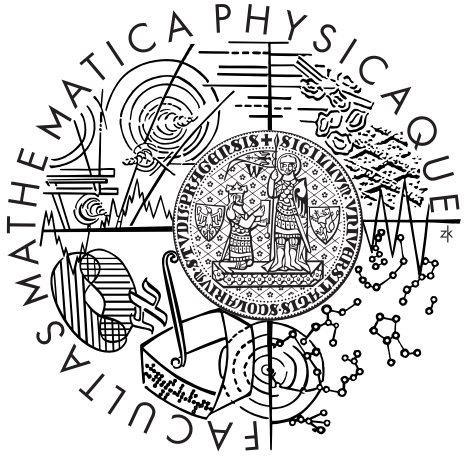
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Stanislav Svoboda

Odhady metodou maximálního součinu mezer

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D.
Studijní Program: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) 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

Zde bych rád poděkoval doc. Ing. Marku Omelkovi, Ph.D. za ochotu, vstřícnost a veškerý věnovaný čas. Také bych velice rád poděkoval své úžasné rodině za neustálou podporu. Tátovi za jeho trpělivost, snaživost a vytvoření ideálních podmínek pro studium. Mamce za její nedocenitelnou pomoc s každou otázkou. Své přítelkyni za to, že tu pro mě vždy byla. Za její optimismus a ujištění, že to studium nemůže být tak těžké, jak říkám, protože by se mi nemohlo dařit, tak jak se mi daří.

Název práce: Odhady metodou maximálního součinu mezer

Autor: Stanislav Svoboda

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V práci se zabýváme odhadem metodou maximálního součinu mezer (MPS). Nejdříve si stručně připomeneme metodu maximální věrohodnosti (ML). Poté si podrobně vysvětlíme metodu MPS. Nakonec si v konkrétních případech ukážeme, jak se odvozují odhady metodou MPS a porovnáme je s odhady metodou ML.

Klíčová slova: maximální věrohodnost, metoda maximálních mezer, odhady

Title: Maximum product spacings estimation

Author: Stanislav Svoboda

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this thesis we study the maximum product spacing (MPS) estimation. First we shortly introduce the maximum likelihood (ML) method. Then, we explain the MPS method in detail. Finally, in specific cases, we demonstrate how to derive the MPS estimation and compare it with the ML estimation.

Keywords: estimators, maximum likelihood, maximum product spacings

Obsah

Úvod

V reálném světě často nasbíráme data a z vlastností problému předpokládáme nějaký model. Následně se snažíme zjistit parametry daného modelu. Jedna z nejrozšířenějších metod na odhadování neznámých parametrů je metoda maximální věrohodnosti (ML). Existují však známé parametrické modely, kdy tato metoda selže, například posunuté log-normální rozdělení, viz motivační příklad 1.2. Reakcí na tento problém vznikla metoda maximálního součinu mezer (MPS). Ta zachovává důležité vlastnosti ML a zároveň nemá stejné problémy, viz kapitola 1.3.

V této práci si v první kapitole představíme metodu ML. Poté si v motivačním příkladu odvodíme odhad metodou ML, který není konzistentní. Nakonec si podrobně vysvětlíme metodu MPS obecně.

V druhé kapitole si v konkrétních případech ukážeme, jak se odvozují odhady metodou MPS. Poté si odvodíme odhad metodou ML a budeme porovnávat získané odhady.

1. Metoda maximálního součinu mezer

Nejdříve si představíme metodu maximální věrohodnosti (ML), abychom poté uviděli podobnost s metodou maximálního součinu mezer (MPS). Dále si na motivačním příkladu ukážeme odhad metodou ML, který není konzistentní. Tento a podobné příklady byly motivací pro vznik metody MPS.

1.1 Metoda maximální věrohodnosti

V této práci budeme pracovat pouze s náhodným výběrem $X_1, \ldots, X_n$ z rozdělení s hustotou $f(x; \boldsymbol{\theta})$ vzhledem k jednorozměrné Lebesgueově míře, kde $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$ je neznámý parametr, který se snažíme odhadnout.

Odhad metodou maximální věrohodnosti (ML) je libovolný bod $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, který maximalizuje funkci

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i; \boldsymbol{\theta}).$$

Funkci $L_n(\boldsymbol{\theta})$ nazýváme věrohodnost.

Protože $f(x; \boldsymbol{\theta})$ je hustota a logaritmus je rostoucí funkce na $(0, \infty)$, tak argument maxima $f(x; \boldsymbol{\theta})$ je stejný jako u $\log f(x; \boldsymbol{\theta})$. Tento poznatek nám umožňuje při hledání odhadu metodou ML maximalizovat logaritmus věrohodnosti

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(X_i; \boldsymbol{\theta})$$

místo věrohodnosti. Funkci $\ell_n(\boldsymbol{\theta})$ nazýváme log-věrohodnost.

Pokud je věrohodnost diferencovatelná vzhledem k $\boldsymbol{\theta}$, víme, že věrohodnost má v bodě libovolného odhadu metodou ML nulové parciální derivace, neboli každý odhad metodou ML musí být řešením tzv. systému věrohodnostních rovnic

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_1} = 0, \ \ldots \ , \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_p} = 0, \quad \text{kde } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p).$$

Toto je nejběžnější způsob hledání odhadu metodou ML.

Většina našich příkladů bude mít nosič závislý na parametru, a tedy naše funkce nebude diferencovatelná vzhledem k $\boldsymbol{\theta}$. V takovém případě budeme maximalizovat věrohodnost, protože se nám bude lépe maximalizovat než log-věrohodnost.

1.2 Motivační příklad

Každý odhad metodou maximální věrohodnosti nemusí být konzistentní. Zde ilustračně uvedeme zajímavý případ, který ukázal Hill (1963).

Mějme posunuté log-normální rozdělení se 3 parametry, jehož hustota je

$$f_L(x, \alpha, \mu, \sigma^2) = \frac{1}{\sigma(x-\alpha)\sqrt{2\pi}} \ \exp\left\{\frac{-(\log(x-\alpha)-\mu)^2}{2\sigma^2}\right\} \ , x \geq \alpha,$$

kde jeden z parametrů, který se snažíme odhadnout, je kraj nosiče hustoty.

Mějme $X_1, \ldots, X_n$ náhodný výběr z rozdělení s hustotou f_L a nechť

$$X_{(1)} < \cdots < X_{(n)}$$

je odpovídající uspořádaný náhodný výběr.

Hustotu zapíšeme pomocí indikátoru. Tedy v následujícím, indikátor $\mathbb{I}[X \geq \alpha]$ se rovná jedné, pokud $X \geq \alpha$. Pokud nerovnost neplatí, tak se indikátor rovná 0.

Nejdříve upravíme věrohodnost

$$\begin{aligned} L_n(\alpha, \mu, \sigma^2) &= \prod_{i=1}^{n} f_L(X_i, \alpha, \mu, \sigma^2) = \prod_{i=1}^{n} f_L(X_{(i)}, \alpha, \mu, \sigma^2) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \prod_{j=1}^{n} (X_{(j)} - \alpha)^{-1} \, \mathbb{I}\left[X_{(1)} \geq \alpha\right] \, \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \mu \Big)^2 \right\}. \end{aligned}$$

Druhá rovnost platí, protože v součinu bereme celý náhodný výběr, a proto se výsledek nezmění přeuspořádáním náhodného výběru. Ve třetí rovnosti jsme u indikátoru využili toho, že tvrzení všechny náhodné veličiny jsou větší než α, je ekvivalentní s tvrzením, že nejmenší náhodná veličina je větší než α.

