VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

NEPARAMETRICKÉ METODY ODHADU PARAMETRŮ ROZDĚLENÍ EXTRÉMNÍHO TYPU

NON-PARAMETRIC ESTIMATION OF PARAMETERS OF EXTREME VALUE DISTRIBUTION

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. VÍT BLACHUT
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. JAROSLAV MICHÁLEK, CSc.
SUPERVISOR

BRNO 2013

Vysoké učení technické v Brně, Fakulta strojního inženýrství

Ústav matematiky
Akademický rok: 2012/2013

ZADÁNÍ DIPLOMOVÉ PRÁCE

student(ka): Bc. Vít Blachut

který/která studuje v **magisterském navazujícím studijním programu**

obor: **Matematické inženýrství (3901T021)**

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Neparametrické metody odhadu praramетrů rozdělení extrémního typu

v anglickém jazyce:

Non-parametric estimation of parameters of extreme value distribution

Stručná charakteristika problematiky úkolu:

V současné době v souvislosti se stále častějším výskytem extrémních jevů (povodně, požáry, jiné nehody a pod.) dále také v souvislosti s analýzou spolehlivosti nejrůznějších technických zařízení nebo při zkoumání životností výrobků různých typů či při zkoumání doby přežití jedinců dané populace se pozornost statistiků soustřeďuje na studium rozdělení extrémních hodnot. Rozvíjejí se metody odhadu parametrů tohoto rozdělení pro různá experimentální uspořádání a konstruují se nové statistické testy pro ověření hypotéz vyslovených o parametrech rozdělení extrémního typu. Řada praktických aplikací vychází z maximálně věrohodných odhadů, které jsou při dostatečně velkém rozsahu výběru asymptoticky normální. Menší pozornost je v aplikacích věnována odhadům neparametrickým, které mohou být úspěšně využity i pro náhodné výběry menšího rozsahu.

Cíle diplomové práce:

Prostudujte neparametricke metody odhadu parametrických funkcí rozdělení extrémního typu podle [2], vybrané metody odhadu algoritmizujte a proved'te jejich počítačovou implementaci. Pomocí získaných programů proved'te statistickou analýzu dešť'ových řad z článku [4], vytřídění srážkových dat můžete provést podle [1]. Zejména se zaměřte na statistickou indukci pro krátké dešť'ové řady (10 - 20 let). Pro tyto krátké řady se pokuste vytvořit nové metody odhadu založené na metodě bootstrap, viz [3]. Výsledné odhady porovnejte pomocí simulací.

Seznam odborné literatury:

[1] Ben-Zvi, A.: Rainfall intensity-duration-frequency relationships derived from large partial duration series. Journal of Hydrology, 367, p. 104-114, 2009

[2] De Haan, L., Ferreira, A. Extreme Value Theory. An Introduction. Springer, 421 p. ISBN 0387239464, 2006

[3] Efron B., and Tibshirani R.:. An Introduction to the Bootstrap. Chapman & Hall/CRC. ISBN 9780412042317 1994

[4] Hellebrand, R., Michálek, J., Fusek, M., Prax, P.: Aktualizace hydrologických podkladů pro potřeby městské hydrologie. In Hydrologické podklady pro potřeby městského odvodnění. 1. Brno: VUT Brno, FAST. s. 10-23. ISBN: 978-80-214-4096- 8, 2010

Vedoucí diplomové práce: doc. RNDr. Jaroslav Michálek, CSc.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2012/2013.

V Brně, dne

L.S.

prof. RNDr. Josef Šlapal, CSc. Ředitel ústavu	prof. RNDr. Miroslav Doupovec, CSc., dr. h. c. Děkan fakulty

Abstrakt
Diplomová práce se zabývá rozdělením extrémních hodnot. V první části je zformulována a dokázána limitní věta pro rozdělení maxim. Dále jsou rozebrány základní vlastnosti rozdělení extrémního typu. Ústřední roli diplomové práce hrají neparametrické odhady indexu extrémní hodnoty. Především je zde odvozen Hillův a momentový odhad, pro které je na základě výsledků z matematické analýzy navržena volba indexu prahové statistiky pomocí metody bootstrap. Odhady indexu extrémní hodnoty jsou srovnány na základě simulací z vhodně vybraných rozdělení, která jsou blízká rozdělení srážkových úhrnů z vybrané dešťové řady. Pro tuto řadu je doporučen vhodný odhad a zvolen index prahové statistiky, což patří mezi nejtěžší úlohy z oblasti extrémních hodnot.

Summary
The concern of this diploma thesis is extreme value distributions. The first part formulates and proves the limit theorem for distribution of maximum. Further there are described basic properties of class of extreme value distributions. The key role of this thesis is on non-parametric estimations of extreme value index. Primarily, Hill and moment estimator are derived, for which is, based on the results of mathematical analysis, suggested an alternative choice of optimal sample fraction using a bootstrap based method. The estimators of extreme value index are compared based on simulations from proper chosen distributions, being close to distribution of given rain-fall data series. This time series is recommended a suitable estimator and suggested choice of optimal sample fraction, which belongs to the most difficult task in the area of extreme value theory.

Klíčová slova
rozdělení extrémních hodnot, index extrémní hodnoty, Hillův odhad, Momentový odhad, bootstrap

Keywords
extreme value distribution, extreme value index, Hill estimator, Moment estimator, bootstrap

BLACHUT, V. *Neparametrické metody odhadu parametru rozdělení extrémního typu*. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2013. 74 s. Vedoucí doc. RNDr. Jaroslav Michálek, CSc.

Prohlašuji, že jsem diplomovou práci „Neparametrické metody odhadu parametru rozdělení extrémního typu“ vypracoval samostatně s použitím odborné literatury a pramenů, uvedených na seznamu, jenž je součástí této práce.

Vít Blachut

Děkuji panu doc. RNDr. Jaroslavu Michálkovi, CSc. za rady, věnovaný čas a odborné vedení při tvorbě této diplomové práce.

Vít Blachut

Obsah

1. Úvod

Teorie extrémních hodnot je jedinečnou statistickou disciplínou pro vytváření technik a modelů k popisu především neobvyklých jevů. Nalezne své uplatnění především v oblastech hydrologie, finančnictví, pojišťovnictví, strojírenství a v mnoha dalších odvětvích. Slouží například k predikci extrémních záplav, povodní, velkých bouří, přírodních požárů, plnění velkých pojistných událostí, akciového rizika, rizika denního trhu, životností výrobků, doby přežití jedinců populace a k mnoha dalším účelům.

Existují dva různé přístupy k analýze extrémních hodnot. První z nich je založen na vytvoření blokových maxim. V mnoha situFacích je zvykem uvažovat roční maxima (metoda AMS-Annual Maxima Series). Druhá takzvaná prahová metoda je založena na výběru největších pozorování z celého pozorovaného výběru (metoda POT-Point Over Threshold).

Analýza blokových maxim vede na parametrické modely, které vycházejí z limitní věty pro rozdělení maxim. Jako model předpokládají třídu rozdělení extrémních hodnot. Parametry rozdělení lze odhadnout například pomocí metody maximální věrohodnosti. Ukazuje se, že konvergence je velmi pomalá. Vzhledem k omezení se na bloková maxima je potřeba velmi rozsáhlých náhodných výběrů a některá významná velká pozorování nemusí být zohledněna. Výhodou může být eliminace cyklické a periodické složky. V některých případech lze pomocí blokových maxim získat nezávislá stejně rozdělená pozorování.

V případě prahových metod lze vzhledem k limitnímu rozdělení velkých pozorování provést statistickou indukci na základě zobecněného Paretova rozdělení. Menší pozornost je v praxi věnována metodám semi-parametrickým, na které je zaměřena tato práce. Statistická indukce je sice oproti blokovým maximám založena na více pozorováních, v praxi se ale často jedná o časové řady, které nepředstavují nezávislá pozorování. Míra vlivu stochastických procesů na semi-parametrické metody je předmětem mnoha jiných studií. Velkým problémem všech prahových metod je volba prahu, která bude navržena v této práci.

2. Základní pojmy

Nejprve budou uvedeny některé základní pojmy z teorie pravděpodobnosti, které budou často užívány v dalších částech textu. Budeme se částečně držet značení podle [15], případně [1].

Předpokládejme, že je pevně dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$, kde Ω je libovolná neprázdná množina elementárních jevů, $\mathcal{A}$ je množinová σ-algebra na Ω a P je pravděpodobnost na jevovém poli $(\Omega, \mathcal{A})$.

Definice 2.1 *Náhodnou veličinou vzhledem k jevovému poli $\mathcal{A}$ rozumíme zobrazení $X : \Omega \to \mathbb{R}$, kdy pro každé $x \in \mathbb{R}$ platí $\{\omega : X(\omega) \leq x\} \in \mathcal{A}$. Ω je množina všech elementárních jevů.*

Definice 2.2 *Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. Funkci $F(x) = P(X \leq x)$ definovanou pro všechna $x \in \mathbb{R}$ nazýváme distribuční funkcí náhodné veličiny X.*

Definice 2.3 *Nechť $X_1, X_2, X_3, \ldots$ je posloupnost náhodných veličin s příslušnou posloupností distribučních funkcí $F_1, F_2, F_3, \ldots$. Pak řekneme, že posloupnost $X_1, X_2, X_3, \ldots$ konverguje v distribuci k náhodné veličině X s distribuční funkcí F, právě když*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

alespoň ve všech bodech spojistosti x distribuční funkce F. Píšeme $X_n \xrightarrow{D} X$. Rozdělení náhodné veličiny X nazýváme limitní nebo asymptotické.

Definice 2.4 *Řekneme, že posloupnost náhodných veličin $X_1, X_2, X_3, \ldots$ konverguje k náhodné veličině X podle pravděpodobnosti, jestliže pro každé $\epsilon > 0$ platí*

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

a píšeme $X_n \xrightarrow{P} X$.

Věta 2.1 *(**Centrální limitní věta.**) Nechť $X_1, X_2, X_3, \ldots$ je posloupnost nezávislých stejně rozdělených náhodných veličin s konečnou střední hodnotou μ a konečným rozptylem $\sigma > 0$. Pak posloupnost*

$$\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma} \xrightarrow{D} \mathrm{N}(0,1).$$

Důkaz. Například v [31] nebo [22].

Definice 2.5 *Nechť F je distribuční funkcí. Pravý koncový bod x^* distribuční funkce F definujeme jako*

$$x^* = \sup\{x : F(x) < 1\}.$$

Poznámka. Připouštíme i možnost $x^* = \infty$

Definice 2.6 *Nechť X je náhodná veličina, která má distribuční funkci F. Řekneme, že náhodná veličina X má degenerované rozdělení, jestliže existuje $x_0 \in \mathbb{R}$, takové, že platí*

$$F(x) = \begin{cases} 0, & \text{pro } x < x_0 \\ 1, & x \geq x_0. \end{cases}$$

Definice 2.7 *(**Normální (Gaussovo) rozdělení**). Řekneme, že náhodná veličina X má normální rozdělení, je-li hustota pravděpodobnosti tvaru*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

kde $\mu \in \mathbb{R}$ a $\sigma^2 > 0$ jsou neznámé parametry. Značíme $X \sim \mathrm{N}(\mu; \sigma^2)$. Střední hodnota X je $\mathrm{E}X = \mu$ a rozptyl $\mathrm{D}X = \sigma^2$. Rozdělení $N(0,1)$ bývá označováno jako normované (nebo standardizované) normální rozdělení.

Definice 2.8 *(**standardizované zobecněné Paretovo rozdělení**). Řekneme, že náhodná veličina X má standardizované zobecněné Paretovo rozdělení, je-li distribuční funkce*

$$H_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma}, & \gamma \neq 0, \\ 1 - e^{-x}, & \gamma = 0. \end{cases} \tag{2.1}$$

kde $x \geq 0$ pro $\gamma \geq 0$ a $0 \leq x \leq \frac{-1}{\gamma}$ pro $\gamma < 0$. Značíme $X \sim \mathrm{GP}(\gamma)$

Definice 2.9 *Nechť h je neklesající funkce na množině reálných čísel. Funkci $h^{\leftarrow}$ určenou předpisem*

$$h^{\leftarrow}(x) = \inf\{y : h(y) \geq x\}$$

nazveme zleva spojitou pseudo-inverzí k funkci h. Speciálně, je-li F distribuční funkcí, pak $F^{\leftarrow}$ nazýváme kvantilovou funkcí k F. Je-li h bijektivní funkcí, pak $h^{\leftarrow}$ je rovněž obyčejnou inverzní funkcí k h.

Lemma 2.10 *Nechť f_n je neklesající posloupnost funkcí, g je neklesající funkce. Dále předpokládejme, že x je bod spojitosti funkce g, $x \in (a,b)$, $a, b \in \mathbb{R}$ a*

$$\lim_{n\to\infty} f_n(x) = g(x).$$

Pak pro každý bod spojitosti $x \in (g(a), g(b))$ platí

$$\lim_{n\to\infty} {f_n}^{\leftarrow}(x) = g^{\leftarrow}(x).$$

Důkaz. Viz. [15] (str. 5).

Definice 2.11 *Nechť F je distribuční funkce. Funkci U určenou předpisem*

$$U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right), \qquad t > 1,$$

nazveme kvantilovou funkcí chvostu k distribuční funkci F.

Poznámka. Položme $t = \frac{1}{1-F(U)}$ a vyjádřeme

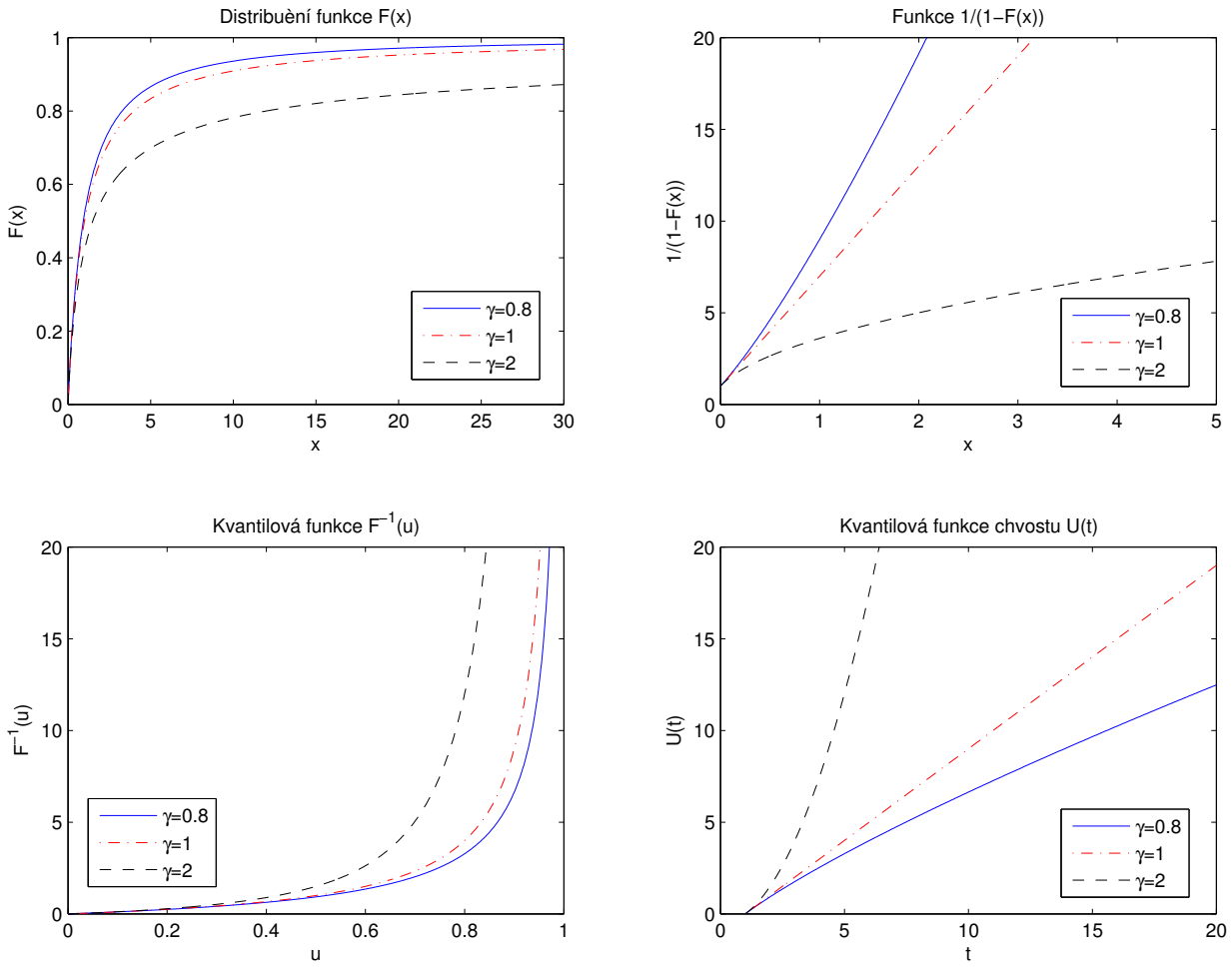
$$\begin{aligned} t - tF(U) &= 1 \\ F(U) &= \frac{t-1}{t} = 1 - \frac{1}{t}. \end{aligned}$$

Protože F je neklesající, dostáváme

$$U(t) = F^{\leftarrow}\left(1 - \frac{1}{t}\right).$$

Tedy $U(t)$ je zleva spojitou pseudo-inverzí k $\frac{1}{1-F(x)}$.

Na obrázku (2.1) je vykreslena distribuční funkce $F(x) = H_\gamma(x)$, funkce $\frac{1}{1-F(x)}$, kvantilová funkce $F^{\leftarrow}(u)$ a kvantilová funkce chvostu $U(t)$ pro zobecněné Paretovo rozdělení GP(γ). Všimněme si, že $x > 0$, $\frac{1}{1-F(x)} \in (1, \infty)$, $t > 1$.



Obrázek 2.1: Zobecněné Paretovo rozdělení GP(γ) s distribuční funkcí $F(x) = H_\gamma(x)$, kvantilovou funkcí $F^{\leftarrow}(u) = \frac{1}{\gamma}\left((1-u)^{-\gamma} - 1\right)$ a kvantilovou funkcí chvostu $U(t) = \frac{1}{\gamma}\left(t^{\gamma} - 1\right)$ pro různé hodnoty γ.

Definice 2.12 *Nechť $X_1, X_2 \ldots, X_n$ je konečná posloupnost nezávislých stejně rozdělených náhodných veličin z rozdělení o distribuční funkci F. Posloupnost $X_1, X_2 \ldots, X_n$ pak nazýváme náhodným výběrem rozsahu n z rozdělení o distribuční funkci F.*

Definice 2.13 *Mějme náhodný výběr* $X_1 = X_1(\omega), X_2 = X_2(\omega) \ldots, X_n = X_n(\omega)$, $\omega \in \Omega$ *z rozdělení o distribuční funkci* F, *který uspořádáme vzestupně podle velikosti. Dostaneme náhodný vektor* $(X_{1,n}(\omega), X_{2,n}(\omega), \ldots, X_{n,n}(\omega))$, *kde*

$$X_{i,n}(\omega) = X_{k_i}(\omega), \text{ pro } 1 \leq i \leq n, \ \{k_1, k_2, \ldots, k_n\} = \{1, 2, \ldots, n\}, \ \omega \in \Omega,$$

$$X_{1,n}(\omega) \leq X_{2,n}(\omega) \ldots \leq X_{n,n}(\omega).$$

Speciálně

$$X_{1,n}(\omega) = \min\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}, \forall \omega \in \Omega,$$
$$X_{n,n}(\omega) = \max\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}, \forall \omega \in \Omega.$$

Veličinám

$$X_{1,n}, X_{2,n}, \ldots, X_{n,n}$$

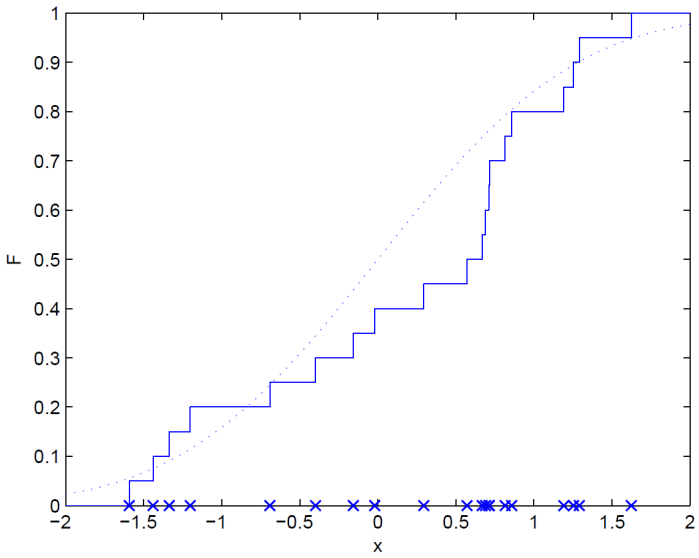
se říká uspořádaný náhodný výběr rozsahu n. *Jeho realizaci pak značíme* $x_{1,n}, x_{2,n}, \ldots, x_{n,n}$.

