



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Bc. Matěj Lebeda

**Postupy pro statistickou kontrolu
náhodných procesů**

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. RNDr. Jaromír Antoch, CSc.

Studijní program: Pravděpodobnost, matematická
statistika a ekonometrie

Studijní obor: Pravděpodobnost, matematická
statistika a ekonometrie

Praha 2024

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V dne

Podpis autora

Na tomto místě bych rád poděkoval panu profesoru Antochovi za ochotný přístup, pádné připomínky a cenné rady. Nesmím zapomenout poděkovat svým nejbližším za podporu nejen při studiu. Velké díky patří především mojí mamince za víru ve mě.

Název práce: Postupy pro statistickou kontrolu náhodných procesů

Autor: Bc. Matěj Lebeda

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. RNDr. Jaromír Antoch, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Postupy pro statistickou kontrolu náhodných procesů jsou z literatury dobře známy. Co ale v literatuře nenacházíme, je porovnání navržených metod. Nejprve zavedeme lineární regresní model, z něhož v celé práci vycházíme. Následně vysvětlíme tři základní typy porušení modelu, přičemž dvěma z nich se v práci budeme věnovat podrobně. V praxi se uplatňují dva základní přístupy k detekci změny při vývoji náhodného procesu: offline a online. Metoda offline spočívá v detekci změny ex-post. Navrhujeme postupy využívající normalitu dat i robustní postupy. Metody online, též sekvenční, fungují na jiném principu. Mezi základní patří Shewhartova metoda a CUSUM metoda. Těm se věnujeme ve čtvrté kapitole. Konečně v poslední páté kapitole představíme již avizované porovnání těchto metod. Hlavní zájmy detekce změny jsou, aby vůbec procedura změnu detekovala, a pokud ano, tak aby ji detekovala co nejrychleji a zároveň ne předčasně. Metody porovnáme právě z těchto hledisek.

Klíčová slova: statistická kontrola kvality, detekce změny, střední hodnota procesu, metody CUSUM

Title: Procedures for statistical control of random processes

Author: Bc. Matěj Lebeda

Department: Department of probability and mathematical statistics

Supervisor: prof. RNDr. Jaromír Antoch, CSc., Department of Probability and Mathematical Statistics

Abstract: Procedures for statistical control of random processes are well known. What we miss, is the comparison of such procedures. In the beginning, we will introduce the linear regression model which will be our assumption throughout the whole thesis. Then we will explain three most common violations of the model whereas two of them will be studied closely. In practice, two fundamental approaches are employed: offline and online approach. The offline methods are performed ex-post. We will propose procedures leaning on the assumption of normality, but robust procedures as well. Online methods (so called sequential) are based on a different principle. The most common are Shewhart's and CUSUM method. Finally, the last fifth chapter will be dedicated to comparison of these methods. Our main interests are to detect as fast as possible but also not before the time of change. The approaches will be compared from these aspects.

Keywords: statistical quality control, change point detection, mean of a random process, CUSUM methods

Obsah

Úvod	2
1 Model	4
2 Možná porušení modelu	5
2.1 Změna parametru posunutí	5
2.2 Postupná změna	6
2.3 Změna modelových křivek	6
3 Offline detekce změny	8
3.1 Detekce změny parametru posunutí	8
3.1.1 Za předpokladu normality dat	8
3.1.2 Bez předpokladu normality dat	17
3.2 Detekce postupné změny	20
4 Základní zastavovací pravidla pro online detekci změny	25
4.1 Shewhartova metoda	26
4.2 Metoda CUSUM	28
4.2.1 Darmois-Koopmanova třída rozdělení	29
5 Simulace	32
5.1 Postup simulace	32
5.2 Výsledky simulací	33
5.2.1 Výsledky offline detekce	33
5.2.2 Výsledky online detekce	42
5.2.3 Porovnání offline metod a metody CUSUM	50
Závěr	52
Seznam použité literatury	54
Seznam obrázků	55

Úvod

Postupům pro statistickou kontrolu náhodných procesů již bylo v literatuře věnováno spoustu pozornosti. Co ale postrádáme, je porovnání těchto metod. Dojde-li při vývoji náhodného procesu ke změně, ať jde třeba o proces výrobní, je v našem přirozeném zájmu tuto změnu zaznamenat co nejrychleji. Fungují navržené metody spolehlivě? V praxi bychom rádi aplikovali tu proceduru, která změnu detekuje co nejdříve, ale zároveň ne dříve, než změna skutečně nastala. Která metoda je tedy neoptimálnější?

Nejprve zadefinujeme lineární regresní model, z něhož budeme v celé práci vycházet. Poté představíme tři základní typy porušení modelu, přičemž dvěma z nich se budeme věnovat podrobně. Následně vyložíme základní přístupy offline detekce, které se liší například dle toho, zda předpokládají normalitu dat, či nikoli, ale i pro normální data se zvolený postup odvíjí dle toho, známe-li počáteční střední hodnotu náhodného procesu, nebo ne. V kapitole věnované online detekci se soustředíme na dva nejznámější postupy: Shewhartovu a CUSUM metodu. Shewhartovu metodu aplikujeme v normálním modelu, metodu CUSUM nejprve v širší, tzv. Darmais-Koopmanově rodině rozdělání, a poté speciálně v normálním modelu. Nakonec budeme prezentovat výsledky simulací, pomocí kterých porovnáme metody dle již zmíněných kritérií.

Použité symboly

- $Z \sim (\mu, \sigma^2)$... pro náhodnou veličinu Z platí $EZ = \mu$, $\text{var}Z = \sigma^2$
- u_α ... α -kvantil standardního normálního rozdělení
- $t_n(\alpha)$... α -kvantil t -rozdělení o n stupních volnosti
- Φ ... distribuční funkce standardního normálního rozdělení

1. Model

Pokud nebude řečeno jinak, budeme všechny vektory uvažovat jako sloupcové. Upozorníme hned zpočátku, že data $(Y_i, \mathbf{x}_i^\top)^\top$ jsou získávána v časech t_i , $i = 1, \dots, n$, takových, že $t_1 < t_2 < \dots < t_n$.

V rámci celého textu budeme předpokládat, že data lze modelovat lineárním regresním modelem s absolutním členem

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde:

- $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ je vektor odezev;
- $\mathbf{1} = (1, \dots, 1)^\top$ je jednotkový vektor;
- $\mathbf{X} = (x_{ij})_{i=1, \dots, n}^{j=1, \dots, p}$ je modelová matice známých konstant s plnou sloupcovou hodnotí p ;
- $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ je i -tý řádek matice \mathbf{X} ;
- $\mu \in \mathbb{R}$ značí absolutní člen;
- $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ značí vektor regresních parametrů (koeficientů);
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ je vektor neznámých chyb.

LM předpokládá, že:

- $(Y_i, \mathbf{x}_i^\top)^\top$ jsou nezávislé,
- $\boldsymbol{\varepsilon}|\mathbf{X} \sim (0, \sigma^2)$, $\sigma^2 > 0$, a tedy $\mathbf{Y}|\mathbf{X} \sim (\mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta}, \sigma^2)$

2. Možná porušení modelu

Při detekování změny se často setkáváme se třemi základními typy porušení modelu. Problém můžeme formulovat v plné obecnosti pro k -rozměrný případ, předmětem našeho zájmu ale vždy bude nejjednodušší situace.

2.1 Změna parametru posunutí

Při statistické kontrole výroby je běžnou praxí, že se v náhodném čase může (ale nemusí) výrazně změnit hmotnost výrobku a hmotnosti následujících výrobků se ustálí kolem této nové hodnoty. V kontextu matematické statistiky je rozhodování o tom, zda vývoj náhodného procesu zůstal po celou dobu stejný, nebo zda při vývoji došlo ke změně specifického druhu, obvykle založeno na testování hypotéz. Popíšme tedy první obvyklý případ změny parametru posunutí jako test následující hypotézy proti alternativě:

$$\begin{aligned} \mathbf{H} : & Y_i = \mu + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n, \\ \mathbf{A} : \exists m \in \{1, \dots, n-1\} : & Y_i = \mu + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, m, \\ & Y_i = \mu + \Delta + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = m+1, \dots, n, \end{aligned}$$

přičemž $\varepsilon_i \sim (0, \sigma^2)$, $\Delta \in \mathbb{R}$.

Obecně říkáme, že za platnosti nulové hypotézy je náhodný proces $\{Y_1, \dots, Y_n\}$ stacionární, zatímco za platnosti alternativy je proces nestacionární a stacionarita je specifickým způsobem porušena.

Parametry našeho zájmu, typicky neznámými, jsou v tomto modelu μ , Δ a m . Rušivým parametrem pak je σ^2 . V praktických úlohách, v nichž vystupují data, nás samozřejmě zajímají i odhady parametrů $\boldsymbol{\beta}$. Z hlediska detekce změny to ale není nutně parametr klíčový.

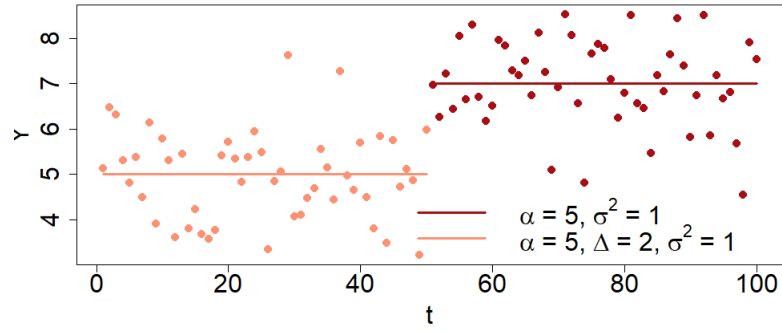
Dále se budeme zabývat jen nejjednodušším případem tak zvaného skoku ve střední hodnotě:

$$\begin{aligned} \mathbf{H} : & Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n, \\ \mathbf{A} : \exists m \in \{1, \dots, n-1\} : & Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, m, \\ & Y_i = \mu + \Delta + \varepsilon_i, \quad i = m+1, \dots, n, \end{aligned} \tag{2.1}$$

tj. alternativa říká, že v nějakém náhodném čase dojde k posunu střední hodnoty o Δ .

Situaci za alternativy můžeme vizualizovat tak, jak ukazuje obr. 2.1.

Avšak i v tomto nejjednodušším případě je třeba rozlišovat dvě možnosti, které mají různá řešení; jednou je známá hodnota μ a druhou je hodnota μ neznámá. Od toho se odvíjejí i odlišné testové statistiky s jinými asymptotickými vlastnostmi. V prvním případě je limitou posloupnosti příslušných testových statistik Wienerův proces, zatímco v případě druhém Brownův můstek. Podrobnosti vizte v Csörgő (1997).



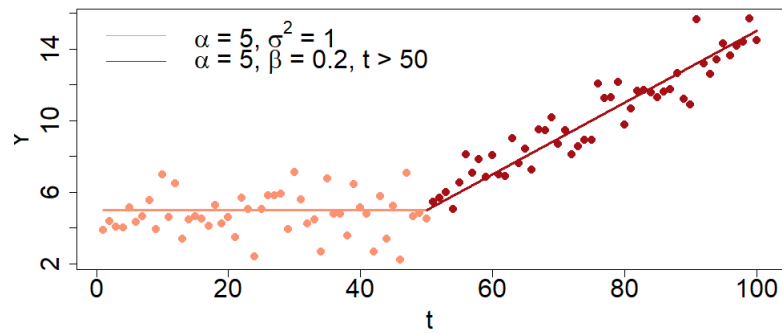
Obrázek 2.1: Model změny skokem; chyby jsou normálně rozdělené

2.2 Postupná změna

Dalším typickým příkladem změny modelu je přechod z konstanty v přímku (či v k -rozměrnou rovinu). Pozornost budeme věnovat jen nejjednoduššímu případu

$$\begin{aligned} \mathbf{H} : & Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, n, \\ \mathbf{A} : \exists m \in \{1, \dots, n-1\} : & Y_i = \mu + \varepsilon_i, \quad i = 1, \dots, m, \\ & Y_i = \mu + \beta(i - m) + \varepsilon_i, \quad i = m + 1, \dots, n. \end{aligned} \quad (2.2)$$

Opět, pro představu situace můžeme problém vizualizovat za pomoci obrázku 2.2.



Obrázek 2.2: Model postupné změny se spojitým přechodem, tzv. hokejka; chyby jsou normálně rozdělené

2.3 Změna modelových křivek

Poslední případ, který v praxi často nastává, je změna modelových křivek. Formulujme nejprve matematicky:

$$\mathbf{H} : Y_i = \mu + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.3)$$

$$\begin{aligned} \mathbf{A} : \exists m \in \{1, \dots, n-1\} : & Y_i = \mu + \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, m, \\ & Y_i = \delta + \mathbf{x}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \quad i = m + 1, \dots, n. \end{aligned} \quad (2.4)$$

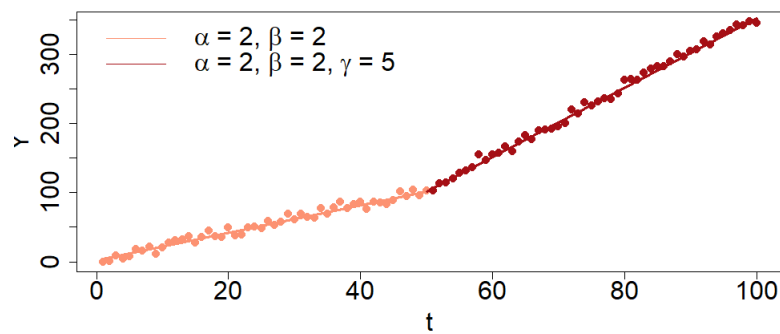
Někdy pro zachování spojitosti požadujeme, aby $\mu + \mathbf{x}_m^\top \boldsymbol{\beta} = \delta + \mathbf{x}_m^\top \boldsymbol{\gamma}$, tj. $\delta = \mu + \mathbf{x}_m^\top (\boldsymbol{\beta} - \boldsymbol{\gamma})$, a pak (2.4) bude tvaru $Y_i = \mu + \mathbf{x}_m^\top (\boldsymbol{\beta} - \boldsymbol{\gamma}) + \mathbf{x}_i^\top \boldsymbol{\gamma} + \varepsilon_i$, $i = m + 1, \dots, n$.

V tomto případě se zaměříme jen na nejjednodušší případ:

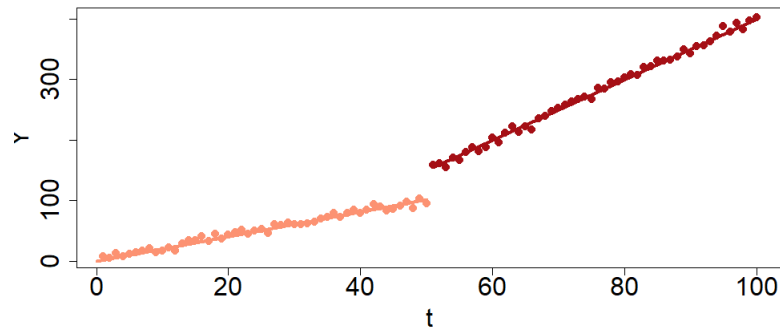
$$\mathbf{H} : Y_i = \mu + \beta i + \varepsilon_i, \quad i = 1, \dots, n, \quad (2.5)$$

$$\mathbf{A} : \exists m \in \{1, \dots, n - 1\} : Y_i = \mu + \beta i + \varepsilon_i, \quad i = 1, \dots, m, \\ Y_i = \mu + \beta m + \gamma(i - m) + \varepsilon_i, \quad i = m + 1, \dots, n.$$

Další typická alternativa je znázorněna na obrázku 2.4, kde přímka nejen změní směr, ale dojde i ke skoku.



Obrázek 2.3: Model postupné změny přímky v přímku; chyby jsou normálně rozdělené



Obrázek 2.4: Model postupné změny přímky v přímku se skokem; chyby jsou normálně rozdělené

3. Offline detekce změny

Předpokládejme, že pozorujeme odezvy $Y_i \sim (\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, $i = 1, \dots, n$, a necht parametry $\boldsymbol{\beta}$ a σ^2 jednoznačně určují rozdělení Y_i . V praxi může dojít v náhodném čase ke změně modelu. Zpravidla jde o změnu střední hodnoty, tj. změnu $\boldsymbol{\beta}$, anebo o změnu rozptylu σ^2 , někdy se může změnit obojí.

Při snaze detekovat změnu ve vývoji náhodného procesu postupujeme zpravidla v následujících krocích:

1. Rozhodnout, zda vůbec došlo ke změně.
2. Pokud ano, tak odhadnout, v jakém čase změna nastala.
3. Rozhodnout, zda změn nebylo více.
4. Pokud ano, tak odhadnout časy všech změn.

Nejprve se soustředíme na první otázku. Budeme testovat změnu parametrů modelu. Offline detekce spočívá v tom, že již máme sesbíraná data a inferenci provádíme ex post. Metody detekce se liší dle toho, jaký klademe předpoklad na rozdělení dat. Popíšeme si metody umožňující testovat tři nejčastější typy porušení modelu popsané v kapitole 2, a to jejich nejjednodušší verze.

3.1 Detekce změny parametru posunutí

3.1.1 Za předpokladu normality dat

Nejprve budeme vycházet z předpokladu $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, $i = 1, \dots, n$, kde σ^2 je rušivý parametr.

Naším cílem bude odvodit test hypotézy popsané modelem (2.1). Obvykle nastávají dvě situace. Počáteční střední hodnota μ je známá a neznámým parametrem je skok Δ a čas změny m . V druhém případě pak neznáme ani počáteční střední hodnotu. Navrhne tedy dva různé testy.

(i)

$$\begin{aligned} \mathbf{H}^1 : & Y_i \sim \mathcal{N}(\mu_0, \sigma^2), \quad i = 1, \dots, n, \\ \mathbf{A}^1 : \exists m \in \{1, \dots, n-1\} : & Y_i \sim \mathcal{N}(\mu_0, \sigma^2), \quad i = 1, \dots, m, \\ & Y_i \sim \mathcal{N}(\mu_0 + \Delta, \sigma^2), \quad i = m+1, \dots, n, \end{aligned} \quad (3.1)$$

přičemž μ_0 je známý parametr a Δ , m jsou neznámé parametry.

(ii)

$$\begin{aligned} \mathbf{H}^2 : & Y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \\ \mathbf{A}^2 : \exists m \in \{1, \dots, n-1\} : & Y_i \sim \mathcal{N}(\mu_1, \sigma^2), \quad i = 1, \dots, m, \\ & Y_i \sim \mathcal{N}(\mu_2, \sigma^2), \quad i = m+1, \dots, n, \end{aligned} \quad (3.2)$$

μ, μ_1, μ_2 a m jsou neznámé parametry.

Vysvětleme nejprve stručně základní myšlenku konstrukce testových statistik pro úlohy s neznámým časem změny. Nejprve čas eventuální změny m považujeme za známý. Jinými slovy vycházíme z toho, že změna proběhla v konkrétním čase m , $1 \leq m < n$. Pak se ale úloha redukuje v klasický dvouvýběrový problém rovnosti středních hodnot, kterému byla v literatuře věnována velká pozornost. Matematicky vyjádřeno, jde o následující testy.

ad (i)

$$\begin{aligned} \mathbf{H}_m^1 : Y_i &\sim \mathcal{N}(\mu_0, \sigma^2), \quad i = 1, \dots, n, \\ \mathbf{A}_m^1 : Y_i &\sim \mathcal{N}(\mu_0, \sigma^2), \quad i = 1, \dots, m, \\ &Y_i \sim \mathcal{N}(\mu_0 + \Delta, \sigma^2), \quad i = m + 1, \dots, n. \end{aligned} \quad (3.3)$$

ad (ii)

$$\begin{aligned} \mathbf{H}_m^2 : Y_i &\sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n, \\ \mathbf{A}_m^2 : Y_i &\sim \mathcal{N}(\mu_1, \sigma^2), \quad i = 1, \dots, m, \\ &Y_i \sim \mathcal{N}(\mu_2, \sigma^2), \quad i = m + 1, \dots, n. \end{aligned} \quad (3.4)$$

Na σ^2 nemusíme klást žádný předpoklad, jak se ukáže v následujícím textu.

Přistupme nejprve ke konstrukci testů (3.3) a (3.4); následně vysvětlíme přechod k (3.1) a (3.2). Postupy se v obou případech mírně liší:

Detekce skoku při známé počáteční střední hodnotě

Pomocí principu maximální věrohodnosti odvodíme testovou statistiku pro případ (3.3). Uvažujme věrohodnostní poměr

$$\begin{aligned} \Lambda_{n,m}(\Delta) &= \frac{\prod_{i=1}^n f_A(Y_i)}{\prod_{i=1}^n f_H(Y_i)} \\ &= \frac{\prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(Y_i - \mu_0)^2}{2\sigma^2}\right\} \prod_{i=m+1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[(Y_i - \mu_0) - \Delta]^2}{2\sigma^2}\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(Y_i - \mu_0)^2}{2\sigma^2}\right\}} \\ &= \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=m+1}^n [(Y_i - \mu_0) - \Delta]^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=m+1}^n (Y_i - \mu_0)^2\right\}} \\ &= \exp\left\{\frac{\Delta}{\sigma^2} \sum_{i=m+1}^n (Y_i - \mu_0) - \frac{n-m}{2\sigma^2} \Delta^2\right\}. \end{aligned} \quad (3.5)$$

Logaritmus věrohodnostního poměru $\Lambda_{n,m}$ má tvar:

$$\log \Lambda_{n,m}(\Delta) = \frac{\Delta}{\sigma^2} \sum_{i=m+1}^n (Y_i - \mu_0) - \frac{n-m}{2\sigma^2} \Delta^2. \quad (3.6)$$

Jako testovou statistiku vezmeme $\Lambda := \sup_{\Delta} 2 \log \Lambda_{n,m}(\Delta)$. Protože výraz (3.6) je parabola v Δ , tak argument maxima nalezneme snadno a dostáváme výslednou testovou statistiku ve tvaru

$$\Lambda = \frac{1}{\sigma^2} (n - m) (\bar{Y}_m^o)^2, \quad (3.7)$$

kde $\bar{Y}_m^o = \frac{1}{n-m} \sum_{i=m+1}^n (Y_i - \mu_0)$ je průměr posledních $n - m$ centrovaných pozorování.

Výraz (3.7) lze upravit na tvar

$$\Lambda = \frac{1}{\sigma^2 (n - m)} \left[\sum_{i=m+1}^n (Y_i - \mu_0) \right]^2 = \left[\frac{1}{\sqrt{n - m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right]^2. \quad (3.8)$$

Je tedy zřejmé, že proti hypotéze budou svědčit velké hodnoty testové statistiky Λ . Poznamenejme, že umíme nalézt přesné rozdělení (3.8), a tedy pokud bychom znali σ^2 , mohli bychom ji použít jako testovou statistiku. V případě, kdy známe σ^2 , je ale jednodušší použít ekvivalentní testovou statistiku

$$T_m^Z = \frac{1}{\sqrt{n - m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \stackrel{H_m^1}{\sim} \mathcal{N}(0, 1).$$

Tedy při známé hodnotě σ^2 zamítáme $\mathbf{H}_m^1 \iff |T_m^Z| \geq u_{1-\frac{\alpha}{2}}$. Pokud nás zajímají jednostranné alternativy $\Delta > 0$, resp. $\Delta < 0$, postupujeme analogicky a testovou statistiku srovnáváme s $u_{1-\alpha}$, resp. u_α .