Začneme tím, že položíme α pevné a podíváme se na věrohodnost jako na funkci proměnných μ a σ^2 a tu maximalizujeme. Tato funkce je diferencovatelná podle μ a σ^2, takže budeme maximalizovat log-věrohodnost

$$\begin{aligned} \ell_n(\mu, \sigma^2) = \log L_n(\alpha, \mu, \sigma^2) &= \frac{-n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \mu \Big)^2 \\ &\quad - \sum_{j=1}^{n} \log(X_{(j)} - \alpha) - \frac{n}{2} \log(2\pi) + \log\Big(\mathbb{I}\left[X_{(1)} \geq \alpha\right] \Big). \end{aligned}$$

Odhady budeme hledat pomocí systému věrohodnostních rovnic, tj. spočítáme parciální derivace $\ell_n(\mu, \sigma^2)$ podle μ a σ^2 a následně je položíme rovné 0. Poté systém rovnic vyřešíme, a tím získáme odhady $\widetilde{\mu}_n$ a $\widetilde{\sigma}_n^2$. Tedy řešíme

$$\begin{aligned} \frac{\partial \ell_n(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \mu \Big) = 0, \\ \frac{\partial \ell_n(\mu, \sigma^2)}{\partial \sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \mu \Big)^2 = 0. \end{aligned}$$

Naše odhady μ a σ^2 jsou

$$\widetilde{\mu}_n(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \log(X_{(i)} - \alpha), \quad \widetilde{\sigma}_n^2(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \widetilde{\mu}_n(\alpha) \Big)^2.$$

Nyní dosadíme naše odhady do věrohodnosti a dostaneme funkci

$$\begin{aligned} L_n^*(\alpha) = L_n(\alpha, \widetilde{\mu}_n(\alpha), \widetilde{\sigma}_n^2(\alpha)) &= (2\pi\widetilde{\sigma}_n^2(\alpha))^{-\frac{n}{2}} \prod_{j=1}^{n} (X_{(j)} - \alpha)^{-1} \\ &\quad \exp\left\{ \frac{-1}{2\widetilde{\sigma}_n^2(\alpha)} \sum_{i=1}^{n} \Big(\log(X_{(i)} - \alpha) - \widetilde{\mu}_n(\alpha) \Big)^2 \right\} \mathbb{I}\left[X_{(1)} \geq \alpha\right] \\ &= \widetilde{\sigma}_n^2(\alpha)^{-\frac{n}{2}} \prod_{j=1}^{n} (X_{(j)} - \alpha)^{-1} \, (2\pi e)^{-\frac{n}{2}} \, \mathbb{I}\left[X_{(1)} \geq \alpha\right]. \end{aligned}$$

Využili jsme toho, že

$$\exp\left\{\frac{-1}{2\widetilde{\sigma}_n^2(\alpha)}\sum_{i=1}^{n}\Big(\log(X_{(i)}-\alpha)-\widetilde{\mu}_n(\alpha)\Big)^2\right\}=\exp\left\{\frac{-1}{2\widetilde{\sigma}_n^2(\alpha)}n\widetilde{\sigma}_n^2(\alpha)\right\}=e^{\frac{-n}{2}}.$$

Na důkazy následujících tvrzení odkazujeme čtenáře na článek Hill (1963). Zde pouze podotkneme, že naše funkce $L_n^*(\alpha)$ se rovná, až na multiplikativní konstantu, tamní funkci $L^{**}(\gamma)$, a tedy tvrzení naleznete formulované pro ni.

Platí, že

$$\lim_{\alpha\to X_{(1)}} L_n^*(\alpha)=+\infty,\quad \lim_{\alpha\to X_{(1)}}\widetilde{\mu}_n(\alpha)=-\infty,\quad \lim_{\alpha\to X_{(1)}}\widetilde{\sigma}_n^2(\alpha)=+\infty.$$

První limita nám dává, že můžeme dosáhnout libovolně velké věrohodnosti po křivce $(\alpha,\widetilde{\mu}_n(\alpha),\widetilde{\sigma}_n^2(\alpha))$ tím, že α půjde k $X_{(1)}$. Takto dostaneme odhad maximální věrohodnosti $(X_{(1)},-\infty,+\infty)$.

Tento výsledek má pár problémů. Zjevný problém je ten, že odhad μ a σ^2 nezávisle na datech dává stejný výsledek. Tedy nemáme konzistentní odhady μ a σ^2. Dále za pozornost stojí to, že pokud si položíme 2 parametry pevné a vezmeme limitu třetího, tak dostaneme

$$\lim_{\alpha\to X_{(1)}} L_n(\alpha,\mu,\sigma^2)=\lim_{\mu\to-\infty} L_n(\alpha,\mu,\sigma^2)=\lim_{\sigma^2\to\infty} L_n(\alpha,\mu,\sigma^2)=0.$$

Vidíme, že situace je velice komplikovaná.

Celkově máme, že náš odhad metodou maximální věrohodnosti v daném případě nemá požadované vlastnosti. Problém spočívá v tom, že existuje křivka v parametrickém prostoru, po které věrohodnost jde do nekonečna, pro α jdoucí k $X_{(1)}$ a výsledné odhady μ a σ^2 nejsou konzistentní.

1.3 Metoda maximálního součinu mezer

Nejdříve uvedeme pomocná tvrzení, která nám pomohou metodu motivovat.

1.3.1 Pomocná tvrzení

Metoda maximálního součinu mezer využívá následující větu (viz Věta 1.4 Anděl, 2007).

Věta 1. *Nechť náhodná veličina X má rostoucí spojitou distribuční funkci F. Pak pro náhodnou veličinu $U=F(X)$ platí*

$$\mathsf{P}(U<u)=\begin{cases}0 & \textit{pro}\quad u\le 0,\\ u & \textit{pro}\quad 0<u<1,\\ 1 & \textit{pro}\quad 1\le u.\end{cases}$$

Neboli $F(X)$ má rovnoměrné rozdělení na intervalu $(0,1)$. Dále ještě uvedeme pomocnou větu (viz Věta 1.2 Anděl, 1998).

Věta 2 (nerovnost geometrického a aritmetického průměru)**.** *Nechť $x_1, \ldots, x_n$ jsou nezáporná čísla. Pak platí*

$$\left\{\prod_{i=1}^{n} x_i\right\}^{\frac{1}{n}} = \bar{x}_G \leq \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

a rovnosti je dosaženo právě tehdy, platí-li $x_1 = \cdots = x_n$.

Nyní si dokážeme jedno pomocné tvrzení.

Lemma 3. *Nechť $x_1, \ldots, x_n$ jsou kladná čísla a $\sum_{i=1}^{n} x_i = 1$.*
Pak $\bar{x}_G = \{\prod_{i=1}^{n} x_i\}^{\frac{1}{n}}$ je maximální právě tehdy, když $x_1 = \cdots = x_n$.

Důkaz. Z věty 2 a předpokladu $\sum_{i=1}^{n} x_i = 1$ máme, že

$$\bar{x}_G = \left\{\prod_{i=1}^{n} x_i\right\}^{\frac{1}{n}} \leq \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}.$$

Dále také z věty 2 víme, že $\bar{x}_G = \frac{1}{n}$ pokud $x_1 = \cdots = x_n$. A také $\bar{x}_G < \frac{1}{n}$, pokud existuje $i,j \in \{1, \ldots, n\}$ tak, že $x_i \neq x_j$. To jsme chtěli dokázat.