Definice 2.14 *Mějme náhodný výběr* $X_1, X_2 \ldots, X_n$ *z rozdělení o distribuční funkci* F. *Funkci*

$$\hat{F}_n(x) = \begin{cases} 0, & x < X_{1,n}. \\ \frac{i}{n}, & X_{i,n} \leq x < X_{i+1,n} \\ 1, & x \geq X_{n,n} \end{cases}$$

nazýváme výběrovou (empirickou) distribuční funkcí náhodného výběru $X_1, X_2 \ldots, X_n$.

Poznámka. Empirická distribuční funkce $\hat{F}_n(x)$ je přirozeně odhadem distribuční funkce F a hraje velmi významnou roli při neparametrických odhadech a u neparametrické metody bootstrap, ke které se dostaneme později. Dále budou uvedeny některé její základní vlastnosti a ukážeme, že $\hat{F}_n(x)$ je nestranným a konzistentním odhadem F. Na obrázku (2.2) je příklad empirické distribuční funkce náhodného výběru z normálního rozdělení $\mathrm{N}(0, 1)$.



Obrázek 2.2: Empirická distribuční funkce $\hat{F}_{20}(x)$ na základě dvaceti pozorování z normálního rozdělení $N(0, 1)$. Tečkovaná čára značí skutečnou distribuční funkci F. Realizace náhodného výběru je znázorněna pomocí symbolu $\times$.

- Empirická distribuční funkce $\hat{F}_n(x)$ je distribuční funkce rovnoměrného diskrétního rozdělení na množině $\{X_{i,n}\}_{i=1}^n$.

- Náhodná veličina $Z \in \mathbb{R}$ s rozdělením o distribuční funkci $\hat{F}_n$ je diskrétní a platí $P(Z = x_{i,n}) = \frac{1}{n}$ pokud hodnoty $\{x_{i,n}\}_{i=1}^n$ jsou různé, což pro výběry ze spojitého rozdělení platí s pravděpodobností 1.

- Zřejmě
$$\lim_{x\to-\infty} \hat{F}_n(x) = \lim_{x\to-\infty} F(x) = 0,$$
$$\lim_{x\to\infty} \hat{F}_n(x) = \lim_{x\to\infty} F(x) = 1,$$

- Označme $R_n(x)$ počet prvků, které jsou menší než x. Zřejmě $R_n(x) = n\hat{F}_n(x)$ a
$$P(R_n(x) = k) = \binom{n}{k} F^n(x)(1 - F(x))^{n-k}, \quad k \in \{0, 1, 2, \ldots, n\}.$$
$R_n(x)$ má tedy binomické rozdělení $Bi(n, F(x))$ se střední hodnotou $nF(x)$ a rozptylem $nF(x)(1 - F(x))$. Odtud
$$\mathrm{E}(\hat{F}_n(x)) = F(x)$$
a
$$\mathrm{D}(\hat{F}_n(x)) = \frac{F(x)(1 - F(x))}{n},$$
a tedy pro každé $x \in \mathbb{R}$ je $\hat{F}_n(x)$ nestranným odhadem $F(x)$.

- Protože $\hat{F}_n(x)$ je nestranným odhadem $F(x)$ a $\mathrm{D}(\hat{F}_n(x)) \to 0$ pro $n \to \infty$, platí (viz. [1] nebo [21]), že $\hat{F}_n(x)$ konverguje v pravděpodobnosti k $F(x)$ tj.
$$\lim_{n\to\infty} P(|\hat{F}_n(x) - F(x)| > \epsilon) = 0, \quad \text{pro všechna } \epsilon > 0,$$
a tedy $\hat{F}_n(x)$ je konzistentním odhadem $F(x)$ pro každé $x \in \mathbb{R}$. Odtud zřejmě
$$\mathrm{E}(\hat{F}_n(x) - F(x))^2 \to 0.$$

- Z centrální limitní věty (2.1) podle [31] str. 265 plyne
$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{D} \mathrm{N}(0, F(x)(1 - F(x))).$$

Poznámka o značení. V práci se vyskytují pouze přirozené logaritmy. Veškerá označení logaritmu představují přirozený logaritmus. Výběrový průměr je obvykle značen vodorovnou čarou, například $\overline{X}$. Využívají se základní pojmy z teorie odhadů, které lze nalézt v [1]. Buď $\hat{\theta}$ odhadem parametru θ. Vychýlení odhadu $\hat{\theta}$ značíme jako $\mathrm{bias}(\hat{\theta})$. Podobně střední kvadratickou chybu odhadu $\hat{\theta}$ značíme $\mathrm{MSE}(\hat{\theta})$.

3. Rozdělení extrémních hodnot

3.1. Úvod

Mějme posloupnost nezávislých stejně rozdělených náhodných veličin $X_1, X_2, \ldots$ s distribuční funkcí F. Centrální limitní věta popisuje limitní chování částečných sum $X_1 + X_2 + \cdots + X_n$ pro $n \to \infty$, zatímco teorie extrémních hodnot popisuje limitní chování extrémů

$$M_n = M_n(\omega) = \max\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}, \omega \in \Omega$$

nebo $\min\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\}, \omega \in \Omega$ pro $n \to \infty$. Omezíme se však pouze na případ výběrového maxima M, protože

$$\min\{X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\} = -\max\{-X_1(\omega), -X_2(\omega), \ldots, -X_n(\omega)\}, \quad \omega \in \Omega.$$

Naším cílem v této kapitole je najít možná rozdělení výběrových maxim M. Platí, že

$$P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = F^n(x). \tag{3.1}$$

Dále

$$\lim_{n\to\infty} F^n(x) = \begin{cases} 0, & \text{pro } x < x^* \\ 1, & \text{jinak,} \end{cases}$$

kde $x^* \leq \infty$ je pravý koncový bod distribuční funkce F.

Degenerovaná limitní distribuční funkce nám ovšem neposkytuje příliš mnoho informace. V centrální limitní větě 2.1 se vyskytuje parametr polohy μ a parametr měřítka $\sigma > 0$. Analogicky se můžeme zabývat standardizovanými maximy, kde parametry polohy a tvaru jsou reálná čísla zavisející na rozsahu výběru n. Dále uvažujme reálné posloupnosti $a_n > 0, b_n, n = 1, 2, \ldots$ takové, že náhodná veličina $\frac{M_n - b_n}{a_n}$ má pro $n \to \infty$ nedegenerovanou limitní distribuční funkci $G(x)$. Tímto z (3.1) plyne

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{3.2}$$

pro každý bod spojitosti x funkce G. Tedy

$$\frac{M_n - b_n}{a_n} \xrightarrow{D} G(x).$$

Distribuční funkce $G(x)$ popisuje chování maxim, proto definujme

Definice 3.1 *Nechť $X_1, X_2, \ldots$ je posloupnost nezávislých stejně rozdělených náhodných veličin s distribuční funkcí F. Nechť $a_n > 0$ a b_n jsou posloupnosti realných čísel, G nedegenerovaná distribuční funkce a platí 3.2. Pak množinu*

$$\mathcal{D}(G) = \{F : \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \text{ pro všechny body spojitosti } x \text{ funkce } G\}$$

nazveme oborem atraktivity rozdělení o distribuční funkci G. Rozdělení s distribuční funkcí $G(x)$ nazveme rozdělení extrémní hodnoty.

Poznámka. Je možné ukázat, že v podstatě všechna v praxi běžně uvažovaná spojitá rozdělení jsou prvky množiny $\mathcal{D}(G)$. Příkladem rozdělení, které neleží v oboru atraktivity funkce G, je von Mises rozdělení s distribuční funkcí $F(x) = 1 - e^{-x-\sin x}, x > 0$ [15].

Věta 3.1 *Jestliže $a_n > 0$ a b_n jsou posloupnosti realných čísel a G nedegenerovaná distribuční funkce, pak jsou následující vztahy ekvivalentní:*

1. $$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{3.3}$$
   pro každý bod spojitosti x funkce G tj. $F \in \mathcal{D}(G)$.

2. $$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -\log G(x) \tag{3.4}$$
   pro každý bod spojitosti x funkce G, pro který $0 < G(x) < 1$.

3. *Je-li U kvantilová funkce chvostu k distribuční funkci F, pak*
   $$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = D(x) \tag{3.5}$$
   pro každý bod spojitosti $x > 0$ funkce $D(x) = G^{\leftarrow}\left(e^{-\frac{1}{x}}\right)$.

Důkaz. Pokud x je bod spojitosti G, pak

$$\log \lim_{n\to\infty} F^n(a_n x + b_n) = \lim_{n\to\infty} \log F^n(a_n x + b_n)$$

a rovnici (3.3) lze po zlogaritmování přepsat jako

$$\lim_{n\to\infty} n\log F(a_n x + b_n) = \log G(x) \quad \forall x : 0 < G(x) < 1. \tag{3.6}$$

Aby posloupnost $n\log F(a_n x + b_n)$ nedivergovala pro $n \to \infty$, musí $F(a_n x + b_n) \to 1$. Platí

$$\lim_{n\to\infty} \frac{-\log(F(a_n x + b_n))}{1 - F(a_n x + b_n)} = 1, \tag{3.7}$$

neboť pomocí L'Hospitalova pravidla je (3.7) ekvivalentní s

$$\lim_{n\to\infty} \frac{1}{F(a_n x + b_n)} = 1.$$

Tímto z (3.7)

$$\lim_{n\to\infty} -\log(F(a_n x + b_n)) = \lim_{n\to\infty} (1 - F(a_n x + b_n)).$$

Do vztahu (3.6) pak lze za levou stranu dosadit výraz $[-(1 - F(a_n x + b_n))]$, tudiž

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -\log G(x),$$

což lze ekvivalentně přepsat na

$$\lim_{n\to\infty} \frac{1}{n(1 - F(a_n x + b_n))} = -\frac{1}{\log G(x)}. \tag{3.8}$$

Nyní sestrojíme zleva spojitou pseudo-inverzi k oběma stranám rovnice (3.8). Využijeme vlastnosti kvantilové funkce chvostu U, která je zleva spojitou pseudo-inverzí k $\frac{1}{1-F}$,

$$U = \left(\frac{1}{1 - F}\right)^{\leftarrow}.$$

i) Položme $x = \frac{1}{n(1-F(a_n y+b_n)}$, pak

$$nx = \frac{1}{1-F(a_n y+b_n)},$$
$$U(nx) = U\left(\frac{1}{1-F(a_n y+b_n)}\right) = a_n y + b_n,$$

a tedy

$$y = \frac{U(nx)-b_n}{a_n}.$$

ii) Dále, je-li $x = -\frac{1}{\log G(y)}$, dostaneme

$$\log G(y) = -\frac{1}{x},$$

a tedy

$$y = G^{\leftarrow}\left(e^{-\frac{1}{x}}\right).$$

Vrátíme-li se ke vztahu (3.8) a využijeme lemma (2.10), (i) a (ii) dostáváme

$$\lim_{n\to\infty} \frac{U(nx)-b_n}{a_n} = G^{\leftarrow}\left(e^{-\frac{1}{x}}\right)$$

pro $x > 0$, protože $x = \frac{1}{n(1-F(a_n y+b_n)} > 0$. ■

Poznámka. Věta platí i pro případ, kdy místo indexu n uvažujeme spojitou proměnnou t. Stačí použít $a(t) = a_{[t]}$ a $b(t) = b_{[t]}$. Index $[t]$ značí celoučíselnou část proměnné t. Viz. [15] str. 6.

Věta 3.2 (Fisher a Tippet (1928), Gnedenko(1943)) *Třída rozdělení extrémních hodnot je tvaru $G_\gamma(ax+b)$, $a > 0, b \in \mathbb{R}$, kde*

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-\frac{1}{\gamma}}), & 1+\gamma x > 0, \gamma \neq 0 \\ \exp\left(-e^{-x}\right), & \gamma = 0. \end{cases} \tag{3.9}$$

Důkaz. Vycházíme ze vztahu (3.5). Předpokládejme, že D je spojitá v bodě 1. Pro její bod spojitosti $x > 0$ položme

$$E(x) = D(x) - D(1) = \lim_{t\to\infty} \frac{U(tx)-b(t)}{a(t)} - \lim_{t\to\infty} \frac{U(t)-b(t)}{a(t)} = \lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)}.$$

Dále pro $y > 0$ dosadíme xy za x a dostaneme po jednoduché úpravě

$$\frac{U(txy)-U(t)}{a(t)} = \frac{U(txy)-U(t)-U(ty)+U(ty)}{a(t)} = \frac{U(txy)-U(ty)}{a(ty)}\frac{a(ty)}{a(t)} + \frac{U(ty)-U(t)}{a(t)}.$$

Z věty (3.1) plyne existence limit $\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)}$, a tedy i $\lim_{t\to\infty} \frac{U(txy)-U(ty)}{a(ty)}$ a $\lim_{t\to\infty} \frac{U(ty)-U(t)}{a(t)}$. Zřejmě pak musí existovat i $\lim_{t\to\infty} \frac{a(ty)}{a(t)}$. Dále položme $A(y) = \lim_{t\to\infty} \frac{a(ty)}{a(t)}$. Z předchozích vztahů dostaneme pro $x, y > 0$

$$E(xy) = \lim_{t\to\infty} \frac{U(txy)-U(ty)}{a(ty)}\frac{a(ty)}{a(t)} + \lim_{t\to\infty} \frac{U(ty)-U(t)}{a(t)} = E(x)A(y) + E(y).$$

Dále pro $(x, y > 0) \wedge (x, y \neq 1)$ položme $s = \log x$, $t = \log y$ a $H(x) = E(e^x)$, pak z předchozího vztahu

$$\begin{aligned} E(e^s e^t) &= E(e^s)A(e^t) + E(e^t) \\ H(t+s) &= H(s)A(e^t) + H(t). \end{aligned} \tag{3.10}$$

Protože $H(0) = E(e^0) = D(1) - D(1) = 0$, můžeme rovnici (3.10) napsat jako

$$\frac{H(t+s) - H(t)}{s} = \frac{H(s) - H(0)}{s} A(e^t). \tag{3.11}$$

Funkce H je monotónní, neboť $H(t) = E(e^t) = D(e^t) - D(1) = G^{\leftarrow}\left(e^{-e^{-t}}\right) - G^{\leftarrow}(e^{-1})$. Proto existuje $t \in \mathbb{R}$, kde funkce H má derivaci $H'(t)$. Podrobněji o derivacích funkcí například v [27]. Z rovnosti funkcí (3.11) plyne, že H má derivaci všude a odtud pro $s \to 0$

$$H'(t) \quad = \quad H'(0)A(e^t). \tag{3.12}$$

Nyní ukážme sporem, že $H'(0) \neq 0$. Předpokládáme-li, že $H'(0) = 0$, pak z (3.12) $H'(t) = 0$ pro $\forall t$, ale H není konstantní funkce.

Z rovnice (3.10) vyjádřeme

$$\frac{H(t+s) - H(t)}{H'(0)} \quad = \quad \frac{H(s)A(e^t)}{H'(0)}$$

a položme $Q(t) = \frac{H(t)}{H'(0)}$. Dostáváme

$$Q(t+s) - Q(t) \quad = \quad Q(s)A(e^t). \tag{3.13}$$

Z (3.12) dostaneme $A(e^t) = \frac{H'(t)}{H'(0)} = Q'(t)$, pak z (3.13)

$$Q(t+s) = Q(s)Q'(t) + Q(t). \tag{3.14}$$

Ve vztahu (3.14) zaměňme proměnné t a s, tedy

$$Q(s+t) = Q(t)Q'(s) + Q(s) \tag{3.15}$$

a dosaďme $Q(s)Q'(t) + Q(t)$ z (3.14) do (3.15). Dostáváme

$$Q(s)Q'(t) + Q(t) = Q(t)Q'(s) + Q(s)$$

a po jednoduché úpravě

$$Q(t)(Q'(s) - 1) = Q(s)(Q'(t) - 1). \tag{3.16}$$

Protože $Q(0) = \frac{H(0)}{H'(0)} = 0$ a $Q'(0) = \frac{H'(0)}{H'(0)} = 1$, dostáváme z (3.16)

$$Q(t)\frac{(Q'(s) - Q'(0))}{s} = \frac{Q(s) - Q(0)}{s}(Q'(t) - 1).$$

Odtud pro $s \to 0$

$$Q(t)Q''(0) = (Q'(t) - 1)$$

a z rovnosti funkcí a existence derivace levé strany plyne existence derivace pravé strany a derivujeme-li ještě jednou, dostaneme

$$Q''(0)Q'(t) = Q''(t). \tag{3.17}$$

Položme $\gamma = Q''(0)$. Z (3.17)

$$\gamma = \frac{Q''(t)}{Q'(t)} = (\log Q'(t))'.$$

Dostáme jednoduchou diferenciální rovnici $\gamma = (\log Q(t)')'$ s podmínkami $Q'(0) = 1$, $Q(0) = 0$. Přímou integrací a využitím podmínky $Q'(0) = 1$ dostaneme

$$Q'(t) = e^{\gamma t} \tag{3.18}$$

s počáteční podmínkou $Q(0) = 0$. Abychom mohli integrovat ještě jednou a použít podmínky $Q(0) = 0$, je třeba rozlišit 2 případy, kdy

a) $\gamma \neq 0$. Řešením (3.18) je

$$Q(t) = \frac{e^{\gamma t} - 1}{\gamma} = \frac{H(t)}{H'(0)} = \frac{D(e^t) - D(1)}{H'(0)}.$$

Odtud

$$D(e^t) = \frac{e^{\gamma t} - 1}{\gamma} H'(0) + D(1).$$

Vyjádřeme

$$D(t) = D(1) + H'(0)\frac{t^\gamma - 1}{\gamma}, \tag{3.19}$$

a k ní zleva spojitou pseudo-inverzi

$$D^{\leftarrow}(x) = \left(1 + \gamma\frac{x - D(1)}{H'(0)}\right)^{\frac{1}{\gamma}}, \quad 1 + \gamma\frac{x - D(1)}{H'(0)} > 0.$$

Protože $D(x) = G^{\leftarrow}(e^{-\frac{1}{x}})$, dostaneme dosazením výrazu $D^{\leftarrow}(x) = -\frac{1}{\log G(x)}$ do předchozí rovnice

$$\log G(x) = -\left(1 + \gamma\frac{x - D(1)}{H'(0)}\right)^{-\frac{1}{\gamma}}, 1 + \gamma\frac{x - D(1)}{H'(0)} > 0$$

a odtud

$$G\left(H'(0)x + D(1)\right) = \exp\left(-(1 + \gamma x)^{-\frac{1}{\gamma}}\right), \quad 1 + \gamma x > 0.$$

b) pro $\gamma = 0$. Řešením (3.18) je

$$Q(t) = t. \tag{3.20}$$

Připomeňme, že $Q(t) = \frac{H(t)}{H'(0)} = \frac{D(e^t) - D(1)}{H'(0)}$. Z (3.20) dostaneme

$$D(t) = D(1) + H'(0)\log t. \tag{3.21}$$

Z označení na (3.5) je $D(t) = G^{\leftarrow}(e^{-\frac{1}{t}})$ a odtud

$$G(D(1) + H'(0)\log t) = e^{-\frac{1}{t}}.$$

Položíme-li $\log t = x$, dostáváme

$$G\left(H'(0)x + D(1)\right) = \exp\left(-e^{-x}\right).$$

■

Definice 3.2 *Parametr γ z věty 3.2 nazýváme index extrémní hodnoty.*

Věta 3.3 *Pro $\gamma \in \mathbb{R}$ jsou následující výroky ekvivalentní [15]:*

1. *Existují posloupnosti realných čísel $a_n > 0$ a b_n takové, že*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & 1+\gamma x > 0,\ \gamma \neq 0 \\ \exp\left(-e^{-x}\right), & \gamma = 0. \end{cases} \tag{3.22}$$

2. *Existuje kladná funkce $a(t)$ taková, že pro všechna $x > 0$*

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = D_\gamma(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0 \\ \log x, & \gamma = 0. \end{cases} \tag{3.23}$$

3. *Existuje kladná funkce $a(t)$ taková, že*

$$\lim_{t\to\infty} t(1 - F(a(t)x + U(t))) = \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}}, & 1+\gamma x > 0,\ \gamma \neq 0 \\ e^{-x} & \gamma = 0. \end{cases} \tag{3.24}$$

4. *Existuje kladná neklesající funkce h taková, že*

$$\lim_{t\uparrow x^*} \frac{1 - F(t + xh(t))}{1 - F(t)} = \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}}, & 1+\gamma x > 0,\ \gamma \neq 0 \\ e^{-x} & \gamma = 0, \end{cases} \tag{3.25}$$

kde x^* je pravý koncový bod distribuční funkce F. Navíc $h(t) = a\left(\frac{1}{1-F(t)}\right)$.