V případě, že σ^2 neznáme, je třeba jej odhadnout. Nabízí se použít jeden ze tří konzistentních odhadů:

$$S^2 = \frac{1}{n - 1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2, \quad (3.9)$$

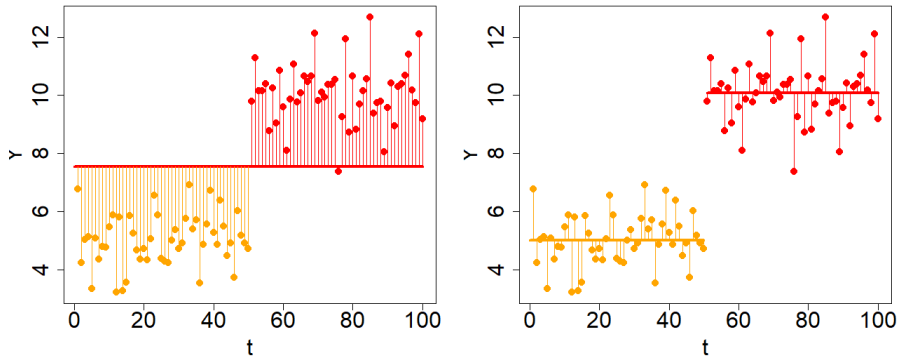
$$\tilde{S}_m^2 = \frac{1}{n - 2} \left[\sum_{i=1}^m (Y_i - \bar{Y}_m)^2 + \sum_{i=m+1}^n (Y_i - \bar{Y}_m^0)^2 \right], \quad (3.10)$$

$$S_{*m}^2 = \frac{1}{m - 1} \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 + \frac{1}{n - m - 1} \sum_{i=m+1}^n (Y_i - \bar{Y}_m^0)^2, \quad (3.11)$$

kde $\bar{Y}_m = \frac{1}{m} \sum_{i=1}^m Y_i$, $\bar{Y}_m^0 = \frac{1}{n-m} \sum_{i=m+1}^n Y_i$.

S_{*m}^2 není nic jiného než součet výběrových rozptylů dvou skupin Y_1, \dots, Y_m a Y_{m+1}, \dots, Y_n . Pro $m = 1$ klademe rozptyl první skupiny roven nule a podobně pro $m = n - 1$ a rozptyl druhé skupiny. V takových případech můžeme tedy teoretický rozptyl odhadovat výběrovým rozptyl druhé, resp. první skupiny. Poznamenejme nicméně, že pro odhad jakéhokoli parametru potřebujeme alespoň několik pozorování, aby byl odhad rozumný.

Za platnosti hypotézy jsou hodnoty S^2 , \tilde{S}_m^2 i S_{*m}^2 zpravidla velmi blízké. Naproti tomu, za platnosti alternativy tomu tak není, jak je zřejmé z obrázku 3.1.



Obrázek 3.1: Porovnání dvou uvažovaných odhadů rozptylu: $S^2 = 7.37$, $S_{*m}^2 = 1.61$, přičemž $\sigma^2 = 1$.

S^2 sčítá rezidua jako vzdálenosti jednotlivých pozorování od celkového průměru. Všimněme si, že za platnosti alternativy mohou být tato rezidua příliš velká, a tedy S^2 bude zmenšovat testovou statistiku více než S_{*m}^2 či \tilde{S}_m^2 . Taková volba by vedla k testu s menší silou. Proto se doporučuje používat odhad S_{*m}^2 (nebo \tilde{S}_m^2) a testovou statistiku tvaru

$$T_m^N = \frac{1}{\sqrt{S_{*m}^2(n-m)}} \sum_{i=m+1}^n (Y_i - \mu_0),$$

popř.

$$\tilde{T}_m^N = \frac{1}{\sqrt{\tilde{S}_m^2(n-m)}} \sum_{i=m+1}^n (Y_i - \mu_0).$$

Platí, že T_m^N i \tilde{T}_m^N mají za nulové hypotézy asymptoticky standardní normální rozdělení, neboť S_{*m}^2 i \tilde{S}_m^2 jsou konzistentními odhady rozptylu σ^2 . Nicméně ve zbytku textu se budeme bavit pouze o statistice T_m^N , neboť pro \tilde{T}_m^N platí vše stejně.

Při neznámé hodnotě σ^2 zamítáme $\mathbf{H}_m^1 \iff |T_m^N| \geq u_{1-\frac{\alpha}{2}}$.

Připomeňme, že doposud jsme si představili testy pro případ známého času změny. Pokud ale tento čas neznáme, tak si uvědomme, že alternativa \mathbf{A}^1 je vlastně sjednocením všech jednotlivých alternativ \mathbf{A}_m^1 , $m = 1, \dots, n-1$. Neboli, nulovou hypotézu (3.1) zamítáme, pokud proti ní svědčí alespoň jedna z alternativ \mathbf{A}_m^1 , $m = 1, \dots, n-1$, tj. pokud alespoň jedna testová statistika T_m^Z , resp. T_m^N je příliš „malá“ nebo příliš „velká“. Toto je ekvivalentní tomu, že pro testování použijeme testovou statistiku

$$T^Z = \max_{1 \leq m < n} |T_m^Z|, \text{ resp. } T^N = \max_{1 \leq m < n} |T_m^N|. \quad (3.12)$$

Poznamenejme, že $\max_m |T_m^N|$ se v literatuře většinou nazývá *vážená statistika maximálního typu*. Vedle statistik maximálního typu existují statistiky součtového typu, a jiné. Těmi se zde zabývat nebudeme.

Nalézt rozdělení statistik (3.12) je velmi obtížné. Jedna z možností, jak stanovit kritické hodnoty, vychází z Bonferonniho metody. Totiž,

$$\begin{aligned}
& \mathbb{P} \left(\max_{1 \leq m \leq n-1} \left| \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq C \right) \\
&= \mathbb{P} \left(\bigcup_{m=1}^{n-1} \left[\left| \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq C \right] \right) \\
&\leq \sum_{m=1}^{n-1} \mathbb{P} \left(\left| \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq C \right) \\
&= (n-1) \mathbb{P} \left(\left| \frac{1}{\sqrt{n-1}} \sum_{i=2}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq C \right). \tag{3.13}
\end{aligned}$$

Evidentně, $\frac{1}{\sqrt{n-1}} \sum_{i=2}^n \left(\frac{Y_i - \mu_0}{\sigma} \right)$ má (za nulové hypotézy) rozdělení $\mathcal{N}(0,1)$. Tedy pro

$C = u_{1-\frac{\alpha}{2(n-1)}}$ je

$$\mathbb{P} \left(\left| \frac{1}{\sqrt{n-1}} \sum_{i=2}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq C \right) = \frac{\alpha}{n-1},$$

a tedy v takovém případě bude horní odhad (3.13) roven α . Takto získané aproximované kritické hodnoty slouží dobře pro malé rozsahy výběru; pro velké n jsou příliš konzervativní. Přesnější aproximaci kritických hodnot nabízí následující věta (Csörgö (1997)):

Věta 1 (Aproximace rozdělení statistiky maximálního typu při známé počáteční střední hodnotě náhodného procesu). *Pro všechna $x \in \mathbb{R}$ a dostatečně velké n platí za nulové hypotézy (3.1)*

$$\mathbb{P} \left(\max_{1 \leq m < n} \left| \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \geq \frac{x + b_n}{a_n} \right) \approx 1 - \exp\{-e^{-x}\}, \tag{3.14}$$

kde $a_n = \sqrt{2 \log \log n}$ a $b_n = 2 \log \log n + \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi$.

Poznámka: Tvrzení věty 1 se také opírá i o určité podmínky regularity, podrobnosti viz Csörgö (1997), paragraf 3.1.1.

Aproximaci (3.14) samozřejmě můžeme použít i v případě statistiky T^N , neboť S_{*m}^2 je konzistentní odhad σ^2 , a tedy dle Cramér-Sluckého věty má T^N stejné asymptotické chování jako T^Z .

Naším požadavkem je, aby příslušný test dodržoval hladinu α , tj. hledáme takovou kritickou mez C_α , že $\mathbb{P}(T^Z > C_\alpha) = \alpha$ za platnosti nulové hypotézy. Díky větě 1 jsme toho schopni docílit alespoň přibližně. Totiž, dá se najít $x_0 \in \mathbb{R}$ takové, které bude řešením exponenciální rovnice

$$1 - e^{-e^{-x_0}} = \alpha.$$

Jednoduchými výpočty získáváme:

$$x_0 = -\log \log \left(\frac{1}{1-\alpha} \right).$$

Kritickou hodnotou pak tedy je $C_\alpha = \frac{x_0 + b_n}{a_n}$.

Poznamenejme, že pro jednostrannou alternativu $\Delta > 0$, resp. $\Delta < 0$, je třeba pro výpočet kritických hodnot použít následující limitní aproximaci, tj.

$$P \left(\max_{1 \leq m < n} \left\{ \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right\} > \frac{x + b_n}{a_n} \right) \approx 1 - \exp\left\{-\frac{1}{2}e^{-x}\right\}, \quad x \in \mathbb{R}. \quad (3.15)$$

V případě jednostranné alternativy $\Delta > 0$ zamítáme nulovou hypotézu, je-li

$$\max_{1 \leq m \leq n-1} \left\{ \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right\} > C_\alpha^+.$$

Kritickou mez spočteme dle (3.15) stejnými výpočty jako pro oboustrannou alternativu. Tedy

$$C_\alpha^+ = \frac{x_0^+ + b_n}{a_n}, \quad \text{kde } x_0^+ = -\log \log \left(\frac{1}{(1-\alpha)^2} \right).$$

V případě opačné jednostranné alternativy $\Delta < 0$ zamítáme nulovou hypotézu, je-li

$$\min_{1 \leq m \leq n-1} \left\{ \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right\} < -C_\alpha^+.$$

Abychom tedy shrnuli výsledky části (i); pro test (3.1) navrhuje dvě testové statistiky, a to podle toho, zda známe, či neznáme σ^2 . Známe-li σ^2 , volíme

$$T^Z = \max_{1 \leq m \leq n-1} \left\{ \left| \frac{1}{\sqrt{n-m}} \sum_{i=m+1}^n \left(\frac{Y_i - \mu_0}{\sigma} \right) \right| \right\},$$

v opačném případě

$$T^N = \max_{1 \leq m \leq n-1} \left\{ \left| \frac{1}{\sqrt{S_{*m}^2(n-m)}} \sum_{i=m+1}^n (Y_i - \mu_0) \right| \right\}.$$

Kritické hodnoty pak aproximujeme dle věty 1.

Detekce skoku při neznámé počáteční hodnotě

Nyní navrhneme test pro úlohu (3.4). Jde přesně o klasický dvouvýběrový test shody středních hodnot.

Přirozeným odhadem střední hodnoty prvního výběru je průměr $\bar{Y}_m = \frac{1}{m} \sum_{i=1}^m Y_i$ a podobně pro druhý výběr a jeho střední hodnotu $\bar{Y}_m^0 = \frac{1}{n-m} \sum_{i=m+1}^n Y_i$. Odpovídající testová statistika má tvar

$$T_m^* = \sqrt{\frac{m(n-m)}{n}} \frac{\bar{Y}_m - \bar{Y}_m^0}{\sigma}, \quad (3.16)$$

která má za platnosti hypotézy $\mathcal{N}(0,1)$ (Anděl (2005), věta 4.26 a její důkaz). Pokud σ neznáme, odhadneme: $\hat{\sigma} = S_{*m} := \sqrt{S_{*m}^2}$.

Všimněme si, že statistiku (3.16) můžeme pomocí elementárních úprav převést na tvar

$$T_m^* = \sqrt{\frac{n}{m(n-m)}} \frac{1}{\sigma} \sum_{i=1}^m (Y_i - \bar{Y}_n). \quad (3.17)$$

Vidíme, že se jedná o vážený kumulativní součet reziduí od celkového průměru, odkud pochází její název *CUSUM statistika*, z anglického „cumulative sum“.

Pro otestování (3.2) bychom tím pádem zvolili testovou statistiku

$$T^* = \max_{1 \leq m \leq n-1} \left\{ \frac{1}{\sigma} \left| \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n) \right| \right\}.$$

Abychom našli aproximaci kritických hodnot, budeme postupovat stejně jako v případě známé počáteční střední hodnoty. Je-li rozsah náhodného výběru malý, můžeme použít Bonferonniho nerovnost, a pro velké n aproximuje pomocí následující věty z Csörgö (1997):

Věta 2 (Aproximace rozdělení statistiky maximálního typu při neznámé počáteční hodnotě náhodného procesu). *Pro všechna $x \in \mathbb{R}$ a dostatečně velké n platí za nulové hypotézy (3.2)*

$$P \left(\max_{1 \leq m < n} \left\{ \frac{1}{\sigma} \left| \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n) \right| \right\} > \frac{x + b_n}{a_n} \right) \approx 1 - \exp\{-2e^{-x}\},$$

kde a_n, b_n jsou tytéž jako ve větě 1.

Poznámka: Opět, stejně jako v případě věty 1, i tvrzení této věty se opírá o jisté podmínky regularity, blíže v Csörgö (1997), paragraf 3.1.1.

Abychom aproximovali kritickou hodnotu v tomto případě, budeme postupovat jako v předchozím odstavci, jen s tou výjimkou, že $x_0^* \in \mathbb{R}$ je nyní řešením rovnice

$$1 - e^{-2e^{-x_0^*}} = \alpha.$$

Stejnými úpravami jako v předchozím odstavci dostaneme řešení

$$x_0^* = -\log \log \left(\frac{1}{\sqrt{1-\alpha}} \right),$$

a tedy jako kritickou hodnotu volíme $C_\alpha = \frac{x_0^* + b_n}{a_n}$.

Nyní přistupme k podrobnému odvození testové statistiky (3.16) metodou maximální věrohodnosti (za předpokladu, že Y_1, \dots, Y_n jsou iid s normálním roz-

dělením a σ^2 je známé). Opět uvažujme věrohodnostní poměr

$$\begin{aligned}
\Lambda_{n,m}^* &= \frac{\prod_{i=1}^n f_A(Y_i)}{\prod_{i=1}^n f_H(Y_i)} \\
&= \frac{\prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu_1)^2}{2\sigma^2}\right\} \prod_{i=m+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu_2)^2}{2\sigma^2}\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^m (Y_i-\mu_1)^2\right\} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=m+1}^n (Y_i-\mu_2)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i-\mu)^2\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^m (Y_i^2 - 2\mu_1 Y_i + \mu_1^2)\right\} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=m+1}^n (Y_i^2 - 2\mu_2 Y_i + \mu_2^2)\right\}}{\exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i^2 - 2\mu Y_i + \mu^2)\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n Y_i^2 - 2\mu_1 \sum_{i=1}^m Y_i + m\mu_1^2 - 2\mu_2 \sum_{i=m+1}^n Y_i + (n-m)\mu_2^2\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n Y_i^2 - 2\mu \sum_{i=1}^n Y_i + n\mu^2\right)\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2} \left(2\mu_1 \sum_{i=1}^m Y_i - m\mu_1^2 + 2\mu_2 \sum_{i=m+1}^n Y_i - (n-m)\mu_2^2 - 2\mu \sum_{i=1}^n Y_i + n\mu^2\right)\right\} \\
&= \exp\left\{\frac{1}{2\sigma^2} \left(2 \sum_{i=1}^m Y_i(\mu_1 - \mu) + 2 \sum_{i=m+1}^n Y_i(\mu_2 - \mu) - m\mu_1^2 - (n-m)\mu_2^2 + n\mu^2\right)\right\}.
\end{aligned} \tag{3.18}$$

Nyní aplikujme dvojnásobek logaritmu:

$$2 \log \Lambda_{n,m}^* = \frac{1}{\sigma^2} \left(2 \sum_{i=1}^m Y_i(\mu_1 - \mu) + 2 \sum_{i=m+1}^n Y_i(\mu_2 - \mu) - m\mu_1^2 - (n-m)\mu_2^2 + n\mu^2 \right)$$

Neznámé parametry μ, μ_1, μ_2 je třeba nahradit jejich maximálně věrohodnými odhady:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i =: \bar{Y}_n, \quad \hat{\mu}_1 = \frac{1}{m} \sum_{i=1}^m Y_i =: \bar{Y}_m, \quad \hat{\mu}_2 = \frac{1}{n-m} \sum_{i=m+1}^n Y_i =: \bar{Y}_m^o.$$

Dosadíme a pokračujeme v úpravách:

$$\begin{aligned}
\Lambda^* &= 2 \log \Lambda_{n,m}^*(\hat{\mu}, \hat{\mu}_1, \hat{\mu}_2) \\
&= \frac{1}{\sigma^2} \left(2 \sum_{i=1}^m Y_i(\bar{Y}_m - \bar{Y}_n) + 2 \sum_{i=m+1}^n Y_i(\bar{Y}_m^o - \bar{Y}_n) - m\bar{Y}_m^2 - (n-m)(\bar{Y}_m^o)^2 + n\bar{Y}_n^2 \right) \\
&= \frac{1}{\sigma^2} \left(2m\bar{Y}_m(\bar{Y}_m - \bar{Y}_n) + 2(n-m)\bar{Y}_m^o(\bar{Y}_m^o - \bar{Y}_n) - m\bar{Y}_m^2 - (n-m)(\bar{Y}_m^o)^2 + n\bar{Y}_n^2 \right) \\
&= \frac{1}{\sigma^2} \left[m\bar{Y}_m^2 + (n-m)(\bar{Y}_m^o)^2 + \bar{Y}_n(-m\bar{Y}_m - (n-m)\bar{Y}_m^o) \right]
\end{aligned} \tag{3.19}$$

Abychom dospěli do finálního tvaru (3.16), je třeba vyjádřit \bar{Y}_n pomocí \bar{Y}_m a \bar{Y}_m^0 :

$$\bar{Y}_n = \frac{1}{n} \left(\sum_{i=1}^m Y_i + \sum_{i=m+1}^n Y_i \right) = \frac{m}{n} \bar{Y}_m + \frac{n-m}{n} \bar{Y}_m^0.$$

Dosadíme toto vyjádření do (3.19) a dostáváme

$$\begin{aligned} \Lambda^* &= \frac{1}{\sigma^2} \left[m \bar{Y}_m^2 + (n-m) (\bar{Y}_m^0)^2 - 2 \frac{m(n-m)}{n} \bar{Y}_m \bar{Y}_m^0 - \frac{m^2}{n} \bar{Y}_m^2 - \frac{(n-m)^2}{n} (\bar{Y}_m^0)^2 \right] \\ &= \frac{1}{\sigma^2} \left[\frac{mn - m^2}{n} \bar{Y}_m^2 + \frac{(n-m)n - (n-m)^2}{n} (\bar{Y}_m^0)^2 - 2 \frac{m(n-m)}{n} \bar{Y}_m \bar{Y}_m^0 \right] \\ &= \frac{1}{\sigma^2} \left[\frac{m(n-m)}{m} \bar{Y}_m^2 + \frac{m(n-m)}{n} (\bar{Y}_m^0)^2 - 2 \frac{m(n-m)}{n} \bar{Y}_m \bar{Y}_m^0 \right] \\ &= \frac{1}{\sigma^2} \left[\sqrt{\frac{m(n-m)}{m}} \bar{Y}_m - \sqrt{\frac{m(n-m)}{n}} \bar{Y}_m^0 \right]^2 \end{aligned}$$

Nakonec Λ^* odmocníme a obdržíme

$$\begin{aligned} \sqrt{\Lambda^*} &= \frac{1}{\sigma} \left[\sqrt{\frac{m(n-m)}{m}} \bar{Y}_m - \sqrt{\frac{m(n-m)}{n}} \bar{Y}_m^0 \right] \\ &= \frac{1}{\sigma} \sqrt{\frac{m(n-m)}{n}} (\bar{Y}_m - \bar{Y}_m^0), \end{aligned}$$

což je přesně statistika (3.16).

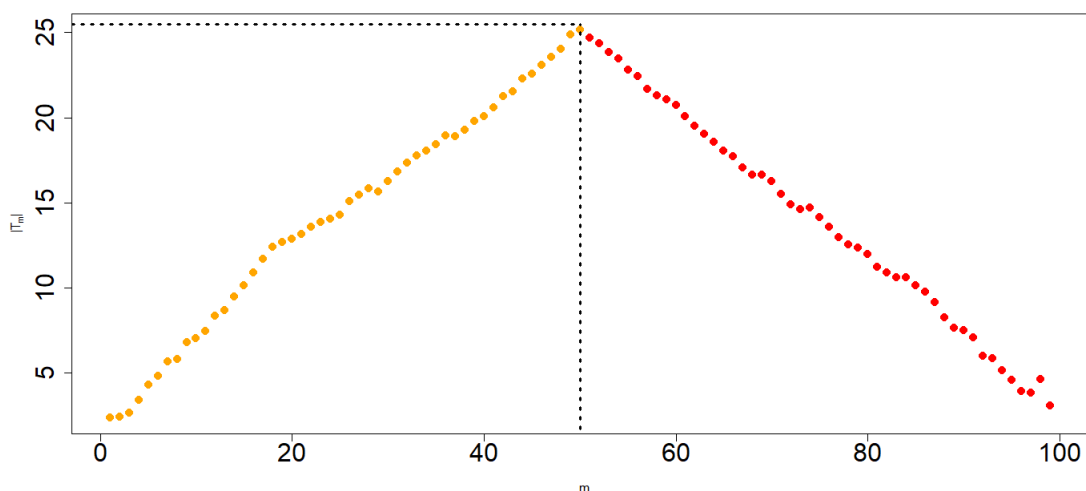
Dospěli jsme tedy do bodu, kdy máme k dispozici testy obou hypotéz (3.1) i (3.2). Shrnutí nabízíme v tabulce 3.1.

počáteční střední hodnota	rozptyl	testová statistika
známá μ_0	známý	$\max_m \left \frac{1}{\sigma} \sqrt{\frac{1}{n-m}} \sum_{i=m+1}^n (Y_i - \mu_0) \right $
známá μ_0	neznámý	$\max_m \left \frac{1}{S_{*m}} \sqrt{\frac{1}{n-m}} \sum_{i=m+1}^n (Y_i - \mu_0) \right $
neznámé μ, μ_1, μ_2	známý	$\max_m \left \frac{1}{\sigma} \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n) \right $
neznámé μ, μ_1, μ_2	neznámý	$\max_m \left \frac{1}{S_{*m}} \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n) \right $

Tabulka 3.1: Testové statistiky pro detekci skoku ve střední hodnotě za předpokladu normality

Vrátíme-li se ale na začátek této kapitoly, vzpomeneme si, že dále je třeba odhadnout čas, v němž eventuálně došlo ke změně. Není překvapením, že za odhad bodu změny se bere ten bod, v němž posloupnost statistik $|T_m^Z|$, resp. $|T_m^N|$ – pro jednoduchost označme souhrnně T_m^Z i T_m^N jako T_m – nabývá svého maxima. To jest,

$$\hat{m} = \operatorname{argmax}\{|T_m|; 1 \leq m \leq n-1\}.$$



Obrázek 3.2: Vývoj posloupnosti absolutních hodnot statistik $T_m = \frac{1}{\sigma} \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n)$ pro čas změny $m = 50$. Generovány byly dva výběry z normálních rozdělení: $Y_1, \dots, Y_{50} \sim \mathcal{N}(5, 1)$, $Y_{51}, \dots, Y_{100} \sim \mathcal{N}(10, 1)$

Všimněme si na obrázku 3.2, že absolutní hodnota testové statistiky T_m vychází skutečně největší v čase změny. Proč tomu tak je? Vysvětleme ideu:

Označme m_0 skutečný čas změny, tj. $Y_1, \dots, Y_{m_0} \sim \mathcal{N}(\mu_1, \sigma^2)$ a $Y_{m_0+1}, \dots, Y_n \sim \mathcal{N}(\mu_2, \sigma^2)$. Nechť bez újmy na obecnosti $\mu_1 < \mu_2$, a tedy očekáváme, že $\mu_1 - \bar{Y}_n < 0$ a $\mu_2 - \bar{Y}_n > 0$. Potom pro $m \leq m_0$

$$\sum_{i=1}^m (Y_i - \bar{Y}_n) = \sum_{i=1}^m Y_i - m\bar{Y}_n \approx m(\mu_1 - \bar{Y}_n) \text{ je klesající v } m$$

a pro $m > m_0$

$$\begin{aligned} \sum_{i=1}^m (Y_i - \bar{Y}_n) &= \sum_{i=1}^{m_0} (Y_i - \bar{Y}_n) + \sum_{i=m_0+1}^m (Y_i - \bar{Y}_n) \\ &\approx m_0(\mu_1 - \bar{Y}_n) + (m - m_0)(\mu_2 - \bar{Y}_n) \end{aligned}$$

nejprve klesá a od času $m_0 + 1$ začíná růst.