□

1.3.2 Popis metody

Na začátek je důležité zmínit, že metoda maximálního součinu mezer (MPS) funguje pouze pro náhodné veličiny, ale pro náhodné vektory nikoliv.

Předpokládejme, že máme hustotu $f(x;\boldsymbol{\theta})$ vzhledem k jednorozměrné Lebesgueově míře s distribuční funkcí $F(x;\boldsymbol{\theta})$. Dále předpokládejme, že nosič hustoty je kladný pouze na intervalu (α_1, α_2), kde α_1, α_2 nejsou nutně známá.

Buď $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ skutečná hodnota parametru. Mějme náhodný výběr $X_1, \ldots, X_n$ z rozdělení s hustotou $f(x;\boldsymbol{\theta}_0)$ o rozsahu n a nechť $X_{(1)} < \cdots < X_{(n)}$ je odpovídající uspořádaný náhodný výběr.

Věta 1 nám říká, že pokud uděláme transformaci $U_{(i)} = F(X_{(i)};\boldsymbol{\theta}_0)$, kde $\boldsymbol{\theta}_0$ je skutečná hodnota parametru, tak dostaneme uspořádaný náhodný výběr $U_{(1)} < \cdots < U_{(n)}$ z rovnoměrného rozdělení na intervalu $(0, 1)$. Dále využijeme znalost střední hodnoty a rozptylu i-té pořádkové statistiky $U_{(i)}$ (viz strana 115 Balakrishan a Nevzorov, 2003)

$$\mathsf{E}\, U_{(i)} = \frac{i}{n+1}, \quad \operatorname{var} U_{(i)} = \frac{i(n-i+1)}{(n+1)^2(n+2)}.$$

Vidíme, že střední hodnota je řádu n^{-1} a rozptyl je řádu n^{-2}. Z toho plyne, že realizace $U_{(i)}$ budou blízko své střední hodnoty. Poslední věc, které si všimneme, je, že rozdíl středních hodnot pořádkových statistik $U_{(i)}$ a $U_{(i+1)}$ je pro všechny i stejný a to $\frac{1}{n+1}$. Zjistili jsme, že pokud transformujeme náhodný výběr pomocí skutečné distribuční funkce, tak můžeme očekávat, že dostaneme body, které dělí interval $(0, 1)$ do přibližně stejně velkých částí.

My se pokusíme zvolit $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ tak, aby transformované body $Y_i(\boldsymbol{\theta}) = F(X_{(i)};\boldsymbol{\theta})$ měly od vedlejších bodů stejnou vzdálenost. Pokud se nám to podaří, tak to

zhruba odpovídá tomu, že se nám podařilo najít skutečný parametr. Tedy se budeme snažit, aby vzdálenosti

$$D_i(\boldsymbol{\theta}) = Y_i(\boldsymbol{\theta}) - Y_{i-1}(\boldsymbol{\theta}) = \int_{X_{(i-1)}}^{X_{(i)}} f(x;\boldsymbol{\theta})\,dx, \quad \text{pro } i = 1,\ldots,n+1,$$

byly stejně velké. Zde jsme dodefinovali $X_{(0)}, X_{(n+1)}$ tak, abychom také mohli využít vzdálenost $X_{(1)}$ (resp. $X_{(n)}$) od počátku (resp. konce) nosiče hustoty. Tedy $Y_0(\boldsymbol{\theta}) = F(X_{(0)};\boldsymbol{\theta}) = 0$, $X_{(0)} = \alpha_1$ a $Y_{n+1}(\boldsymbol{\theta}) = F(X_{(n+1)};\boldsymbol{\theta}) = 1$, $X_{(n+1)} = \alpha_2$. Je zřejmé, že součet vzdáleností je roven 1.

Metoda spočívá v tom, že se snažíme maximalizovat geometrický průměr vzdáleností

$$G_n(\boldsymbol{\theta}) = \left\{\prod_{i=1}^{n+1} D_i(\boldsymbol{\theta})\right\}^{\frac{1}{n+1}}.$$

Myšlenka je ta, že za podmínky $\sum_{i=1}^{n+1} D_i(\boldsymbol{\theta}) = 1$ geometrický průměr maximalizujeme právě tehdy, když všechny $D_i(\boldsymbol{\theta})$ jsou stejně velké, viz lemma 3.

Zde stejně jako u ML nás zajímá pouze argument maxima, a protože $D_i(\boldsymbol{\theta})$ pochází z intervalu $(0,1)$, můžeme maximalizovat jeho logaritmus

$$H_n(\boldsymbol{\theta}) = \log G_n(\boldsymbol{\theta}) = \frac{1}{n+1}\sum_{i=1}^{n+1} \log D_i(\boldsymbol{\theta}),$$

protože argument se nezmění.

Celkově máme, že odhad $\boldsymbol{\theta}_0$ metodou MPS je

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n+1}\sum_{i=1}^{n+1} \log D_i(\boldsymbol{\theta}), \quad \text{kde } D_i(\boldsymbol{\theta}) = F(X_{(i)};\boldsymbol{\theta}) - F(X_{(i-1)};\boldsymbol{\theta}).$$

V motivačním příkladu jsme viděli, že u metody maximální věrohodnosti může nastat, že pro nějaké pevné n metoda nebude fungovat, jelikož věrohodnost nemusí být shora omezená. Naopak u metody maximálního součinu mezer můžeme díky rovnosti $\sum_{i=1}^{n+1} D_i(\boldsymbol{\theta}) = 1$ a pomocí nerovnosti geometrického a aritmetického průměru (věta 2) odhadnout $H_n(\boldsymbol{\theta})$ seshora:

$$H_n(\boldsymbol{\theta}) = \log\left\{\prod_{i=1}^{n+1} D_i(\boldsymbol{\theta})\right\}^{\frac{1}{n+1}} \le \log\frac{\sum_{i=1}^{n+1} D_i(\boldsymbol{\theta})}{n+1} = \log\frac{1}{n+1} = -\log(n+1).$$

Tento horní odhad nám dává, že ačkoliv věrohodnost může být nekonečná, tak $H_n(\boldsymbol{\theta})$ bude vždy konečná. Díky tomu může být odhad metodou MPS konzistentní i v situaci, kdy odhad metodou ML konzistentní není. V Cheng a Amin (1983) zmiňují, že v příkladu, kde máme náhodný výběr z posunutého log-normálního rozdělení se 3 parametry, jsou MPS odhady α, μ a σ^2 konzistentní a navíc odhad (μ, σ^2) je asymptoticky normální.

2. Příklady

Ukážeme si, jak se hledá odhad metodou MPS. Začneme s náhodným výběrem z rozdělení, jehož nosič hustoty závisí na neznámém parametru. Důvod je ten, že tato metoda vznikla, protože v tomto případě občas selhávaly jiné metody, viz motivační příklad. Chceme se tedy podívat na to, jak se nám s metodou v daném případě bude pracovat. V každém příkladu nejdříve nalezneme odhad metodou MPS a následně odhad metodou ML. Poté budeme porovnávat některé vlastnosti získaných odhadů.

2.1 Posunuté exponenciální rozdělení

Mějme $X_1, \ldots, X_n$ náhodný výběr z posunutého exponenciálního rozdělení s hustotou $f(x;\alpha) = e^{-(x-\alpha)}$, $x \geq \alpha$, kde α je neznámé. Nechť $X_{(1)} < \cdots < X_{(n)}$ je odpovídající uspořádaný náhodný výběr.