Důkaz. Ze vztahu (3.19) plyne rovnost $D_\gamma(x) = \frac{x^\gamma - 1}{\gamma}$ pro $\gamma \neq 0$. Podobně z (3.21) plyne $D_\gamma(x) = e^{-x}$ pro $\gamma = 0$. Ekvivalence výrazu (1), (2) a (3) byly rozebrány ve větě (3.1). Důkaz, že výrazy (2) a (4) jsou ekvivalentní, je technického rázu a lze nalézt v [15] str. 11.

■

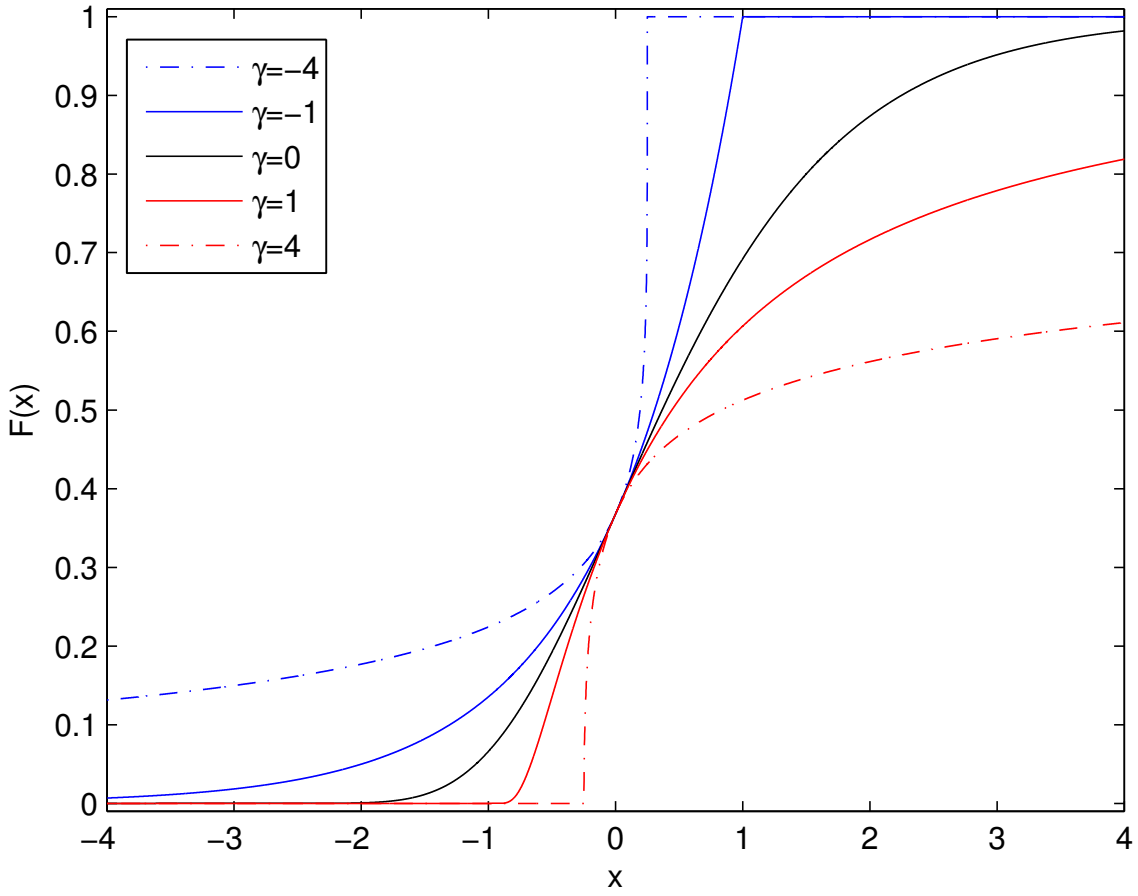
3.2. Základní typy rozdělení extrémních hodnot

V této kapitole budou rezebrány některé základní vlastnosti rozdělení extrémního typu. Vyjdeme z věty 3.2, v které vystupuje parameter γ. Situaci trochu zjednodušíme tím, že budeme uvažovat rozdělení ve standardizovaných tvarech.

Obrázek 3.1 znázorňuje distribuční funkce z třídy rozdělení extrémních hodnot pro různé hodnoty indexu extrémní hodnoty γ dané předpisem

$$G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & 1+\gamma x > 0, \gamma \neq 0 \\ \exp\left(-\mathrm{e}^{-x}\right), & \gamma = 0. \end{cases} \tag{3.26}$$

Je patrné, že rozdělení jsou rozdílného typu pro $\gamma > 0$, $\gamma = 0$, $\gamma < 0$.



Obrázek 3.1: Distribuční funkce G_γ z třídy rozdělení extrémních hodnot

Uvážíme-li možnosti $\gamma > 0$, $\gamma = 0$, $\gamma < 0$ odděleně dostaneme:

a) $\gamma > 0$: položíme-li x rovno $\frac{x-1}{\gamma}$ a dosadíme do (3.26) dostaneme rozdělení takzvaného *Fréchetova* typu s distribuční funkcí

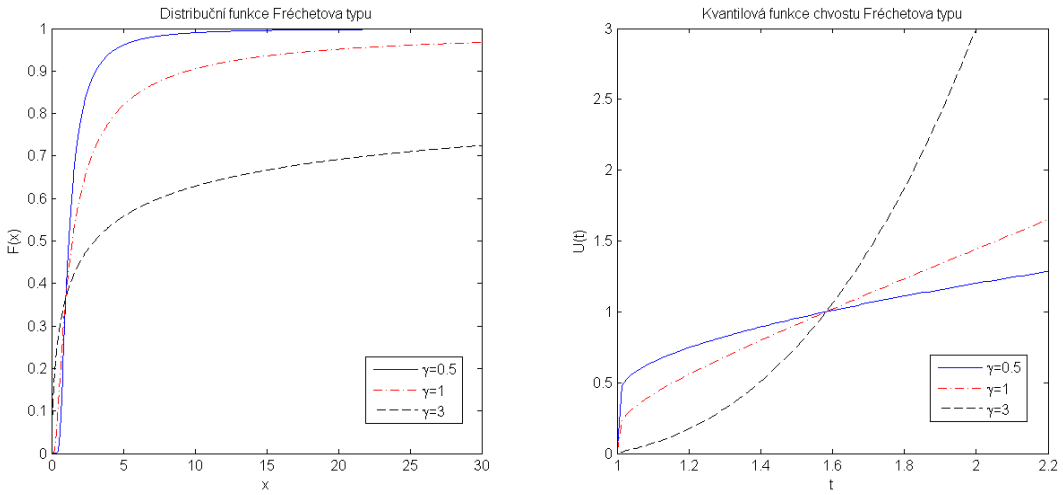
$$\Phi_\alpha(x) = \begin{cases} 0, & x \leq 0 \\ e^{-x^{-\alpha}}, & x > 0, \end{cases} \tag{3.27}$$

kde $\alpha = \frac{1}{\gamma}$. Pravý koncový bod je zde roven nekonečnu, tedy $\Phi_\alpha(x) < 1$, pro všechna $x \in \mathbb{R}$. Tato skupina rozdělení má spíše těžké chvosty, neboť velké hodnoty x jsou více pravděpodobné. Distribuční funkce $\Phi_\alpha(x)$ konverguje „velmi pomalu" k 1 pro $x \to \infty$.

Ukažme, že pro $p \geq \alpha$, $p > 0$ neexistuje p-tý obecný moment.

$$\begin{aligned} \mathrm{E}(X^p) &= \int_{-\infty}^{\infty} x^p f(x)\mathrm{d}x = \int_0^{\infty} x^p \left(e^{-x^{-\alpha}}\right)' \mathrm{d}x = \int_0^{\infty} \alpha x^{p-\alpha-1} e^{-x^{-\alpha}} \mathrm{d}x \\ &= \left| t = x^{-\alpha};\ \mathrm{d}t = -\alpha x^{-\alpha-1}\mathrm{d}x;\ x = t^{-\frac{1}{\alpha}} \right| = \int_0^{\infty} t \frac{t^{-\frac{1}{\alpha}(p-1)}}{t^{-\frac{1}{\alpha}(-\alpha-1)}} e^{-t}\mathrm{d}t \qquad (3.28) \\ &= \int_0^{\infty} t^{-\frac{p}{\alpha}} e^{-t}\mathrm{d}t = \int_0^{\infty} t^{\left(1-\frac{p}{\alpha}\right)} \frac{e^{-t}}{t}\mathrm{d}t = \Gamma\left(1 - \frac{p}{\alpha}\right), \end{aligned}$$

kde gamma funkce a integrál daný gamma funkcí, konverguje absolutně pro $1 - \frac{p}{\alpha} > 0$. Tedy p-tý moment existuje pouze pro $p < \alpha$.



Obrázek 3.2: Distribuční funkce $\Phi_\alpha(x)$ a k ní kvantilová funkce chvostu $U(t) = \log(\frac{t}{t-1})^{-\frac{1}{\alpha}}$, $\alpha = \frac{1}{\gamma}$.

Průběhy distribučních funkcí a kvantilových funkcí chvostu pro rozdělení Frechétova typu jsou na obrázku 3.2. Vzhledem k dosazení $\frac{x-1}{\gamma}$ za x se změní poloha a měřítko distribuční funkce G_γ. V uvedeném obrázku jsou červeně ($\gamma = \frac{1}{\alpha} = 1$) a černě ($\gamma = \frac{1}{\alpha} = 3$) znázorněna rozdělení, která nemají konečnou střední hodnotu. Modře je znázorněno rozdělení, které již má konečnou střední hodnotu, ale nemá konečný rozptyl.

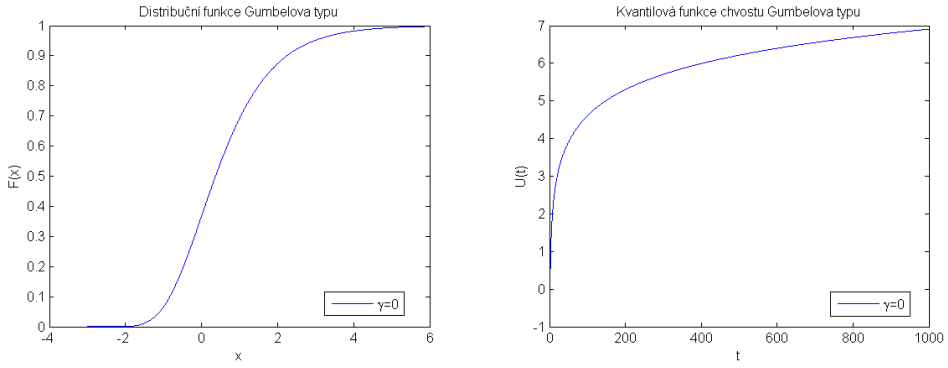
b) $\gamma = 0$: rozdělení *Gumbelova* typu s distribuční funkcí

$$G_0(x) = e^{-e^{-x}}.$$

Pravý koncový bod je zde roven nekonečnu, tedy $G_0(x) < 1$, pro všechna $x \in \mathbb{R}$. Jedná se o rozdělení s lehčím chvostem, neboť velké hodnoty x jsou méně pravděpodobné. Všechny p-té momenty, $p > 0$, existují, neboť

$$\begin{aligned} \mathrm{E}(X^p) &= \int_{-\infty}^{\infty} x^p f(x)\mathrm{d}x = \int_{-\infty}^{\infty} x^p \left(e^{-e^{-x}}\right)' \mathrm{d}x = \int_{-\infty}^{\infty} x^p e^{-x} e^{-e^{-x}} \mathrm{d}x \\ &= \left| x = -\log t;\ \mathrm{d}x = -\frac{1}{t}\mathrm{d}t \right| = \int_0^{\infty} (-1)^p (\log t)^p e^{-t} \mathrm{d}t < \infty. \end{aligned} \tag{3.29}$$

V případě $p = 1$ je $\mathrm{E}(X^p) = \mathrm{E}(X)$ a integrál (3.29) je roven Euler–Mascheroniho konstantě $\gamma_{eul} \approx 0.577$. V případě p=2 je $\mathrm{E}(X^2) = \gamma_{eul}^2 + \frac{\pi^2}{6}$. Momenty pro Gumblovo rozdělení je také vhodné počítat pomocí charakteristické funkce, tj. Fourierovy transformace hustoty. Viz [17] str. 11.



Obrázek 3.3: Distribuční funkce $G_0(x)$ a k ní kvantilová funkce chvostu $U(t) = -\log(-\log(1 - \frac{1}{t}))$.

c) $\gamma < 0$: pravý koncový bod je roven $\frac{-1}{\gamma}$, takže rozdělení má kratké chvosty. Položíme-li $-\frac{x+1}{\gamma}$ místo x a dosadíme do (3.26), dostaneme takzvané extremální (reverzní) Weibullovo rozdělení s distribuční funkcí

$$\Psi_\alpha(x) = \begin{cases} e^{-(-x)^\alpha}, & x < 0 \\ 1, & x \geq 0, \end{cases} \tag{3.30}$$

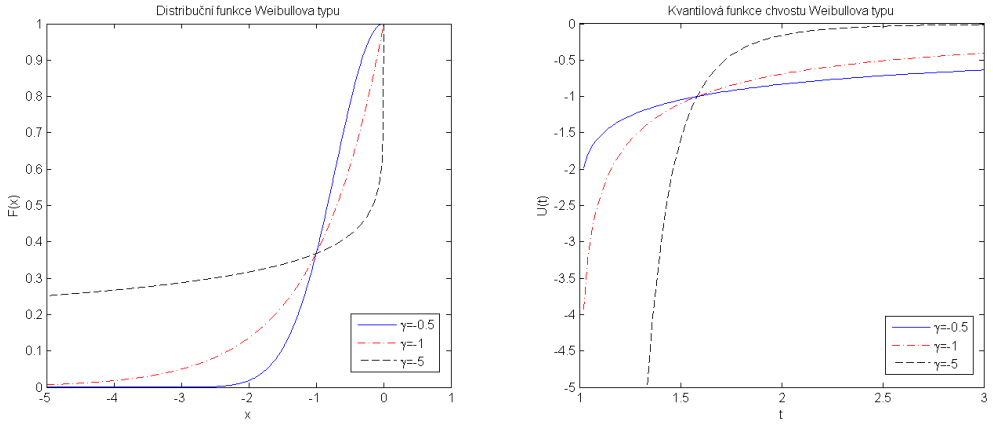
kde $\alpha = -\frac{1}{\gamma} > 0$. Po této transformaci je pravý koncový bod roven 0. Ukažme, že pro $p > 0$ všechny obecné momenty existují:

$$\begin{aligned} \mathrm{E}(X^p) &= \int_{-\infty}^{\infty} x^p f(x)\mathrm{d}x = \int_{-\infty}^{0} x^p \left(e^{-(-x)^\alpha}\right)' \mathrm{d}x = \int_{-\infty}^{0} \alpha x^p (-x)^{\alpha-1} e^{-(-x)^\alpha} \mathrm{d}x \\ &= \left| t = (-x)^\alpha;\ \mathrm{d}t = -\alpha(-x)^{\alpha-1}\mathrm{d}x;\ x = -t^{\frac{1}{\alpha}} \right| = -\int_0^{\infty} (-1)^p t^{\frac{p}{\alpha}} t \frac{t^{-\frac{1}{\alpha}}}{t^{\frac{1}{\alpha}(\alpha-1)}} e^{-t}\mathrm{d}t \\ &= (-1)^{p+1} \int_0^{\infty} t^{\frac{p}{\alpha}} e^{-t}\mathrm{d}t = (-1)^{p+1} \int_0^{\infty} t^{\left(1+\frac{p}{\alpha}\right)} \frac{e^{-t}}{t}\mathrm{d}t = (-1)^{p+1}\Gamma\left(1 + \frac{p}{\alpha}\right). \end{aligned}$$

Protože integrál, který gamma funkce představuje, konverguje absolutně v případě $1 + \frac{p}{\alpha} > 0$, tak p-tý moment existuje pro všechna $p > 0$.

Na obrázku 3.4 vidíme, že kvantilová funkce chvostu jde k nule, když $t \to \infty$. Pravý koncový bod je skutečně roven nule. Pro malé hodnoty γ konverguje distribuční

funkce ,,pomalu" k nule, když $x \to -\infty$. Malé hodnoty x se tak vyskytují častěji. Ty mají velký význam, pokud pracujeme s minimem.



Obrázek 3.4: Distribuční funkce $\Psi_{-\frac{1}{\gamma}}(x)$ s kvantilovou funkcí chvostu $U(t) = -\left(-\log\left(1-\frac{1}{t}\right)\right)^{-\gamma}$.

3.3. Obor atraktivity

V nadcházející části textu se budeme zabývat oborem atraktivity pro rozdělení Fréchetova, Gumbelova a Weibullova typu a volbou reálných posloupností $a_n > 0$ a b_n. Uvedeme některé důležité podmínky na distribuční funkci náhodného výběru, popřípadě kvantilovou funkci chvostu. Podle [5] uvedeme několik základních rozdělení, která leží v oboru atraktivity rozdělení extrémních hodnot.

Nejprve uveďme dvě následující věty, kde první má spíše teoretický charakter a druhá je potřeba při odhadu indexu extrémní hodnoty γ a v jiných situacích.

Věta 3.4 *Nechť F je distribuční funkce s pravým koncovým bodem x^*. Předpokládejme, že $F''(x)$ existuje a $F'(x) \geq 0$ pro všechna x ležící v libovolném levém okolí bodu x^*. Jestliže*

$$\lim_{t \uparrow x^*} \left(\frac{1-F}{F'}\right)'(t) = \gamma,$$

pak F leží v oboru atraktivity G_γ.

Důkaz. Viz. [15].

Věta 3.5 *Distribuční funkce F s pravým koncovým bodem x^* a s kvantilovou funkcí chvostu $U(t)$ je v oboru atraktivity distribuční funkce G_γ, právě když*

1. *pro $\gamma \in \mathbb{R}$:*

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = D_\gamma(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0 \\ \log x, & \gamma = 0, \end{cases} \tag{3.31}$$

pro vhodnou kladnou funkci $a(t)$ a pro všechna $x > 0$.