Maximum nemusí být jednoznačné; pak zpravidla bereme ten čas, kdy bylo dosaženo maxima poprvé. Poznamenejme, že tento případ spíše indikuje více změn.

3.1.2 Bez předpokladu normality dat

Nyní je nutné pozastavit se nad předpokladem normality. Jak jsme uvedli na počátku kapitoly 2, data, jež máme k dispozici, jsou zatížena mnoha zdroji šumu. Tento šum je prakticky nemožné odfiltrovat. To mj. vyžaduje opatrnost při používání testových statistik odvozených za předpokladu normality chyb, jež jsme odvodili v předchozím odstavci.

Ke slovu se tak dostávají tak zvané robustní postupy. V našem případě se jedná např. o modifikace, které místo výběrového průměru jako odhadu parametru polohy používají jiné odhady, méně citlivé na odlehlá pozorování, apod. Nejvíce pozornosti bylo v literatuře věnováno M-odhadům.

V této práci se nebudeme věnovat M-odhadům, nýbrž se zaměříme na postupy založené na pořadích. V tomto případě jsou testové statistiky typu CUSUM založeny na jednoduchých lineárních pořadových statistikách tvaru

$$V_m = \sum_{i=1}^m (a_n(R_i) - \bar{a}_n), \quad m = 1, \dots, n,$$

kde

1. (R_1, \dots, R_n) je vektor pořadí odpovídajících pozorováním Y_1, \dots, Y_n ,
2. $a_n(1), \dots, a_n(n)$ jsou vhodné skóry,
3. $\bar{a}_n = \frac{1}{n} \sum_{i=1}^n a_n(i)$.

Nejčastěji používanými skóry jsou

- Wilcoxonovy skóry: $a_n(i) = \frac{i}{n+1}$, $i = 1, \dots, n$,
- van der Waerdenovy skóry: $a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$, $i = 1, \dots, n$.

Srovnáme-li V_m a T_m^* , můžeme si všimnout, že roli reziduí hrají rozdíly $a_n(R_i) - \bar{a}_n$, $i = 1, \dots, n$. Dále, je-li σ^2 neznámé, zpravidla jej odhadujeme pomocí

$$\hat{\sigma}_{n,R}^2 = \frac{1}{n-1} \sum_{i=1}^n (a_n(R_i) - \bar{a}_n)^2. \quad (3.20)$$

Podobně jako v předchozích odstavcích se častěji pracuje s váženými statistikami

$$V_m^* = w_m V_m, \quad m = 1, \dots, n-1, \quad (3.21)$$

kde váhy w_m se nejčastěji volí jako $\sqrt{\frac{n}{m(n-m)}}$.

Další postup je analogický tomu popsanému v paragrafu vycházejícího z předpokladu normality. Nejprve předpokládáme, že čas změny m je známý. Pro takto vzniklý dvouvýběrový problém napočítáme V_m či V_m^* , o nichž víme, že za jistých podmínek regularity jsou asymptoticky normální. Nejdůležitějším předpokladem je, že distribuční funkce pozorování je spojitá.

Pokud bod změny neznáme (jako v našem případě), nulovou hypotézu zamítneme tehdy, když alespoň jedna testová statistika je „velká“, což je ekvivalentní tomu, že statistika

$$\begin{aligned} & \max_{1 \leq m < n} \left\{ \frac{1}{\hat{\sigma}_{n,R}} \sqrt{\frac{n}{m(n-m)}} |V_m| \right\} \\ &= \max_{1 \leq m < n} \left\{ \frac{1}{\hat{\sigma}_{n,R}} \sqrt{\frac{n}{m(n-m)}} \left| \sum_{i=1}^m (a_n(R_i) - \bar{a}_n) \right| \right\} \end{aligned} \quad (3.22)$$

překročí příslušnou kritickou mez.

Nalézt asymptotické rozdělení statistiky (3.22) je opět velmi obtížné. Podrobnosti lze najít např. v Antoch a kol. (2001), kde se uvádí, že za vhodných podmínek regularity pro velké n a $x \in \mathbb{R}$ platí

$$\mathbb{P} \left[\max_{1 \leq m < n} \left\{ \frac{1}{\hat{\sigma}_{n,R}} \sqrt{\frac{n}{m(n-m)}} \left| \sum_{i=1}^m (a_n(R_i) - \bar{a}_n) \right| \right\} > \frac{x + b_n}{a_n} \right] \approx 1 - \exp\{-2e^{-x}\}, \quad (3.23)$$

kde $a_n = \sqrt{2 \log \log n}$ a $b_n = 2 \log \log n + \frac{1}{2} \log \log \log n - \frac{1}{2} \log \pi$.

Metoda založená na dvouvýběrovém Wilcoxonově testu

Odprostíme-li se od předpokladu na rozdělení dat a uvažujeme-li opět nejprve úlohu se známým časem změny m , pak se naše úloha redukuje na dvouvýběrový problém v neparametrickém modelu. Nejznámějším přístupem je v takovém případě Wilcoxonův test, který navrhuje použít testovou statistiku

$$W_m^* = \frac{W_m - \mathbb{E}_0 W_m}{\sqrt{\text{var}_0 W_m}}, \quad (3.24)$$

kde $W_m = \sum_{i=1}^m R_i$ je součet pořadí R_1, \dots, R_m náhodných veličin Y_1, \dots, Y_m v rámci sdruženého náhodného výběru $Y_1, \dots, Y_m, Y_{m+1}, \dots, Y_n$; $\mathbb{E}_0 W_m$, $\text{var}_0 W_m$ značí pořadě střední hodnotu a rozptyl statistiky W_m za platnosti nulové hypotézy, že rozdělení obou náhodných výběrů Y_1, \dots, Y_m a Y_{m+1}, \dots, Y_n jsou stejná.

Poznámka: Z teorie pořadových testů je známo, že pro lineární pořadovou statistiku $Z_n = \sum_{i=1}^n c_i a_n(R_i)$ platí

$$\mathbb{E} Z_n = n \bar{a}_n \bar{c}, \quad \text{var} Z_n = \frac{1}{n-1} \sum_{i=1}^n (a_n(i) - \bar{a}_n)^2 \sum_{i=1}^n (c_i - \bar{c})^2.$$

Podrobnosti viz Anděl (2005), tvrzení a důkaz věty 11.6.

Dle poznámky platí

$$\mathbb{E}_0 W_m = \frac{1}{2} m(n+1), \quad \text{var}_0 W_m = \frac{1}{12} m(n-m)(n+1).$$

Výpočet rozptylu je mírně komplikovaný, nicméně abychom spočetli $\mathbb{E}_0 W_m$ za platnosti hypotézy, postačí nám jednoduchá heuristika. Je jasné, že

$$\sum_{i=1}^n R_i = \sum_{i=1}^n i = \frac{1}{2} n(n+1),$$

a tedy, má-li výběr Y_1, \dots, Y_m stejné rozdělení jako Y_{m+1}, \dots, Y_n , pak by očekávaný součet pořadí mohl být

$$\mathbb{E}_0 W_m = \frac{m}{n} \frac{1}{2} n(n+1) = \frac{1}{2} m(n+1).$$

Anděl (2005) také uvádí, že testová statistika (3.24) má asymptoticky rozdělení $\mathcal{N}(0,1)$.

V literatuře byly představeny různé modifikace základního Wilcoxonova testu (3.24), těmi se zde ale zabývat nebudeme.

Stejně jako v předchozích odstavcích nás ale u úloh s neznámým časem změny nezajímá toliko rozdělení statistik W_m^* , $m = 1, \dots, n-1$, nýbrž rozdělení maxima jejich absolutních hodnot

$$W^* = \max_{1 \leq m \leq n-1} |W_m^*| = \max_{1 \leq m \leq n-1} \frac{1}{\sqrt{\frac{1}{12}m(n-m)(n+1)}} \left| \sum_{i=1}^m R_i - \frac{1}{2}m(n+1) \right|. \quad (3.25)$$

Metodu založenou na této statistice budeme krátce označovat jako Wilcoxonova metoda. Na předchozí stránce jsme představili aproximaci rozdělení maxima statistik tvaru (3.22). Dá se ale snadno ukázat, že statistika (3.25) odpovídá statistice (3.22), volíme-li již představené Wilcoxonovy skóry. Tím pádem můžeme používat aproximaci kritických hodnot dle (3.23).

3.2 Detekce postupné změny

Analogicky k části 3.1, předpokládejme nejprve lineární model. Již v části 2.2 jsme úlohu zformulovali jako problém klasického statistického testování hypotéz (2.2). Zároveň jde ale o regresní úlohu. Totiž, model pro pevné m za platnosti alternativy lze zapsat maticově:

$$\mathbf{Y} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & n-m \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \\ \varepsilon_{m+1} \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad (3.26)$$

takže úloha (2.2) není nic jiného než test nulovosti směrnice v lineárním modelu (3.26):

$$\begin{aligned} \mathbf{H} : \beta &= 0 \\ \mathbf{A} : \beta &\neq 0. \end{aligned}$$

Pro test této hypotézy volíme t-statistiku

$$t = \frac{\hat{\beta}}{\sqrt{MS_{e22}}}, \quad (3.27)$$

viz Appendix, věta 4.

Nyní se ale vraťme k původní formulaci (2.2). Opět nejprve navrhne test pro m pevné: testovou statistiku zkonstruujeme nám již známým způsobem - metodou poměru věrhodností. Nejprve odvodme test pro μ známé.

Detekce postupné změny při známé počáteční střední hodnotě

Uvažujme věrohodnostní poměr

$$\begin{aligned}
 \Lambda_{m,n}(\beta) &= \frac{\prod_{i=1}^n f_A(Y_i; \beta)}{\prod_{i=1}^n f_H(Y_i)} \\
 &= \frac{\prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu)^2}{2\sigma^2}\right\} \prod_{i=m+1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[(Y_i-\mu)-\beta(i-m)]^2}{2\sigma^2}\right\}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i-\mu)^2}{2\sigma^2}\right\}} \\
 &= \frac{\prod_{i=m+1}^n \exp\left\{-\frac{[(Y_i-\mu)-\beta(i-m)]^2}{2\sigma^2}\right\}}{\prod_{i=m+1}^n \exp\left\{-\frac{(Y_i-\mu)^2}{2\sigma^2}\right\}} \\
 &= \frac{\prod_{i=m+1}^n \exp\left\{-\frac{1}{2\sigma^2}[(Y_i-\mu)^2 - 2\beta(Y_i-\mu)(i-m) + \beta^2(i-m)^2]\right\}}{\prod_{i=m+1}^n \exp\left\{-\frac{1}{2\sigma^2}(Y_i-\mu)^2\right\}} \\
 &= \exp\left\{\frac{\beta}{\sigma^2} \sum_{i=m+1}^n (Y_i-\mu)(i-m) - \frac{\beta^2}{2\sigma^2} \sum_{i=m+1}^n (i-m)^2\right\}. \tag{3.28}
 \end{aligned}$$

Logaritmus věrohodnostního poměru (3.28) je roven:

$$\log \Lambda_{m,n}(\beta) = \frac{\beta}{\sigma^2} \sum_{i=m+1}^n (Y_i-\mu)(i-m) - \frac{\beta^2}{2\sigma^2} \sum_{i=m+1}^n (i-m)^2. \tag{3.29}$$

Opět se jedná o parabolu v β , tedy argument maxima nalezneme snadno a je roven

$$\hat{\beta} = \frac{\sum_{i=m+1}^n (Y_i-\mu)(i-m)}{\sum_{i=m+1}^n (i-m)^2} = 6 \frac{\sum_{i=m+1}^n (Y_i-\mu)(i-m)}{(n-m)(n-m+1)(2n-2m+1)},$$

neboť $\sum_{i=1}^k i^2 = \frac{k(k+1)(2k+1)}{6}$.

Výslednou testovou statistikou je tedy

$$\begin{aligned}
 \Lambda &= \sup_{\beta} 2 \log \Lambda_{m,n}(\beta) \\
 &= 2 \cdot \frac{6}{\sigma^2} \frac{\sum_{i=m+1}^n (Y_i-\mu)(i-m)}{(n-m)(n-m+1)(2n-2m+1)} \sum_{i=m+1}^n (Y_i-\mu)(i-m) \\
 &\quad - \frac{36}{\sigma^2} \frac{[\sum_{i=m+1}^n (Y_i-\mu)(i-m)]^2}{[(n-m)(n-m+1)(2n-2m+1)]^2} \cdot \frac{1}{6} (n-m)(n-m+1)(2n-2m+1) \\
 &= \frac{12}{\sigma^2} \frac{[\sum_{i=m+1}^n (Y_i-\mu)(i-m)]^2}{(n-m)(n-m+1)(2n-2m+1)} - \frac{6}{\sigma^2} \frac{[\sum_{i=m+1}^n (Y_i-\mu)(i-m)]^2}{(n-m)(n-m+1)(2n-2m+1)} \\
 &= \frac{6}{\sigma^2} \frac{[\sum_{i=m+1}^n (Y_i-\mu)(i-m)]^2}{(n-m)(n-m+1)(2n-2m+1)}.
 \end{aligned}$$

Jako ekvivalentní testovou statistiku vezmeme odmocninu z Λ :

$$\mathcal{K}_m = \frac{1}{\sigma} \frac{\sum_{i=m+1}^n (Y_i-\mu)(i-m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6}}}. \tag{3.30}$$

Odvodit rozdělení \mathcal{K}_m za platnosti nulové hypotézy je snadné. Rozepišme sumu z čitatele:

$$\sum_{i=m+1}^n (Y_i - \mu)(i - m) = (Y_{m+1} - \mu) \cdot 1 + (Y_{m+2} - \mu) \cdot 2 + \dots + (Y_n - \mu) \cdot (n - m).$$

Jde o lineární kombinaci nezávislých, stejně rozdělených centrovaných náhodných veličin $Y_{m+1} - \mu, \dots, Y_n - \mu \sim \mathcal{N}(0, \sigma^2)$, jejich součet je tedy opět normálně rozdělená náhodná veličina s nulovou střední hodnotou. Její rozptyl je roven:

$$\begin{aligned} \text{var} \sum_{i=m+1}^n (Y_i - \mu)(i - m) &= \sigma^2 \cdot 1 + \sigma^2 \cdot 2^2 + \dots + \sigma^2 \cdot (n - m)^2 \\ &= \sigma^2 \sum_{k=1}^{n-m} k^2 = \sigma^2 \frac{(n - m)(n - m + 1)(2n - 2m + 1)}{6}, \end{aligned}$$

tedy $\mathcal{K}_m \stackrel{H_0}{\sim} \mathcal{N}(0, 1)$.

Za pozornost stojí si povšimnout, že t -statistika (3.27) a statistika \mathcal{K}_m jsou ekvivalentní; to na první pohled není zřejmé, proto bude třeba t -statistiku vhodně rozepsat a upravit.

Protože stále vycházíme z předpokladu, že μ je známý parametr, můžeme bez újmy na obecnosti předpokládat, že $\mu = 0$. Pak se lineární model (3.26) redukuje na model

$$\mathbf{Y} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ n - m \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \\ \varepsilon_{m+1} \\ \vdots \\ \varepsilon_n \end{pmatrix}. \quad (3.31)$$

V modelu (3.31) je

$$\mathbf{X}^T \mathbf{X} = \sum_{i=1}^{n-m} i^2, \quad \mathbf{X}^T \mathbf{Y} = \sum_{i=m+1}^n Y_i (i - m).$$

Tedy

$$\begin{aligned} t &= \frac{1}{\sqrt{MSe}} \frac{\hat{\beta}}{\sqrt{v}} = \frac{1}{\sqrt{MSe}} \frac{(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}}{\sqrt{(\mathbf{X}^T \mathbf{X})^{-1}}} = \frac{1}{\sqrt{MSe}} \frac{\sum_{i=m+1}^n Y_i (i - m)}{\sqrt{\sum_{i=1}^{n-m} i^2}} \\ &= \frac{1}{\sqrt{MSe}} \frac{\sum_{i=m+1}^n Y_i (i - m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6}}}. \end{aligned}$$

Samozřejmě, pokud bychom znali rozptyl dat, nahradili bychom MSe přímo pomocí σ^2 a t -statistika se dokonce přímo shoduje se statistikou \mathcal{K}_m (položíme-li $\mu = 0$). V případě neznámé hodnoty σ^2 můžeme použít jakýkoliv konzistentní odhad, v kontextu lineární regrese je zvykem používat MSe , v tradiční statistice S_{*m}^2 . I v takovém případě jsou ale statistiky pochopitelně ekvivalentní.

Vrátíme-li se k úloze s neznámým časem změny (2.2), tak ze stejného důvodu jako v sekci 3.1 budeme uvažovat testovou statistiku tvaru

$$\mathcal{H}^Z = \max_{1 \leq m \leq n-1} \frac{1}{\sigma} \left| \frac{\sum_{i=m+1}^n (Y_i - \mu)(i - m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6}}} \right| \quad (3.32)$$

a při neznámé hodnotě parametru σ jej nahradíme konzistentním odhadem, tj. \sqrt{MSe} nebo S_{*m} .

Detekcí postupné změny se mezi jinými zabývala Jarušková (1998). Stejně jako my nejprve uvažuje situaci, kdy jsou μ a σ^2 známé. Dostává statistiku shodnou s (3.32). Na stranách 270-271 ukazuje, že za určitých podmínek regularity pro velké n a $x \in \mathbb{R}$ platí

$$\mathbb{P} \left[\max_{1 \leq m \leq n-1} \frac{1}{\sigma} \left| \frac{\sum_{i=m+1}^n (Y_i - \mu)(i - m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6}}} \right| > u_n \right] \approx 1 - \exp\{-2e^{-x}\}, \quad (3.33)$$

kde

$$u_n = \sqrt{2 \log \log n} + \frac{1}{\sqrt{2 \log \log n}} \left(\log \frac{\sqrt{3}}{4\pi} + x \right).$$

Pro případ jednostranné alternativy $\beta > 0$ bychom uvažovali pouze maximum \mathcal{K}_m (bez absolutní hodnoty) a aproximovali bychom pomocí $1 - \exp\{-e^{-x}\}$.

Detekce postupné změny při neznámé počáteční střední hodnotě

Je-li nyní parametr μ neznámý, přirozeně by nás napadlo nahradit jej ve výrazu (3.32) jeho maximálně věrohodným odhadem za platnosti nulové hypotézy - výběrovým průměrem všech dat \bar{Y}_n . To se ale ukazuje, že nestačí, jak vysvětluje Jarušková (1998). Je totiž třeba použít jinou standardizaci. Správnou statistiku můžeme odvodit třeba tak, že opět rozepíšeme t -statistiku. Nyní platí

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} 1 & \dots & 1 & 1 & \dots & 1 \\ 0 & \dots & 0 & 1 & \dots & n-m \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & n-m \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n-m} i \\ \sum_{i=1}^{n-m} i & \sum_{i=1}^{n-m} i^2 \end{pmatrix}$$

$$\text{a } \mathbf{X}^T \mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=m+1}^n Y_i(i-m) \end{pmatrix}. \text{ Dále}$$

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{\det \mathbf{X}^T \mathbf{X}} \text{adj}(\mathbf{X}^T \mathbf{X}) = \frac{1}{\det \mathbf{X}^T \mathbf{X}} \begin{pmatrix} \sum_{i=1}^{n-m} i^2 & -\sum_{i=1}^{n-m} i \\ -\sum_{i=1}^{n-m} i & n \end{pmatrix}.$$

Tedy

$$\hat{\beta} = \frac{1}{\det(\mathbf{X}^T \mathbf{X})} \left(-\sum_{i=1}^{n-m} i \sum_{i=1}^n Y_i + n \sum_{i=m+1}^n Y_i(i-m) \right).$$

Dále t -statistika zahrnuje hodnotu v_{22} , což je druhý diagonální prvek matice $(\mathbf{X}^T \mathbf{X})^{-1}$, tedy

$$v_{22} = \frac{n}{\det(\mathbf{X}^T \mathbf{X})}.$$

Determinant matice $\mathbf{X}^T \mathbf{X}$ je roven

$$\begin{aligned} \det(\mathbf{X}^T \mathbf{X}) &= n \sum_{i=1}^{n-m} i^2 - \left(\sum_{i=1}^{n-m} i \right)^2 \\ &= n \left(\frac{(n-m)(n-m+1)(2n-2m+1)}{6} - \frac{1}{n} \frac{(n-m)^2(n-m+1)^2}{4} \right). \end{aligned}$$

Uvažujme nejprve podíl

$$\begin{aligned} \frac{\hat{\beta}}{\sqrt{v_{22}}} &= \frac{\frac{1}{\det(\mathbf{X}^T \mathbf{X})} \left(- \sum_{i=1}^{n-m} i \sum_{i=1}^n Y_i + n \sum_{i=m+1}^n Y_i (i-m) \right)}{\sqrt{\frac{n}{\det(\mathbf{X}^T \mathbf{X})}}} \\ &= \frac{- \sum_{i=1}^{n-m} i \sum_{i=1}^n Y_i + n \sum_{i=m+1}^n Y_i (i-m)}{\sqrt{n \det(\mathbf{X}^T \mathbf{X})}} \\ &= \frac{\sum_{i=m+1}^n Y_i (i-m) - \bar{Y}_n \sum_{i=m+1}^n (i-m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6} - \frac{1}{n} \frac{(n-m)^2(n-m+1)^2}{4}}} \\ &= \frac{\sum_{i=m+1}^n Y_i (i-m) - \sum_{i=m+1}^n \bar{Y}_n (i-m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6} - \frac{1}{n} \frac{(n-m)^2(n-m+1)^2}{4}}} \\ &= \frac{\sum_{i=m+1}^n (Y_i - \bar{Y}_n) (i-m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6} - \frac{1}{n} \frac{(n-m)^2(n-m+1)^2}{4}}}, \end{aligned}$$

což je přesně statistika jako v článku Jarušková (1998), strana 265.

Opět pro test s neznámým časem změny volíme jako testovou statistiku maximum

$$\mathcal{H}^N = \max_{1 \leq m \leq n-1} \frac{1}{\sigma} \left| \frac{\sum_{i=m+1}^n (Y_i - \bar{Y}_n) (i-m)}{\sqrt{\frac{(n-m)(n-m+1)(2n-2m+1)}{6} - \frac{1}{n} \frac{(n-m)^2(n-m+1)^2}{4}}} \right|.$$

Jarušková (1998) dále ukazuje (strana 272), že můžeme použít stejnou aproximaci kritických hodnot jako v případě známé μ .