Nejdříve si ukážeme odhad metodou MPS. Zde naše $\theta = \alpha$, $\Theta = \mathbb{R}$. Dále $X_{(0)} = \alpha$ a $X_{(n+1)} = \infty$. Spočítáme vzdálenosti

$$\begin{aligned} D_i(\alpha) &= \int_{X_{(i-1)}}^{X_{(i)}} f(x;\alpha)\,dx = \int_{X_{(i-1)}}^{X_{(i)}} e^{-(x-\alpha)}\,dx = e^{-\left(X_{(i-1)}-\alpha\right)} - e^{-\left(X_{(i)}-\alpha\right)} \\ &= e^{\alpha}\left(e^{-X_{(i-1)}} - e^{-X_{(i)}}\right). \end{aligned}$$

Vidíme, že logaritmus vzdáleností je

$$\log D_i(\alpha) = \alpha + \log\left(e^{-X_{(i-1)}} - e^{-X_{(i)}}\right).$$

Nyní si musíme uvědomit, které $X_{(i)}$ závisí na neznámém parametru α. V tomto případě máme $X_{(0)} = \alpha$ a X_j nezávisí na α pro $j \neq 0$. Dostáváme tedy (viz kapitola 1.3)

$$H_n(\alpha) = \frac{1}{n+1}\left(\sum_{i=1}^{n+1}\alpha + \log\left(e^{-\alpha} - e^{-X_{(1)}}\right) + \sum_{i=2}^{n+1}\log\left(e^{-X_{(i-1)}} - e^{-X_{(i)}}\right)\right).$$

Hledáme maximum $H_n(\alpha)$ tím, že derivaci položíme rovnou 0, tj.

$$H_n'(\alpha) = 1 + \frac{1}{n+1}\,\frac{-e^{-\alpha}}{e^{-\alpha} - e^{-X_{(1)}}} = 0.$$

Jednoduchými úpravami dostaneme, že odhad metodou MPS je

$$\widehat{\alpha}_n = X_{(1)} - \log\left(1 + \frac{1}{n}\right).$$

Nyní si ukážeme odhad metodou ML. Protože indikátor závisí na α a není diferencovatelný vzhledem k α, budeme maximalizovat věrohodnost

$$L_n(\alpha) = \prod_{i=1}^{n} f(X_i;\alpha) = \prod_{i=1}^{n} e^{-(X_i-\alpha)}\,\mathbb{I}\left[X_i \geq \alpha\right] = e^{-\sum_{i=1}^{n} X_i}\, e^{n\alpha}\,\mathbb{I}\left[\alpha \leq X_{(1)}\right].$$

Exponenciála je rostoucí funkce na $\mathbb{R}$ a indikátor je konstantní funkce pro $\alpha \leq X_{(1)}$. Věrohodnost $L_n(\alpha)$ maximalizujeme v α tak, že maximalizujeme α. Jediný požadavek na α je $\alpha \leq X_{(1)}$. Vidíme tedy, že odhad metodou ML je

$$\widetilde{\alpha}_n = X_{(1)}.$$

Porovnáme některé vlastnosti získaných odhadů. Nejdříve zjistíme, zda jsou odhady nestranné. Protože oba odhady závisí pouze na $X_{(1)}$, nejdříve spočítáme střední hodnotu $X_{(1)}$, a pak dopočítáme střední hodnotu odhadů. Před výpočtem střední hodnoty zjistíme rozdělení $X_{(1)}$ pomocí distribuční funkce

$$\begin{aligned}\mathsf{P}(X_{(1)} \leq x) &= 1 - \mathsf{P}(X_{(1)} > x) = 1 - \mathsf{P}(\forall i, X_i > x) \\ &= 1 - \mathsf{P}(X_1 > x)^n = 1 - e^{(x-\alpha)n}.\end{aligned}$$

Zde jsme využili toho, že X_i jsou nezávislé stejně rozdělené náhodné veličiny z posunutého exponenciálního rozdělení. Můžeme si všimnout, že $X_{(1)}$ má jiné posunuté exponenciální rozdělení s hustotou $f_{X_{(1)}}(x) = n\, e^{-(x-\alpha)n}, x \geq \alpha$.

Již jsme schopni spočítat střední hodnotu

$$\begin{aligned}\mathsf{E}\ X_{(1)} &= \int_{-\infty}^{\infty} x f_{X_{(1)}}(x)\, dx = \int_{\alpha}^{\infty} xn\, e^{-(x-\alpha)n}\, dx = \int_0^{\infty} (y+\alpha) n\, e^{-yn}\, dy \\ &= \alpha \int_0^{\infty} e^{-yn}\, n\, dy + \frac{1}{n} \int_0^{\infty} yn\, e^{-yn}\, n\, dy = \alpha + \frac{1}{n}.\end{aligned}$$

Zde jsme ve třetí rovnosti použili substituci $y = x - \alpha$. Na vypočítání integrálu jsme využili znalost, že $\int_0^{\infty} x^n\, e^{-x} = n!$ pro $n \in \mathbb{N}$.

Stačí nám už jenom dopočítat střední hodnotu odhadů

$$\begin{aligned}\mathsf{E}\ \widehat{\alpha}_n &= \mathsf{E}\left[X_{(1)} - \log\left(1 + \frac{1}{n}\right)\right] = \alpha + \frac{1}{n} - \log\left(1 + \frac{1}{n}\right), \\ \mathsf{E}\ \widetilde{\alpha}_n &= \mathsf{E}\ X_{(1)} = \alpha + \frac{1}{n}.\end{aligned}$$

Vidíme, že žádný odhad není nestranný, takže ani nemůže být nejlepší nestranný odhad (NNO).

Nicméně si ukážeme, že $X_{(1)}$ je úplná statistika a z druhé Lehmann-Sheffého věty (viz kapitola 7.4 Anděl, 2007) lehce získáme NNO. Tuto vlastnost budeme dokazovat z definice.

Statistika $X_{(1)}$ je úplná, pokud pro každou reálnou měřitelnou funkci ω platí implikace

$$\left[\mathsf{E}\ \omega\left(X_{(1)}\right) = 0 \ \text{ pro každé } \alpha \in \mathbb{R}\right] \implies \left[\omega\left(X_{(1)}\right) = 0 \ \text{ sj. pro každé } \alpha \in \mathbb{R}\right].$$

Použijeme již vypočítanou hustotu $X_{(1)}$ při výpočtu střední hodnoty

$$\begin{aligned}\mathsf{E}\ \omega\left(X_{(1)}\right) &= \int_{-\infty}^{\infty} \omega(x) f_{X_{(1)}}(x)\, dx = \int_{\alpha}^{\infty} \omega(x) n\, e^{-(x-\alpha)n}\, dx \\ &= \int_{\alpha}^{\infty} \omega(x) n\, e^{n\alpha}\, e^{-xn}\, dx = 0 \ \text{ pro každé } \alpha \in \mathbb{R}.\end{aligned}$$

Protože $n\, e^{n\alpha}$ je kladná funkce pro každé $\alpha \in \mathbb{R}$, tak rovnici můžeme vydělit a řešit

$$F(\alpha) = \int_{\alpha}^{\infty} \omega(x)\, e^{-xn}\, dx = 0 \ \text{ pro každé } \alpha \in \mathbb{R}.$$

Pokud na daný integrál budeme nahlížet jako na funkci α, naše funkce $F(\alpha)$ bude nulová (konstantní). Také její derivace bude všude nulová. Dále využijeme základní větu kalkulu a zderivujeme rovnici a dostaneme

$$F'(\alpha) = -\omega(\alpha)\, e^{-n\alpha} = 0 \;\text{ pro každé } \alpha \in \mathbb{R}.$$

Protože $-e^{-n\alpha}$ je záporná funkce pro každé $\alpha \in \mathbb{R}$, tak s ní také můžeme rovnici vydělit. Nakonec dostaneme, že

$$\omega(\alpha) = 0 \;\text{ pro každé } \alpha \in \mathbb{R}.$$

A tím, že $X_{(1)}$ je úplná statistika.