2. *pro* $\gamma > 0$*:* $x^* = \infty$ *a*

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{\frac{-1}{\gamma}}, \tag{3.32}$$

pro každé $x > 0$*.*

3. *pro* $\gamma > 0$*:* $x^* = \infty$*,* $\int_1^\infty \frac{(1-F(x))}{x} dx < 0$ *a*

$$\lim_{t\to\infty} \frac{\int_t^\infty (1-F(x))\frac{dx}{x}}{1-F(t)} = \gamma; \tag{3.33}$$

4. *pro* $\gamma > 0$*:* $x^* = \infty$ *a*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma, \quad x > 0,, \tag{3.34}$$

navíc pro $a(t)$ *splňující (3.31) platí*

$$\lim_{t\to\infty} \frac{a(t)}{U(t)} = \gamma. \tag{3.35}$$

5. *pro* $\gamma < 0$*:* $x^* < \infty$ *a*

$$\lim_{t\to\infty} \frac{x^* - U(tx)}{x^* - U(t)} = x^\gamma, \quad x > 0. \tag{3.36}$$

6. *pro* $\gamma \leq 0$*:*

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = 1, \quad x > 0. \tag{3.37}$$

Důkaz. Vztah (3.31) přímo plyne z věty 3.3. Podrobnosti a důkaz zbývajících tvrzení lze nalézt v [15] sekce 1.2. ■

Důsledek věty 3.5. Je-li $F \in \mathcal{D}_\gamma$, $\gamma > 0$, pak pro náhodnou veličinu X s distribuční funkcí F platí (vzhledem k vztahu (3.33)):

$$\lim_{t\to\infty} \mathrm{E}(\log X - \log t | X > t) = \gamma.$$

Důkaz. Rozepišme

$$\begin{aligned} \lim_{t\to\infty} \mathrm{E}(\log X - \log t | X > t) &= \lim_{t\to\infty} \frac{\int_t^\infty (\log x - \log t) f(x) \mathrm{d}x}{P(X > t)} \\ &= \lim_{t\to\infty} \frac{\int_t^\infty f(x) \log x \mathrm{d}x - \log t \int_t^\infty f(x) \mathrm{d}x}{1 - P(X \leq t)} \end{aligned} \tag{3.38}$$

a použijeme-li integraci per partes

$$\begin{aligned} (3.38) &= \lim_{t\to\infty} \frac{\lim_{u\to\infty}[F(x)\log x]_t^u - \int_t^\infty F(x)\frac{1}{x}\mathrm{d}x - \log t \lim_{u\to\infty}[F(x)]_t^u}{1-F(t)} = \\ &= \lim_{t\to\infty} \frac{\lim_{u\to\infty}(F(u)\log u) - \log t - \int_t^\infty \frac{F(x)}{x}\mathrm{d}x}{1-F(t)} = \lim_{t\to\infty} \frac{\int_t^\infty \frac{1}{x}\mathrm{d}x - \int_t^\infty \frac{F(x)}{x}\mathrm{d}x}{1-F(t)} \\ &= \lim_{t\to\infty} \frac{\int_t^\infty (1-F(x))\frac{\mathrm{d}x}{x}}{1-F(t)}. \end{aligned}$$

Z (3.33) pak vidíme, že

$$\gamma = \lim_{t\to\infty} \mathrm{E}(\log X - \log t | X > t) = \lim_{t\to\infty} \frac{\int_t^\infty (\log x - \log t) f(x)\mathrm{d}x}{1 - F(t))}. \tag{3.39}$$

■

Předchozí důsledek je základem pro Hillův odhad parametru γ. Tímto odhadem se budeme zabývat v kapitole 5.2.

Někdy samotný požadavek $F \in \mathcal{D}(G_\gamma)$ nestačí. Je třeba klást silnější podmínku na distribuční funkci, tzn. podmínku druhého druhu. Nejdříve pomocí (3.31) definujme:

Definice 3.3 *Řekneme, že kvantilová funkce chvostu U (popřípadě k ní příslušná distribuční funkce F) splňuje podmínku druhého druhu, jestliže existují funkce $a(t) > 0$ a funkce A taková, že nemění znaménko a $\lim_{t\to\infty} A(t) = 0$, a platí*

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = H(x), \quad x > 0,$$

kde

$$D_\gamma(x) = \left\{ \begin{array}{ll} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0 \\ \log x, & \gamma = 0 \end{array} \right\}$$

a H je funkce, která není násobkem funkce $D_\gamma(x)$. Zejména H není identicky rovna nule. Paramter ρ nazýváme parametrem druhého druhu.

Pokud kvantilová funkce chvostu U splňuje podmínku druhého druhu, pak je splněna i podmínka atraktivity, tedy $F \in \mathcal{D}_\gamma$. Obrácená implikace ale neplatí. Příkladem, kdy neplatí podmínka druhého druhu, je rozdělení s distribuční funkcí $F(x) = 1 + x^{-1}\exp(\sin\log x)$, $x > 0$ [8].

V následující větě je uveden speciální případ podmínky druhého druhu, kdy se omezujeme pouze na obor atraktivity Fréchetova typu.

Věta 3.6 *Předpokládejme, že distribuční funkce F splňuje podmínku druhého druhu. Potom pro všechna $x > 0$ platí*

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{3.40}$$

kde $\gamma > 0$, $\rho \leq 0$ a A je funkce, která nemění znaménko, přičemž $\lim_{t\to\infty} A(t) = 0$.

Důkaz. Viz. [15].

Definice 3.4 *Lebesgueovsky měřitelnou funkci $l : \mathbb{R}^+ \to \mathbb{R}$ na $(0, \infty)$ nazveme regulárně se měnící (v nekonečnu), jestliže pro nějaké $\alpha \in \mathbb{R}$ platí*

$$\lim_{t\to\infty} \frac{l(tx)}{l(t)} = x^\alpha, \quad x > 0. \tag{3.41}$$

Číslo α se nazývá index změny. Funkci, která pro $\alpha = 0$ splňuje (3.41), nazýváme pomalu se měnící.

Vztah (3.34) znamená, že kvantilová funkce chvostu U je regulárně se měnící s indexem γ. Proto se u Fréchetovy třídy často klade požadavek, aby U byla regulárně se měnící.

Uveďme ještě tvrzení spíše teoretického charakteru, protože v praxi většinou neznáme distribuční funkci F.

Tvrzení 3.5 *Nechť F je distribuční funkcí, U kvantilovou funkcí chvostu k F, a nechť $F \in \mathcal{D}(G)$, kde G je distribuce Fréchetova typu. Pak existují regulárně se měnící funkce $l_F(x)$ a $l_U(t)$ takové, že*

$$1 - F(x) = x^{-\frac{1}{\gamma}} l_F(x) \quad a \quad U(t) = t^{\gamma} l_U(t), \tag{3.42}$$

přičemž mezi funkcemi l_F a l_U existuje jednoznačný vztah.

Důkaz. Viz. [5] sekce 2.3.2.
Poznámka. Zavést rozdělení z Fréchetova oboru atraktivity je možné nejen pomocí distribuční funkce, ale také pomocí kvantilové funkce chvostu U. Pro snažší odvození tvaru funkce l_F pomocí kvantilové funkce U lze z (3.42) vyvodit vztah

$$t^{\gamma} = U(t) l_F^{-\gamma}\left(U(t)\right).$$

Podrobnější analýzu regulárně se měnících funkcí lze najít v [5],[15].

3.3.1. Obor atraktivity pro rozdělení Fréchetova typu

V tabulce 1 jsou některá rozdělení patřící do Fréchetova oboru atraktivity se vztahy pro distribuční funkci F a kvantilovou funkcí chvostu U. Kvantilovou funkci $F^{\leftarrow}$ pak dostaneme snadno ze vztahu $F^{\leftarrow}(u) = U\left(\frac{1}{1-u}\right)$, $0 \le u \le 1$. Všimněme si, že je zde obsaženo také Fréchetovo rozdělení s distribuční funkcí $\Phi_\alpha(x)$ podle (3.27). Cauchyho rozdělení je speciálním případem Studentova rozdělení s jedním stupněm volnosti. Velký význam pro parametrické modelování chvostu má zobecněné Paretovo rozdělení, na kterém je založena prahová metoda maximální věrohodnoti v sekci (5.1). Zobecněné Paretovo rozdělení obecně může mít parametr tvaru γ také nulový nebo záporný. Pro $\gamma \in \mathbb{R}$ můžeme napsat distribuční funkci zobecněného Paretova rozdělení H_γ jako

$$H_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma}, & 0 < x < (\sup(0, -\gamma))^{-1}, \gamma \neq 0, \\ 1 - e^{-x}, & 0 < x < \infty, \quad \gamma = 0. \end{cases} \tag{3.43}$$

Odtud, podobně jako u extrémálního rozdělení, rozlišujeme u zobecněného Paretova rozdělení tři případy v závislosti na γ:

$$\begin{array}{lrll} \gamma > 0: & H_\gamma\left(\frac{(x-1)}{\gamma}\right) & = 1 - x^{-\frac{1}{\gamma}}, & x > 0, \quad \text{rozdělení je Paretovo} \\ \gamma = 0: & H_0(x) & = 1 - e^{-x}, & x > 0, \quad \text{rozdělení je exponenciální} \\ \gamma < 0: & H_\gamma\left(\frac{-(x+1)}{\gamma}\right) & = 1 - (-x)^{-\frac{1}{\gamma}} & 0 < x < \frac{-1}{\gamma}. \end{array}$$

V případě, kdy parametr $\gamma = 0$, dostaneme exponenciální rozdělení, které leží v Gumbelově oboru atraktivity. Dalším speciálním případem je i samotné Paretovo rozdělení, které dostaneme pro $\gamma > 0$. Toto rozdělení bylo původně používáno k popisu rozdělení bohatství mezi jednotlivce. Pokud zlogaritmujeme náhodnou veličinu s Paretový rozdělením, dostaneme transformovanou náhodnou veličinu, která má exponenciální rozdělení.

Tabulka 1: Rozdělení patřící do Fréchetova oboru atraktivity ($\gamma > 0$)

Rozdělení	Distribuční funkce $F(x)$ a kvantilová funkce chvostu $U(t)$	Index extrémní hodnoty
Pa(γ)	$F(x) = 1 - x^{-\frac{1}{\gamma}}, \quad \gamma > 0;\ x \geq 1$	γ
(Paretovo)	$U(t) = t^{\gamma}, \quad t > 1$	
GP(γ)	$F(x) = 1 - (1+\gamma x)^{-\frac{1}{\gamma}}, \quad \gamma > 0;\ x > 0$	γ
(Zobecněné Pa)	$U(t) = \frac{1}{\gamma}(t^{\gamma} - 1), \quad t > 1$	
Burr(η, τ, λ)	$F(x) = 1 - \left(\frac{\eta}{\eta + x^{\tau}}\right)^{\lambda}, \quad \eta, \tau, \lambda > 0;\ x > 0$	$\frac{1}{\lambda\tau}$
(Typ XII)	$U(t) = \left(\eta(t^{\frac{1}{\lambda}} - 1)\right)^{\frac{1}{\tau}}, \quad t > 1$	
Burr(η, τ, λ)	$F(x) = \left(\frac{\eta}{\eta + x^{-\tau}}\right)^{\lambda}, \quad \eta, \tau, \lambda > 0;\ x > 0$	$\frac{1}{\tau}$
(Typ III)	$U(t) = \left(\frac{\eta}{(1-t^{-1})^{\frac{1}{\lambda}}} - \eta\right)^{-\frac{1}{\tau}}, \quad t > 1$	
F(m, n)	$F(x) = \int_0^x \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2} \mathrm{d}w, \quad m, n > 0;\ x > 0$	$\frac{2}{n}$
(Fisher-Snedecor)	předpis pro $U(t)$ není znám	
Inv$\Gamma(\lambda, \alpha)$	$F(x) = \int_0^x \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda/w} w^{-\alpha-1} \mathrm{d}w,\ \lambda, \alpha > 0;\ x > 0$	$\frac{1}{\alpha}$
	předpis pro $U(t)$ není znám	
log$\Gamma(\lambda, \alpha)$	$F(x) = \int_1^x \frac{\lambda^{\alpha}}{\Gamma(\alpha)} w^{-\lambda-1} (\log w)^{\alpha-1} dw, \lambda, \alpha > 0;\ x > 1$	$\frac{1}{\lambda}$
	předpis pro $U(t)$ není znám	
Fréchet(γ)	$F(x) = \exp\left(-x^{-\frac{1}{\gamma}}\right), \gamma > 0;$ x¿0	γ
	$U(t) = \left(\log \frac{t}{t-1}\right)^{-\gamma}, \quad t > 1$	
t_n	$F(x) = \int_{-\infty}^x \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{w^2}{n}\right)^{-\frac{n+1}{2}} \mathrm{d}w,\ n > 0;\ x \in \mathbb{R}$	$\frac{1}{n}$
(Studentovo)	předpis pro $U(t)$ není znám	
Cauchy	$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x;\ x \in \mathbb{R}$	1
(t_1)	$U(t) = \tan\left(\pi(\frac{t-1}{t} - \frac{1}{2})\right), \quad t > 1$	

V následujícím tvrzení bude ukázáno, jak lze pro Fréchetovu třídu rozdělení vhodně standardizovat maxima.

Tvrzení 3.6 *Nechť distribuční funkce F patří do Fréchetova oboru atraktivity, pak*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-x^{-\frac{1}{\gamma}}\right)$$

pro $x > 0$ při volbě

$$a_n = U(n) \quad a \quad b_n = 0.$$

Důkaz. Vyjdeme z věty 3.31. Pro $\gamma > 0$ a $x > 0$ platí (3.34):

$$\lim_{n\to\infty} \frac{U(nx)}{U(n)} = x^{\gamma}. \tag{3.44}$$

Uvažujme lemma 2.10. Připomeňme, že z definice kvantilové funkce chvostu plyne $U = \left(\frac{1}{1-F}\right)^{\leftarrow}$. Sestrojme inverzi k $\frac{U(nx)}{U(n)}$ tak, že položíme $x = \frac{U(ny)}{U(n)}$. Pak

$$xU(n) = U\left(\frac{1}{1 - F(xU(n))}\right) = U(ny).$$

Odtud

$$\frac{1}{1 - F(xU(n))} = ny$$

a použijeme-li lemma 2.10, můžeme (3.44) přepsat na

$$\lim_{n\to\infty} \frac{1}{y(n)} = \lim_{n\to\infty} n\{1 - F(xU(n))\} = x^{-\frac{1}{\gamma}}.$$

Vzhledem k větě 3.1 je $\lim_{n\to\infty} n\{1 - F(xU(n))\} = -\log G(x)$. Odtud dostáváme distribuční funkci Fréchetova typu $G(x) = \exp\left(-x^{-\frac{1}{\gamma}}\right)$, $x > 0$. Dále z 3.1 je

$$\lim_{n\to\infty} F^n(xU(n)) = \exp\left(-x^{-\frac{1}{\gamma}}\right), \quad x > 0.$$

Odtud vidíme, že $a_n = U(n)$ a $b_n = 0$. ■

3.3.2. Obor atraktivity pro rozdělení Gumbelova typu

Vybraná rozdělení patřící do Gumblova oboru atraktivity jsou uvedena v tabulce 2. Rozdělení gamma představuje širokou třídu rozdělení, pod kterou spadá i exponenciální. To dostaneme snadno, pokud položíme parametr rozdělení gamma $m = 1$. Pokud zlogaritmujeme náhodnou veličinu s log-normálním rozdělením, dostaneme, jak samotný název napovídá, náhodnou veličinu, která má normální rozdělení. Pomocí log-normálního rozdělení můžeme v ekonomice modelovat například návratnost. Pomocí rozdělení gamma, log-normálního či Weibullova můžeme modelovat dešťové úhrny.

Tabulka 2: Některá rozdělení patřící do Gumbelova oboru atraktivity ($\gamma = 0$)	
Rozdělení	Distribuční funkce $F(x)$ a kvantilová funkce chvostu $U(t)$
Weibull(λ, α)	$F(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right), \quad \lambda, \alpha > 0;\ x > 0$
	$U(t) = \lambda \left(-\log \frac{1}{t}\right)^{\frac{1}{\alpha}}, \quad t > 1$
exp(λ)	$F(x) = 1 - \exp\left(-\lambda x\right), \quad \lambda > 0 \ \ x > 0$
(Exponenciální)	$U(t) = \frac{-\log \frac{1}{t}}{\lambda}, \quad t > 1$
Log-normal(σ, μ)	$F(x) = \int_0^x \frac{1}{\sqrt{2\pi}\sigma u} \exp\left(-\frac{1}{2\sigma^2}(\log(u) - \mu)^2\right) \mathrm{d}u, \quad \sigma > 0, \mu \in \mathbb{R};\ \ x > 0$
(Logaritmicko-normální)	předpis pro $U(t)$ není znám
Gamma(m, λ)	$F(x) = \int_0^x \frac{\lambda^m}{\Gamma(m)} \exp(-\lambda u) u^{m-1} \mathrm{d}u, \quad \lambda, m > 0;\ \ x > 0$
	předpis pro $U(t)$ není znám
Logistické	$F(x) = \frac{1}{1+e^{-x}} = \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{x}{2}\right), \quad x \in \mathbb{R}$
	$U(t) = \log(t-1), \quad t > 1$

Z následujících tvrzení plyne, jak pro Gumbelovou třídu rozdělení vhodně standardizovat maxima.

Tvrzení 3.7 *Nechť distribuční funkce F patří do Gumbelova oboru atraktivity, pak*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_0(x) = e^{-e^{-x}}$$

platí pro $x \in \mathbb{R}$ při volbě

$$a_n = h(U(n)) \quad a \quad b_n = U(n),$$

kde h je neklesající funkce splňující (3.25).

Důkaz. [15] str. 21.

Příklad 3.8 *Nechť F je distribuční funkce standardizovaného normálního rozdělení a*

$$b_n = (2\log n - \log\log n - \log(4\pi))^{\frac{1}{2}}, \quad a_n = \frac{1}{b_n} \tag{3.45}$$

reálné posloupnosti. Pak pro $x > 0$

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = \text{-log}G(x) = e^{-x}.$$

$G(x)$ je tedy Gumbelova distribuční funkce s parametrem $\gamma = 0$.

Odvození. Standardizované normální rozdělení má hustotu $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$. Derivováním dostaneme

$$\begin{aligned}
-\frac{d}{dx} n(1 - F(a_n x + b_n)) &= n a_n f(a_n x + b_n) = \\
&= \frac{n}{b_n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(\frac{x}{b_n} + b_n\right)^2}{2} \right) = \\
&= \exp\left(\log \frac{n}{b_n \sqrt{2\pi}} \right) \exp\left(-\frac{x^2}{2{b_n}^2} - x - \frac{{b_n}^2}{2} \right) = \\
&= e^{(\log n - \log b_n - \frac{1}{2}\log(2\pi))} e^{-\frac{{b_n}^2}{2}} e^{-\frac{x^2}{2{b_n}^2}} e^{-x} = \\
&= e^{-(\frac{{b_n}^2}{2} + \log b_n - \log n + \frac{1}{2}\log(2\pi))} e^{-\frac{x^2}{2{b_n}^2}} e^{-x} \\
&\to e^{-x}, \quad \text{pro } n \to \infty
\end{aligned}$$

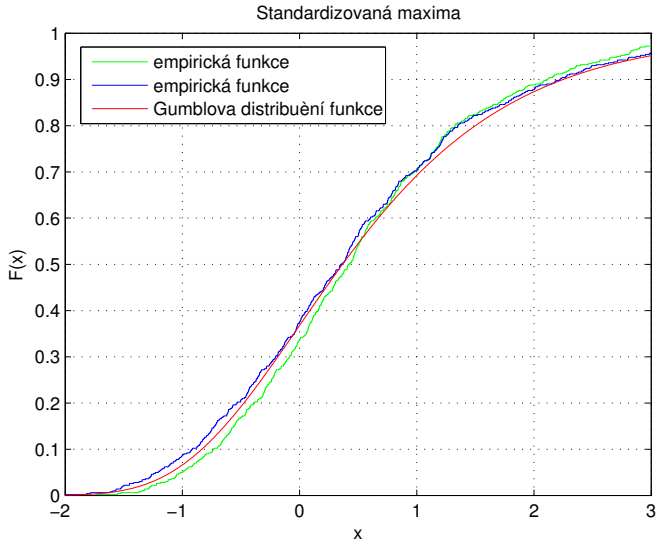
protože

$$\frac{{b_n}^2}{2} + \log b_n - \log n + \frac{1}{2}\log(2\pi) \to 0, \quad -\frac{1}{2{b_n}^2} \to 0, \quad \text{pro } n \to \infty.$$

Tímto na základě Lebesgueovy věty o záměně limity a integrálu

$$\begin{aligned}
n(1 - F(a_n x + b_n)) &= 1 - e^{-(\frac{{b_n}^2}{2} + \log b_n - \log n + \frac{1}{2}\log(2\pi))} \int_{-\infty}^{x} e^{-\frac{u^2}{2{b_n}^2}} e^{-u} \mathrm{d}u \\
\lim_{n\to\infty} n(1 - F(a_n x + b_n)) &= \int_x^\infty e^{-u} \mathrm{d}u = e^{-x}.
\end{aligned}$$

Na obrázku 3.5 je znázorněná konvergence standardizovaných maxim ke Gumblově distribuční funkci, která je vyznačena červeně. Bylo nasimulováno $M_n^{(1)}, M_n^{(2)} \dots M_n^{(500)}$ blokových maxim ze standardizového normálního rozdělení rozsahu $n = 1000$. Empirická distribuční funkce byla sestavena ze standardizovaných maxim $\frac{M_n^{(1)}-b_n}{a_n}, \frac{M_n^{(2)}-b_n}{a_n}, \dots, \frac{M_n^{(500)}-b_n}{a_n}$. Zelená barva pak představuje empirickou distribuční funkci pro volbu a_n a b_n podle (3.45). Volíme-li $a_n = h(U(n))$ a $b_n = U(n) = F^{\rightarrow}\left(1 - \frac{1}{n}\right)$ podle tvrzení 3.7, pak můžeme podle [15] str. 22 zvolit $h(U(n)) = \frac{1-F(U(n))}{F'(U(n))}$. DeHaan namísto h používá značení f. Pro takto zvolené konstanty je empirická distribuční funkce znázorňena modře.