4. Základní zastavovací pravidla pro online detekci změny

Představme si například výrobní proces, který je schopen potenciálně poskytnout nekonečnou posloupnost pozorování X_1, X_2, \dots . Zpočátku je proces tak zvaně „pod kontrolou“, což interpretujeme tak, že výroba postupuje, jak byla zpočátku nastavena, a tedy pozorovatel je s jejím chodem spokojen a nijak nezasahuje. V nějakém čase τ se proces ale změní, a pak je třeba provést co nejdříve nějaký zásah. Říkáme, že se proces dostal „mimo kontrolu“. Základní úlohou pro pozorovatele je tak odhalit případnou změnu na základě dostupných dat co nejdříve a zasáhnout.

Pro jednoduchost našich úvah budeme předpokládat, že náhodné veličiny X_1, X_2, \dots mají absolutně spojitě rozdělení vzhledem k Lebesgueově míře (a tedy existuje hustota). Tuto hustotu před změnou budeme označovat f_0 , zatímco po změně f_1 .

Zatímco myšlenku online (též sekvenční) detekce si představíme pro obecná rozdělení, v našich aplikacích se soustředíme na hustoty z tzv. Darmois-Koopmanova systému hustot, tj. na jednoparametrickou exponenciální třídu hustot.

Označme P_0 rozdělení, za něž jsou X_1, X_2, \dots iid s hustotou f_0 . Dále označme P_ν , $\nu = 1, 2, 3, \dots$, rozdělení, za něž jsou $X_1, \dots, X_{\nu-1}$ iid s hustotou f_0 a $X_\nu, X_{\nu+1}, \dots$ iid s hustotou f_1 . Označme ještě E_ν odpovídající střední hodnoty.

Naším hlavním cílem bude nalézt takové zastavovací pravidlo (angl. stopping time) τ , že pokud ke změně došlo v čase ν , tak aby zpoždění odhalení změny $(\tau - \nu)^+$ bylo co nejmenší. Na druhé straně nechceme detekovat změnu, která nenastala, čemuž říkáme „vyvolávat falešné poplachy“. Jinými slovy, pokud by ke změně nedošlo, τ by mělo být co nejdelší, ideálně nekonečné. Procedura pak vlastně bude výsledkem optimalizační úlohy, kdy bychom rádi minimalizovali střední dobu do odhalení skutečné změny a maximalizovali střední dobu mezi falešnými poplachy.

Z matematického hlediska se jedná o postupy tzv. sekvenční analýzy, navržené A. Waldem během druhé světové války a popsané v jeho monografii Wald (1947). Pro naše potřeby je vhodným čtením např. monografie Wetherill a Brown (1991).

V této sekci se soustředíme na dvě základní zastavovací pravidla, tj. na Shewhartovu metodu a CUSUM metodu. Pro naše rozhodování budeme uvažovat posloupnost sekvenčních testů, kde

$$\begin{aligned} H : X_i &\sim f_0, \quad i = 1, 2, \dots \\ A_\nu : X_i &\sim f_0, \quad i = 1, \dots, \nu - 1, \\ &X_i \sim f_1, \quad i = \nu, \nu + 1, \dots \end{aligned} \tag{4.1}$$

4.1 Shewhartova metoda

Tradičním a nejstarším přístupem je Shewhartova metoda (a její modifikace), která odhaduje čas změny pomocí zastavovacího pravidla tvaru

$$\tau_S \stackrel{\text{def.}}{=} \inf\{m : \log \frac{f_1(X_m)}{f_0(X_m)} > h\}, \quad (4.2)$$

tj. proces výroby zastavujeme přesně v tom čase m , kdy poprvé logaritmus věrohodnostního poměru pro poslední pozorování, které máme k dispozici, překročí vhodnou úroveň h .

Ekvivalentně se dá vyjádřit ve tvaru

$$\tau_S = \inf\{m : X_m > b\}, \quad (4.3)$$

neboť $\log \frac{f_1(X_m)}{f_0(X_m)} > h \iff g(X_m) > e^h \iff X_m > g^{-1}(e^h) =: b$, kde $g = \frac{f_1}{f_0}$ a g^{-1} je zobecněná inverze.

V praxi může dojít ke dvěma chybným závěrům:

1. Rozhodneme ve prospěch alternativy, zatímco stále platí hypotéza, tj. vyvoláme falešný poplach.
2. Nezamítneme hypotézu, přestože došlo k porušení.

Při detekci postupujeme následovně:

1. Předepíšeme střední dobu mezi falešnými poplasy - označme ji $E_0\tau_S$ - což je střední doba do zastavení, platí-li nulová hypotéza (4.1). Tím bude jednoznačně určeno požadované rozdělení τ_S , jak uvidíme v následujícím odstavci.
2. K ní spočteme h , resp. b .
3. Posoudíme sílu takové procedury pomocí $E_1\tau_S$.

Abychom tedy mohli přistoupit k bodu 1., je třeba odvodit rozdělení zastavovacího pravidla τ_S . Podíváme-li se na (4.2) blíže, vidíme, že τ_S se řídí geometrickým rozdělením, neboť čekáme na první vadný výrobek. Přesněji:

$$\tau_S \sim Ge(p), \text{ kde } p = \mathbb{P}\left(\log \frac{f_1(X_m)}{f_0(X_m)} > h\right), \text{ resp. } p = \mathbb{P}(X_m > b).$$

Z výrazu (4.2) je také zřejmé, že se jedná o geometrické rozdělení jak za H , tak za A_1 . To tedy znamená, že

$$\mathbb{P}(\tau_S = k) = (1 - p)^{k-1}p, \quad k = 1, 2, \dots, \quad (4.4)$$

a tedy $E\tau_S = \frac{1}{p}$ a $\text{var}\tau_S = \frac{1}{p^2}$. Z toho tedy mj. vidíme, že střední hodnota určuje rozdělení τ_S jednoznačně, jak již bylo avizováno. Vztah (4.4) platí jak při P_0 , tak P_1 .

Příklad: Situaci ilustrujme na příkladu normálního rozdělení $\mathcal{N}(\mu, \sigma^2)$, kdy nám jde o odhalení případné změny parametru polohy μ . Pro jednoduchost budeme předpokládat, že σ^2 je známý a, bez újmy na obecnosti, že je-li proces pod (statistickou) kontrolou, tak $\mu = 0$.

Označme P_μ rozdělení, za něž jsou X_i iid a řídí se normálním rozdělením $\mathcal{N}(\mu, \sigma^2)$. Abychom objasnili značení P_μ : dle (4.1) uvažujeme dvě distribuce f_0 a f_1 a mezi nimi se rozhodujeme. Zde $f_0 \sim \mathcal{N}(0, \sigma^2)$ a $f_1 \sim \mathcal{N}(\mu, \sigma^2)$. Aby bylo jasné, jak velký je skok μ , místo P_1 budeme psát P_μ .

Odvoďme tedy zastavovací pravidlo (4.2). Logaritmus věrohodnostního podílu je roven

$$\begin{aligned} \log \frac{f_1(X_m)}{f_0(X_m)} &= \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(X_m - \mu)^2\right\}}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}X_m^2\right\}} \\ &= -\frac{1}{2\sigma^2}(X_m - \mu)^2 + \frac{1}{2\sigma^2}X_m^2 \\ &= -\frac{1}{2\sigma^2}(X_m^2 - 2\mu X_m + \mu^2) + \frac{1}{2\sigma^2}X_m^2 \\ &= \frac{1}{\sigma^2} \left(\mu X_m - \frac{\mu^2}{2} \right). \end{aligned}$$

Zastavovací pravidlo je výhodnější uvádět ve tvaru (4.3), neboť pak se snáze spočte parametr p . Tvar tohoto zastavovacího pravidla se ale liší v závislosti na uvažované alternativě; my si jej odvodíme pro případ jednostranné alternativy $\mu > 0$:

- Proces je pod kontrolou, je-li $\mu = 0$,
- Proces je mimo kontrolu, jestliže $\mu = \mu^* > 0$.

Poznámka: Tady tedy nastává zásadní změna oproti klasickému (offline) testování hypotéz. Zatímco v kapitole 3 jsme se snažili detekovat jakékoli porušení nulové hypotézy, zde detekujeme konkrétní porušení, tj. konkrétní posun střední hodnoty.

Jednoduchými úpravami dostaneme, že

$$\frac{1}{\sigma^2} \left(\mu^* X_m - \frac{\mu^{*2}}{2} \right) \geq h \iff X_m \geq \frac{\sigma^2}{\mu^*} h + \frac{\mu^*}{2} =: b.$$

Abychom našli optimální hodnotu b , musíme spočítat $\mathbf{E}_0 \tau_S^+$. Za P_0 má τ_S^+ geometrické rozdělení

$$P_0(\tau_S^+ = k) = (1 - p_0)^{k-1} p_0, \quad k = 1, 2, \dots,$$

kde $p_0 = P_0(X_m \geq b) = 1 - \Phi(b/\sigma)$, takže

$$\mathbf{E}_0 \tau_S^+ = \frac{1}{p_0} = \frac{1}{1 - \Phi(b/\sigma)}.$$

Podobně za P_{μ^*} má τ_S^+ opět geometrické rozdělení a platí

$$\mathbf{E}_1 \tau_S^+ = \frac{1}{p_{\mu^*}} = \frac{1}{\mathbf{P}(X_m \geq b)} = \frac{1}{1 - \Phi\left(\frac{b - \mu^*}{\sigma}\right)}.$$

Obvyklým pravidlem je zvolit pevnou hodnotu $E_0\tau_S^+$ a k ní dopočítat $E_1\tau_S^+$. Standardně značíme $ARL_0 := E_0\tau_S^+ = \frac{1}{1-\Phi(b/\sigma)}$. Zkratka *ARL* pochází z anglického „average run length“, tj. průměrná délka jednoho běhu, a 0 symbolizuje, že za platnosti nulové hypotézy. ARL_0 tedy stanovíme my a pomocí ní vyjádříme b : $b = \sigma\Phi^{-1}(1 - \frac{1}{ARL_0})$.

V praxi většinou usilujeme o to detekovat posun ve střední hodnotě o nějaký násobek σ , tj. $\mu = k\sigma$ pro nějaké $k \in \mathbb{R}^+$. Potom jsme schopni spočítat $E_1\tau_S^+$ pro libovolnou hodnotu $\sigma > 0$:

$$E_1\tau_S^+ = \frac{1}{1 - \Phi\left(\frac{b-\mu}{\sigma}\right)} = \frac{1}{1 - \Phi\left(\frac{\sigma\Phi^{-1}(1-1/ARL_0)-k\sigma}{\sigma}\right)} = \frac{1}{1 - \Phi(\Phi^{-1}(1 - 1/ARL_0) - k)}.$$

Spočtěme tuto hodnotu pro posuny střední hodnoty o vybrané násobky σ při $ARL_0 = 1000$:

μ	0	$\frac{\sigma}{2}$	σ	$\frac{3}{2}\sigma$	2σ	3σ
$E_\mu\tau_S^+$	1000	208	55	18	7	2

Tabulka 4.1: Střední doby zastavení za platnosti alternativy pro vybrané posuny ve střední hodnotě, zaokrouhlené na celé jednotky.

Poznámka: Pokud se chceme bránit proti alternativě, kdy $\mu < 0$, resp. $\mu \neq 0$, postupujeme analogicky. Zastavovací pravidla pak pro větší přehlednost budeme označovat τ_S^- , resp. τ_S .

4.2 Metoda CUSUM

Další metody sekvenční detekce jsou založeny na tak zvaných kumulativních součtových statistikách $\{S_n\}_{n \in \mathbb{N}}$. Tyto metody označujeme zkratkou CUSUM (z anglického „cumulative sum“). My představíme kumulativní součtové statistiky vycházející z logaritmu věrohodnostního poměru:

$$S_n = \log \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)} = \sum_{i=1}^n \log \frac{f_1(X_i)}{f_0(X_i)}, \quad n \in \mathbb{N}, \quad S_0 = 0,$$

kde $f_1(X_i)$ je hustota X_i za alternativy a analogicky $f_0(X_i)$ je hustota X_i za hypotézy.

Za povšimnutí stojí, že posloupnost $\{S_n\}_{n \in \mathbb{N}}$ tvoří náhodnou procházku, neboť se jedná o postupné součty náhodných veličin $\{Y_i\}$, $Y_i := \log \frac{f_1(X_i)}{f_0(X_i)}$.

Popíšme si myšlenku zastavovacího pravidla.

Nejprve předpokládejme, že máme k dispozici konečnou posloupnost nezávislých náhodných veličin X_1, \dots, X_n a že čas změny je známý. To jest, navrhujeme test

$$\begin{aligned} H &: X_1, \dots, X_n \sim f_0, \\ A_m &: X_1, \dots, X_m \sim f_0, \\ & \quad X_{m+1}, \dots, X_n \sim f_1. \end{aligned}$$

Test založíme standardně na poměru věrohodností:

$$\begin{aligned} S_n &= \log \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)} = \sum_{i=1}^n \log \frac{f_1(X_i)}{f_0(X_i)} = \sum_{i=1}^m \log \frac{f_0(X_i)}{f_0(X_i)} + \sum_{i=m+1}^n \log \frac{f_1(X_i)}{f_0(X_i)} \\ &= \sum_{i=m+1}^n \log \frac{f_1(X_i)}{f_0(X_i)} = S_n - S_m, n = 1, 2, \dots \end{aligned}$$

Nulovou hypotézu bychom zamítali pro velké hodnoty $S_n - S_m$.

Nyní přistupme k obecnějšímu testu s neznámým časem změny, nicméně se stále pevným rozsahem výběru. Myšlenka takového testu již byla popsána v kapitole 3; hypotézu zamítneme tehdy, je-li $\max_{0 \leq m \leq n-1} (S_n - S_m)$ velké. Všimněme si, že

$$\max_{0 \leq m \leq n-1} (S_n - S_m) = S_n - \min_{0 \leq m \leq n-1} S_m.$$

Protože hledáme první čas τ_C , ve kterém došlo ke změně, tak zastavovací pravidlo musí být dáno následovně:

$$\tau_C = \inf\{n \in \mathbb{N} : S_n - \min_{0 \leq m \leq n-1} S_m \geq \tilde{h}\}. \quad (4.5)$$

V sekvenčním postupu tedy musíme test zkonstruovat pro každé nové pozorování znovu, čímž dostáváme sekvenční posloupnost testů poměrem věrohodnosti, v angličtině SPRT, což je zkratka pro „sequential probability ratio tests“.

Zatímco jsme úlohu sekvenční analýzy 4.1 definovali v plné obecnosti, pro praktické aplikace je metoda CUSUM navržena tak, aby detekovala konkrétní porušení nulové hypotézy. Úloha se tak redukuje na problém, kdy se rozhodujeme mezi dvěma hypotézami. V praxi tak zpravidla jde o detekci konkrétní změny parametru rozdělení uvažované posloupnosti $\{X_i\}$, který dále budeme označovat θ .

Pro metodu CUSUM je dále nezbytné nastavit hladinu \tilde{h} v (4.5). Zvolíme-li posun parametru θ , který si přejeme detekovat, nastavíme délku střední doby mezi falešnými poplachy, díky čemuž dopočteme \tilde{h} . V následujících odstavcích si vysvětlíme, jak se dá zastavovací pravidlo CUSUM zkonstruovat ve speciální rodině rozdělení.

4.2.1 Darmois-Koopmanova třída rozdělení

Ve zbytku kapitoly nechť pozorování X_1, X_2, \dots pocházejí z rozdělení z jednoparametrické exponenciální třídy s hustotou tvaru

$$f(x) = h(x) \exp\{\theta T(x) - b(\theta)\}, \quad (4.6)$$

kde $\theta \in \mathbb{R}$ je jednorozměrný reálný parametr, $h(x), T(x)$ jsou reálné funkce a $b(\theta)$ je dvakrát spojitě diferencovatelná reálná funkce, splňující $b(0) = 0$.

Do této rodiny rozdělení patří například normální rozdělení, gama rozdělení (uvažujeme-li jeden z parametrů známý), speciálně tedy exponenciální rozdělení aj. Naopak, například Cauchyho rozdělení do této třídy nepatří.

Stejně jako v případě Shewhartovy metody, tvar zastavovacího pravidla se liší dle volené alternativy. Bez újmy na obecnosti nechť

- proces je pod kontrolou, je-li $\theta = 0$;
- proces je mimo kontrolu, je-li $\theta = \theta^* \neq 0$, nebo $\theta = \theta^* > 0$, nebo $\theta = \theta^* < 0$.

V Darmois-Koopmanově rodině rozdělení platí

$$\begin{aligned} S_n &= \sum_{i=1}^n \log \frac{f_1(X_i)}{f_0(X_i)} \\ &= \sum_{i=1}^n \log \frac{h(X_i) \exp\{\theta^* T(X_i) - b(\theta^*)\}}{h(X_i)} \\ &= \sum_{i=1}^n (\theta^* T(X_i) - b(\theta^*)) \end{aligned}$$

a tím pádem zastavovací pravidlo (4.5) je tvaru

$$\begin{aligned} \tau_C &= \inf\{n : \max_{0 \leq m < n} (S_n - S_m) \geq \tilde{h}\} \\ &= \inf\{n : \max_{0 \leq m < n} \sum_{i=m+1}^n (\theta^* T(X_i) - b(\theta^*)) \geq \tilde{h}\}. \end{aligned}$$

Pro jednostrannou alternativu $\theta = \theta^* > 0$ upravíme jednoduše zastavovací pravidlo do tvaru

$$\begin{aligned} \tau_C^+ &= \inf \left\{ n : \max_{0 \leq m < n} \sum_{i=m+1}^n \left(T(X_i) - \frac{b(\theta^*)}{\theta^*} \right) \geq \frac{\tilde{h}}{\theta^*} \right\} \\ &= \inf \left\{ n : \max_{0 \leq m < n} \sum_{i=m+1}^n (T(X_i) - k) \geq h \right\}, \end{aligned} \quad (4.7)$$

kde značíme $k := \frac{b(\theta^*)}{\theta^*}$, $h = \frac{\tilde{h}}{\theta^*}$.

Pro opačnou alternativu $\theta = \theta^* < 0$ musíme postupovat opatrněji:

$$\begin{aligned} \tau_C^- &= \inf \left\{ n : \max_{0 \leq m < n} \theta^* \sum_{i=m+1}^n \left(T(X_i) - \frac{b(\theta^*)}{\theta^*} \right) \geq \tilde{h} \right\} \\ &= \inf \left\{ n : \theta^* \min_{0 \leq m < n} \sum_{i=m+1}^n \left(T(X_i) - \frac{b(\theta^*)}{\theta^*} \right) \geq \tilde{h} \right\} \\ &= \inf \left\{ n : \min_{0 \leq m < n} \sum_{i=m+1}^n \left(T(X_i) - \frac{b(\theta^*)}{\theta^*} \right) \leq \frac{\tilde{h}}{\theta^*} \right\}. \end{aligned} \quad (4.8)$$

Očividně, zastavovacím pravidlem při oboustranné alternativě $\theta = \theta^* \neq 0$ pak bude

$$\tau_C = \min\{\tau_C^-, \tau_C^+\}.$$

Jak jsme již předeslali v úvodním odstavci, zvolená hladina \tilde{h} závisí na střední době mezi falešnými poplachy, kterou si sami zvolíme. Aproximaci vztahu mezi ARL a \tilde{h} nabízí Wald (1947)

$$ARL_+(\theta) \approx \frac{h(\theta - \tilde{\theta}) - 1 + e^{-(\theta - \tilde{\theta})h}}{(\theta - \tilde{\theta}) \left(b'(\theta) - \frac{b(\theta) - b(\tilde{\theta})}{\theta - \tilde{\theta}} \right)},$$

kde $\tilde{\theta}$ je definováno vztahem

$$\frac{b(\theta) - b(\tilde{\theta})}{\theta - \tilde{\theta}} = \frac{b(\theta^*)}{\theta^*} = k.$$

CUSUM v normálním modelu

Všimněme si, že normální rozdělení spadá do Darmois-Koopmanovy třídy rozdělení. Totiž,

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\} \\ &= h(x) \exp\{\theta T(x) - b(\theta)\}, \end{aligned}$$

volíme-li

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}, \quad \theta = \frac{\mu}{\sigma}, \quad T(x) = \frac{x}{\sigma} \quad \text{a} \quad b(\theta) = \frac{\theta^2}{2}.$$

Dle (4.7), (4.8) jsou zastavovací pravidla pro jednostranné alternativy $\mu = \mu^* > 0$, $\mu = \mu^* < 0$ po řadě rovny

$$\begin{aligned} \tau_C^+ &= \inf \left\{ n : \max_{0 \leq m < n} \sum_{i=m+1}^n \left(\frac{X_i}{\sigma} - \frac{\mu^*}{2\sigma} \right) \geq \frac{\tilde{h}}{\mu^*} \sigma \right\}, \\ \tau_C^- &= \inf \left\{ n : \min_{0 \leq m < n} \sum_{i=m+1}^n \left(\frac{X_i}{\sigma} - \frac{\mu^*}{2\sigma} \right) \leq \frac{\tilde{h}}{\mu^*} \sigma \right\}. \end{aligned}$$

a konečně pro oboustrannou alternativu $\mu = \mu^* \neq 0$ je zastavovací pravidlo tvaru

$$\tau_C = \min\{\tau_C^-, \tau_C^+\}.$$

ARL můžeme opět aproximovat dle Walda

$$\begin{aligned} ARL_+(\mu) = \mathbb{E}_\mu \tau_C^+ &\approx \frac{2h(\mu - k) - 1 + e^{-2h(\mu - k)}}{2(\mu - k)^2}, \quad \mu \neq k, \\ &\approx h^2, \quad \mu = k. \end{aligned}$$

Bohužel, tato aproximace není dostatečně přesná. V literatuře byly navrženy různé modifikace, s jednou z nich přišel například David Siegmund (1985):

$$\begin{aligned} ARL_+(\mu) &\approx \frac{2(h + 1,166)(\mu - k) - 1 + e^{-2(h+1,166)(\mu - k)}}{2(\mu - k)^2}, \quad \mu \neq k, \\ &\approx (h + 1,166)^2, \quad \mu = k. \end{aligned} \tag{4.9}$$

5. Simulace

Postupy pro online či offline detekci jsou z literatury běžně známé. Co ale chybí, je porovnání těchto dvou přístupů.

V přirozeném zájmu výrobce je detekovat poruchu při výrobní dávce co nejdříve, neboť čím déle probíhá vadná výrobní dávka, tím k větším ztrátám dochází.

V předchozích kapitolách jsme pro různé typy porušení modelu popsali různé metody detekce. Která detekce je rychlejší? Fungují spolehlivě různé metody detekce i pro zanedbatelné změny modelu?

Cílem této simulace bude porovnat metody pro detekci změny parametru posunutí z hlediska zpoždění zastavení a počtu předčasných detekcí. Změnou v tomto případě bude pokles střední hodnoty. Budeme zaznamenávat čas zastavení a odhadovat jeho rozdělení. Nejprve porovnáme časy zastavení v rámci offline metod, tj. metod využívajících normality dat a robustních metod, popsaných v sekci 3.1. Zadruhé porovnáme detekce offline a online přístupem, popsaným v kapitole 4.

Třída offline postupů využívajících normality dat bude reprezentována metodou postavenou na testové statistice

$$T^* = \max_{1 \leq m \leq n-1} \sqrt{\frac{n}{m(n-m)}} \frac{1}{\sigma} \left| \sum_{i=1}^m (Y_i - \bar{Y}_n) \right|;$$

zvoleným robustním postupem pak bude Wilcoxonova metoda

$$W^* = \max_{1 \leq m \leq n-1} \frac{1}{\sqrt{\frac{1}{12}m(n-m)(n+1)}} \left| \sum_{i=1}^m R_i - \frac{1}{2}m(n+1) \right|.$$

Konečně, pro online detekci zvolíme metodu CUSUM se zastavovacím pravidlem

$$\tau_C = \inf\{n \in \mathbb{N} : S_n - \min_{0 \leq m \leq n-1} S_m \geq h\}.$$

Všechny tři metody budou aplikovány na ta stejná data.