Předtím jsme vypočítali střední hodnotu $\mathsf{E}\, X_{(1)} = \alpha + \frac{1}{n}$. Nyní položíme $g(x) = x - \frac{1}{n}$. Vidíme, že $g(X_{(1)})$ je nestranný odhad α.

Poté, co si spočítáme druhý moment $X_{(1)}$ a uvidíme, že je konečný. Z vyjádření

$$\mathsf{E}\left[g(X_{(1)})\right]^2 = \mathsf{E}\left[X_{(1)}\right]^2 - \frac{2}{n}\,\mathsf{E}\, X_{(1)} + \frac{1}{n^2}$$

vidíme, že druhý moment $g(X_{(1)})$ je také konečný.

Nyní obdobně jako u střední hodnoty spočítáme druhý moment $X_{(1)}$

$$\begin{aligned}
\mathsf{E}\,[X_{(1)}]^2 &= \int_{-\infty}^{\infty} x^2 f_{X_{(1)}}(x)\, dx = \int_{\alpha}^{\infty} x^2 n\, e^{-(x-\alpha)n}\, dx = \int_0^{\infty} (y+\alpha)^2 n\, e^{-yn}\, dy \\
&= \frac{1}{n^2}\int_0^{\infty} (yn)^2\, e^{-yn}\, n\, dy + \frac{2\alpha}{n}\int_0^{\infty} yn\, e^{-yn}\, n\, dy + \alpha^2 \int_0^{\infty} e^{-yn}\, n\, dy \\
&= \frac{2}{n^2} + \frac{2\alpha}{n} + \alpha^2.
\end{aligned}$$

Ověřili jsme všechny předpoklady druhé Lehmann-Sheffého věty a dostali jsme, že nejlepší nestranný odhad α je

$$X_{(1)} - \frac{1}{n}.$$

Další vlastnost, kterou porovnáme, je střední čtvercová chyba. Můžeme si všimnout, že

$$\operatorname{var}(\widehat{\alpha}_n) = \operatorname{var}\left(X_{(1)} - \log\left(1 + \frac{1}{n}\right)\right) = \operatorname{var}(X_{(1)}) = \operatorname{var}(\widetilde{\alpha}_n).$$

Již víme, že druhý moment $X_{(1)}$ je konečný, takže si můžeme vyjádřit střední čtvercovou chybu

$$\operatorname{MSE}(\widehat{\alpha}_n) = \operatorname{var}(\widehat{\alpha}_n) + [\operatorname{bias}(\widehat{\alpha}_n)]^2.$$

Zde $\operatorname{bias}(\widehat{\alpha}_n)$ značí vychýlení odhadu, tj. $\mathsf{E}\,\widehat{\alpha}_n - \alpha$.

Z předešlého poznatku lehce spočítáme rozdíl středních čtvercových chyb odhadů

$$\begin{aligned}
\operatorname{MSE}(\widehat{\alpha}_n) - \operatorname{MSE}(\widetilde{\alpha}_n) &= \operatorname{var}(\widehat{\alpha}_n) + [\operatorname{bias}(\widehat{\alpha}_n)]^2 - \left(\operatorname{var}(\widetilde{\alpha}_n) + [\operatorname{bias}(\widetilde{\alpha}_n)]^2\right) \\
&= [\operatorname{bias}(\widehat{\alpha}_n)]^2 - [\operatorname{bias}(\widetilde{\alpha}_n)]^2 = \left(\frac{1}{n} - \log\left(1 + \frac{1}{n}\right)\right)^2 - \left(\frac{1}{n}\right)^2 \\
&= \log\left(1 + \frac{1}{n}\right)\left(\log\left(1 + \frac{1}{n}\right) - \frac{2}{n}\right).
\end{aligned}$$

Zde jsme ve třetí rovnosti využili to, že jsme již spočítali střední hodnotu odhadů.

Protože $\log\left(1+\frac{1}{n}\right)$ je kladný pro všechna $n \in \mathbb{N}$, tak nás zajímá pouze rozdíl $\log\left(1+\frac{1}{n}\right) - \frac{2}{n}$. Využijeme nerovnost $\log(1+x) \leq x$, pro $x > 0$ a dostaneme, že

$$\log\left(1+\frac{1}{n}\right) - \frac{2}{n} \leq \frac{1}{n} - \frac{2}{n} = \frac{-1}{n} < 0, \quad \text{pro všechna } n \in \mathbb{N}.$$

Rozdíl je vždy záporný, takže $\mathrm{MSE}(\widehat{\alpha}_n) < \mathrm{MSE}(\widetilde{\alpha}_n)$. Dostali jsme, že z pohledu střední čtvercové chyby je odhad metodou MPS lepší.

Poslední vlastnost, kterou porovnáme je konzistence odhadů. Ukážeme si, že $\widetilde{\alpha}_n = X_{(1)} \xrightarrow[n\to\infty]{\mathsf{P}} \alpha$. Potom pro každé $\epsilon > 0$

$$\begin{aligned}
&\mathsf{P}(|X_{(1)} - \alpha| > \epsilon) = \mathsf{P}(X_{(1)} > \alpha + \epsilon) + \mathsf{P}(X_{(1)} < \alpha - \epsilon) \\
&= \mathsf{P}(X_{(1)} > \alpha + \epsilon) = \int_{\alpha+\epsilon}^{\infty} f_{X_{(1)}}(x)\,dx = \int_{\alpha+\epsilon}^{\infty} n\,e^{-(x-\alpha)n}\,dx \\
&= \int_{\epsilon}^{\infty} n\,e^{-ny}\,dy = e^{-n\epsilon} \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

Zde jsme ve druhé rovnosti využili to, že $X_{(1)} \geq \alpha$ skoro jistě a v páté rovnosti jsme využili substituci $y = x - \alpha$.

Víme, že $\log(1+\frac{1}{n}) \xrightarrow[n\to\infty]{} 0$, zároveň je nenáhodný, a proto triviálně splňuje konvergenci s.j., a tedy i konvergenci v pravděpodobnosti. Z toho dostáváme, že

$$\widehat{\alpha}_n = X_{(1)} - \log\left(1+\frac{1}{n}\right) \xrightarrow[n\to\infty]{\mathsf{P}} \alpha - 0 = \alpha.$$

Celkově máme, že oba odhady jsou konzistentní.

2.2 Rovnoměrné rozdělení I

Mějme $X_1, \ldots, X_n$ náhodný výběr z rovnoměrného rozdělení na intervalu $[\alpha_1, \alpha_2]$ s hustotou $f(x;\boldsymbol{\theta}) = \frac{1}{\alpha_2-\alpha_1}$, $x \in [\alpha_1, \alpha_2]$, kde $\boldsymbol{\theta} = (\alpha_1, \alpha_2)$ je neznámé. Nechť $X_{(1)} < \cdots < X_{(n)}$ je odpovídající uspořádaný náhodný výběr.