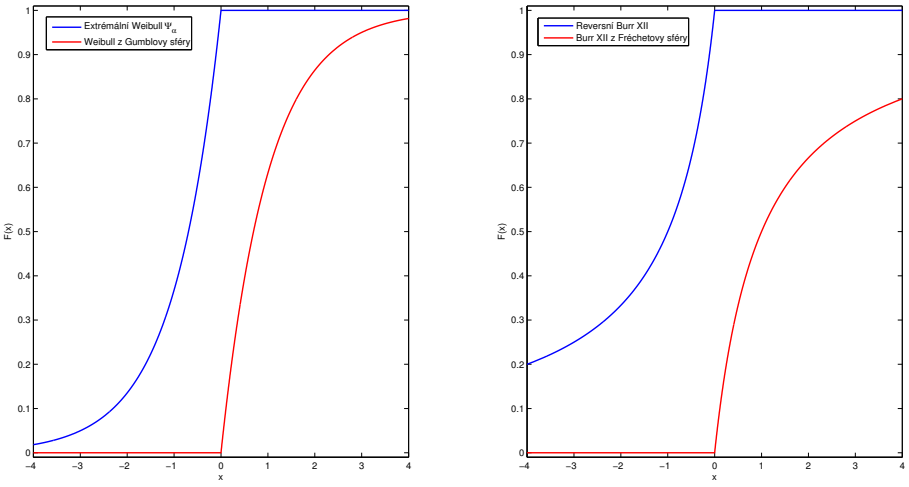
Obrázek 3.5: Gumblova distribuční funkce $G_0(x)$ (červeně). Empirická distribuční funkce ze standardizovaných maxim pro volbu a_n a b_n podle (3.45) - zeleně nebo podle tvrzení 3.7 - modře.

3.3.3. Obor atraktivity pro rozdělení Weibullova typu

Některá rozdělení patřící do Weibullovy třídy jsou uvedena v tabulce 3. Extremální Weibullovo rozdělení, někdy také nazývané reverzní Weibullovo rozdělení, je zde totožné s rozdělením s distribuční funkcí Ψ_α podle (3.30). Obyčejné Weibullovo rozdělení, které se používá ve spolehlivosti, má však nenulovou distribuční funkci pro $x > 0$, respektive $F(x) = 1 - e^{-x^\alpha}$, $x > 0, \alpha > 0$. Častěji totiž ve spolehlivosti pracujeme s minimem než maximem. Toto rozdělení vzhledem k nekonečnému pravému koncovému bodu spadá do Gumblovy sféry. Obdobně je tomu u reverzního Burrova rozdělení. Z obrázku 3.6 je patrná určitá symetrie. Distribuční funkce jsou vůči sobě pouze pootočeny.

Tabulka 3: Rozdělení patřící do Weibullova oboru atraktivity ($\gamma < 0$)

Rozdělení	Distribuční funkce $F(x)$ a kvantilová funkce chvostu $U(t)$	Index extrémní hodnoty
R(a, b)	$F(x) = \frac{x-a}{b-a}, \quad -\infty < a < b < \infty;\ a \le x \le b$	-1
(Rovnoměrné)	$U(t) = \frac{t-1}{t}(b-a) + a, \quad t > 1$	
Beta(p, q)	$F(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x u^{p-1}(1-u)^{q-1}\mathrm{d}u,\ p, q > 0;\ 0 \le x \le 1$	$-\frac{1}{q}$
	předpis pro $U(t)$ není znám	
Reverzní Burr(η, τ, λ)	$F(x) = 1 - \left(\frac{\eta}{\eta + (-x)^{-\tau}}\right)^{\lambda}, \eta, \tau, \lambda > 0; \quad x \le 0$	$-\frac{1}{\lambda\tau}$
(Typ XII)	$U(t) = -\left(\eta(t^{\frac{1}{\lambda}} - 1)\right)^{-\frac{1}{\tau}}, \quad t > 1$	
Weibull(α)	$F(x) = e^{-(-x)^{\alpha}}, \quad \alpha > 0;\ x \le 0$	$-\frac{1}{\alpha}$
(extremální)	$U(t) = -\left(-\log\left(1 - \frac{1}{t}\right)\right)^{\frac{1}{\alpha}}, \quad t > 1$	



Obrázek 3.6: Srovnání distribučních funkcí. Vlevo extremální ($\gamma = -1$) a obyčejné ($\gamma = 0$) Weibullovo rozdělení; $\alpha = 1$. Vpravo reversní ($\gamma = -1$) a obyčejné ($\gamma = 1$) Burrovo rozdělení typu XII; $\lambda = \tau = 1$.

Speciálním případem Beta rozdělení je rovnoměrné rozdělení: Beta$(1,1)$ $\equiv$R$(0,1)$. Zajímavý případ nastává pro Beta$(1/2, 1/2)$, který vede na arcsinus rozdělení. Toto rozdělení má význam pro teorii náhodné procházky. Například rozdělení doby, kdy symetrická náhodná procházka nabývá vyšších hodnot než počáteční, má základ v Beta$(1/2, 1/2)$ rozdělení.

Z následujícího tvrzení plyne, jak pro Weibullovu třídu rozdělení vhodně standardizovat maxima.

Tvrzení 3.9 *Nechť distribuční funkce F patří do Weibullova oboru atraktivity, pak*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = e^{-(-x)^{-\frac{1}{\gamma}}}$$

platí pro $x < 0$ při volbě

$$a_n = x^* - U(n) \quad a \quad b_n = x^*.$$

Důkaz. Podobně jako u důkazu tvrzení 3.6 vyjdeme z věty 3.31. Pro $\gamma < 0$ a $x > 0$ platí (3.36):

$$\lim_{t\to\infty} \frac{x^* - U(tx)}{x^* - U(t)} = x^{\gamma}. \tag{3.46}$$

Sestrojme inverzi k $\frac{x^*-U(tx)}{x^*-U(t)}$ tak, že položíme $x = \frac{x^*-U(ty)}{x^*-U(t)}$. Pak

$$x^* - x(x^* - U(n)) = U\left(\frac{1}{1 - F(x^* - x(x^* - U(n)))}\right) = U(ny).$$

Odtud

$$\frac{1}{1 - F(x^* - x(x^* - U(n)))} = ny$$

a použijeme-li lemma 2.10, můžeme (3.46) přepsat na

$$\lim_{n\to\infty} \frac{1}{y(n)} = \lim_{n\to\infty} n\{1 - F(x^* - x(x^* - U(n)))\} = x^{-\frac{1}{\gamma}}.$$

Zřejmě $x^* - U(n) \geq 0$ pro všechna n. Vzhledem k větě 3.1 je

$$\lim_{n\to\infty} n\{1 - F(x^* + x(x^* - U(n)))\} = -\log G(-x)$$

pro $x < 0$. Odtud dostáváme distribuční funkci Weibullova typu $G(x) = \exp\left(-(-x)^{-\frac{1}{\gamma}}\right)$, $x < 0$. Dále z 3.1 je

$$\lim_{n\to\infty} F^n(x^* + x(x^* - U(n))) = \exp\left(-(-x)^{-\frac{1}{\gamma}}\right), \quad x < 0.$$

Odtud vidíme, že $a_n = x^* - U(n)$ a $b_n = x^*$. ■

4. Metoda bootstrap

Metoda bootstrap je intenzivní numerickou počítačovou metodou pro statistickou analýzu dat. Jako první přišel s touto metodou Bradley Efron v roce 1979, který se inspiroval metodou jacknife. Jeho první článek o bootstrapu [11] vyvolal velký ohlas a našel uplatnění v mnoha oblastech aplikací matematické statistiky. V prvotních počátcích byla metoda v praxi obtížně využitelná kvůli časové náročnosti. Postupem času s vývojem informačních technologií dochází k jejímu velkému rozvoji. Booststrap je v současnosti stále atraktivním tématem.

Metoda bootstrap kombinuje tzv. substituční princip a metodu Monte Carlo. Nejprve stručně uvedeme tyto metody. V této kapitole bylo čerpáno z [2].

4.1. Metoda Monte Carlo

Metoda Monte Carlo je numerická metoda využívající pseudonáhodná čísla. Základy této metody se objevují v mnoha článcích a knihách jako je [30]. Vychází ze zákona velkých čísel. Základní myšlenkou je odhad střední hodnoty na základě opakování, respektive simulací. Snad nejjednodušším prvotním příkladem je Buffonova úloha o jehle, kdy se snažíme vyčíslit hodnotu π na základě opakovaného házení jehlou. Pokus byl časově náročný a nepřesný. S počtem pozorování klesá chyba a pomocí Čebyševovy nerovnosti lze ukázat, že chyba je úměrná $\frac{1}{\sqrt{n}}$. V dnešní době s rozvojem informačních technologií můžeme dosáhnout daleko vyšších přesností.

Metoda má široké pole uplatnění. Typickým využitím metody Monte Carlo je výpočet integrálů, které jsou analyticky velmi obtížně řešitelné. Pomocí této metody také můžeme simulovat náhodné výběry z některých rozdělení, kde je neznámý tvar kvantilové funkce. Jde například o metody založené na Markovských řetězcích viz. algoritmy Gibbs sampling, Metropolis Hastings a další. Tyto algoritmy lze najít například v [2], [28]. Také ve stochastickém programování a v mnoha dalších odvětví optimalizace najde Monte Carlo své uplatnění. Například můžeme stanovit interval spolehlivosti nalezeného optimalního řešení viz. [24].

4.2. Substituční metoda ("Plug-in principle")

Předpokládejme náhodný výběr $X_1, X_2, \dots X_n$ o distribuční funkci F, která je z nějaké třídy rozdělení pravděpodobnosti $\mathcal{P}$. Nejširší uvažovaná třída může být množina všech distribučních funkcí tj. $\mathcal{P} = \mathcal{F}$. Obvykle předpokládáme, že F je prvkem parametrické třídy rozdělení $\mathcal{P} = \{F_\theta : \theta \in \Theta\}$, kde parametr θ je konečné dimenze. Každé $\theta \in \Theta$ koresponduje jednoznačně s $F_\theta \in \mathcal{P}$.

- Pokud $\theta \in \mathbb{R}^n$, $n \in \mathbb{N}$, mluvíme o parametrických modelech. Například $F_\theta = \{$všechna rozdělení extrémního typu $G_\gamma, \gamma \in \mathbb{R}\}$.
- Pokud $\theta \in \mathbb{R}^n$, kde n značí nekonečnou dimenzi, mluvíme o neparametrických modelech. Například množina všech spojitých rozdělení.
- Jestliže θ může být rozdělena na $\theta_1 \in \mathbb{R}^m$ a $\theta_2 \in \mathbb{R}^n$, kde n značí konečnou dimenzi a m nekonečnou dimenzi, a chceme činit statistickou indukci o θ_1, pak mluvíme o semi-

parametrických modelech. Například množina všech distribučních funkcí ležících v oboru atraktivity rozdělení extrémních hodnot.

Mějme zobrazení $\tau : \Theta \to \mathbb{R}$. Jestliže $\theta = \tau(F)$ je nějaká charakteristika rozdělení, kterou chceme odhadnut na základě realizace náhodného výběru, substituční metoda spočívá v nahrazení F odhadem $\hat{F}$. Dostáváme odhad

$$\hat{\theta} = \tau(\hat{F}).$$

Substituční metoda také funguje s empirickou distribuční funkcí $\hat{F}_n$ jako odhadem F. Například střední hodnotu

$$\theta = \tau(F) = \mathrm{E}(X) = \int x f(x) \mathrm{d}x = \int x \mathrm{d}F(x)$$

odhadneme vzhledem ke schodovitosti empirické distribuční funkce jako

$$\hat{\theta} = \tau(\hat{F}) = \int x \mathrm{d}\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

kde uvedený integrál je Lebesgue-Stieltjesův integrál.

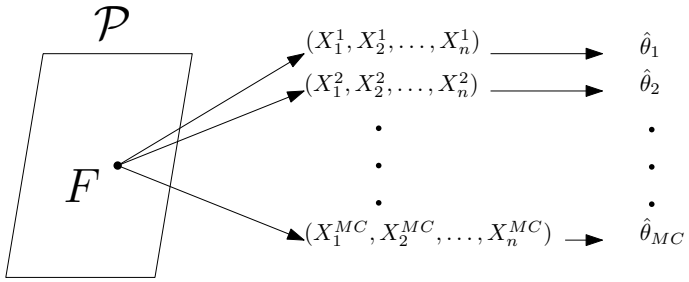
4.3. Princip metody bootstrap

Naším cílem je přistoupit ke kvalitě odhadu $\hat{\theta}$. Zajímáme se tedy o rozptyl odhadu $\mathrm{D}(\hat{\theta})$, interval spolehlivosti I_θ, vychýlení $\mathrm{E}\hat{\theta} - \theta$. V praxi tyto hodnoty většinou nejsou známy a musí být odhadnuty.

Dále ukážeme Monte Carlo aproximaci k vyhodnocení kvality odhadu $\hat{\theta}$. Předpokládejme, že máme velký počet MC nezávislých náhodných výběrů

$$X_1^i, X_2^i, \ldots, X_n^i, \quad i = 1, 2, \ldots, MC$$

ze stejného rozdělení o stejném rozsahu n s příslušnými odhady $\hat{\theta}_i = \hat{\theta}_i(X_1^i, X_2^i, \ldots, X_n^i)$.



Obrázek 4.1: Schéma Monte Carlo

Pro dostatečně velké MC bychom mohli aproximovat hustotu rozdělení θ pomocí histogramu ze všech $\hat{\theta}_i$, $i = 1, 2, \ldots, MC$. Platí

$$\mathrm{D}(\hat{\theta}) = \lim_{MC \to \infty} \frac{1}{MC - 1} \sum_{i=1}^{MC} (\hat{\theta}_i - \overline{\theta}_{MC})^2,$$

$$\text{bias}_{\hat{\theta}} = \text{E}\hat{\theta} - \theta = \lim_{MC\to\infty}(\overline{\theta}_{MC} - \theta),$$

kde $\overline{\theta}_{MC} = \frac{1}{MC}\sum_{i=1}^{MC}\hat{\theta}_i$.

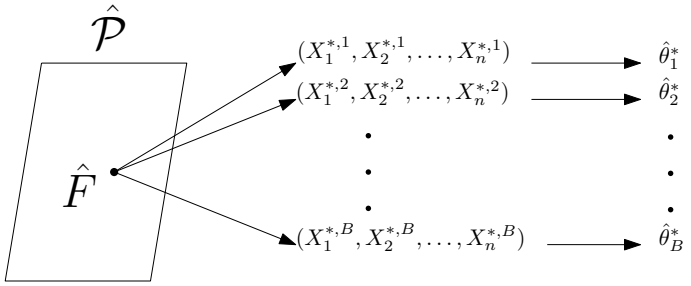
Problémem ovšem je, že distribuční funkci F neznáme. Proto definujme bootstrapový náhodný výběr:

Definice 4.1 *Mějme náhodný výběr $X_1, X_2, \ldots, X_n$ z rozdělení o distribuční funkci F. Nechť $X_1^*, X_2^*, \ldots, X_n^*$ je náhodný výběr z empirické distribuční funkce $\hat{F}_n$, tj. při daných pozorování jsou $X_1^*, X_2^*, \ldots, X_n^*$ nezávislé (podmíněně), stejně rozdělené náhodné veličiny, z nichž každá nabývá hodnot $X_1, X_2, \ldots, X_n$ s pravděpodobností $\frac{1}{n}$. Soubor $X_1^*, X_2^*, \ldots, X_n^*$ nazýváme bootstrapový výběr.*

Abychom dostali počítatelnou verzi, neznámou distibuční funkci F nahradíme empirickou distribuční funkcí $\hat{F}_n$ a z ní vygenerujeme B nezávislých bootstrapových náhodných výběrů $X_1^{*,b^*}, X_2^{*,b^*}, \ldots, X_n^{*,b^*}$, $b^* = 1, 2, \ldots, B$. Dostaneme parametery

$$\theta_{b^*}^* = \hat{\theta}_{b^*}(X_1^{*,b^*}, X_2^{*,b^*}, \ldots, X_n^{*,b^*}), \quad b^* = 1, 2, \ldots, B.$$

Viz. obrázek 4.2.



Obrázek 4.2: Schéma bootstrap

Počet všech různých bootstrapových výběrů je $\binom{2n-1}{n}$. Jedná se o kombinace s opakováním. Například již pro $n = 20$ dostáváme $6.8923 \cdot 10^{10}$ možností. Proto uvažovat všechny bootstrapové výběry je možné jen pro výběry velmi malého rozsahu. Pro rozsáhlejší výběry je potřeba aplikovat metodu Monte Carlo, kdy generujeme libovolný počet B bootstrapových náhodných výběrů. Dostaneme odhady

$$\hat{\text{D}}(\hat{\theta}) = \lim_{B\to\infty}\frac{1}{B-1}\sum_{b^*=1}^{B}(\hat{\theta}_{b^*}^* - \overline{\theta_B^*})^2 = \text{D}_*(\hat{\theta}^*),$$

$$\widehat{\text{bias}_{\hat{\theta}}} = \lim_{B\to\infty}(\overline{\theta_B^*} - \hat{\theta}) = \text{bias}^*(\hat{\theta}^*),$$

kde $\overline{\theta_B^*} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$.

	Původní model	Bootstrap
Model	$\mathcal{P}$	$\hat{\mathcal{P}}$
Rozdělení	F	$\hat{F}_n$
Náhodný výběr	$X_1, \ldots, X_n$	$X_1^*, \ldots, X_n^*$
Parametr	$\theta = \theta(F)$	$\hat{\theta} = \theta(\hat{F})$
Odhad	$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) = \theta(\hat{F}_n)$	$\hat{\theta} = \hat{\theta}(X_1^*, \ldots, X_n^*) = \theta(\hat{F}_n^*)$

Tabulka 4.1: Vztah mezi původním modelem a jeho aproximací metodou bootstrap. $\hat{F}_n^*$ zde představuje empirickou distribuční funkci bootstrapového náhodného výběru.

Výše popsaná metoda je neparametrická. Pokud bychom předpokládali parametrický model

$$\mathcal{P} = \{F_\theta : \theta \in \Theta\},$$

pak je vhodnější opomenout empirickou distribuční funkci a pomocí Monte Carla simulovat náhodné výběry přímo z této třídy rozdělení. Mluvíme o takzvaném parametrickém bootstrapu. Zajisté, pokud je analytický výsledek znám, bootstrap postrádá smysl.

Pomocí metody bootstrap můžeme dále například konstruovat intervaly spolehlivostí odhadů a neparametricky testovat některé statistické hypotézy. Podrobněji v [12].

5. Odhady indexu extrémní hodnoty

Hlavním předmětem statistiky extrémních hodnot je předpovídání vzácných (zřídka se vyskytujících) jevů, které mohou být dokonce za hranicí dostupných pozorování. K tomu je primárně potřeba co nejkvalitnější odhad parametru tvaru tj. indexu extrémní hodnoty γ.

Nejprve bude uvedena prahová metoda maximální věrohodnosti, která vychází z maximalizace věrohodnostní funkce zobecněného Paretova rozdělení. Dále uvedeme několik semi-parametrických metod včetně Pickandsova odhadu, který bude popsán pouze stručně. Metody nepředpokládají konkrétní rozdělení náhodného výběru, ale rozsáhlejší třídu rozdělení pravděpodobnosti tj. $F \in \mathcal{D}(G)$. Všechny tyto metody jsou založeny na úvahách o největších pořadových statistikách a substituční metodě popsané v sekci 4.2. Sdílí ovšem jeden velký společný problém, kterým je správná volba takzvané prahové statistiky. Otázkou je, která pozorování považovat za velká.