Základním nástrojem pro posouzení rozdělení časů zastavení budou histogramy a empirické distribuční funkce. Za kvalitnější považujeme takovou metodu, která odhalí změnu dříve. Je třeba si naopak uvědomit, že za nežádoucí považujeme detekci předčasnou, tj. před skutečným časem změny. V případě offline postupů budeme mluvit o chybě předčasné detekce, v případě sekvenčních postupů o vyvolání falešného poplachu. Pravděpodobnost chyby předčasné detekce přirozeně odhadneme relativním počtem předčasných detekcí, v online případě pravděpodobnost falešného poplachu odhadneme analogicky.

Kvalitu offline testů posuzujeme též pomocí jejích síly, kterou budeme odhadovat relativním počtem detekcí provedených po čase změny. To stejné spočteme i pro CUSUM metodu.

5.1 Postup simulace

Inspirací budiž třeba výrobní dávka piva. Generované hodnoty si tedy pro názornost můžeme představovat jako objemy stočených piv v mililitrech. V praxi

se v rámci jedné výrobní dávky stočí několik tisíc lahví, my z důvodu časové náročnosti zvolíme dávku o 300 produktech. Jedna simulace nechť tedy odpovídá jedné výrobní dávce a takových simulací provedeme 1000. Data budou generována z normálního rozdělení se střední hodnotou 500 a směrodatnou odchylkou 1, kterou budeme považovat za známou. Simulace bude nastavena tak, že ke změně - poklesu - střední hodnoty dojde při 151. výrobku, tj. přesně v polovině stáčení. V rámci každé simulace se s každým novým výrobkem spočte odpovídající testová statistika a porovná se s kritickou hodnotou.

Normy stanovují toleranci změny objemu poměrně přísně. Například euro-norma stanovuje 1 ml na půl litr piva, Pilsner Urquell dokonce svého času určoval 0,5 ml. Nás tedy přirozeně bude zajímat, jak zmíněné postupy dokáží reagovat i vůči tak zanedbatelným změnám.

Člověk by očekával, že čím významnější bude změna ve střední hodnotě, tím rychleji budou procedury detekovat. Abychom tuto hypotézu mohli ověřit, budeme snižovat objem stočeného piva postupně o 0,5, 1, 1,5, 2 a 3 ml a pro každý posun provedeme 1000 simulací. V případě offline procedur se zaměříme i na chování, pokud k poruše při výrobní dávce vůbec nedojde, tj. bude-li se držet hladina stále na úrovni 500 ml.

Připomeňme, že zatímco offline metody jsou navrženy tak, aby detekovaly jakékoli porušení hypotézy, metodu CUSUM je třeba nastavit pro detekci konkrétního posunu parametru posunutí. Dále musíme zvolit střední dobu mezi falešnými poplasy, abychom spočetli hladinu h . Střední dobu budeme volit postupně 300, 500, 700 a 1000. Zastavovací pravidlo CUSUM bude pak tvaru

$$\tau_C^- = \inf\{n : \min_{0 \leq m < n} \sum_{i=m+1}^n (X_i - 500 + k^*) < -h\},$$

kde k^* je rovno polovině velikosti detekovaného poklesu a h je dáno Siegmundovou aproximací ARL_+ (4.9) (za platnosti nulové hypotézy).

5.2 Výsledky simulací

5.2.1 Výsledky offline detekce

Jak již bylo řečeno, představu o rozdělení časů zastavení si můžeme udělat na základě histogramů. Vedle sebe postavíme histogramy časů detekce metodou pro normálně rozdělená data a Wilcoxonovou metodou. Představujeme histogramy pro všechny uvažované posuny ve střední hodnotě, abychom mohli porovnat, jak zpoždění detekce (či počet předčasných detekcí) závisí na velikosti posunu.

Prohlédněme si nejprve obrázky histogramů 5.1 až 5.6. Podíváme-li se na histogram 5.1 udávající četnosti časů zastavení, pokud výrobní dávka celou dobu běžela podle hypotézy, může se nám zdát zvláštní, že součet četností časů zastavení není stejný pro obě dvě metody. To je ale dáno tím, že metoda normality (mylně) detekovala častěji než Wilcoxonova metoda. Na obrázcích 5.1 až 5.6 si můžeme všimnout dalších skutečností. Zprvce, porovnáme-li rozdělení časů zastavení metodou normality a Wilcoxonovou metodou, pak, nehledě na velikost posunu ve střední hodnotě, můžeme říct, že první metoda předčasně zastavuje častěji než metoda robustní. Na první pohled by se mohlo zdát, že četnosti předčasných detekcí jsou nejvyšší při žádném posunu a při posunu o $\frac{1}{2}\sigma$; musíme si

ale uvědomit, že histogramy pro různé posuny mají různě škálovanou vertikální osu. Proto si člověk lepší představu udělá z obrázku 5.13 pravděpodobností chyby předčasné detekce, ze kterého pro obě dvě metody můžeme vyčíst závislost pravděpodobnosti předčasné detekce na velikosti poklesu. Tam se ukazuje, že sice, nedošlo-li k žádnému posunu, pak četnosti časů předčasné detekce obou metod jsou podstatně větší, než když posouváme o nějaký nenulový násobek σ . V rámci posunů o libovolný nenulový násobek σ jsou ale tyto četnosti již podobné.

Při posunu o $\frac{1}{2}\sigma$ jsou histogramy napravo od času změny velmi podobné, jen možná histogram pro robustní metodu má zastoupenější pravý chvost. Tedy u robustní metody pozorujeme častěji pozdnější časy zastavení.

Když střední hodnotu posuneme o σ , pozorujeme, že drtivá většina detekce metodou normality proběhne mezi časy 150 a 170. Srovnáme-li s pravým histogramem, opět můžeme vidět, že robustní metoda reaguje na změnu trochu později. To samé platí pro posun o $\frac{3}{2}\sigma$.

Histogramy (jejich pravé části) pro posun o 2σ a 3σ jsou prakticky stejné.

Závěrem můžeme říct, že obě dvě metody reagují mnohem rychleji na větší posuny než na menší, robustní metoda je mírně pomalejší.

Následující obrázky s empirickými distribučními funkcemi 5.7 až 5.12; pro každý z volených poklesů střední hodnoty prezentujeme jeden obrázek porovnávající distribuční funkce obou offline metod. Zde je nutné podotknout, že v případě, kdy k poklesu hladiny nedošlo, tj. výrobní dávka celou dobu běžela na stejném nastavení, v nadpoloviční části případů ani jedna z metod nedetekovala, což je požadovaná vlastnost. Z toho vyplývá, že o rozdělení času zastavení získáváme představu jen do času 300, poté výroba skončí. Nicméně chybějící čas zastavení neznamena, že nemáme žádnou informaci. Chybí-li čas zastavení, můžeme z toho vyvodit, že by čas zastavení byl větší než 300. Odhad úplného rozdělení času zastavení by tak mohl být předmětem analýzy cenzorovaných dat.

At jsme střední hodnotu posouvali o libovolný nenulový násobek σ , vidíme, že graf ECDF časů zastavení stanovených metodou normality se nachází nad grafem ECDF časů zastavení robustní metodou. To tedy jednak znamená, že při robustní metodě je nižší pravděpodobnost předčasné detekce, což koneckonců koresponduje s komentářem v předchozím odstavci, ale na druhou stranu je vyšší pravděpodobnost pozdnější detekce.

Shrneme-li dosavadní pozorování, můžeme stanovit závěr, že metoda využívající normality dat obecně detekuje rychleji, což tedy jednak znamená, že vadnou výrobní dávku zastaví dříve, ale zároveň častěji zastavuje i dávku se správným nastavením. Má-li dojít k větším posunům (aspoň $\frac{3}{2}\sigma$), doporučujeme volit Wilcoxonovu metodu. Pravděpodobnost předčasné detekce je 0,05, což je o celou jednu desetinu nižší, než je pravděpodobnost předčasné detekce pro metodu normality. Navíc při takto velkých posunech je zpoždění obou metod velmi srovnatelné (a minimální). Dojde-li k menším posunům, závisí na úhlu pohledu. Metoda normality je rychlejší, nicméně stále platí, že častěji detekuje předčasně.

Mluvíme-li o offline metodách, pozornost budeme věnovat již zmíněné chybě předčasné detekce a síle těchto metod. Pravděpodobnost chyby předčasné detekce bude odhadnuta pomocí jednoduchého vzorce

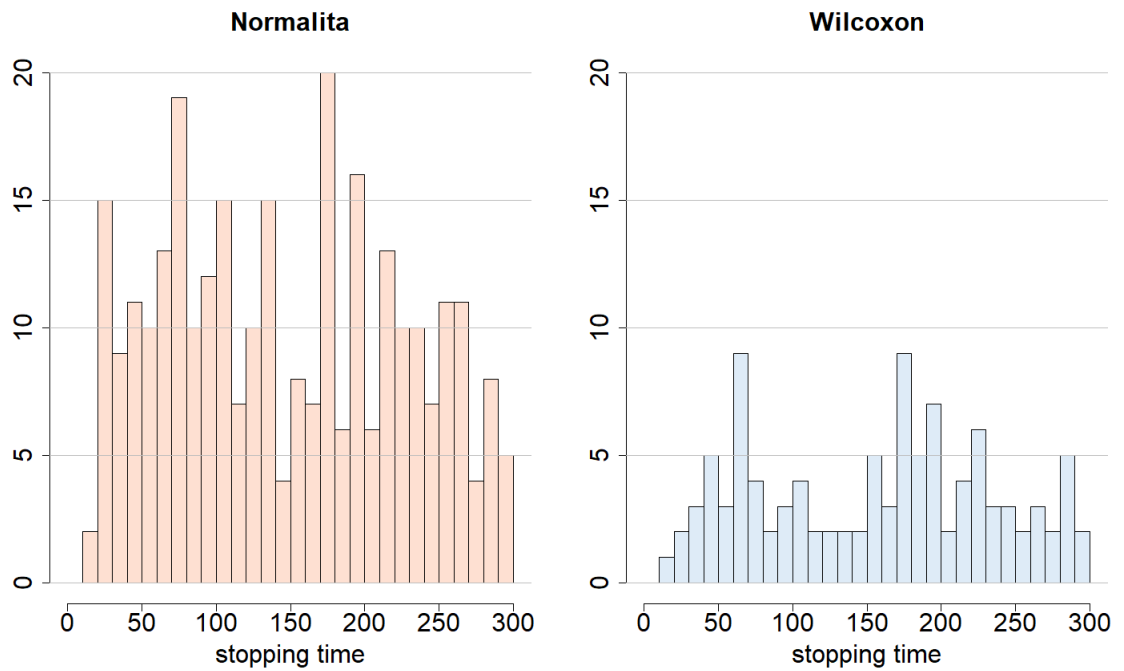
$$\hat{p} = \frac{\text{počet detekcí před časem změny } m = 151}{1000}.$$

Podíváme-li se na obrázek 5.13, tak vidíme vysokou pravděpodobnost zastavení výrobní dávky, která celou dobu běží správně, u obou metod; u metody normality je ale zhruba o dvě desetiny vyšší. Žádný významný trend ve vývoji chyb předčasné detekce v závislosti na velikosti posunu jinak nevidíme. Zajímavé je, že navzdory mnohonásobnému testování se pravděpodobnost předčasného detekování robustní metodou drží kolem 0,05, u druhé metody je asi o jednu desetinu vyšší.

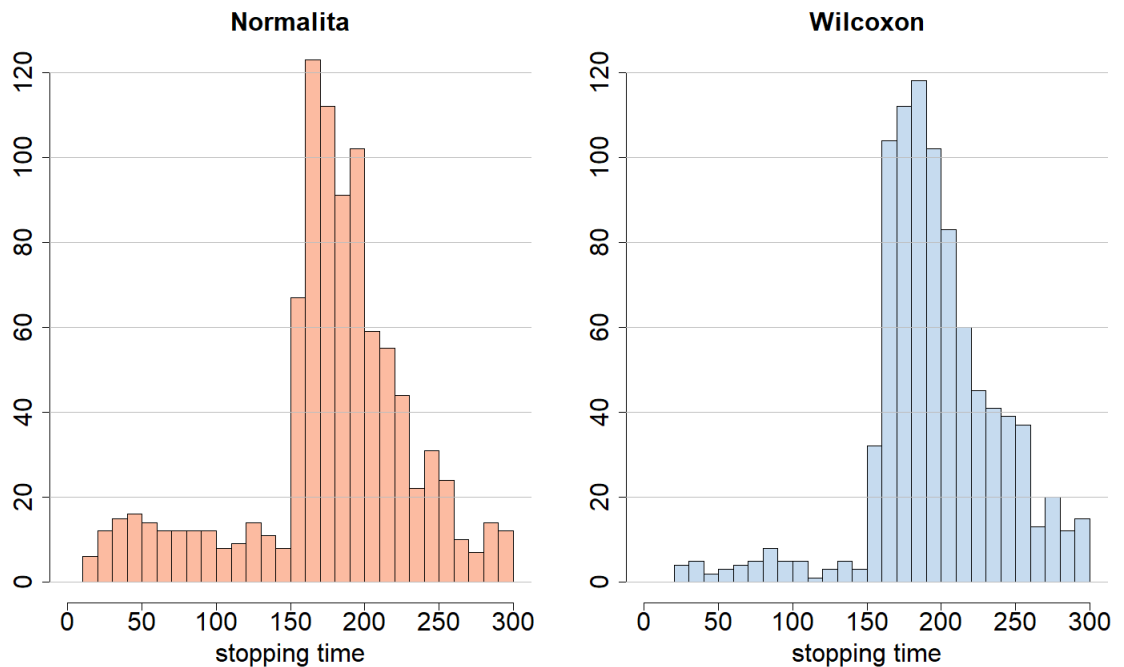
Co se týče síly testu, jde o pravděpodobnost zamítnutí neplatné hypotézy, tedy v našem případě o pravděpodobnost detekce po čase změny. Samozřejmě, rozsah výrobní dávky je omezen 300 výrobky, tedy nedošlo-li k detekci do 300. výrobku, musíme k tomu přistoupit tak, jako kdyby k detekci nedošlo vůbec. Sílu offline testů tedy odhadneme vzorcem

$$\hat{\beta} = \frac{\text{počet detekcí po čase změny } m = 151}{1000}.$$

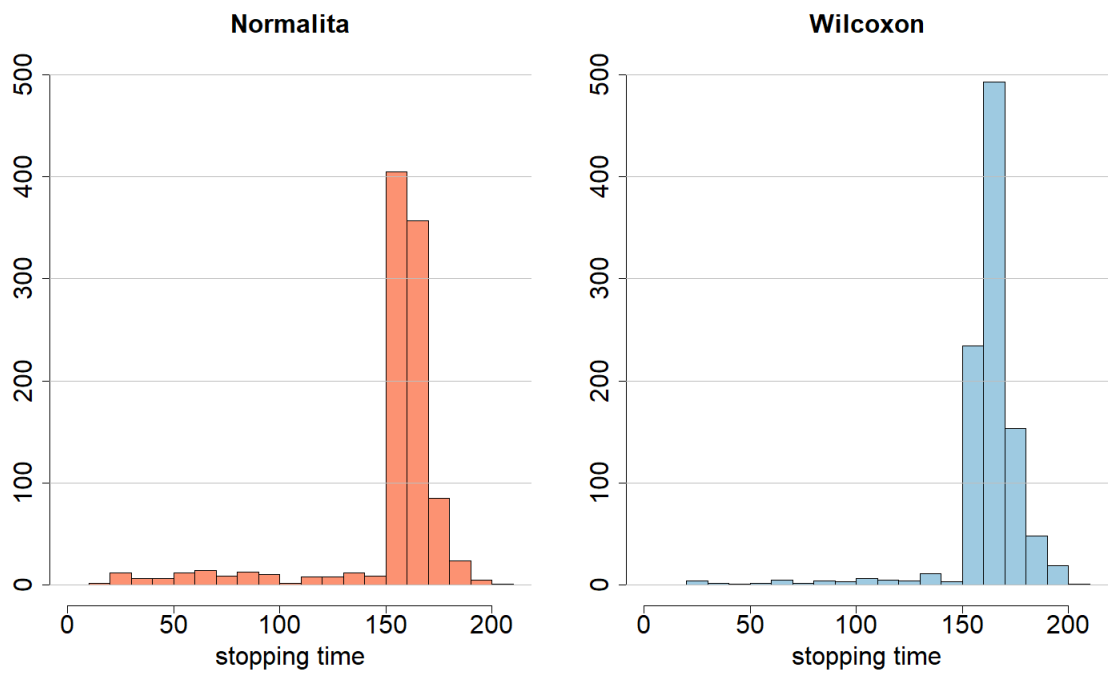
Pro případ simulace, kdy k poklesu hladiny střední hodnoty nedošlo, sílu nebudeme uvažovat. Silnější metodou je opět robustní metoda, nicméně zpoždění časů detekce obrázek 5.14 neukazuje. My už ale víme, že metoda za normality detekuje změnu rychleji než robustní metoda. U obou metod pozorujeme mírně rostoucí trend ve vývoji síly v závislosti na velikosti posunu, což bychom očekávali.



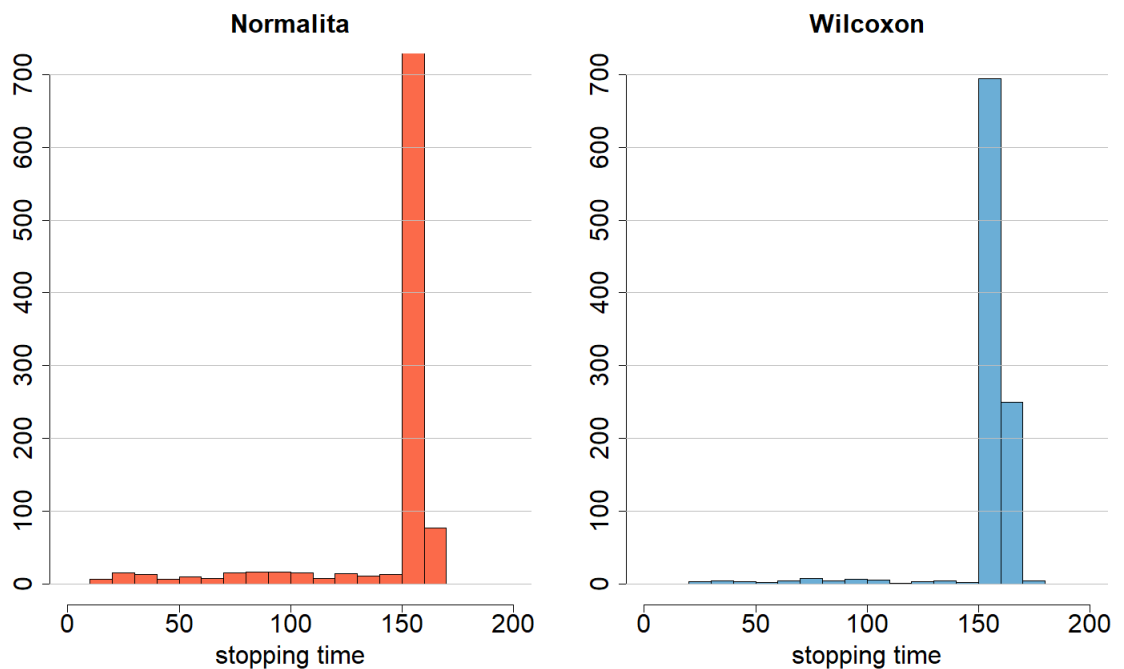
Obrázek 5.1: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), když ke změně nedošlo.



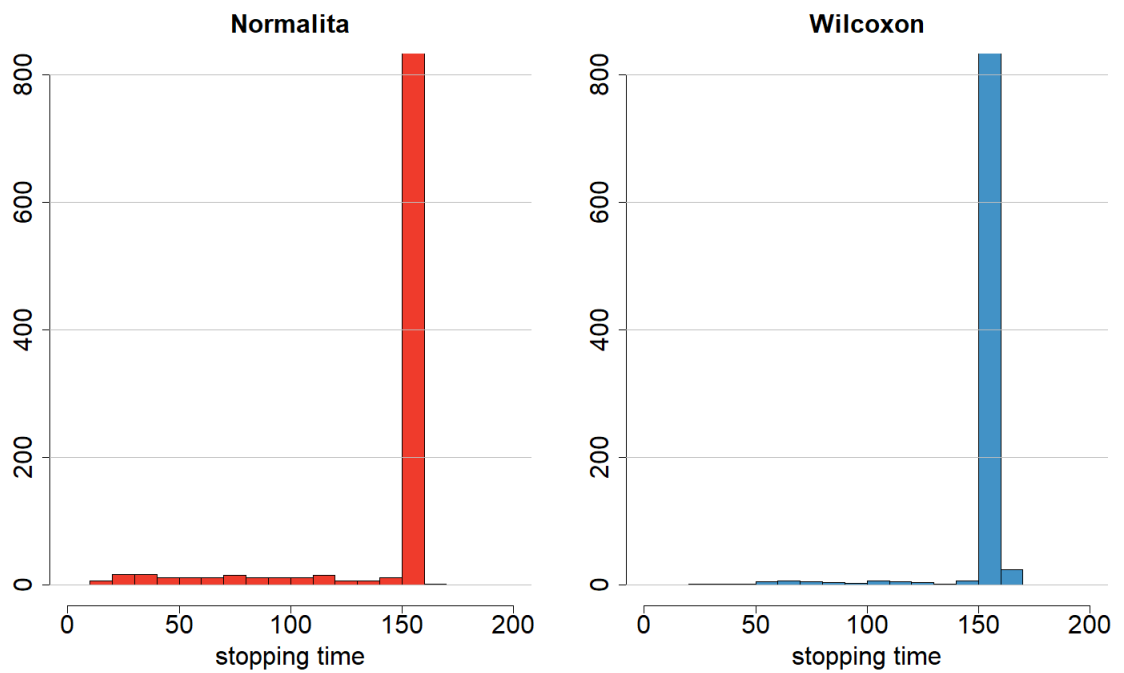
Obrázek 5.2: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o $\frac{1}{2}\sigma$.



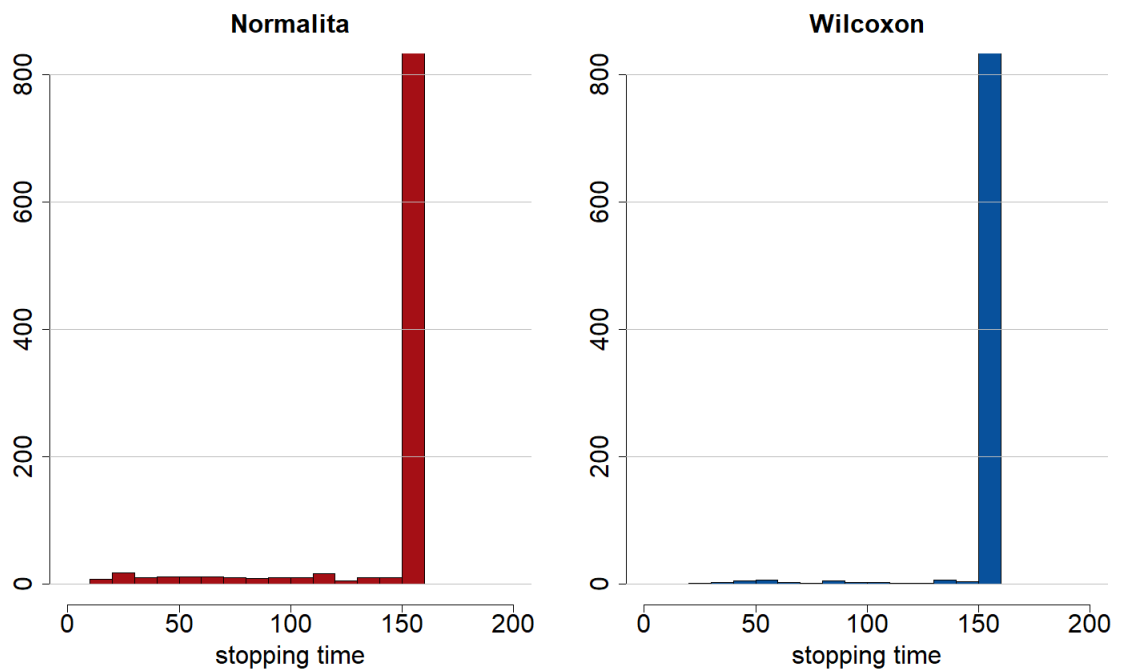
Obrázek 5.3: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o σ .



