá, metoda CUSUM Věta1 - fialová, metoda CUSUM Věta2 - modrá).

metoda	emp. síla	prům. odchylka
Aprox.	0.955	4.880
Bootstrap	0.980	4.760

Tabulka: Porovnání empirické síly a průměrné odchylky s kvantily spočtenými dle Věty 1 pro $n = 100$, $\delta = 2$, $s_0 = 0.5$ a $\alpha = 0.05$ na základě aproximativní a bootsrapové kritické hodnoty.

Informace jsme čerpali z diplomové práce s názvem Některé postupy pro detekce změn ve statistických modelech od autorky Lindy Marešové
https://dspace.cuni.cz/handle/20.500.11956/85944