5.1. Metoda maximální věrohodnosti ($\gamma > -\frac{1}{2}$)

Uvažujme třídu distribučních funkcí splňujících $F \in \mathcal{D}(G_\gamma), \gamma \in \mathbb{R}$ a posloupnost stejně rozdělených náhodných veličin $X_1, X_2, \ldots, X_n$ o této distribuční funkci F. Pro stanovení odhadu γ pomocí metody maximální věrohodnosti je potřebné následující tvrzení.

Tvrzení 5.1 *Nechť h je kladná neklesající funkce splňující (3.25), pak náhodná veličina $\frac{X-t}{h(t)}$ má za podmínky $X > t$ pro $t \uparrow x^*$ zobecněné Paretovo rozdělení pravděpodobnosti s distribuční funkcí*

$$H_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma}, & 0 < x < (\sup(0, -\gamma))^{-1}, \gamma \neq 0, \\ 1 - e^{-x}, & 0 < x < \infty, \quad \gamma = 0. \end{cases}$$

Důkaz. Z (3.25) dostaneme

$$\lim_{t \uparrow x^*} \frac{1 - F(t + xh(t))}{1 - F(t)} = \lim_{t \uparrow x^*} \frac{P(X > t + xh(t))}{P(X > t)} = (1 + \gamma x)^{-\frac{1}{\gamma}} \tag{5.1}$$

pro $\gamma \neq 0$ a pro všechna x splňující nerovnost $1 + \gamma x > 0$. Pro $xh(t) > 0$ zřejmě platí

$$P(X > t + xh(t)) = P(X > t + xh(t), X > t).$$

Tímto z (5.1)

$$\lim_{t \uparrow x^*} \frac{P\left(\frac{X-t}{h(t)} > x, X > t\right)}{P(X > t)} = \lim_{t \uparrow x^*} P\left(\frac{X-t}{h(t)} > x \middle| X > t\right) = (1 + \gamma x)^{-\frac{1}{\gamma}} = 1 - H_\gamma(x)$$

pro všechna $x : 1 + \gamma x > 0 \wedge x > 0$, respektive

- $\gamma > 0 \Rightarrow$

$$\left(x > -\frac{1}{\gamma} \wedge x > 0\right) \Rightarrow x \in (0, \infty),$$

- $\gamma < 0 \Rightarrow$

$$\left(x < -\frac{1}{\gamma} \wedge x > 0\right) \Rightarrow x \in \left(0, -\frac{1}{\gamma}\right),$$

V případě, že $\gamma = 0$, je pravá strana rovnice (5.1) nahrazena výrazem e^{-x} pro $x \in (0, \infty)$. ■

Tvrzení (5.1) nám poskytuje informaci o tom, že velká pozorování mají přibližně zobecněné Paretovo (GP) rozdělení. Tato třída rozdělení je parametrizována jedním parametrem γ. K odvození jeho odhadu se nabízí možnost použít na upravená největší pozorování metodu maximální věrohodnosti a využít přitom zobecněného Paretova rozdělení jako model.

Lemma 5.2 *Nechť $X, X_1, X_2, \ldots, X_n$ je náhodný výběr z rozdělení s distribuční funkcí F a $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$ je jeho uspořádaný náhodný výběr rozsahu n. Pak pro $X_{n-k,n} = t$, kde $k = 1, \ldots, n-1$, je sdružené rozdělení veličin $(X_{n-k+1,n}, X_{n-k+2,n}, \ldots, X_{n,n})$ rovno sdruženému rozdělení uspořádanému náhodnému výběru $(X^*_{1,k}, X^*_{2,k}, \ldots, X^*_{k,k})$ s distribuční funkcí*

$$F_{X^*}(x) = P(X \leq x | X > t) = \frac{P(X \leq x, X > t)}{P(X > t)} = \frac{F(x) - F(t)}{1 - F(t)}, \quad x > t. \tag{5.2}$$

Důkaz. Viz. [15].

Na základě lemmatu 5.2 jsme schopni ekvivalentně nahradit největší pozorování

$$(X_{n-k,n}, X_{n-k+1,n}, \ldots, X_{n,n})$$

nezávislými veličinami

$$(Z_1, Z_2, \ldots, Z_k) = (X_{n-k+1,n} - X_{n-k,n}, X_{n-k+2,n} - X_{n-k,n}, \ldots, X_{n,n} - X_{n-k,n}),$$

které vzhledem k (5.2) a $t = X_{n-k,n}$ mají distribuční funkci

$$F_Z(t + x) = P(X - t \leq x | X > t) = \frac{F(x + t) - F(t)}{1 - F(t)}, \quad x > 0.$$

Nyní sestavíme věrohodnostní funkci pro $Z_1, Z_2, \ldots, Z_k$ a následně ji maximalizujeme pro reálné parametry $\sigma > 0, \gamma$. Rozlišme případy, kdy

a) $\gamma \neq 0$

K sestavení věrohodnostní funkce potřebujeme hustotu zobecněného Paretova rozdělení

$$h_{\gamma,\sigma}(z) = \frac{\partial H_\gamma(\frac{z}{\sigma})}{\partial z} = \frac{\partial(1 - (1 + \frac{\gamma}{\sigma}z)^{-1/\gamma})}{\partial z} = \frac{1}{\sigma}\left(1 + \frac{\gamma}{\sigma}z\right)^{-\frac{1}{\gamma}-1},$$

kde $\sigma > 0$. Sdružené rozdělení veličin $Z_1, Z_2, \ldots, Z_k$ je pak tvaru

$$h_{\gamma,\sigma}(\boldsymbol{z}) = \prod_{i=1}^{k} h_{\gamma,\sigma}(z_i, \gamma, \sigma) = \prod_{i=1}^{k} \frac{1}{\sigma}\left(1 + \frac{\gamma}{\sigma}z_i\right)^{-\frac{1}{\gamma}-1}.$$

Všimněme si, že pro $\frac{\gamma}{\sigma} \to -\frac{1}{z_i}$ je mocněnec $\left(1 + \frac{\gamma}{\sigma}z_i\right) = 0$, a pro $\gamma < -1$ mocnitel $\left(-\frac{1}{\gamma} - 1\right) < 0$.Proto se dále omezme pouze na oblast $(\gamma, \sigma) \in \left(-\frac{1}{2}, \infty\right) \times (0, \infty)$. Hodnoty $\gamma \in (-1, -1/2]$ nebudeme uvažovat kvůli numerické stabilitě.

Dále určíme logaritmickou věrohodnostní funkci

$$l(\boldsymbol{z},\gamma,\sigma) = \log h_{\gamma,\sigma}(\boldsymbol{z}) = \sum_{i=1}^{k} \log\left(\frac{1}{\sigma}\left(1+\frac{\gamma}{\sigma}z_i\right)^{-\frac{1}{\gamma}-1}\right) = \\ = \sum_{i=1}^{k} \log\frac{1}{\sigma} + \left(-\frac{1}{\gamma}-1\right)\log\left(1+\frac{\gamma}{\sigma}z_i\right).$$

Protože logaritmus je rostoucí funkce, logaritmická věrohodnostní funkce $l(\boldsymbol{z},\gamma,\sigma)$ dosahuje svého maxima pro stejná γ a σ jako sdružená hustota $h_{\gamma,\sigma}(\boldsymbol{z})$. Tedy $h_{\gamma,\sigma}(\boldsymbol{z})$ je maximální právě tehdy, když $\frac{\partial l(\boldsymbol{z},\gamma,\sigma)}{\partial\gamma} = 0$, a $\frac{\partial l(\boldsymbol{z},\gamma,\sigma)}{\partial\sigma} = 0$. Derivací logaritmické věrohodnostní funkce dostáváme

$$\frac{\partial l(\boldsymbol{z},\gamma,\sigma)}{\partial\gamma} = \sum_{i=1}^{k}\frac{1}{\gamma^2}\log\left(1+\frac{\gamma}{\sigma}z_i\right) - \frac{1}{\gamma}\frac{\sigma}{\gamma z_i+\sigma}\frac{z_i}{\sigma} - \frac{\sigma}{\gamma z_i+\sigma}\frac{z_i}{\sigma} = \\ = \sum_{i=1}^{k}\frac{1}{\gamma^2}\log\left(\frac{\gamma}{\sigma}z_i+1\right) - \left(\frac{1}{\gamma}+1\right)\left(\frac{\frac{z_i}{\sigma}}{1+\frac{\gamma}{\sigma}z_i}\right) = 0,$$

$$\frac{\partial l(\boldsymbol{z},\gamma,\sigma)}{\partial\sigma} = -\frac{1}{\sigma} + \sum_{i=1}^{k} -\frac{1}{\gamma}\left(\frac{\sigma}{\gamma z_i+\sigma}\cdot\frac{-\gamma}{\sigma^2}z_i\right) + \frac{\sigma}{\gamma z_i+\sigma}\frac{\gamma}{\sigma^2}z_i = \\ = -\frac{1}{\sigma} + \sum_{i=1}^{k}\left(\frac{1}{\gamma}+1\right)\frac{\frac{\gamma}{\sigma^2}z_i}{\frac{\gamma}{\sigma}z_i+1} = 0.$$

Odtud

$$\sum_{i=1}^{k}\frac{1}{\gamma^2}\log\left(\frac{\gamma}{\sigma}z_i+1\right) - \sum_{i=1}^{k}\left(\frac{1}{\gamma}+1\right)\left(\frac{\frac{z_i}{\sigma}}{1+\frac{\gamma}{\sigma}z_i}\right) = 0 \tag{5.3}$$

$$\sum_{i=1}^{k}\left(\frac{1}{\gamma}+1\right)\frac{\frac{\gamma}{\sigma^2}z_i}{\frac{\gamma}{\sigma}z_i+1} = \frac{1}{\sigma} \tag{5.4}$$

Vynásobíme-li (5.4) výrazem $\frac{\sigma}{\gamma}$ a dosadíme do (5.3) dostáváme

$$\sum_{i=1}^{k}\frac{1}{\gamma^2}\log\left(\frac{\gamma}{\sigma}z_i+1\right) - \frac{1}{\gamma} = 0,$$

$$\sum_{i=1}^{k}\left(\frac{1}{\gamma}+1\right)\frac{\frac{z_i}{\sigma}}{\frac{\gamma}{\sigma}z_i+1} = \frac{1}{\gamma}.$$

Tímto dostáváme soustavu dvou nelinárních rovnic

$$\sum_{i=1}^{k}\log\left(\frac{\gamma}{\sigma}z_i+1\right) = \gamma, \tag{5.5}$$

$$\sum_{i=1}^{k}\frac{\frac{z_i}{\sigma}}{\frac{\gamma}{\sigma}z_i+1} = \frac{1}{\gamma+1}, \tag{5.6}$$

které je potřeba numericky řešit pro neznámé parametry γ,σ. Nalezené řešení této soustavy vede na maximálně věrohodné odhady a značíme $\hat{\gamma}_{MLE},\hat{\sigma}_{MLE}$.

b) $\gamma = 0$. Sdružená hustota

$$h_\sigma(\boldsymbol{z}) = \prod_{i=1}^{k} h(z_i, \sigma) = \prod_{i=1}^{k} \frac{1}{\sigma} e^{-\frac{z_i}{\sigma}}$$

a k ní logaritmická věrohodnostní funkce

$$l(\boldsymbol{z}, \sigma) = \sum_{i=1}^{k} \log \frac{1}{\sigma} + \log e^{-\frac{z_i}{\sigma}} = k \log \frac{1}{\sigma} - \sum_{i=1}^{k} \frac{z_i}{\sigma}.$$

Řešením

$$\frac{\partial l(\boldsymbol{z}, \sigma)}{\partial \sigma} = -k \frac{1}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{k} z_i = 0,$$

dostáváme

$$\sigma = \frac{1}{k} \sum_{i=1}^{k} z_i = \frac{1}{k} \sum_{i=1}^{k} x_{n-i+1,n} - x_{n-k,n}.$$

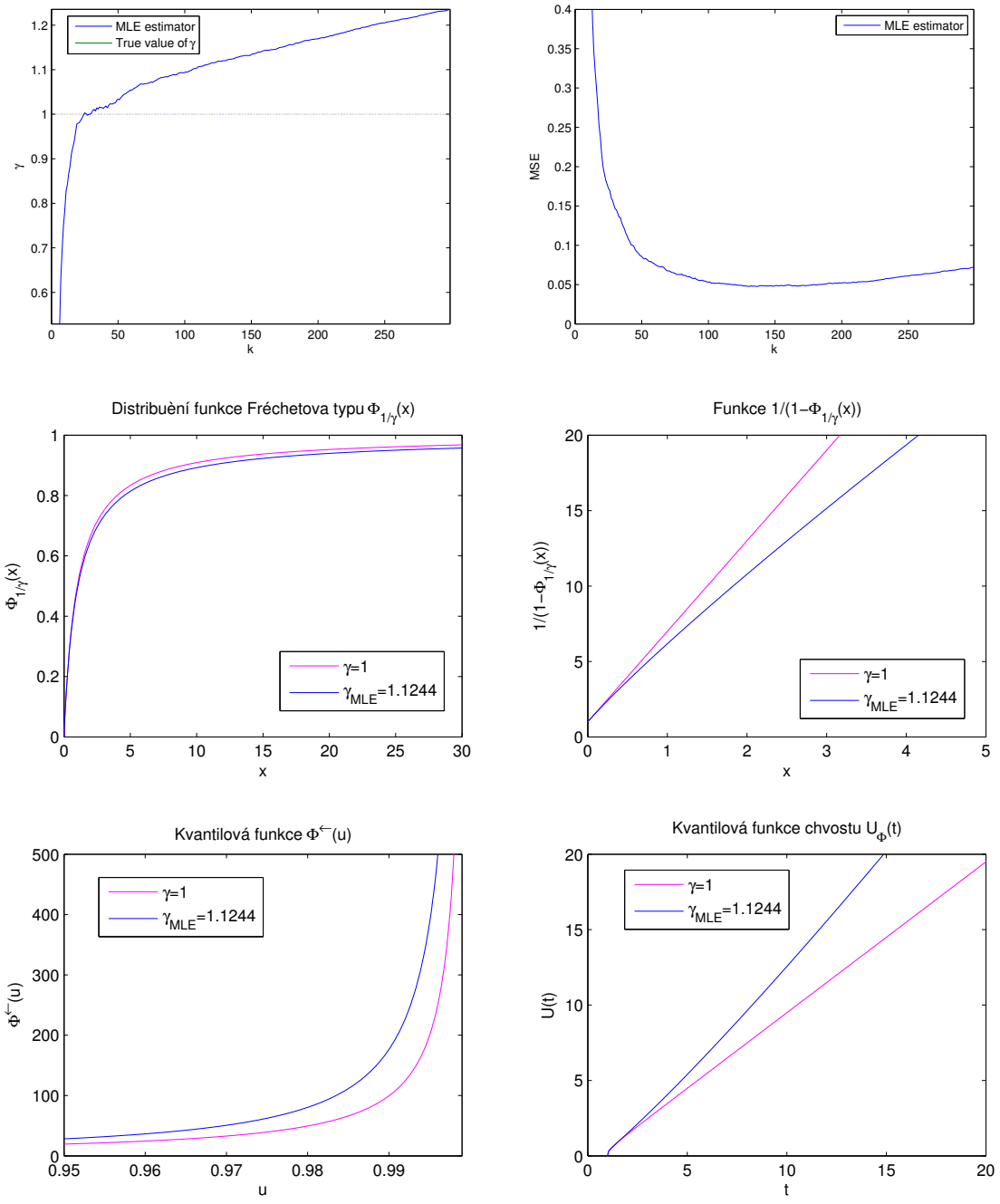
5.1.1. Vlastnosti maximálně věrohodného odhadu

Soustavu nelineárních rovnic (5.5) je v praxi potřeba vyřešit pro každé $k \in \{1, 2, \ldots, n-1\}$. Jednou z možností řešení je využití matlabovské funkce fsolve pro systém nelineárních rovnic. Tato iterační metoda vyžaduje vhodné určení počátečního bodu a mohou nastat problémy s konvergencí k řešení. Mnohem výhodnější je využít matlabovskou funkci gpfit, která je určena přímo pro odhady parametrů γ, σ zobecněného Paretova rozdělení. Tato funkce využívá metodu nelineární optimalizace, která je obecně bez omezení účelové funkce. Jmenovitě se jedná o simplex search metodu, pro kterou není potřeba numericky nebo analyticky počítat gradient. Jako startovací bod algoritmu se využívá momentových odhadů parametrů γ, σ pro zobecněné Paretovo rozdělení. Podrobněji o těchto funkcích a metodách v [20].

Odhad je závislý na volbě prahového indexu k. Správná volba tohoto indexu je podstatná pro praktické použití tohoto odhadu. Volíme-li k příliš malé, roste rozptyl odhadu. Naopak volíme-li k příliš velké, klesá rozptyl, ale roste vychýlení odhadu. Navíc konvergence k zobecněnému Paretovu rozdělení může být někdy velmi pomalá. Více o volbě k pro věrohodný odhad například v [6].

Z obrázku 5.1 lze pozorovat rostoucí vychýlení. Nejmenší střední kvadratická chyba byla vypočtena pro $k = 136$. Dolní část obrázku znázorňuje přesnost odhadu indexu extrémní hodnoty na Fréchetově rozdělení.

Věrohodný odhad je konzistentním odhadem parametru γ pro $k, n \to \infty, \frac{k}{n} \to 0$. Rovněž je invariantní vůči změně měřítka i posunutí. Za určitých předpokladů je asymptoticky normální viz. [15] str. 92.



Obrázek 5.1: Průměrný maximálně věrohodný odhad parametru γ na simulovaných datech o rozsahu $n = 1000$ z rozdělení Burr$(1, 0.5, 2)$ (typ XII) s parametrem tvaru $\gamma = 1$ (vlevo nahoře). Střední kvadratická chyba odhadu (vpravo nahoře). Charakteristiky rozdělení Fréchetova typu pro $\hat{\gamma}_{MLE}(k) = 1.124$, $k = 136$ a teoretickou hodnotu $\gamma = 1$ (dole). Počet opakování $MC = 500$.

5.2. Hillův odhad

Nechť $X_1, X_2, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí splňující $F \in \mathcal{D}(G_\gamma), \gamma > 0$, a cílem je odhadnout index extrémní hodnoty γ. Pak nejznámějším odhadem γ je Hillův odhad $\hat{\gamma}_H$ z roku 1975 [16]:

$$\hat{\gamma}_H(k) = \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}. \tag{5.7}$$

Zavést Hillův odhad lze více způsoby. Můžeme například vyjít z kvantilového grafu pro Paretovo rozdělení viz. [5] str. 101. Zde vyjdeme z (3.39):

$$\gamma = \lim_{t\to\infty} \mathrm{E}(\log X - \log t | X > t) = \lim_{t\to\infty} \frac{\int_t^\infty (\log x - \log t)\mathrm{d}F(x)}{1 - F(t))}, \tag{5.8}$$

kde uvedený integrál je Lebesgue–Stieltjesův integrál viz. [29] str. 163. Nahraďme parametr t k-tou nejvyšší pořadovou statistikou $X_{n-k,n}$, přičemž volíme $k = k(n)$ tak, aby $k \to \infty$, $\frac{k}{n} \to 0$ pro $n \to \infty$. Dále, když F nahradíme empirickou distribuční funkcí $\hat{F}_n(x)$ podle definice (2.14), z (5.8) dostaneme

$$\begin{aligned}
\hat{\gamma}_H :&= \frac{\int_{X_{n-k,n}}^\infty (\log x - \log X_{n-k,n})\mathrm{d}\hat{F}_n(x)}{1 - \hat{F}_n(X_{n-k,n}))} = \\
&= \frac{\int_{X_{n-k,n}}^\infty \log x \mathrm{d}\hat{F}_n(x) - \log X_{n-k,n} \int_{X_{n-k,n}}^\infty \mathrm{d}\hat{F}_n(x)}{1 - \frac{n-k}{n}} = \\
&= \frac{(\int_{X_{n-k,n}}^{X_{n,n}} \log x \mathrm{d}\hat{F}_n(x) + \int_{X_{n,n}}^\infty \log x \mathrm{d}\hat{F}_n(x)) - \log X_{n-k,n}(\int_{X_{n-k,n}}^{X_{n,n}} \mathrm{d}\hat{F}_n(x) + \int_{X_{n,n}}^\infty d\hat{F}_n(x))}{\frac{n-n+k}{n}} = \\
&= \frac{(\sum_{i=1}^k \frac{1}{n} \log X_{n-i+1,n} + 0) - \log X_{n-k,n}(\frac{k}{n} + 0)}{\frac{k}{n}} = \\
&= \frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}.
\end{aligned}$$

5.2.1. Vlastnosti Hillova odhadu

Věta 5.1 *Předpokládejme náhodný výběr $X_1, X_2, \ldots, X_n$ z distribuční funkce $F \in \mathcal{D}(G_\gamma)$, $\gamma > 0$, potom pro $n \to \infty$, $k \to \infty$, $\frac{k}{n} \to 0$ platí*

$$\hat{\gamma}_H \xrightarrow{P} \gamma$$

Důkaz [15] str. 70.