Nejdříve spočítáme odhad metodou MPS. Vidíme, že $X_{(0)} = \alpha_1$ a $X_{(n+1)} = \alpha_2$. Spočítáme vzdálenosti

$$D_i(\boldsymbol{\theta}) = \int_{X_{(i-1)}}^{X_{(i)}} f(x;\boldsymbol{\theta})\,dx = \int_{X_{(i-1)}}^{X_{(i)}} \frac{1}{\alpha_2-\alpha_1}\,dx = \frac{X_{(i)} - X_{(i-1)}}{\alpha_2-\alpha_1}.$$

Vidíme, že logaritmus vzdáleností je

$$\log D_i(\boldsymbol{\theta}) = \log\left(X_{(i)} - X_{(i-1)}\right) - \log(\alpha_2 - \alpha_1).$$

Pouze $X_{(0)}$ a $X_{(n+1)}$ závisí na $\boldsymbol{\theta}$, tudíž

$$\begin{aligned}
H_n(\boldsymbol{\theta}) = \frac{1}{n+1}\Bigg(&\log\left(\alpha_2 - X_{(n)}\right) + \log\left(X_{(1)} - \alpha_1\right) \\
&- \sum_{i=1}^{n+1} \log(\alpha_2 - \alpha_1) + \sum_{i=2}^{n} \log\left(X_{(i)} - X_{(i-1)}\right)\Bigg).
\end{aligned}$$

Hledáme maximum $H_n(\boldsymbol{\theta})$ tím, že parciální derivace položíme rovné 0, tj.

$$\frac{\partial H_n(\boldsymbol{\theta})}{\partial \alpha_1} = \frac{-1}{X_{(1)} - \alpha_1} + \frac{n+1}{\alpha_2 - \alpha_1} = 0,$$
$$\frac{\partial H_n(\boldsymbol{\theta})}{\partial \alpha_2} = \frac{1}{\alpha_2 - X_{(n)}} - \frac{n+1}{\alpha_2 - \alpha_1} = 0.$$

Sečteme rovnice a následně vyjádříme

$$\alpha_2 = X_{(1)} + X_{(n)} - \alpha_1.$$

Už jenom dosadíme do první rovnice, dopočítáme a dostaneme odhad α_1, α_2 metodou MPS, který je

$$\widehat{\alpha}_{n,1} = \frac{n\,X_{(1)} - X_{(n)}}{n-1},\ \ \widehat{\alpha}_{n,2} = \frac{n\,X_{(n)} - X_{(1)}}{n-1}.$$

Nyní spočítáme odhad metodou ML. Nosič závisí na $\boldsymbol{\theta}$, takže budeme maximalizovat věrohodnost

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i;\, \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\alpha_2 - \alpha_1}\, \mathbb{I}\left[\alpha_1 \le X_i \le \alpha_2\right]$$
$$= \frac{1}{(\alpha_2 - \alpha_1)^n}\, \mathbb{I}\left[\alpha_1 \le X_{(1)} \le X_{(n)} \le \alpha_2\right].$$

Máme 2 podmínky na parametr: $\alpha_1 \le X_{(1)}$ a $\alpha_2 \ge X_{(n)}$.

Nejdříve se na věrohodnost podíváme jako na funkci α_1. Jelikož funkce $\frac{1}{(\alpha_2-\alpha_1)^n}$ je rostoucí v α_1, tak maximalizujeme věrohodnost $L_n(\alpha_1)$ tím, že maximalizujeme α_1. Z podmínky na parametr dostáváme odhad $\widetilde{\alpha}_{n,1} = X_{(1)}$.

Nyní se podíváme na věrohodnost jako na funkci α_2. Analogicky vidíme, že $L_n(\alpha_2)$ je klesající v α_2, takže věrohodnost maximalizujeme tím, že minimalizujeme α_2. A z podmínky na parametr dostaneme odhad $\widetilde{\alpha}_{n,2} = X_{(n)}$.

Získali jsme odhad α_1, α_2 metodou ML, který je

$$\widetilde{\alpha}_{n,1} = X_{(1)},\ \ \widetilde{\alpha}_{n,2} = X_{(n)}.$$

Nyní zjistíme, zda některý z našich odhadů je nestranný. Nechť $Y_1, \ldots, Y_n$ je náhodný výběr z rovnoměrného rozdělení na intervalu $[0,1]$ a $Y_{(1)} < \cdots < Y_{(n)}$ je odpovídající uspořádaný náhodný výběr. Na vypočítání střední hodnoty odhadů využijeme již uvedené (viz kapitola 1.3) střední hodnoty $\mathsf{E}\, Y_{(1)} = \frac{1}{n+1}$ a $\mathsf{E}\, Y_{(n)} = \frac{n}{n+1}$ a následující myšlenku. Pokud je náhodná veličina Y_i z rovnoměrného rozdělení na intervalu $[0,1]$ a položíme náhodnou veličinu

$$Z_i = \alpha_1 + (\alpha_2 - \alpha_1) Y_i,\ \ \text{pro } i = 1, \ldots, n+1,$$

potom Z_i má rovnoměrné rozdělení na intervalu $[\alpha_1, \alpha_2]$, tedy stejné jako X_i. Dále si můžeme uvědomit, že náhodná veličina

$$Z_{(i)} = \alpha_1 + (\alpha_2 - \alpha_1) Y_{(i)},\ \ \text{pro } i = 1, \ldots, n+1,$$

má stejné rozdělení jako $X_{(i)}$. Z vyjádření výše můžeme spočítat střední hodnoty

$$\mathsf{E}\, X_{(1)} = \mathsf{E}\, Z_{(1)} = \alpha_1 + (\alpha_2 - \alpha_1)\, \mathsf{E}\, Y_{(1)} = \alpha_1 + \frac{\alpha_2 - \alpha_1}{n+1} = \frac{n}{n+1}\alpha_1 + \frac{1}{n+1}\alpha_2,$$
$$\mathsf{E}\, X_{(n)} = \mathsf{E}\, Z_{(n)} = \alpha_1 + (\alpha_2 - \alpha_1)\, \mathsf{E}\, Y_{(n)} = \alpha_1 + n\frac{\alpha_2 - \alpha_1}{n+1} = \frac{1}{n+1}\alpha_1 + \frac{n}{n+1}\alpha_2.$$

Vidíme, že odhad metodou ML není nestranný. Nicméně po kratším počítání dostaneme, že

$$\mathsf{E}\ \widehat{\alpha}_{n,1} = \mathsf{E}\ \frac{n\,X_{(1)} - X_{(n)}}{n-1} = \frac{n\ \mathsf{E}\ X_{(1)} - \mathsf{E}\ X_{(n)}}{n-1} = \alpha_1.$$

Analogicky dostaneme, že $\mathsf{E}\ \widehat{\alpha}_{n,2} = \alpha_2$. Tedy odhad metodou MPS je nestranný.

2.3 Rovnoměrné rozdělení II

Mějme $X_1, \ldots, X_n$ náhodný výběr z rovnoměrného rozdělení na intervalu $[k\theta, (k+1)\theta]$ s hustotou $f(x;\theta) = \theta^{-1}$, $x \in [k\theta, (k+1)\theta]$, kde $k > 0$ známé a $\theta > 0$ neznámé. Nechť $X_{(1)} < \cdots < X_{(n)}$ je odpovídající uspořádaný náhodný výběr.