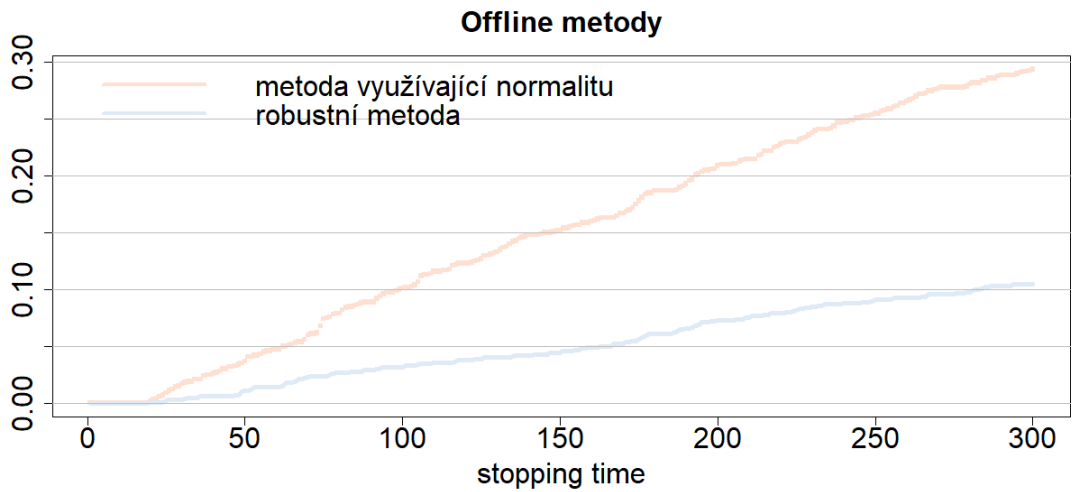
Obrázek 5.4: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o $\frac{3}{2}\sigma$.



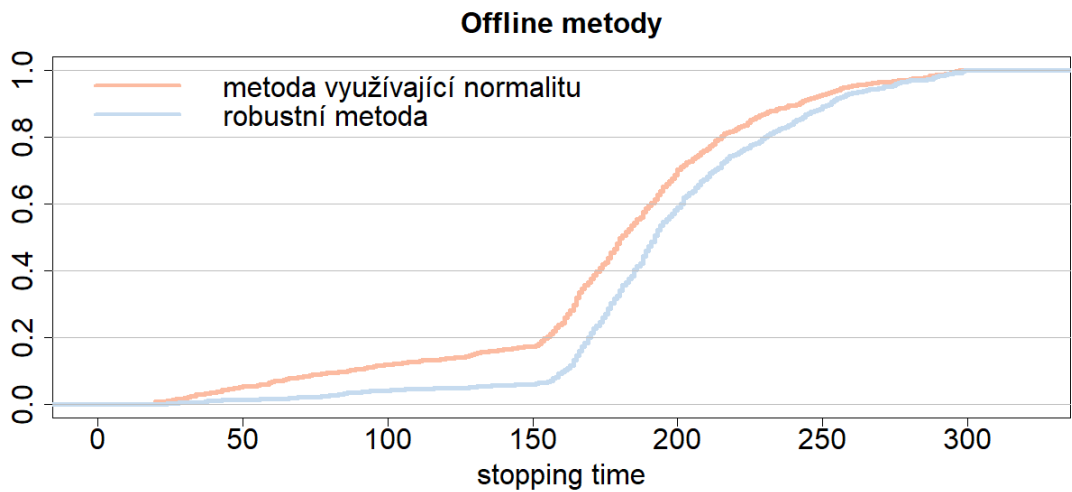
Obrázek 5.5: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o 2σ .



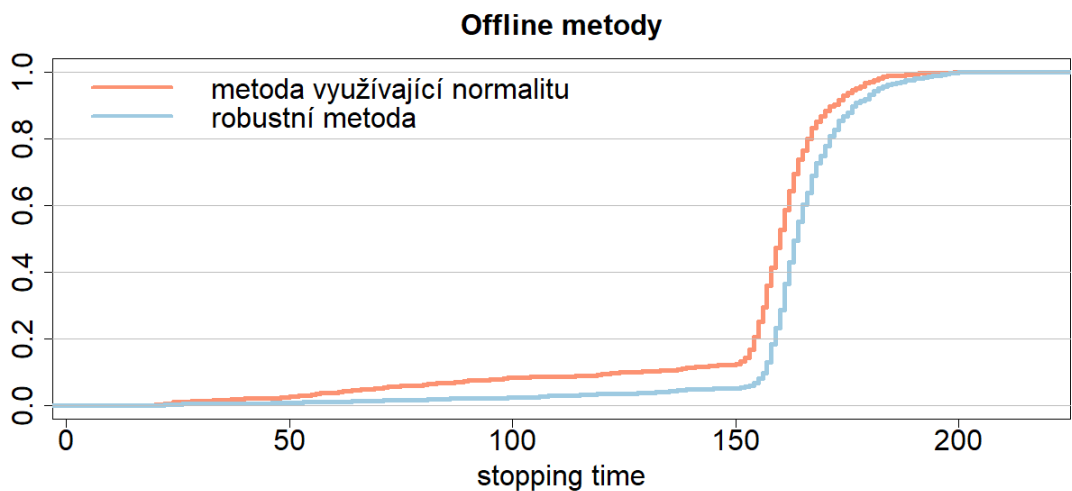
Obrázek 5.6: Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o 3σ .



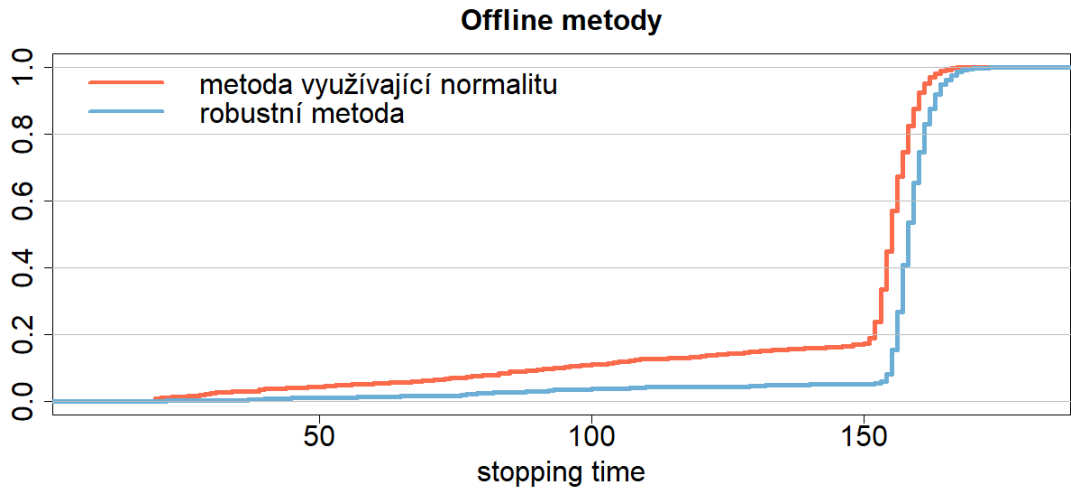
Obrázek 5.7: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, pokud nedošlo ke změně střední hodnoty.



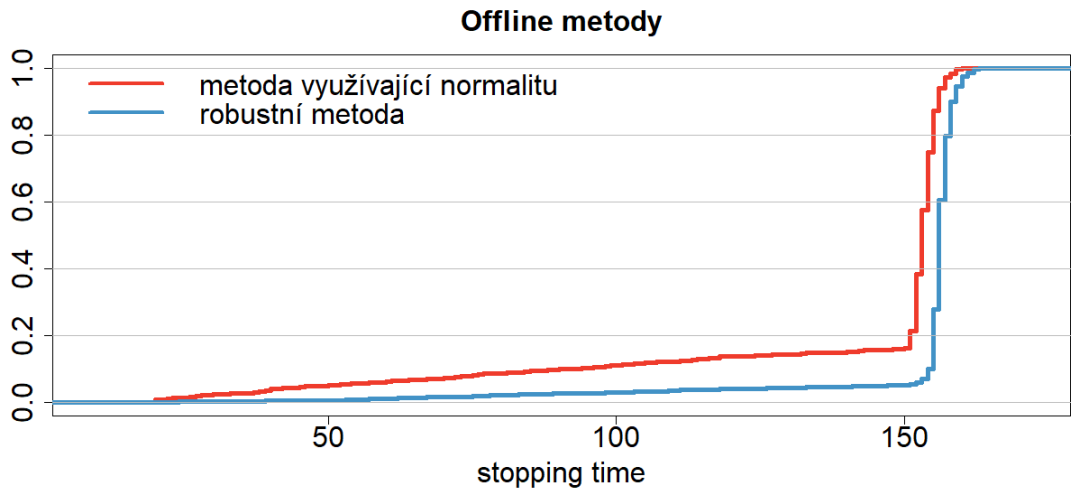
Obrázek 5.8: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o $\frac{1}{2}\sigma$.



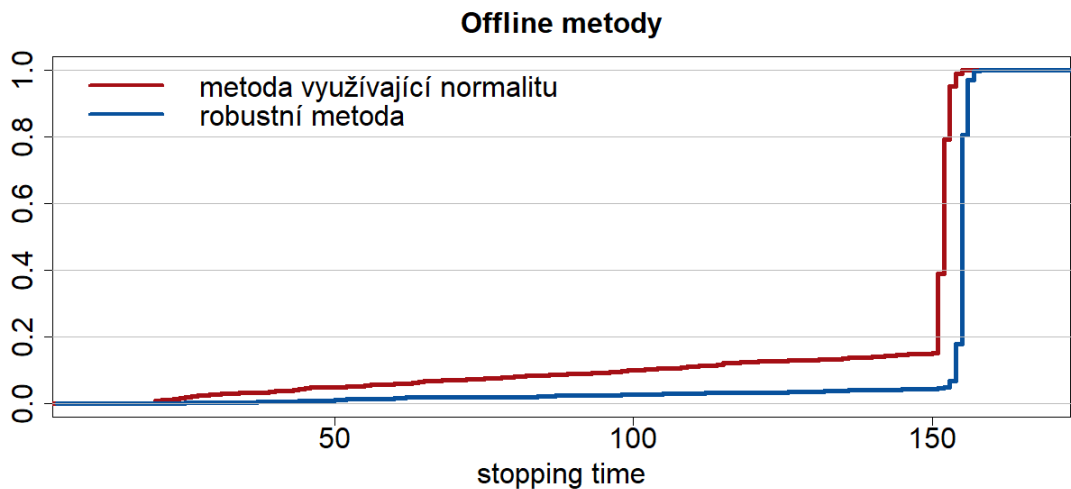
Obrázek 5.9: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o σ .



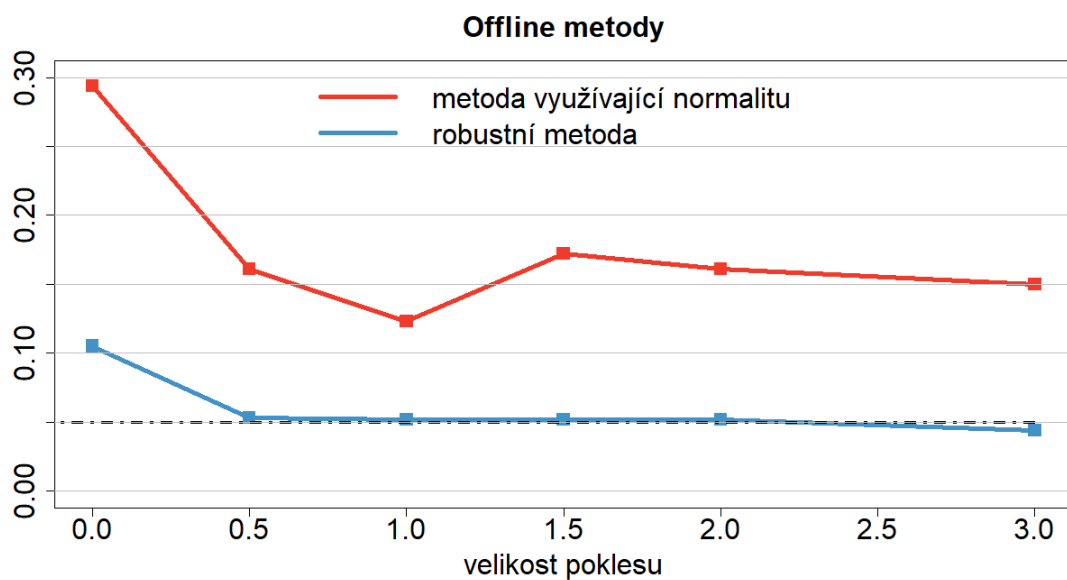
Obrázek 5.10: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o $\frac{3}{2}\sigma$.



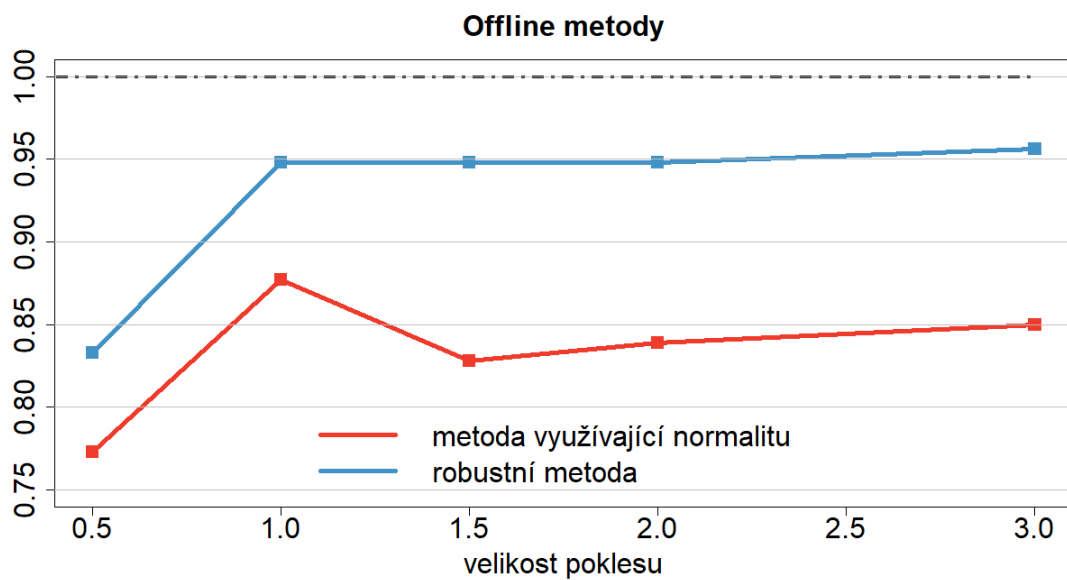
Obrázek 5.11: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o 2σ .



Obrázek 5.12: Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o 3σ .



Obrázek 5.13: Pravděpodobnost chyby předčasné detekce obou offline metod v závislosti na posunu ve střední hodnotě.



Obrázek 5.14: Síla obou offline metod v závislosti na posunu ve střední hodnotě.

5.2.2 Výsledky online detekce

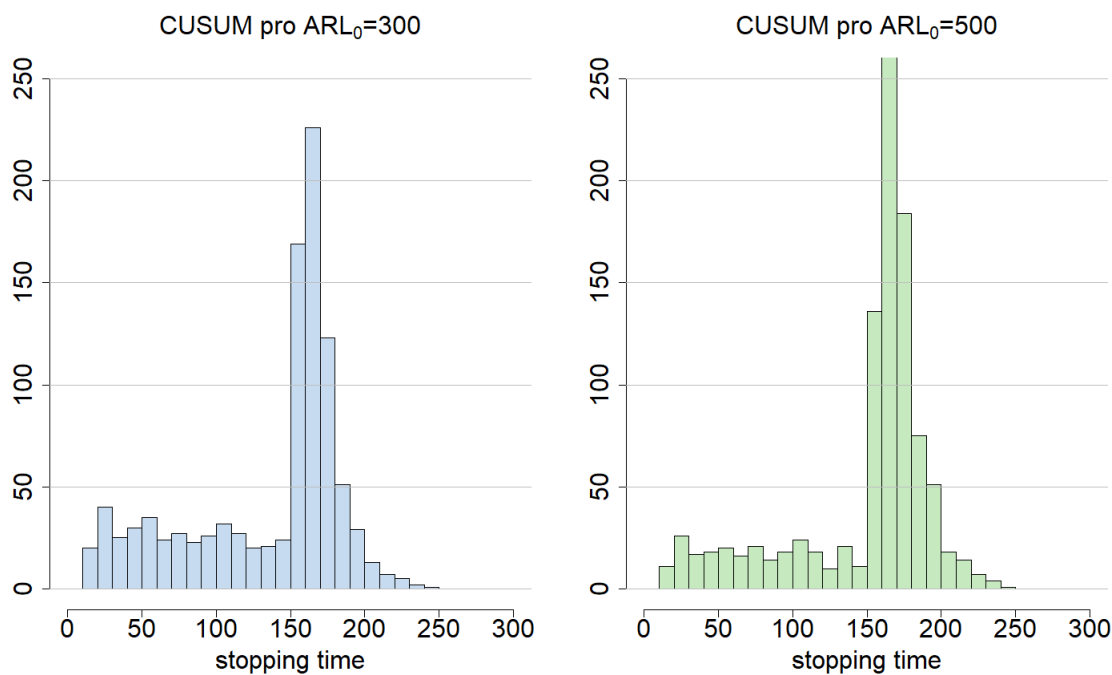
Opět pro popis rozdělení časů zastavení metodou CUSUM nejprve využijeme histogramy, viz obrázky 5.15 až 5.24. Nyní nás bude zajímat závislost času zastavení nejen na velikosti detekovaného poklesu, ale také na délce střední doby mezi falešnými poplachy. Střední dobu do zastavení, je-li proces nastaven na hladině μ , značíme v kapitole 4 $ARL(\mu)$, resp. $ARL_+(\mu)$ v případě „kladné“ jednostranné alternativy a $ARL_-(\mu)$ v případě jednostranné alternativy „záporné.“ Pro stanovení hladiny pro detekci poklesu je klíčové nastavit střední dobu do zastavení za platnosti nulové hypotézy, v našem případě $ARL_-(500)$. V popisících obrázků budeme tuto hodnotu označovat jednodušeji zkratkou ARL_0 .

Již na první pohled si můžeme všimnout několika zjevných trendů. Nejprve se pojdme podívat, jak časy zastavení závisejí na ARL_0 . Nehledě na velikost detekovaného posunu, počet předčasných detekcí klesá s prodlužující se ARL_0 . To je v souladu s přirozeným úsudkem. Totiž, ARL_0 je střední doba do zastavení za platnosti nulové hypotézy a předčasná detekce je detekce, zatímco výrobní dávka ještě pokračuje dle hypotézy. Protože při našich simulacích dojde ke změně již v čase 150 a nejkratší ARL_0 uvažujeme 300, tak by k předčasné detekci mělo docházet jen zřídka, a tím méně, čím delší je ARL_0 . Na druhou stranu, s vyšší ARL_0 zaznamenáváme delší doby do zastavení, tj. větší zpoždění. Nicméně to je patrné, pouze pokud detekujeme menší poklesy, a ani v tomto případě nijak závažné. Pro poklesy o $\frac{3}{2}\sigma$ a větší z hlediska zpoždění již prakticky nezáleží na délce ARL_0 ; a při $ARL_0 = 1000$ proběhne drtivá většina detekce mezi časy 150 a 160.

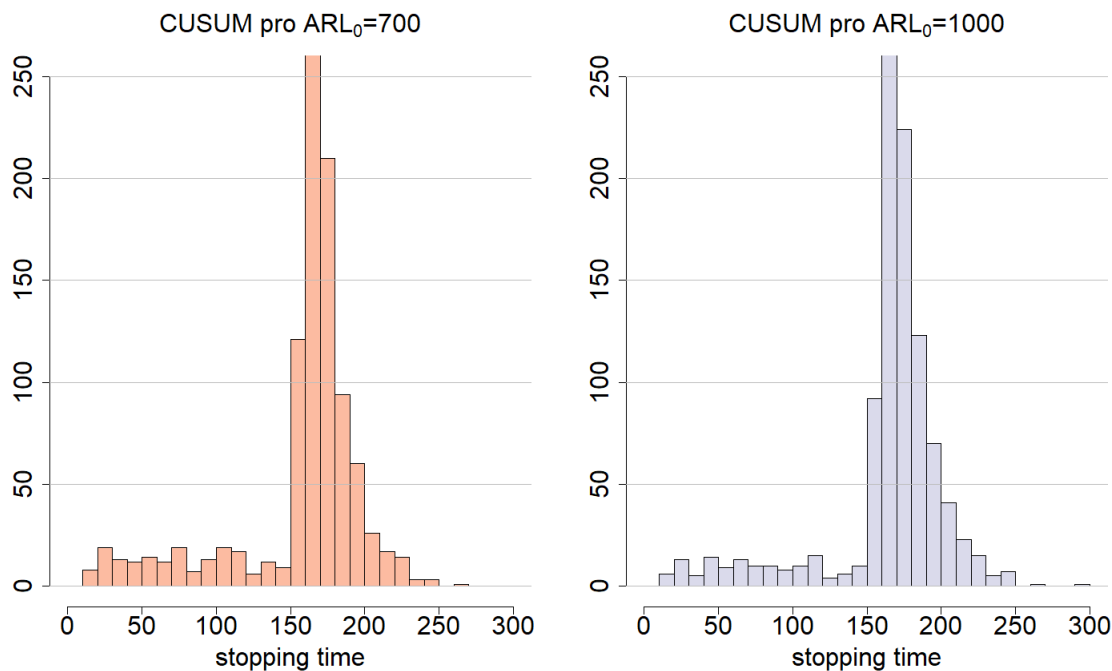
Nyní je otázkou, jak reaguje metoda CUSUM vůči velikosti detekovaného posunu při fixní hodnotě ARL_0 . Počty falešných poplachů teď není rozumné porovnávat z histogramů, neboť pro různé posuny jsou různě naškálované. Místo toho se odrazíme od grafu 5.30 závislosti relativních počtů falešných poplachů na velikosti posunu pro každou volbu ARL_0 . Graf přináší poměrně zajímavé výsledky. Pro každou volbu ARL_0 je nejmenší počet falešných poplachů při detekci poklesu o $\frac{3}{2}\sigma$, potom se zvyšuje a úplně největší je při detekci poklesu o 3σ .

Obrázky empirických distribučních funkcí netřeba příliš komentovat, pouze v jiné podobě přinášejí ty informace, které jsme již komentovali na základě histogramů.