Věta 5.2 *Předpokládejme náhodný výběr $X_1, X_2, \ldots, X_n$ z distribuční funkce $F \in \mathcal{D}(G_\gamma)$, $\gamma > 0$. Dále předpokládejme, že kvantilová funkce U k funkci F splňuje pro nějaké $\rho \leq \leq 0$ a funkci A, která nemění znaménko a $\lim_{t\to\infty} A(t) = 0$, podmínku (3.40). Potom pro $n \to \infty$, $k \to \infty$, $\frac{k}{n} \to 0$ platí*

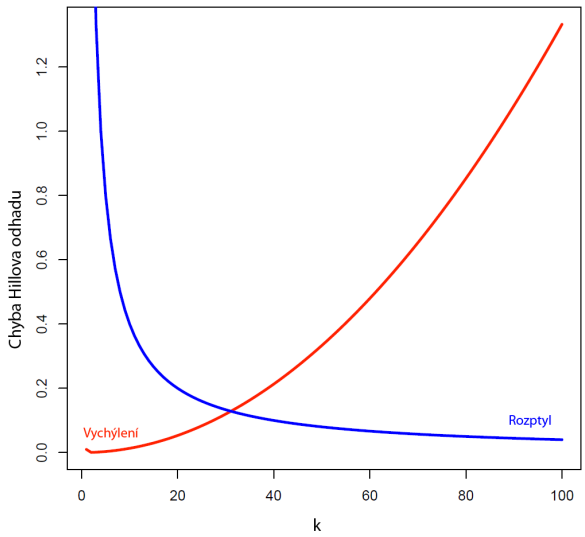
$$\sqrt{k}(\hat{\gamma}_H - \gamma) \xrightarrow{D} N\left(\frac{\lambda}{1-\rho}, \gamma^2\right),$$

kde

$$\lambda = \lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right).$$

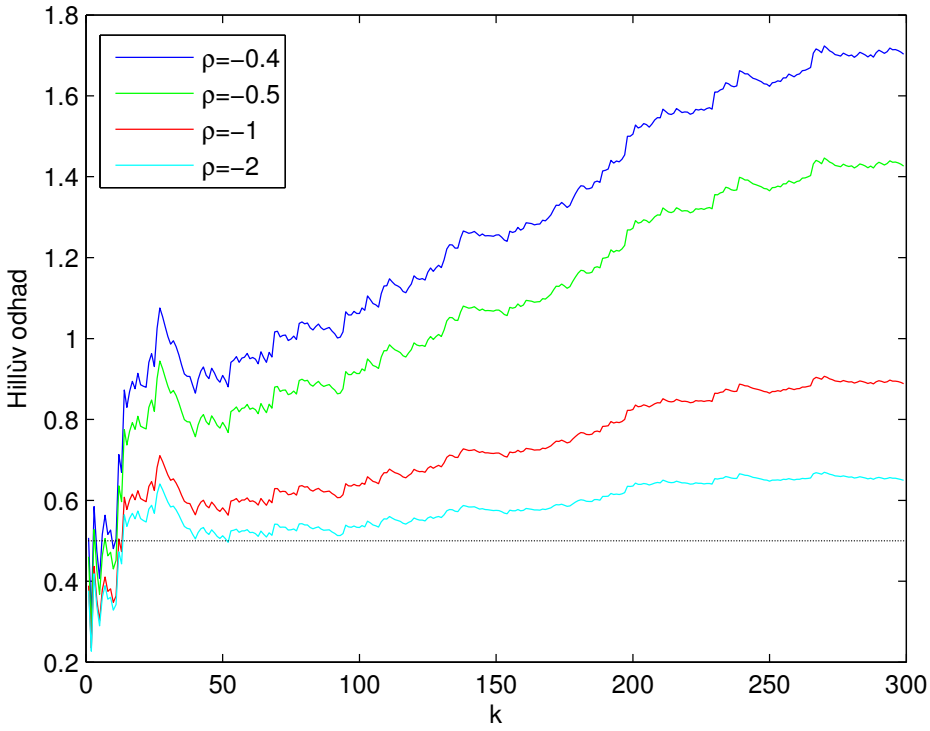
Důkaz [15] str. 74.

- Z věty 5.1 plyne, že za uvedených předpokladů je Hillův odhad konzistentním odhadem γ.
- Pokud na distribuční funkci náhodného výběru navíc klademe podmínku druhého druhu (tj. za předpokladu věty 3.6), pak vzhledem k větě 5.2 má $\sqrt{k}(\hat{\gamma}_H - \gamma)$ asymptoticky normální rozdělení. Tato vlastnost bývá základem pro sestavení intervalu spolehlivosti pro $\hat{\gamma}_H(k)$ a optimální volbu prahu k.
- Pro každou volbu k dostáváme jiný odhad γ. Otázka správné volby k je podstatná pro praktické použití tohoto odhadu. Volíme-li k příliš malé, odhad je počítan z mála pozorování a roste rozptyl odhadu. Naopak volíme-li k příliš velké, klesá rozptyl, ale roste vychýlení odhadu viz. ilustrativní obrázek 5.2.



Obrázek 5.2: Příklad výchýlení a rozptylu Hillova odhadu v závislosti na k [8].

- Zatímco rozptyl statistiky $\sqrt{k}(\hat{\gamma}_H - \gamma)$ je asymptoticky roven γ^2, asymptotické vychýlení závisí na parametru druhého druhu ρ. Obrázek 5.3 ilustruje vychýlení Hillova odhadu v závislosti na parametru ρ pro Burrovo rozdělení typu XII. Byly vygenerovány čtyři náhodné výběry rozsahu 1000 pro různé hodnoty ρ. V případě rovnosti parametru ρ dostaneme identické odhady pro $k = 1, 2, \ldots n-1$.



Obrázek 5.3: Hillův odhad pro různé hodnoty ρ na simulovaných datech o rozsahu $n = 1000$ z rozdělení Burr$(1, -\rho, \frac{1}{-\rho\gamma})$ (typ XII) s parametrem tvaru $\gamma = \frac{1}{2}$.

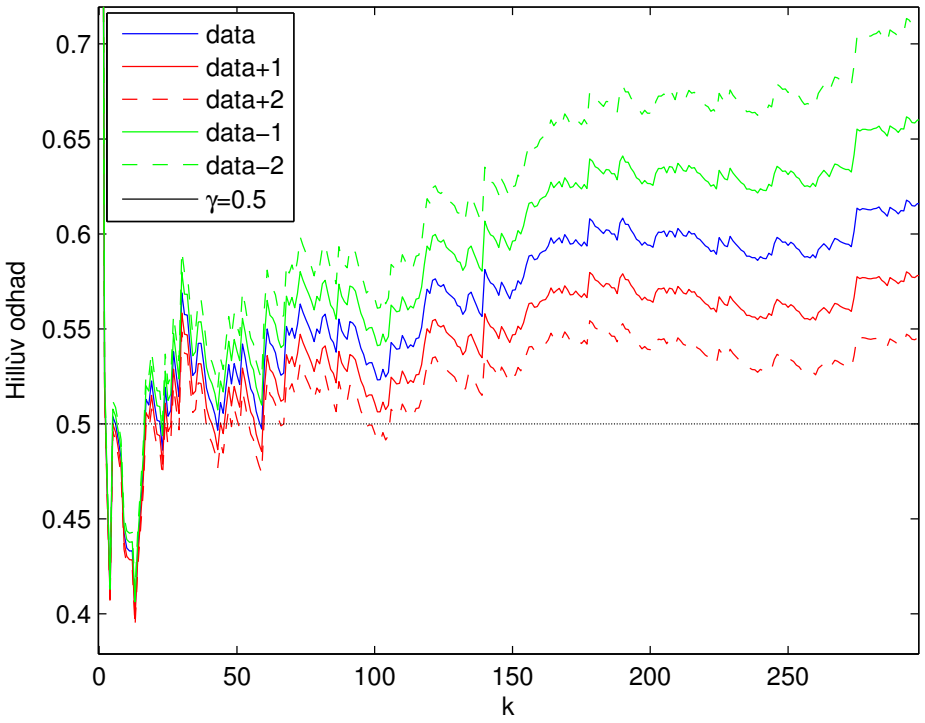
- Hillův odhad je invariantní vůči změně měřítka. Není však invariantní vůči posunutí stejně jako mnoho dalších odhadů založených na logaritmické transformaci dat. Na obrázku 5.4 je Hillův odhad pro simulovaná data posunuta o konstanty +1,+2 (červeně), případně -1,-2 (zeleně). Odhad bez posunutí je vyznačen modře.

  Pokud chceme zajistit invarianci vůči posunutí, je potřeba odhad modifikovat. Například podle [13] se zavádí další pořadový index k_{in}, $k_{in} < k$, který vede na modifikovaný Hillův odhad

  $$\hat{\gamma}_{H_m}(k_{in}, k) = \frac{1}{k_{in}} \sum_{j=1}^{k_{in}} \log \frac{X_{n-j+1,n} - \log X_{n-k,n}}{X_{n-k_{in},n} - \log X_{n-k,n}}.$$

  Lze ukázat, že pro $n \to \infty$, $k \to \infty$, $k_{in} \to \infty$, $\frac{k}{n} \to 0$, $\frac{k_{in}}{n} \to 0$ je $\hat{\gamma}_{H_m}(k_{in}, k)$ konzistentní. Podrobněji jak volit k_{in} je uvedeno v [13].

- Podle [3] (Corollary 1) je Hillův odhad konzistentní i pro závislá data jako je ARMA a ARCH proces. Hillův odhad v souvislosti s ARMA procesem je studován v [18]. Obecně o stochastických procesech v [4].



Obrázek 5.4: Hillův odhad na různě posunutých simulovaných datech ze zobecněného Paretova rozdělení o rozsahu $n = 10000$ s parametrem tvaru $\gamma = \frac{1}{2}$.

5.2.2. Optimální prahový index k_0

Optimální prah

Obvyklým přístupem k optimalizačnímu problému ve statistice je požadavek minimální střední kvadratické chyby. V našem případě bychom rádi volili

$$k_0(n) = \arg \min_{1 \le k < n} \mathrm{E}(\hat{\gamma}_H(k) - \gamma)^2$$

neboli

$$\{k_0 = k \in [1, n) : \mathrm{MSE}(\hat{\gamma}_H(k)) = \mathrm{bias}^2(\hat{\gamma}_H(k)) - \mathrm{D}(\hat{\gamma}_H(k)) \text{ je minimální}\}. \tag{5.9}$$

Poznámka. V následujícím tvrzení ukážeme, že střední kvadratickou chybu lze rozložit na vychýlení a rozptyl.

Tvrzení 5.3 *Buď $\hat{\theta}$ odhadem parametru θ, potom střední kvadratická chyba*

$$\mathrm{E}(\hat{\theta} - \gamma)^2 = \mathrm{D}(\hat{\theta}) + (\mathrm{E}\hat{\theta} - \theta)^2.$$

Důkaz.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) = \mathrm{E}(\hat{\theta} - \theta)^2 &= \mathrm{E}(\hat{\theta} - \mathrm{E}\hat{\theta} + \mathrm{E}\hat{\theta} - \theta)^2 = \\
&= \mathrm{E}[(\hat{\theta} - \mathrm{E}\hat{\theta})^2 + 2(\hat{\theta} - \mathrm{E}\hat{\theta})(\mathrm{E}\hat{\theta} - \theta) + (\mathrm{E}\hat{\theta} - \theta)^2] = \\
&= \mathrm{E}(\hat{\theta} - \mathrm{E}\hat{\theta})^2 + 2\mathrm{E}(\hat{\theta} - \mathrm{E}\hat{\theta})(\mathrm{E}\hat{\theta} - \theta) + \mathrm{E}(E\hat{\theta} - \theta)^2 = \\
&= \mathrm{D}(\hat{\theta}) + 2\mathrm{E}[(\hat{\theta} - \mathrm{E}\hat{\theta})(\mathrm{E}\hat{\theta} - \theta)] + \mathrm{E}(\mathrm{E}\hat{\theta} - \theta)^2 = \mathrm{D}(\hat{\theta}) + (\mathrm{E}\hat{\theta} - \theta)^2 = \\
&= \mathrm{D}(\hat{\theta}) + \mathrm{bias}^2(\hat{\theta}).
\end{aligned}$$

Přitom

$$\begin{aligned}\mathrm{E}[(\hat{\theta}-\mathrm{E}\hat{\theta})(\mathrm{E}\hat{\theta}-\theta)] &= \mathrm{E}(\hat{\theta}\mathrm{E}\hat{\theta}-\theta\hat{\theta}-\mathrm{E}\hat{\theta}\mathrm{E}\hat{\theta}+\theta\mathrm{E}\hat{\theta}) = \\ &= \mathrm{E}\hat{\theta}\mathrm{E}\hat{\theta}-\theta\mathrm{E}\hat{\theta}-\mathrm{E}\hat{\theta}\mathrm{E}\hat{\theta}+\theta\mathrm{E}\hat{\theta}=0.\end{aligned}$$

■

Střední kvadratická chyba MSE (podle (5.9)) obecně nemusí existovat. Připomeňme, že předpokládáme $\gamma > 0$ a podmínku atraktivity $F \in \mathcal{D}(G_\gamma)$. Předpokládejme navíc, že pro nějaké $\rho \leq 0$ a funkci A, která nemění znaménko a $\lim_{t\to\infty} A(t) = 0$, je splněna podmínka druhého druhu (podmínka (3.40)). Potom konvergenci z věty 5.2 můžeme pomocí lineární transformace standardizovaného normálního rozdělení $N \sim \mathrm{N}(0,1)$ přepsat na

$$\sqrt{k}(\hat{\gamma}_H - \gamma) \xrightarrow{D} \frac{\lambda}{1-\rho} + \gamma N,$$

kde $\lambda = \lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right)$. Odtud se nabízí ve dvou krocích aproximovat

$$\hat{\gamma}_H - \gamma \approx \frac{\gamma N}{\sqrt{k}} + \frac{\lambda}{(1-\rho)\sqrt{k}} \approx \frac{\gamma N}{\sqrt{k}} + \frac{A\left(\frac{n}{k}\right)}{(1-\rho)}. \tag{5.10}$$

Nyní můžeme odhadnout vychýlení

$$\mathrm{bias}(k,n) = \mathrm{E}(\hat{\gamma}_H - \gamma) \approx \mathrm{E}\left(\frac{\gamma N}{\sqrt{k}} + \frac{A\left(\frac{n}{k}\right)}{(1-\rho)}\right) = \frac{A\left(\frac{n}{k}\right)}{1-\rho}, \tag{5.11}$$

a asymptotickou střední kvadratickou chybu

$$\mathrm{E}(\hat{\gamma}_H - \gamma)^2 \approx \mathrm{E}\left(\frac{\gamma N}{\sqrt{k}} + \frac{A\left(\frac{n}{k}\right)}{(1-\rho)}\right)^2 = \mathrm{E}\frac{\gamma^2 N^2}{k} + 2\frac{\gamma}{\sqrt{k}}\frac{A\left(\frac{n}{k}\right)}{(1-\rho)}\mathrm{E}(N) + \mathrm{E}\frac{A^2\left(\frac{n}{k}\right)}{(1-\rho)^2}. \tag{5.12}$$

Prostřední člen se vyruší, neboť střední hodnota standardizovaného normálního rozdělení $\mathrm{E}(N)$ je rovna nule. Střední hodnota $\mathrm{E}(N^2) = \mathrm{D}(N)$. Z (5.12) dostáváme odhad asymptotické střední kvadratické chyby

$$\mathrm{AMSE}(k,n) = \mathrm{E}(\hat{\gamma}_H - \gamma)^2 \approx \left(\frac{\gamma^2}{k} + \frac{A^2\left(\frac{n}{k}\right)}{(1-\rho)^2}\right),$$

která je závislá na velikosti náhodného výběru n a volbě indexu prahové statistiky k. Nyní konečně můžeme definovat optimální prahový index jako

$$k_0(n) = \arg\min_{1\leq k<n} \mathrm{AMSE}(k,n).$$

Analytický vztah pro $k_0(n)$ je komplikovaný a v praxi nepoužitelný, protože závisí na několika neznámých parametrech. Lze jej nalézt například v [10]. Než přejdeme k samotné bootstrap proceduře, musíme navíc předpokládat speciální tvar funkce

$$A(t) = Ct^\rho, \quad C \neq 0, \rho < 0, \tag{5.13}$$

která vyhovuje podmínce druhého druhu (tj. platí (3.40)). Tento předpoklad je velmi častý v aplikacích teorie extrémních hodnot viz [15]. Toto zjednodušení ale není přípustné pro $\rho = 0$.

Odhad optimálního prahového indexu k_0 pomocí metody bootstrap

Existuje více přístupů, jak k_0 odhadnout. Uvedeme jeden z možných algoritmů popsaný v [14].

Distribuční funkci F náhodného výběru $X_1, X_2, \ldots, X_n$ neznáme. Použijeme empirickou distribuční funkci $\hat{F}$ jako odhad F. Uvažujme bootstrapový výběr $X_1^*, X_2^*, \ldots, X_{n_1}^*$, $n_1 < n$, z rozdělení o empirické distribuční funkci $\hat{F}$. Hillův odhad z tohoto bootstrapového výběru označme $\hat{\gamma}^*_{n_1,H}$.

Dále označme nějakou počáteční hodnotu prahového indexu k jako k_{aux}. Index aux značí, že se jedná o pomocnou hodnotu. Požadujeme, aby odhad $\hat{\gamma}_H(k_{aux})$ byl konzistentní tj. $k_{aux} \to \infty$, $\frac{k_{aux}}{n} \to 0$. Pro $\hat{\gamma}^*_{n_1,H}$ dostaneme bootstrapový odhad střední kvadratické chyby

$$\mathrm{MSE}^*(n_1, k_1) = \mathrm{E}(\hat{\gamma}^*_{n_1,H}(k_1) - \hat{\gamma}_H(k_{aux}))^2 \tag{5.14}$$

pro $1 \le k_1 < n_1$. Statistika $\hat{\gamma}_H(k_{aux})$ zde hraje jakousi referenční roli.

Z článku [7] za platnosti (5.13) pro $k_1 \to \infty$, $\frac{k_1}{n_1} \to 0$ platí

$$\frac{k^*_{0,n_1}}{k_0(n)} \left(\frac{n_1}{n}\right)^{\frac{-2\rho}{2\rho-1}} \xrightarrow{P} 1, \tag{5.15}$$

kde k^*_{0,n_1} minimalizuje bootstrapový odhad střední kvadratické chyby tj.

$$k^*_{0,n_1} = \arg \min_{1 \le k_1 < n_1} \mathrm{MSE}^*(n_1, k_1). \tag{5.16}$$

Z (5.15) můžeme odhadnout

$$k_0(n) \approx k^*_{0,n_1} \left(\frac{n_1}{n}\right)^{\frac{-2\rho}{2\rho-1}}.$$

Odtud dostaneme odhad, který je závislý na parametru druhého druhu ρ. Abychom ji odstranili, budeme uvažovat další bootstrapový náhodný výběr rozsahu $n_2 < n$ z rozdělení o empirické distribuční funkci $\hat{F}$. Dostaneme

$$k^*_{0,n_2} = \arg \min_{1 \le k_2 < n_2} \mathrm{MSE}^*(n_2, k_2), \quad k_0(n) \approx k^*_{0,n_2} \left(\frac{n_2}{n}\right)^{\frac{-2\rho}{2\rho-1}}.$$

Odtud

$$k_0(n) = \frac{[k_0(n)]^2}{k_0(n)} \approx \frac{[k^*_{0,n_1}]^2}{k^*_{0,n_2}} \frac{\left(\frac{n_1^2}{n^2}\right)^{\frac{-2\rho}{2\rho-1}}}{\left(\frac{n_2}{n}\right)^{\frac{-2\rho}{2\rho-1}}} = \frac{[k^*_{0,n_1}]^2}{k^*_{0,n_2}} \left(\frac{n_1^2}{n^2}\frac{n}{n_2}\right)^{\frac{-2\rho}{2\rho-1}}.$$

Pokud zvolíme $n_2 = \frac{n_1^2}{n}$, $k_0(n)$ na ρ nebude záviset. Konečně pomocí dvojtého bootstrapu dostáváme odhad optimálního indexu prahové statistiky

$$\widehat{k_0}(n) = \widehat{k_0}(n; k_{aux}, n_1) = \frac{[k^*_{0,n_1}]^2}{k^*_{0,n_1^2/n}}, \tag{5.17}$$

který závisí na volbě n_1 a k_{aux}.