Na začátek si uvědomíme, že z nosiče hustoty plynou nerovnosti

$$X_{(1)} \geq k\theta \ \text{ a } \ X_{(n)} \leq (k+1)\theta.$$

Pokud by tomu tak nebylo, znamenalo by to, že náhodný výběr nemohl pocházet z předpokládaného rozdělení. To nám dává, že odhad θ by měl splňovat nerovnosti

$$\widehat{\theta}_n \leq \frac{X_{(1)}}{k} \ \text{ a } \ \widehat{\theta}_n \geq \frac{X_{(n)}}{k+1}.$$

Nejdříve spočítáme odhad metodou MPS. Zde máme $X_{(0)} = k\theta$, $X_{(n+1)} = (k+1)\theta$. První spočítáme vzdálenosti

$$D_i(\theta) = \int_{X_{(i-1)}}^{X_{(i)}} f(x;\theta)\,dx = \int_{X_{(i-1)}}^{X_{(i)}} \frac{1}{\theta}\,dx = \frac{X_{(i)} - X_{(i-1)}}{\theta}.$$

Dále dostáváme logaritmus vzdáleností

$$\log D_i(\theta) = \log(X_{(i)} - X_{(i-1)}) - \log(\theta).$$

Pouze $X_{(0)}$ a $X_{(n+1)}$ závisí na θ, tudíž

$$H_n(\theta) = \frac{1}{n+1}\Bigg(\log\left(X_{(1)} - k\theta\right) + \log\left((k+1)\theta - X_{(n)}\right) - \sum_{i=1}^{n+1}\log(\theta) + \sum_{i=2}^{n}\log\left(X_{(i)} - X_{(i-1)}\right)\Bigg).$$

Pokusíme se najít maximum tím, že položíme derivaci rovnou 0, tj.

$$H_n'(\theta) = \frac{-1}{\theta} + \frac{1}{n+1}\left(\frac{-k}{X_{(1)} - k\theta} + \frac{k+1}{(k+1)\theta - X_{(n)}}\right) = 0.$$

Po delším počítání dostaneme rovnici

$$\theta^2 k(k+1)(n-1) - \theta(n(k+1)X_{(1)} + nkX_{(n)}) + (n+1)X_{(1)}X_{(n)} = 0.$$

Zde si označíme

$$\begin{aligned} a &= \ \ k(k+1)(n-1), \\ b &= -n(k+1)X_{(1)} + nkX_{(n)}, \\ c &= \ \ (n+1)X_{(1)}X_{(n)}. \end{aligned}$$

Řešení kvadratické rovnice jsou

$$\theta_{1,2} = \frac{-b}{2a} \pm \sqrt{\left(\frac{b}{2a}\right)^2 - \frac{c}{a}}.$$

Pokud si označíme $A = \frac{-b}{2a}$ dostaneme:

$$A = \frac{-b}{2a} = \frac{n}{2(n-1)} \left(\frac{X_{(1)}}{k} + \frac{X_{(n)}}{k+1}\right),$$
$$\frac{-c}{a} = \frac{-(n+1)X_{(1)}X_{(n)}}{k(k+1)(n-1)} = -\frac{n+1}{n-1} \frac{X_{(1)}}{k} \frac{X_{(n)}}{k+1}.$$

Naše možné odhady jsou

$$\theta_{1,2} = A \pm \left\{A^2 - \frac{n+1}{n-1} \frac{X_{(1)}}{k} \frac{X_{(n)}}{k+1}\right\}^{\frac{1}{2}}.$$

Díky naší numerické zkušenosti jsme zjistili, že první řešení nesplňuje nerovnost $\theta \leq \frac{X_{(1)}}{k}$. Celkově dostáváme, že odhad metodou MPS je

$$\widehat{\theta}_n = A - \left\{A^2 - \frac{n+1}{n-1} \frac{X_{(1)}}{k} \frac{X_{(n)}}{k+1}\right\}^{\frac{1}{2}}, \text{ kde } A = \frac{n}{2(n-1)} \left(\frac{X_{(1)}}{k} + \frac{X_{(n)}}{k+1}\right).$$

Tento výsledek je v souladu s předlohou, kde je tento odhad uveden bez odvození.

Nyní spočítáme odhad metodou ML. Nosič závisí na θ, takže budeme maximalizovat věrohodnost

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\, \theta) = \prod_{i=1}^{n} \frac{1}{\theta}\, \mathbb{I}\left[k\theta \leq X_i \leq (k+1)\theta\right]$$
$$= \frac{1}{\theta^n}\, \mathbb{I}\left[k\theta \leq X_{(1)} \leq X_{(n)} \leq (k+1)\theta\right].$$

Dostáváme podmínku na parametr: $\frac{X_{(n)}}{k+1} \leq \theta \leq \frac{X_{(1)}}{k}$. Funkci $\frac{1}{\theta^n}$ maximalizujeme tím, že minimalizujeme θ. Z podmínky na parametr dostáváme, že odhad θ metodou ML je

$$\widetilde{\theta}_n = \frac{X_{(n)}}{k+1}.$$

Zde nebudeme porovnávat odhady, ale pouze podotkneme, že odhad metodou ML závisí pouze na $X_{(n)}$, ale odhad metodou MPS závisí na $X_{(n)}$ i $X_{(1)}$. Intuice nám říká, že odhad metodou MPS by měl být v nějakém smyslu lepší. Skutečně, v Cheng a Amin (1983) je uvedeno, že pro velká k je asymptotický rozptyl odhadu metodou MPS přibližně poloviční oproti odhadu metodou ML.

2.4 Exponenciální rozdělení

Již jsme si ukázali metodu MPS v případě, kde jsme měli náhodný výběr z rozdělení, jehož nosič závisel na neznámém parametru. Ukazovalo se, že v takové situaci metoda MPS funguje lépe. Nyní si ukážeme, jak se situace změní, pokud nosič nebude závislý na parametru.

Mějme $X_1, \dots, X_n$ náhodný výběr z exponenciálního rozdělení s hustotou $f(x;\lambda) = \lambda e^{-\lambda x}$, $x \geq 0$, kde $\lambda > 0$ je neznámé. Nechť $X_{(1)} < \cdots < X_{(n)}$ je odpovídající uspořádaný náhodný výběr.

Nejdříve spočítáme odhad metodou ML. Vidíme, že log-věrohodnost

$$\ell_n(\lambda) = \sum_{i=1}^{n} \log f(X_i;\lambda) = \sum_{i=1}^{n} \log\left(\lambda e^{-\lambda X_i}\right) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i$$

je diferencovatelná vzhledem k λ, proto budeme hledat odhad pomocí věrohodnostní rovnice

$$\ell_n'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

Získali jsme odhad metodou ML

$$\widetilde{\lambda}_n = 1/\overline{X}_n, \quad \text{kde } \overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Nyní spočítáme odhad metodou MPS. Položíme $X_{(0)} = 0$ a $X_{(n+1)} = +\infty$ a spočítáme vzdálenosti

$$D_i(\lambda) = \int_{X_{(i-1)}}^{X_{(i)}} f(x;\lambda)\,dx = \int_{X_{(i-1)}}^{X_{(i)}} \lambda e^{-\lambda x}\,dx = e^{-\lambda X_{(i-1)}} - e^{-\lambda X_{(i)}}.$$

Dále spočítáme derivaci logaritmu vzdáleností

$$\Big(\log(D_i(\lambda))\Big)' = \Big(\log\left(e^{-\lambda X_{(i-1)}} - e^{-\lambda X_{(i)}}\right)\Big)' = \frac{X_{(i)}e^{-\lambda X_{(i)}} - X_{(i-1)}e^{-\lambda X_{(i-1)}}}{e^{-\lambda X_{(i-1)}} - e^{-\lambda X_{(i)}}}.$$

Odhad budeme hledat pomocí derivace, tj.

$$H_n'(\lambda) = \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{X_{(i)}e^{-\lambda X_{(i)}} - X_{(i-1)}e^{-\lambda X_{(i-1)}}}{e^{-\lambda X_{(i-1)}} - e^{-\lambda X_{(i)}}} = 0.$$

Bohužel pro obecné n odhad nedokážeme najít analyticky.