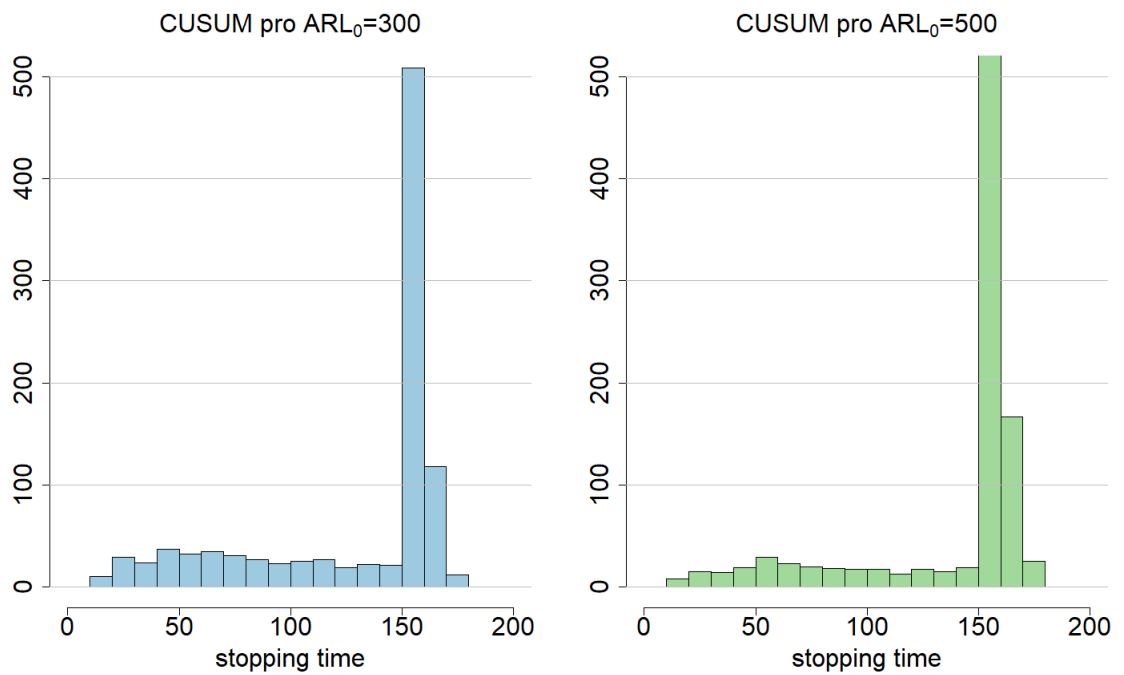
Shrnutí budiž následující. Obecně kratší ARL_0 vede k rychlejší detekci, ale zároveň k většímu počtu falešných poplachů. Z praktického hlediska však zpoždění do odhalení změny není ani při velké ARL_0 nijak zásadní, obzvláště pak detekujeme-li větší posuny. Např. při detekci posunu o 2σ nebo 3σ se drtivá většina správné detekce odehrává mezi časy 150 a 160, a to pro všechny volby ARL_0 . Na zpoždění tedy větší ARL_0 nemá zásadně negativní dopad. Na co ale hodnota ARL_0 má závažný dopad, je počet falešných poplachů. Při $ARL_0 = 300$ dokonce CUSUM metoda detekuje předčasně zhruba ve $\frac{2}{5}$ případů, jak si můžeme všimnout z obrázku 5.30. Ani pro $ARL_0 = 500$ není situace příznivá. Proto doporučujeme nastavit střední dobu mezi falešnými poplachy alespoň na 700, ideálně 1000. Při $ARL_0 = 1000$ je relativní počet falešných poplachů ještě o cca 0,05 nižší než při $ARL_0 = 700$ a zároveň po čase změny detekuje metoda CUSUM nastavená na $ARL_0 = 1000$ téměř stejně rychle.



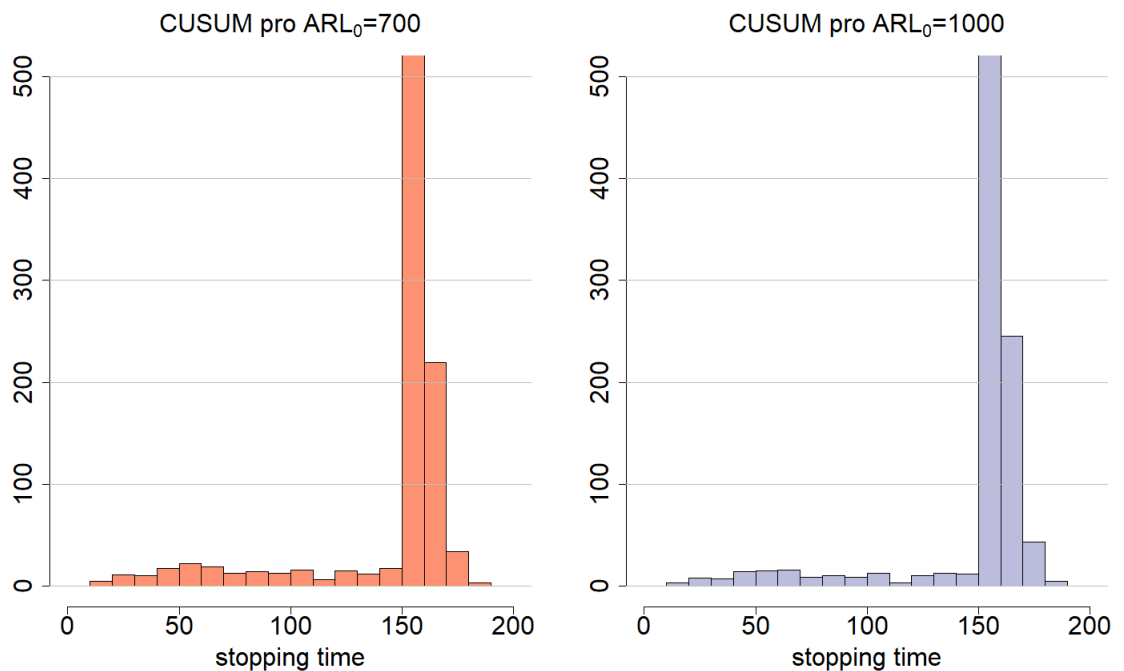
Obrázek 5.15: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{1}{2}\sigma$ při volbách ARL_0 300 (vlevo) a 500 (vpravo).



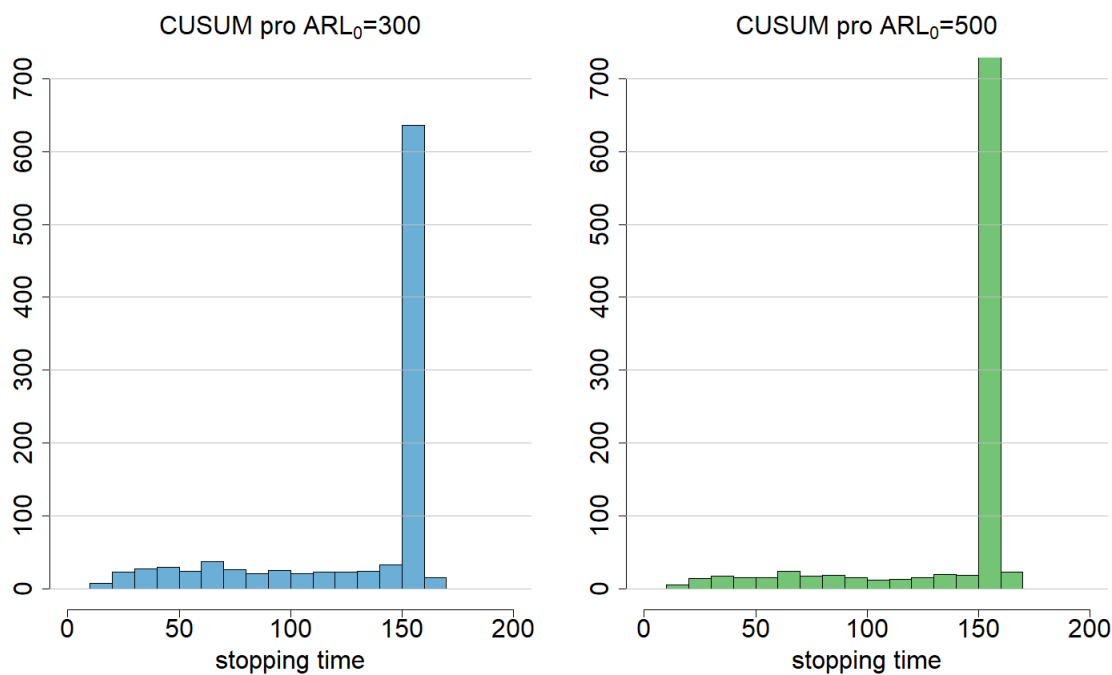
Obrázek 5.16: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{1}{2}\sigma$ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).



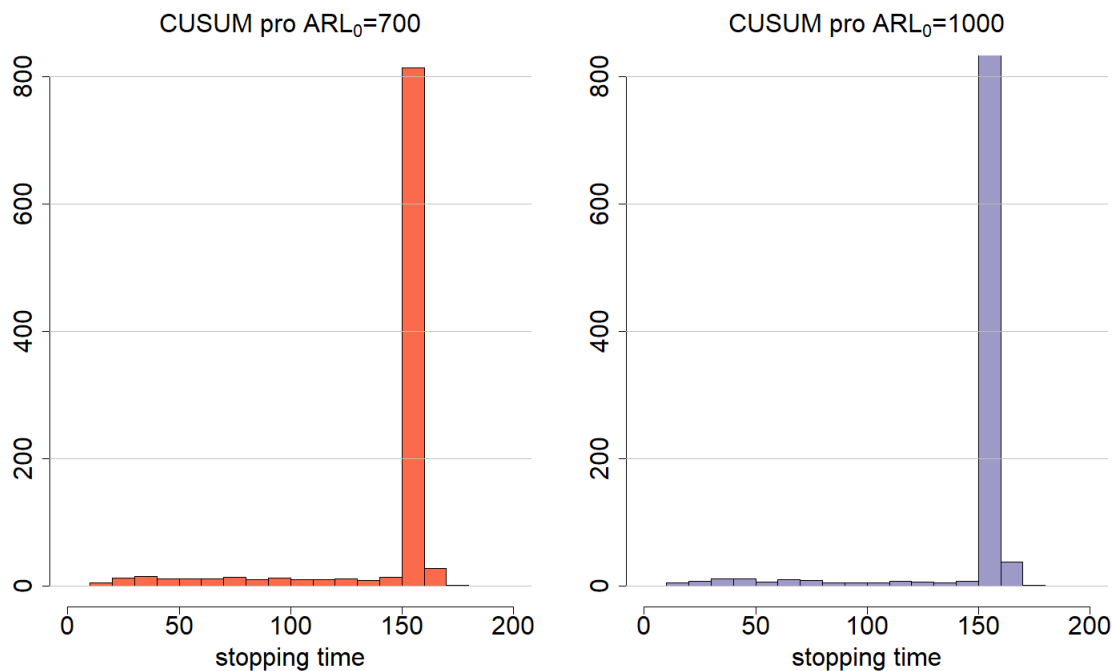
Obrázek 5.17: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty μ při volbách ARL_0 300 (vlevo) a 500 (vpravo).



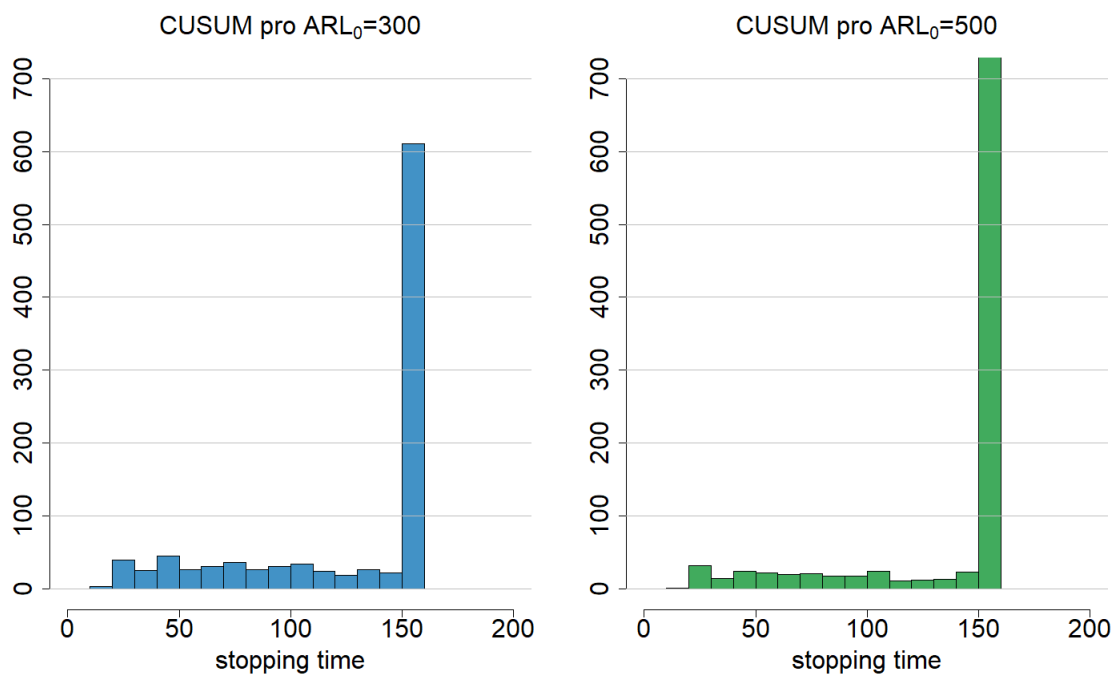
Obrázek 5.18: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty μ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).



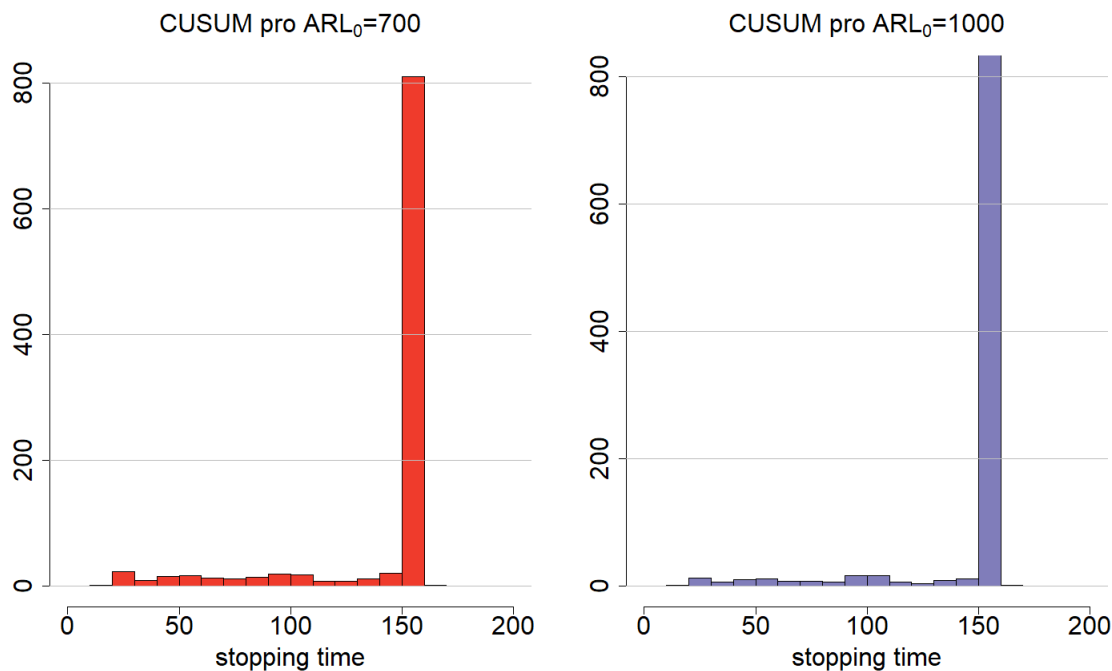
Obrázek 5.19: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{3}{2}\sigma$ při volbách ARL_0 300 (vlevo) a 500 (vpravo).



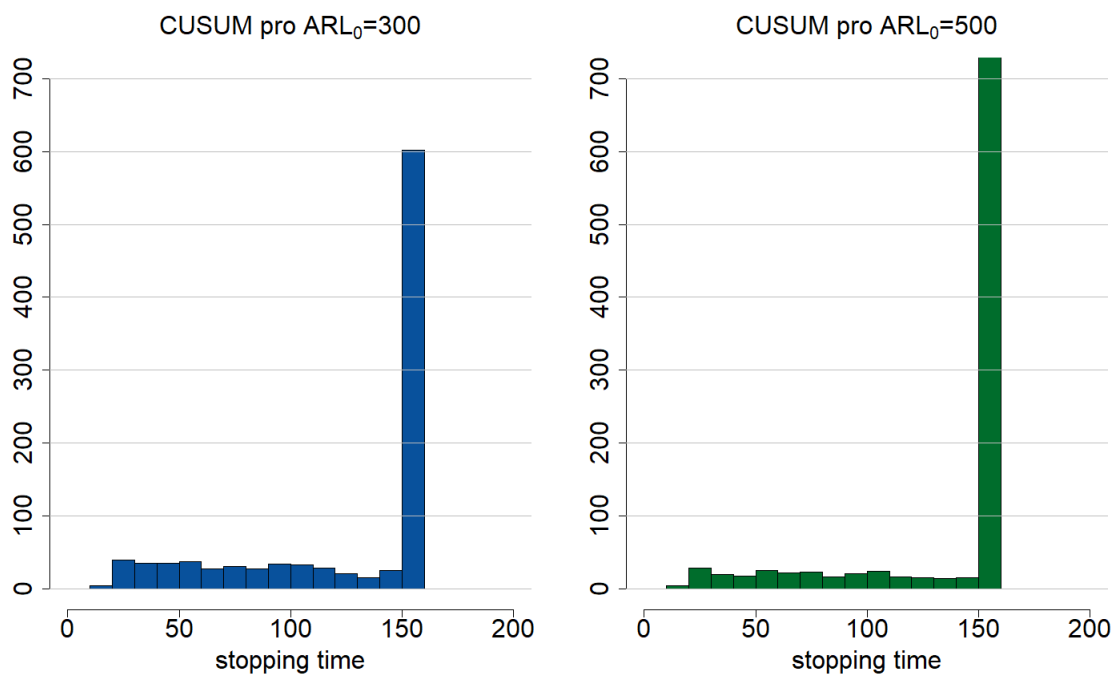
Obrázek 5.20: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{3}{2}\sigma$ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).



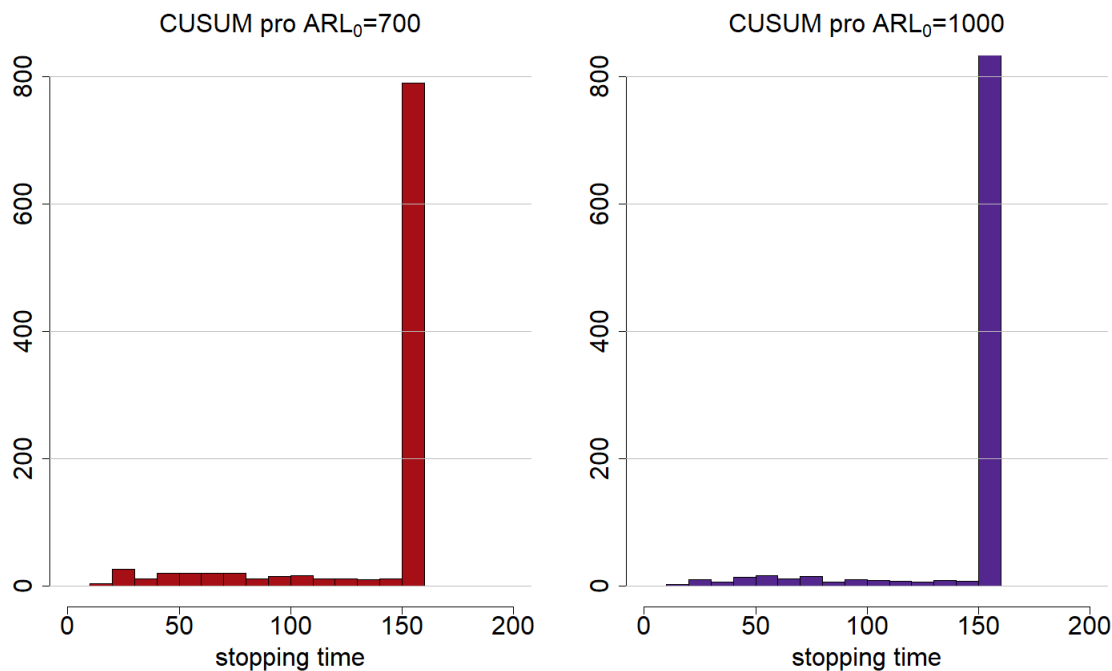
Obrázek 5.21: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 2σ při volbách ARL_0 300 (vlevo) a 500 (vpravo).



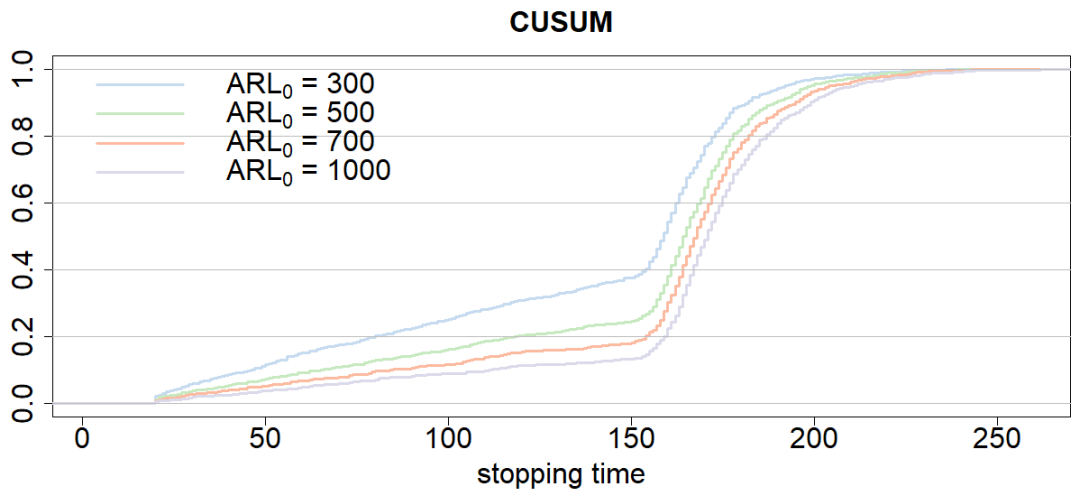
Obrázek 5.22: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 2σ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).



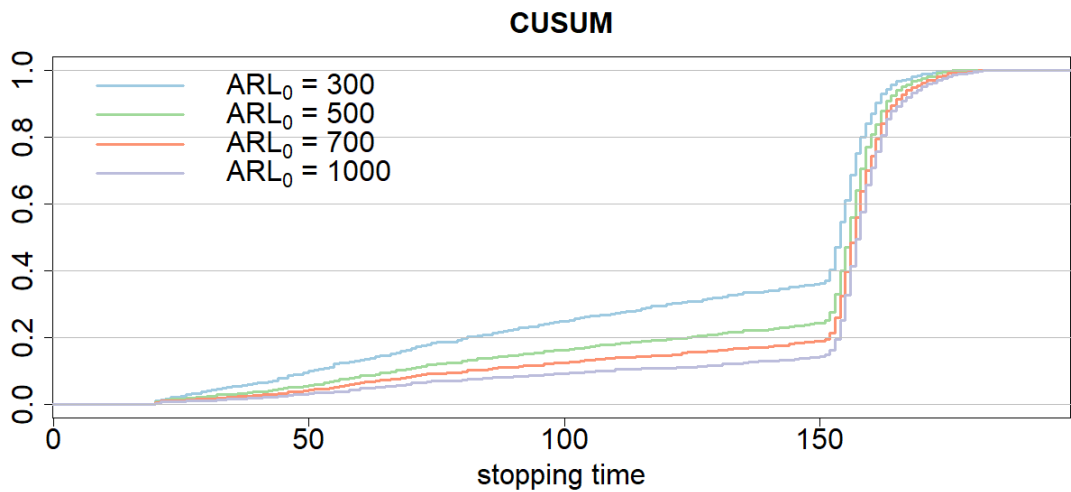
Obrázek 5.23: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 3σ při volbách ARL_0 300 (vlevo) a 500 (vpravo).



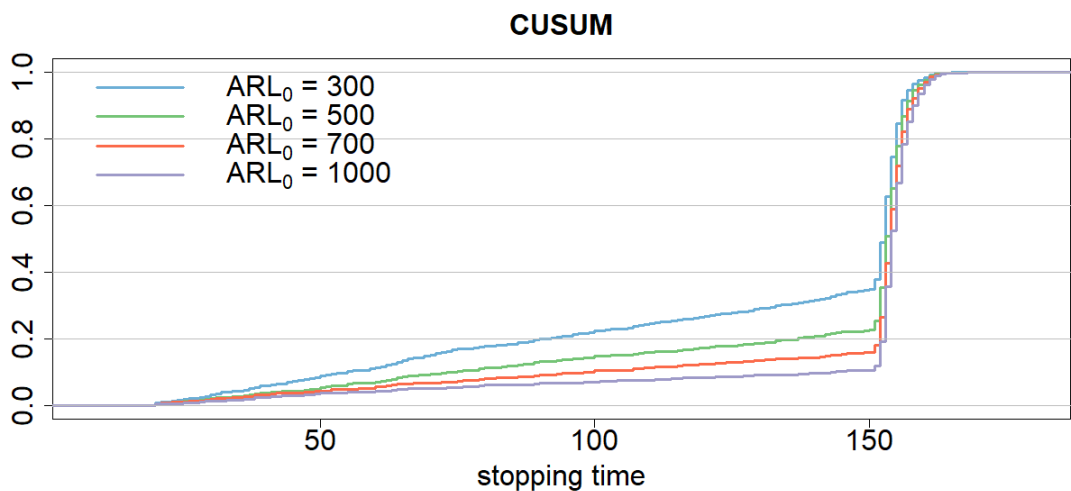
Obrázek 5.24: Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 3σ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).



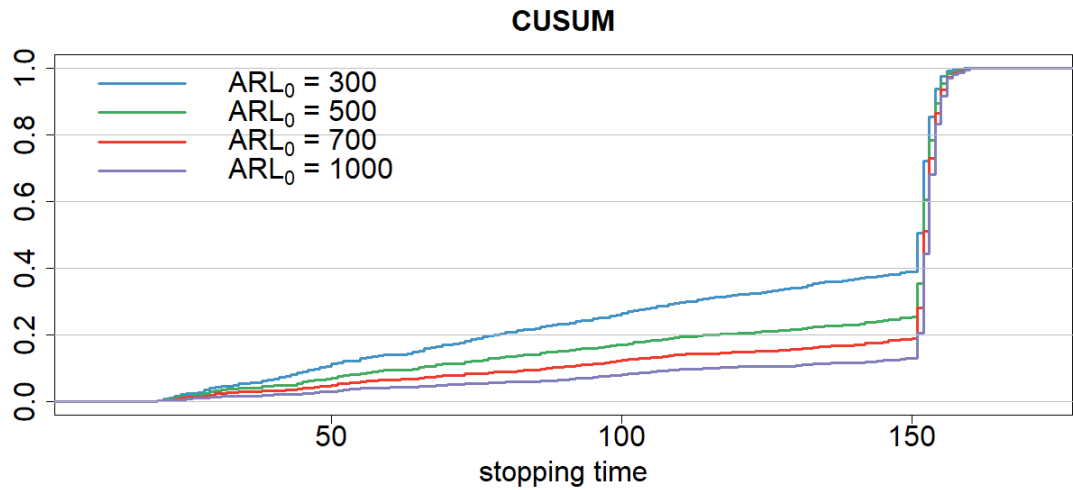
Obrázek 5.25: Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o $\frac{1}{2}\sigma$ pro různé volby ARL_0 .



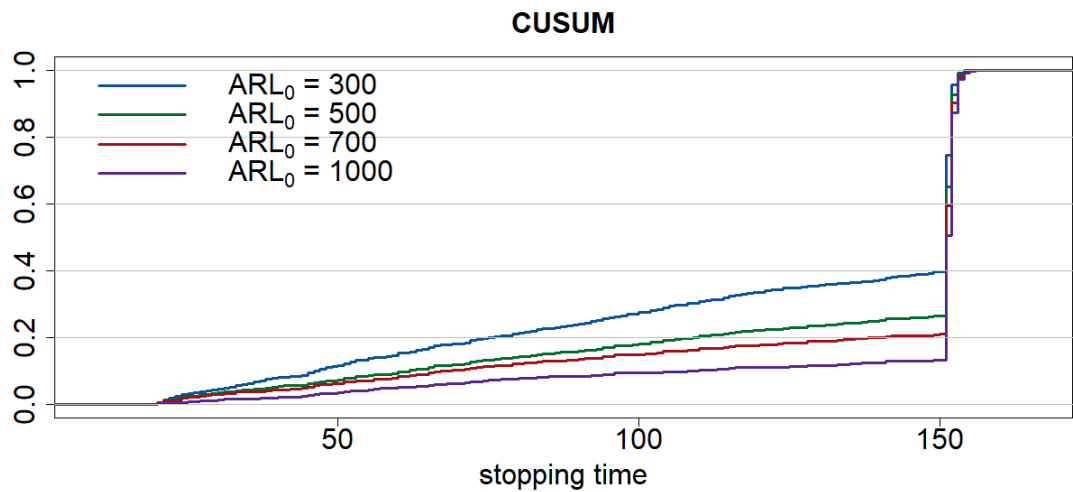
Obrázek 5.26: Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o σ pro různé volby ARL_0 .



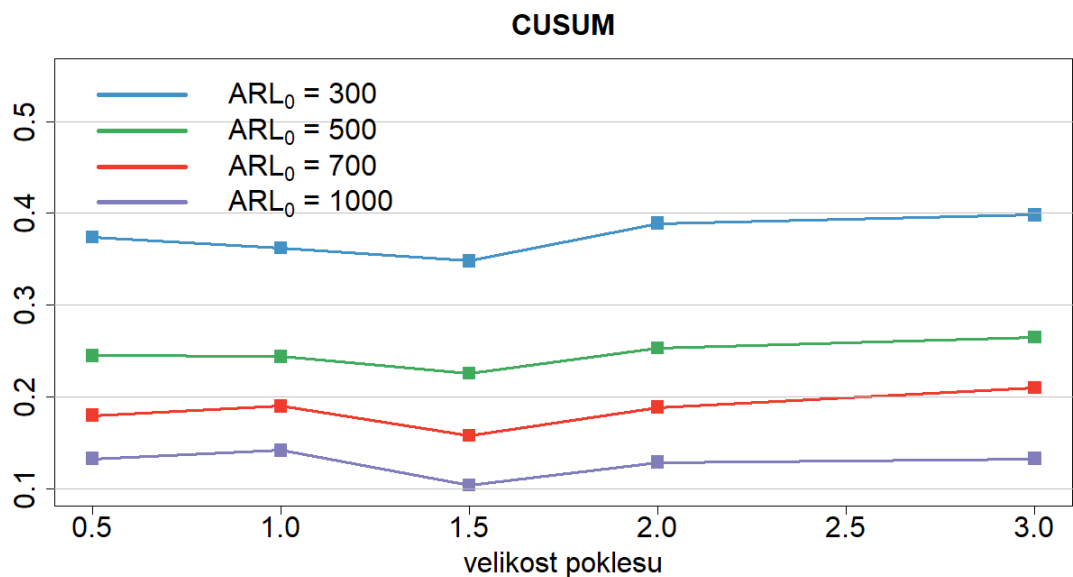
Obrázek 5.27: Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o $\frac{3}{2}\sigma$ pro různé volby ARL_0 .



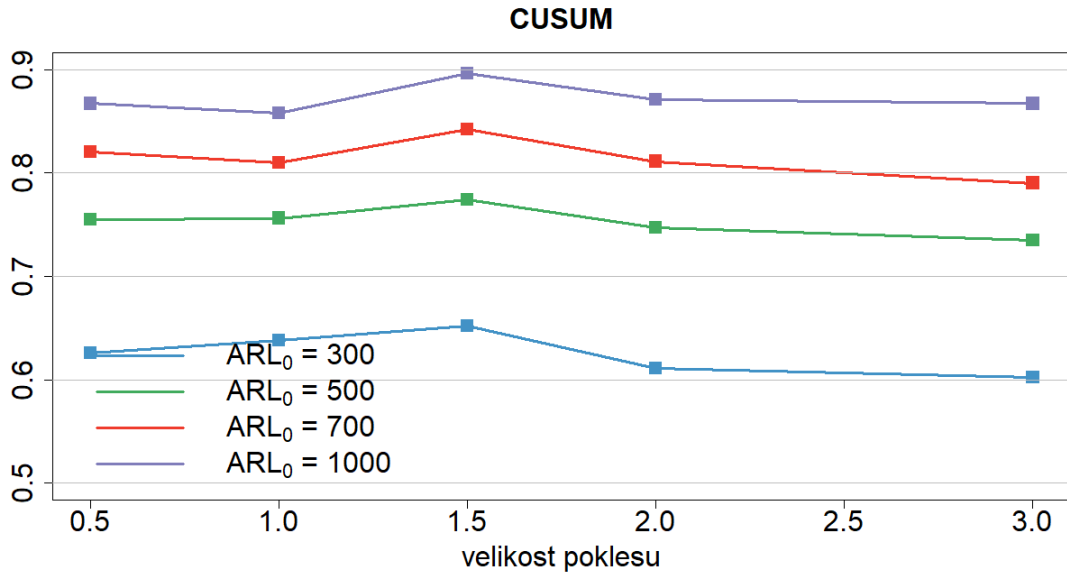
Obrázek 5.28: Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o 2σ pro různé volby ARL_0 .



Obrázek 5.29: Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o 3σ pro různé volby ARL_0 .



Obrázek 5.30: Relativní počty falešných poplachů vyvolaných CUSUM metodou v závislosti na velikosti detekovaného poklesu střední hodnoty pro různé volby ARL_0 .



Obrázek 5.31: Relativní počty správných detekcí CUSUM metodou v závislosti na velikosti detekovaného poklesu střední hodnoty pro různé volby ARL_0 .

5.2.3 Porovnání offline metod a metody CUSUM

Jak již zmiňujeme v úvodu této práce, za optimálnější považujeme tu proceduru, která má co nejmenší zpoždění zastavení a zároveň co nejmenší počet předčasných detekcí (falešných poplachů). Za nejefektivnější nastavení pro CUSUM metodu považujeme $ARL_0 = 1000$, výsledky tohoto nastavení tedy porovnáme s offline metodami.

Co se týče počtu předčasných detekcí, porovnejme obrázky 5.13 a 5.30. Zde jednoznačně vyhrává robustní Wilcoxonova metoda, která předčasně detekuje jen v 5 % případů. Následuje CUSUM metoda, která vyvolá falešný poplach zhruba v 10 % případů a na posledním místě je metoda využívající normality dat, tam je pravděpodobnost předčasné detekce asi 0,15. Nicméně se zdá, že počínaje velikostí posunu $\frac{3}{2}\sigma$ začíná počet falešných poplachů metody CUSUM narůstat, zatímco pro metodu normality počet předčasných detekcí začíná vykazovat tendenci klesající.

Zbývá srovnat metody z hlediska zpoždění zastavení. Zde se situace liší podle velikosti poklesu střední hodnoty. Detekujeme-li větší posuny o 2 nebo 3 σ , pak drtivá většina správné detekce proběhne do času 160 pro všechny tři metody. Zde tedy rozdíl nepozorujeme. Zásadní rozdíl nevidíme ani při posunu o $\frac{3}{2}\sigma$. Detekujeme-li malé posuny o σ a $\frac{1}{2}\sigma$, situace už se ale liší. Většina správné detekce posunu o $\frac{1}{2}\sigma$ metodou CUSUM se odehrává mezi časy 160 a 190. V případě obou offline metod ale s nezanedbatelnou pravděpodobností (cca 0,31 pro metodu normality a 0,37 pro Wilcoxonovu metodu) dojde k detekci mezi časy 190 až 250. Obě offline metody jsou tedy pomalejší.

Při detekci změny o σ metoda CUSUM detekuje v polovině případů mezi časy 150-160, ve čtvrtině případů mezi časy 160-170. Metoda normality zastaví mezi časy 150-160 kolem 400 výrobních dávek, 350 výrobních dávek pak mezi časy 160-170. Konečně Wilcoxonova metoda detekuje mezi časy 150=160 asi ve čtvrt-

tině případů, v polovině případů pak mezi časy 160-170 a ještě asi 15 % všech výrobních dávek zastaví mezi časy 170 a 180. Je tedy pomalejší, ale ne významně.

Pro větší posuny (alespoň $\frac{3}{2}\sigma$) tedy doporučujeme offline Wilcoxonovu metodu. Pro menší posuny by volba metody závisela na uvážení výrobce. Je třeba pamatovat na to, že Wilcoxonova metoda má významně nižší podíl předčasných detekcí, nicméně vůči malým posunům reaguje pomaleji než zbývající dvě metody. Pokud by výrobce považoval za prioritní rychlost do odhalení změny, pak bychom volili mezi metodou normality a CUSUM metodou. Při posunu o σ je pravděpodobnost předčasné detekce mírně nižší pro metodu normality, zatímco obě metody správně detekují převážně mezi časy 150-170. Zde bychom tedy mohli zvolit metodu normality. Při posunu o $\frac{1}{2}\sigma$ je naopak pravděpodobnost předčasně detekce nižší pro CUSUM metodu a zároveň je tato metoda i rychlejší.

Závěr

V této diplomové práci jsme shrnuli standardní postupy pro detekci změny ve vývoji náhodných procesů. Co se týká offline detekce, věnovali jsme se jak změně parametru posunutí, tak přechodu konstanty v přímku. V obou případech jsme podrobně odvodili příslušné testové statistiky a uvedli jsme jejich limitní rozdělení. Přitom ale bylo třeba dát pozor na to, zda známe počáteční střední hodnotu procesu, či nikoli. Pro detekci změny parametru posunutí jsme také navrhli postup, jak detekovat v neparametrickém modelu.

Hlavním přínosem této diplomové práce je ale kapitola přinášející výsledky simulací. Zde jsme došli k několika závěrům.

Zaměříme-li se nejprve na postupy offline detekce, Wilcoxonovu metodu a metodu pro normální data, dojdeme k tomu, že metoda normality sice detekuje rychleji, ale má podstatně vyšší pravděpodobnost předčasné detekce. Očekáváme-li větší posuny ve střední hodnotě (alespoň $\frac{3}{2}\sigma$), pak doporučujeme použít Wilcoxonovu metodu. Pro menší posuny záleží na výrobci, zda upřednostní rychlejší metodu, nebo metodu s nižší pravděpodobností předčasné detekce. Nicméně i v případě malých posunů je z hlediska předčasné detekce metoda normality podstatně chybovější.

Co se týká online metody, došli jsme k tomu, že ideální nastavení ARL je kolem 1000. Srovnáme-li ji s offline metodami, dojdeme k následujícím doporučením. Pro posuny o alespoň $\frac{3}{2}\sigma$ volíme Wilcoxonovu metodu. Při menších posunech Wilcoxonova metoda reaguje pomaleji než zbylé dvě metody, nicméně pravděpodobnost předčasné detekce si stále drží podstatně nižší. Pokud bychom v tomto případě měli volit mezi metodou normality a CUSUM metodou, tak pro posun σ volíme metodu normality a pro posun $\frac{1}{2}\sigma$ metodu CUSUM.

Prostor pro porovnání postupů statistické kontroly samozřejmě nebyl vyčerpán. Zatímco robustní metody mají tu výhodu, že jde o metody neparametrické, metody statistik maximálního typu byly primárně navrženy pro normální data, a v jiných modelech by mohly selhávat. Předmětem dalšího výzkumu by tedy mohla být chybovost těchto metod v jiných třídách spojitých rozdělení a porovnání s robustními metodami.

Appendix

V appendixu jsou shrnuty vybrané definice a výsledky z přednášek, abychom se na ně mohli snáze odkazovat.

Definice 1 (Odhad metodou nejmenších čtverců). *Nechť platí lineární regresní model $\mathbf{Y}|\mathbf{X} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. Potom odhad $\hat{\boldsymbol{\beta}}$ parametru $\boldsymbol{\beta}$ metodou nejmenších čtverců (LSE) definujeme jako*

$$\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

Věta 3 (Odhad metodou nejmenších čtverců). *Budiž $\mathbf{Y}|\mathbf{X} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2)$ a necht matice \mathbf{X} má plnou hodnost. Pak LSE je tvaru*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Věta 4 (Gauss-Markovova). *Nechť platí $\mathbf{Y}|\mathbf{X} \sim (\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. Potom vektor $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ tzv. vyrovnaných hodnot (častěji ale nepřekládáme a hovoříme o „fitted values“) je nejlepším lineárním nestranným odhadem střední hodnoty $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Nejlepším v tom smyslu, že pro libovolný lineární nestranný odhad $\tilde{\mathbf{Y}}$ platí*

$$\operatorname{var}[\tilde{\mathbf{Y}}|\mathbf{X}] - \operatorname{var}[\hat{\mathbf{Y}}|\mathbf{X}]$$

je pozitivně semidefinitní matice. Mimo jiné,

$$\operatorname{var}[\hat{\mathbf{Y}}|\mathbf{X}] = \sigma^2 \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top.$$

Poznámka (Momenty LSE). Platí: $\mathbb{E}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta}$, $\operatorname{var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

Věta 5. *Testy o regresních koeficientech (Anděl (1985), věta 8 str. 101.) Necht $\mathbf{Y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$. Potom testová statistika*

$$t_j = \frac{\hat{\beta}_j - \beta_j}{MS_e v_{jj}}$$

má t_{n-k} rozdělení, kde k je hodnost matice \mathbf{X} , $MS_e = \frac{1}{n-k} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ a v_{jj} je j -tý diagonální prvek matice $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Seznam použité literatury

- ANDĚL, J. (2005). *Základy matematické statistiky*. Matfyzpress, Praha.
- ANDĚL, J. (1985). *Matematická statistika*. Státní nakladatelství technické literatury, Praha.
- ANTOCH, J., HUŠKOVÁ, M. a JARUŠKOVÁ, D. (2001). Off-line statistical process control. *Multivariate Total Quality Control: Foundation and Recent Advances*, pages 1–86.
- CSÖRGÖ, MIKLÓS A HORVÁTH, L. (1997). *Limit theorems in change-point analysis*. J. Wiley, New York.
- JARUŠKOVÁ, D. (1998). Testing appearance of linear trend. *Journal of Statistical Planning and Inference*, **70**(2), 263–276.
- SIEGMUND, D. (1985). *Sequential Analysis: Tests and confidence intervals*. Springer.
- STEIN, C. (prosinec 1946). A note on cumulative sums. *The Annals of Mathematical Statistics*, **17**(4), 498–499.
- WALD, A. (1947). *Sequential Analysis*. John Wiley and Sons, New York.
- WETHERILL, G. B. a BROWN, D. W. (1991). *Statistical Process Control: Theory and Practice*. Chapman and Hall, London.

Seznam obrázků

2.1	Model změny skokem; chyby jsou normálně rozdělené	6
2.2	Model postupné změny se spojitým přechodem, tzv. hokejka; chyby jsou normálně rozdělené	6
2.3	Model postupné změny přímky v přímku; chyby jsou normálně rozdělené	7
2.4	Model postupné změny přímky v přímku se skokem; chyby jsou normálně rozdělené	7
3.1	Porovnání dvou uvažovaných odhadů rozptylu: $S^2 = 7.37$, $S_{*m}^2 = 1.61$, přičemž $\sigma^2 = 1$	11
3.2	Vývoj posloupnosti absolutních hodnot statistik $T_m = \frac{1}{\sigma} \sqrt{\frac{n}{m(n-m)}} \sum_{i=1}^m (Y_i - \bar{Y}_n)$ pro čas změny $m = 50$. Generovány byly dva výběry z normálních rozdělení: $Y_1, \dots, Y_{50} \sim \mathcal{N}(5, 1)$, $Y_{51}, \dots, Y_{100} \sim \mathcal{N}(10, 1)$	17
5.1	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), když ke změně nedošlo.	36
5.2	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o $\frac{1}{2}\sigma$	36
5.3	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o σ	37
5.4	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o $\frac{3}{2}\sigma$	37
5.5	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o 2σ	38
5.6	Histogramy časů zastavení stanovených metodou normality dat (vlevo) a robustní metodou (vpravo), došlo-li k poklesu střední hodnoty o 3σ	38
5.7	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, pokud nedošlo ke změně střední hodnoty.	39
5.8	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o $\frac{1}{2}\sigma$	39
5.9	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o σ	39
5.10	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o $\frac{3}{2}\sigma$	40

5.11	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o 2σ	40
5.12	Empirické distribuční funkce časů zastavení stanovených metodou normality dat a robustní metodou, došlo-li k poklesu střední hodnoty o 3σ	40
5.13	Pravděpodobnost chyby předčasné detekce obou offline metod v závislosti na posunu ve střední hodnotě.	41
5.14	Síla obou offline metod v závislosti na posunu ve střední hodnotě.	41
5.15	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{1}{2}\sigma$ při volbách ARL_0 300 (vlevo) a 500 (vpravo).	43
5.16	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{1}{2}\sigma$ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).	43
5.17	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o σ při volbách ARL_0 300 (vlevo) a 500 (vpravo).	44
5.18	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o σ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).	44
5.19	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{3}{2}\sigma$ při volbách ARL_0 300 (vlevo) a 500 (vpravo).	45
5.20	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o $\frac{3}{2}\sigma$ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).	45
5.21	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 2σ při volbách ARL_0 300 (vlevo) a 500 (vpravo).	46
5.22	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 2σ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).	46
5.23	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 3σ při volbách ARL_0 300 (vlevo) a 500 (vpravo).	47
5.24	Histogramy časů zastavení stanovených metodou CUSUM pro detekci poklesu střední hodnoty o 3σ při volbách ARL_0 700 (vlevo) a 1000 (vpravo).	47
5.25	Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o $\frac{1}{2}\sigma$ pro různé volby ARL_0	48
5.26	Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o σ pro různé volby ARL_0	48
5.27	Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o $\frac{3}{2}\sigma$ pro různé volby ARL_0	48
5.28	Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o 2σ pro různé volby ARL_0	49

5.29	Empirické distribuční funkce časů zastavení stanovených CUSUM metodou pro detekci poklesu o 3σ pro různé volby ARL_0	49
5.30	Relativní počty falešných poplachů vyvolaných CUSUM metodou v závislosti na velikosti detekovaného poklesu střední hodnoty pro různé volby ARL_0	49
5.31	Relativní počty správných detekcí CUSUM metodou v závislosti na velikosti detekovaného poklesu střední hodnoty pro různé volby ARL_0	50