Pro zajímavost si zvolíme náhodný výběr o rozsahu 1, tedy máme pouze jedno pozorování X_1. Máme 3 veličiny $X_0 = 0$, X_1 a $X_2 = \infty$. Rovnici $H_n'(\lambda) = 0$ převedeme na stejného jmenovatele, vynásobíme jmenovatelem a dostaneme

$$2X_1e^{-2\lambda X_1} - (X_0+X_1)e^{-\lambda(X_0+X_1)} - (X_1+X_2)e^{-\lambda(X_1+X_2)} + (X_0+X_2)e^{-\lambda(X_0+X_2)} = 0.$$

Po dosazení za X_0 a X_2 dostaneme

$$2X_1e^{-2\lambda X_1} - X_1e^{-\lambda X_1} = 0.$$

A nakonec po menší úpravě dostane odhad metodou MPS

$$\widehat{\lambda}_1 = \frac{\log(2)}{X_1}, \quad \widetilde{\lambda}_1 = \frac{1}{X_1}.$$

Zde jsme navíc pro srovnání uvedli odhad metodou ML.

Mohlo by nás zajímat pro jaké n ještě dokážeme nalézt odhad metodou MPS explicitně. Stejným postupem pro uspořádaný náhodný výběr o 2 pozorováních $X_{(1)}$ a $X_{(2)}$ bychom dostali

$$\left(2X_{(1)} + X_{(2)}\right) e^{-\lambda\left(2X_{(1)}+X_{(2)}\right)} - \left(X_{(1)} + 2X_{(2)}\right) e^{-\lambda\left(X_{(1)}+2X_{(2)}\right)}$$
$$- \left(X_{(1)} + X_{(2)}\right) e^{-\lambda\left(X_{(1)}+X_{(2)}\right)} + 2X_{(2)} e^{-\lambda 2X_{(2)}} = 0.$$

Tedy vidíme, že explicitní vzorec pro odhad metodou MPS existuje pouze pro náhodný výběr o jednom pozorování.

Protože nemáme explicitní vzorec pro odhad metodou MPS, nemůžeme porovnat naše odhady stejně jako v příkladu Posunuté exponenciální rozdělení (viz kapitola 2.1). Naše odhady porovnáme následovně:

i) Zvolíme si $\lambda = 1$ a vygenerujeme náhodné výběry o rozsahu 10 a 100 z exponenciálního rozdělení s parametrem 1.

ii) Napočítáme pro oba náhodné výběry odhad metodou ML pomocí výše zjištěného $\widetilde{\lambda}_n$ a odhad metodou MPS numericky z rovnice $H_n'(\lambda) = 0$.

iii) Nakonec to zopakujeme 1000× a empiricky odhadneme vychýlení, rozptyl a střední čtvercovou chybu odhadů pro dané rozsahy výběrů.

Výsledky můžeme nalézt v tabulce 2.1. Odhadnuté vychýlení značíme BIAS, odhadnutý rozptyl značíme VAR a odhadnutou střední čtvercovou chybu značíme MSE. Tabulku jsme rozdělili do dvou částí podle rozsahu náhodného výběru. U názvu metody značíme rozsah výběru spodním indexem.

	BIAS	VAR	MSE
MPS_{10}	−0,0228	1,1171	0,1176
ML_{10}	0,1015	1,1482	0,1585
MPS_{100}	−0,0122	1,0104	0,0105
ML_{100}	0,0111	1,0109	0,0110

Tabulka 2.1: Vlastnosti odhadů MPS a ML pro rozsahy výběrů 10 a 100.

Můžeme si všimnout, že pro oba uvažované rozsahy výběrů metoda MPS podhodnocuje a metoda ML nadhodnocuje. Také si můžeme všimnout, že pro rozsah výběru 10 metoda MPS má odhadnuté vychýlení menší než metoda ML. Dále si můžeme všimnout, že pro oba uvažované rozsahy výběrů také odhadnutý rozptyl i odhadnutá střední čtvercová chyba je menší u metody MPS, a tedy je tato metoda o něco přesnější. Ale také vidíme, že u náhodného výběru o rozsahu 100 jsou tyto hodnoty skoro nerozeznatelné.

Celkově jsme zjistili, že odhad metodou MPS je o něco přesnější než odhad metodou ML při malém rozsahu výběru, ale při velkém rozsahu výběru jsou odhady skoro nerozeznatelné.

Pro malý rozsah výběru doporučujeme metodu MPS, protože výpočetní náročnost nebude problém. Pro velký rozsah výběru kvůli výpočetní náročnosti a tomu, že odhady budou skoro nerozeznatelné doporučujeme metodu ML.

Závěr

V první kapitole jsme si představili metodu maximální věrohodnosti. Následně jsme si v motivačním příkladu odvodili odhad metodou maximální věrohodnosti v případě, kde jsme měli posunuté log-normální rozdělení a viděli, že tento odhad není konzistentní. Dále jsme si podrobně vysvětlili odhad metodou maximálního součinu mezer.

V druhé kapitole jsme si ukázali, jak se získá odhad metodou MPS z konkrétního náhodného výběru. Dále jsme ukázali odhad metodou ML, abychom mohli dané odhady porovnat. Na náhodném výběru z posunutého exponenciálního rozdělení jsme zjistili, že ani jedna metoda nám nedala nestranný odhad, a tak jsme si ukázali nejlepší nestranný odhad. Poté jsme vypočítali střední čtvercovou chybu obou odhadů a zjistili, že metoda MPS ji měla menší. Na náhodném výběru z exponenciálního rozdělení jsme zjistili, že odhad metodou MPS se nedal získat analyticky, a proto jsme nemohli porovnat metody standardním způsobem. Tento problém jsme vyřešili tím, že jsme odhady porovnali empiricky.

Bylo by zajímavé se dále zabývat obecnou konzistencí odhadu MPS, ale to je nad rámec této práce. Dále bychom mohli přidat porovnání metod MPS a ML pro další konkrétní rozdělení, například pro normální rozdělení.

Seznam použité literatury

Anděl, J. (1998). *Statistické metody*. Druhé přepracované vydání. Matfyzpress, Praha. ISBN 80-85863-27-8.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Balakrishan, N. a Nevzorov, V. N. (2003). *A Primer on Statistical Distributions*. John Wiley & Sons, Ltd New York. ISBN 9780471722229.

Cheng, R. C. H. a Amin, N. A. K. (1983). Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, **45**(3), 394–403.

Hill, B. M. (1963). The Three-Parameter Lognormal Distribution and Bayesian Analysis of a Point-Source Epidemic. *Journal of the American Statistical Association*, **58**(301), 72–84.