Na začátku byl problém volby pouze jednoho parametru k. Problém jsme převedli na dva nové neznámé parametry k_{aux} a n_1. Otázka je, zda se tímto úloha ještě více

nezkomplikovala. Ještě před tím než navrhneme volbu těchto parametrů, ukážeme jejich vliv na $\widehat{k_0}(n_1, k_{aux})$ na simulovaných datech.

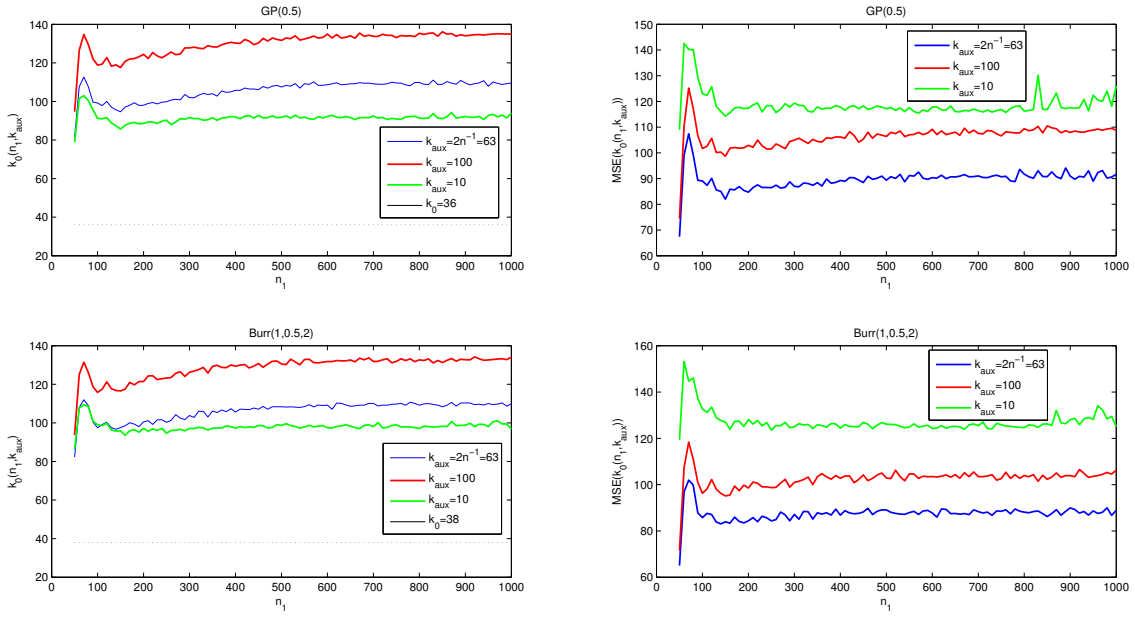
Na obrázku 5.5 je znázorněna citlivost odhadu $\widehat{k_0}(n)$ na volbu k_{aux} a n_1 pro Burrovo rozdělení s parametry $\gamma = 1$, $\rho = -0.5$ a pro zobecněné Paretovo rozdělení s parametrem tvaru $\gamma = 0.5$. Byl zvolen rozsah náhodných výběrů $n = 1000$, počet bootstrapových výběrů $n^* = 250$, hodnoty indexu počáteční prahové statistiky $k_{aux} = 10$, $k_{aux} = 2\sqrt{n} = 63$, $k_{aux} = 100$ a rozsah bootstrapového náhodného výběru $n_1 = 50, 60, 70, \ldots, 1000$. Aby byla snížena variabilita odhadu, simulace byla opakována 1000 krát. Do grafu jsou vynášeny průměrné hodnoty

$$\overline{\widehat{k_0}} = \frac{1}{1000} \sum_{i=1}^{1000} \widehat{k_{0_i}}$$

a odmocnina ze střední kvadratické chyby

$$\mathrm{RMSE} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\widehat{k_{0_i}} - k_0)^2}.$$

Správné hodnoty k_0 byly zjištěny rovněž na základě simulace Monte Carlo. Viz. simulační program uvedený v příloze.



Obrázek 5.5: $\overline{\widehat{k_0}}(n)$ v závislosti na k_{aux} a n_1 (vlevo) na datech z: a) Burr$(1, -\rho, \frac{1}{-\rho\gamma})$ (typ XII) s $\rho = -0.5$ a parametrem tvaru $\gamma = 1$ (nahoře) b) GP(0.5) s parametrem tvaru $\gamma = 0.5$ (dole). Střední kvadratická chyba RMSE (vpravo).

Obdobně byla studována další rozdělení s těžšími chvosty. Na základě těchto studií můžeme usoudit, že volba hodnoty n_1 na odhad téměř nemá vliv. Střední kvadratická chyba je totiž přibližně konstantní pro různé hodnoty n_1. Závisí ale na volbě k_{aux}. Drees a Kaufmann volí v [10] $k_{aux} = 2\sqrt{n}$. Dalším řešením by mohlo být zavedení další metody, která by opět volila k_{aux} adaptivně nebo uvažovat jiný tvar výrazu (5.16) viz. [7].

Na základě 5.11 můžeme podle [14] pomocí bootstrapu odhadnout vychýlení:

$$\widehat{\text{bias}}_{n_1}(\hat{\gamma}_H(k)) = \frac{[\widehat{\text{bias}}^*_{n_1}(k)]^2}{\widehat{\text{bias}}^*_{n_2}(k)}, \quad 1 \le k < n_2,\ n_2 = \frac{n_1^2}{n},$$

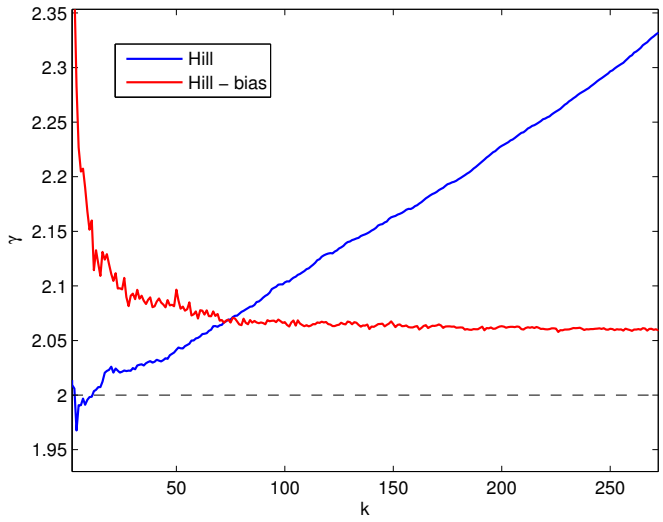
kde

$$\widehat{\text{bias}}^*_{n_i}(k) = \text{E}(\hat{\gamma}^*_{n_i,H}(k) - \hat{\gamma}_H(k_{aux})), \quad i = 1, 2.$$

Hillův odhad pak lze opravit o vychýlení tak, že položíme

$$\hat{\gamma}^C_H(k) = \hat{\gamma}_H(k) - \widehat{\text{bias}}_{n_1}(\hat{\gamma}_H(k)), \quad 1 \le k < n_2.$$

Na obrázku 5.6 je znázorněna oprava vychýlení pomocí metody bootstrap pro Burrovo rozdělení. Byl zvolen rozsah náhodných výběrů $n = 1000$, počet bootstrapových výběrů $n^* = 250$, $k_{aux} = 63$ a rozsah bootstrapového náhodného výběru $n_1 = 975$. Simulace byla opakována 1000 krát. Lze pozorovat zlepšení vychýlení, ale výsledek stále není ideální.



Obrázek 5.6: Průměrný Hillův odhad a průměrný opravený Hillův odhad $\overline{\hat{\gamma}^C_H}$ na simulovaných datech rozsahu $n = 1000$ z rozdělení $\text{Burr}(1, -\rho, \frac{1}{-\rho\gamma})$ (typ XII) s parametrem tvaru $\gamma = 2$ a $\rho = -0.5$.

5.3. Momentový odhad

Se samotným Hillovým odhadem nevystačíme, pokud obecně uvažujeme $\gamma \in \mathbb{R}$. V případě, kdy $\gamma \leq 0$, je pravý koncový bod $x^* \leq 0$ a logaritmus pro takové hodnoty není definovan. Tento problém bychom mohli řešit jednoduchým posunutím pozorování o kladnou konstantu. Je ovšem třeba jisté opatrnosti, protože samotný Hillův odhad není invariantní vůči posunutí. V tomto případě je třeba zavést nový obecnější odhad, který vychází z Hillova odhadu.

Abychom mohli zavést momentový odhad (vzah (5.26)), musíme klást obdobné podmínky jako pro Hillův odhad. Pro náhodný výběr s distribuční funkcí F požadujeme, aby pravý koncový bod $x^* > 0$ a byla splněna podmínka atraktivity tj. $F \in \mathcal{D}(G_\gamma)$, $\gamma \in \mathbb{R}$. Potom z věty 3.3 pro $x > 0$ a vhodnou kladnou funkci $a(t)$ platí

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = D_\gamma(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0, \\ \log x, & \gamma = 0. \end{cases} \tag{5.18}$$

Nyní rozlišme dva případy:

a) Uvážíme-li $\gamma \leq 0$, pak z (3.37)

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = 1 \quad x > 0$$

a použijeme-li logaritmus, můžeme ekvivalentně psát

$$\lim_{t\to\infty} (\log U(tx) - \log U(t)) = 0.$$

Limitu z (5.18) pro $\gamma \leq 0$ upravíme na

$$\lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - 1}{\frac{a(t)}{U(t)}} = \lim_{t\to\infty} \frac{\log U(tx) - \log U(t)}{\frac{a(t)}{U(t)}} = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma < 0, \\ \log x, & \gamma = 0. \end{cases} \tag{5.19}$$

b) Uvážíme-li $\gamma > 0$, pak z (3.34)

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma, \quad x > 0$$

a použijeme-li logaritmus

$$\lim_{t\to\infty} \frac{\log U(tx) - \log U(t)}{\gamma} = \log x. \tag{5.20}$$

Pokud ještě uvážíme (vztah (3.35)) $\gamma = \lim_{t\to\infty} \frac{a(t)}{U(t)}$, přepíšeme (5.20) na

$$\lim_{t\to\infty} \frac{\log U(tx) - \log U(t)}{\frac{a(t)}{U(t)}} = \log x. \tag{5.21}$$

Ze vztahů (5.19) a (5.21) dostaneme pro nějakou ,,velkou" hodnotu t aproximaci

$$\log U(tx) - \log U(t) \approx \begin{cases} \frac{a(t)}{U(t)} \frac{x^\gamma - 1}{\gamma}, & \gamma < 0, \\ \frac{a(t)}{U(t)} \log x, & \gamma \geq 0. \end{cases} \tag{5.22}$$

Dále definujme j-tý moment $M_n^{(j)}$ jako odhad $\lim_{t\to\infty} \mathrm{E}((\log X - \log t)^j | X > t)$ pro $j = 1, 2, \ldots$

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n})^j. \tag{5.23}$$

Vidíme, že pro $j = 1$ je $M_n^{(1)}(k)$ roven Hillovu odhadu $\hat{\gamma}_H$. Dále podle [5] vzhledem k vlastnostem empirické distribuční funkce můžeme odhadnout $\hat{U}\left(\frac{n+1}{k+1}\right) = X_{n-k,n}$ a $\hat{U}\left(\frac{n+1}{i}\right) = X_{n-i+1,n}$ pro nějaké $i \in \{1, 2, \ldots, k\}$. Tímto lze na

$$\log X_{n-i+1,n} - \log X_{n-k,n} = \log \hat{U}\left(\frac{n+1}{i}\right) - \log \hat{U}\left(\frac{n+1}{k+1}\right)$$

pohlížet jako na odhad výrazu

$$\log U\left(\frac{n+1}{i}\right) - \log U\left(\frac{n+1}{k+1}\right) = \log U\left(\left(\frac{n+1}{k+1}\right)\left(\frac{k+1}{i}\right)\right) - \log U\left(\frac{n+1}{k+1}\right).$$

Pokud dosadíme $t = \frac{n+1}{k+1}$, $x = \frac{k+1}{i}$ do (5.22) a uvážíme $\frac{n}{k} \to \infty$, pak pro nějaké $i \in \{1, 2, \ldots, k\}$

$$\log X_{n-i+1,n} - \log X_{n-k,n} \approx \begin{cases} \frac{a\left(\frac{n+1}{k+1}\right)}{U\left(\frac{n+1}{k+1}\right)} \frac{\left(\frac{k+1}{i}\right)^\gamma - 1}{\gamma}, & \gamma < 0, \\ \frac{a\left(\frac{n+1}{k+1}\right)}{U\left(\frac{n+1}{k+1}\right)} \log\left(\frac{k+1}{i}\right), & \gamma \geq 0. \end{cases} \tag{5.24}$$

Dalším cílem je eliminovat členy $a\left(\frac{n+1}{k+1}\right)$ a $U\left(\frac{n+1}{k+1}\right)$ z (5.24). Toho docílíme, pokud položíme

$$\frac{(M_n^{(1)}(k))^2}{M_n^{(2)}(k)} = \frac{(\frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n})^2}{\frac{1}{k}\sum_{i=1}^{k} (\log X_{n-i+1,n} - \log X_{n-k,n})^2} \approx \begin{cases} \frac{\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{k+1}{i}\right)^\gamma - 1\right)^2}{\frac{1}{k}\sum_{i=1}^{k}\left(\left(\frac{k+1}{i}\right)^\gamma - 1\right)^2}, & \gamma < 0, \\ \frac{\left(\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{k+1}{i}\right)\right)^2}{\frac{1}{k}\sum_{i=1}^{k}\left(\log\left(\frac{k+1}{i}\right)\right)^2}, & \gamma \geq 0. \end{cases} \tag{5.25}$$

Jednotlivé sumy dále blíže rozebereme. Jestliže $k \to \infty$, pak viz. [15]

a) pro $\gamma < 0$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \left(\left(\frac{i}{k+1}\right)^{-\gamma} - 1\right) = \int_0^1 (u^{-\gamma} - 1)\mathrm{d}u = \frac{\gamma}{1-\gamma},$$

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \left(\left(\frac{i}{k+1}\right)^{-\gamma} - 1\right)^2 = \int_0^1 (u^{-\gamma} - 1)^2 \mathrm{d}u = \left[\frac{u^{1-2\gamma}}{1-2\gamma}\right]_{u=0}^{1} - 2\left[\frac{u^{1-\gamma}}{1-\gamma}\right]_{u=0}^{1} - 1 =$$

$$= \frac{2\gamma^2}{(1-2\gamma)(1-\gamma)}.$$

b) Obdobně pro $\gamma \geq 0$ s pomocí integrace per partes

$$\frac{1}{k}\sum_{i=2}^{k}\log\left(\frac{k+1}{i}\right) = -\int_0^1 \log u \mathrm{d}u = 1,$$

$$\frac{1}{k}\sum_{i=2}^{k}\left(\log\left(\frac{k+1}{i}\right)\right)^2 = \int_0^1 (\log u)^2 \mathrm{d}u = 2.$$

Odtud s využitím (5.25) pro $k, n \to \infty$, $\frac{k}{n} \to 0$

$$\frac{(M_n^{(1)})^2}{M_n^{(2)}} \xrightarrow{P} \begin{cases} \frac{1-2\gamma}{2(1-\gamma)}, & \gamma < 0, \\ \frac{1}{2}, & \gamma \geq 0. \end{cases}$$

Odtud po úpravě a s uvážením principu spojitosti pro konvergenci podle pravděpodobnosti (viz. věta B9 v [1]) dostaneme

$$\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1} \xrightarrow{P} \begin{cases} \left(1 - \frac{1-2\gamma}{2(1-\gamma)}\right)^{-1}, & \gamma < 0, \\ 2, & \gamma \geq 0, \end{cases}$$

$$1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1} \xrightarrow{P} \begin{cases} \gamma, & \gamma < 0, \\ 0, & \gamma \geq 0. \end{cases}$$

Pomocí posledního vztahu lze získat momentový odhad pro $\gamma \in \mathbb{R}$. Zajisté celou dobu pracujeme s kladnými hodnotami pozorování, pro která je logaritmus definován. V praxi je proto třeba dbát určité opatrnosti.

Jednou z vlastností Hillova odhad je

$$\hat{\gamma}_H \xrightarrow{P} \begin{cases} 0, & \gamma < 0, \\ \gamma, & \gamma \geq 0. \end{cases}$$

Konečně dostáváme momentový odhad pro $\gamma \in \mathbb{R}$ jako

$$\hat{\gamma}_M(k) = \hat{\gamma}_H + \hat{\gamma}_- = M_n^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}}\right)^{-1}. \tag{5.26}$$

Momentový odhad je tedy tvořen odhadem kladné části pomocí Hillova odhadu $\hat{\gamma}_H$ a odhadem záporné části pomocí prvního a druhého momentu.

5.3.1. Vlastnosti momentového odhadu

Momentový odhad sdílí s Hillovým spoustu vlastností, protože se jedná o jeho zobecnění. Opět máme odhad, který je invariantní na změnu měřítka, ale ne na posunutí. Pro různé hodnoty k dostáváme různé odhady. Při zavedení odhadu bylo ukázano, že za podmínek $x^* > 0$, $F \in \mathcal{D}(G_\gamma)$, $\gamma \in \mathbb{R}$ pro $n \to \infty$, $k \to \infty$, $\frac{k}{n} \to 0$ platí

$$\hat{\gamma}_M \xrightarrow{P} \gamma.$$

Momentový odhad je tedy konzistentním odhadem γ. Pokud navíc na distribuční funkci klademe podmínku druhého druhu, pak podle [15] str. 104

$$\sqrt{k}(\hat{\gamma}_M - \gamma) \xrightarrow{D} \mathrm{N}(\lambda b_{\rho,\gamma}, \mathrm{D}(\gamma)),$$

kde $\lambda b_{\rho,\gamma}$ je vychýlení závislé na parametrech ρ a γ. Momentový odhad je tedy za určitých předpokladů asymptoticky normální.

5.3.2. Odhad optimálního prahového indexu k_0 pomocí metody bootstrap

Existuje více přístupů, jak odhadnout optimální index prahové statistiky k_0 vzhledem k minimalizaci variability a vychýlení. V této části vyjdeme ze stejného pricipu popsaného v sekci 5.2.2.

Zaveďme alternativní odhad parametru γ:

$$\hat{\gamma}_{M_{aux}}(k) = \sqrt{0.5 M_n^{(2)}(k)} + 1 - \frac{2}{3}\left(1 - \frac{M_n^{(1)}(k) M_n^{(2)}(k)}{M_n^{(3)}(k)}\right)^{-1}.$$

Stejně jako v 5.2.2 předpokládáme náhodný výběr $X_1, X_2, \ldots, X_n$. Použijeme empirickou distribuční funkci $\hat{F}$ jako odhad F. Uvažujeme bootstrapový výběr $X_1^*, X_2^*, \ldots, X_{n_1}^*$, $n_1 < n$, z rozdělení o empirické distribuční funkci $\hat{F}$. Momentový odhad z tohoto bootstrapového výběru označme $\hat{\gamma}^*_{n_1,M}$. Alternativní odhad z tohoto bootstrapového výběru označme odhad $\hat{\gamma}^*_{n_1,M_{aux}}(k)$.

Draisma v článku [9] navrhuje minimalizovat bootstrapový odhad střední kvadratické chyby

$$\mathrm{MSE}^*(n_1, k_1) = \mathrm{E}(\hat{\gamma}^*_{n_1,M}(k_1) - \hat{\gamma}^*_{n_1,M_{aux}}(k_1))^2. \tag{5.27}$$

pro $1 \le k_1 < n_1$. Pokud předpokládáme $F \in \mathcal{D}_\gamma$ a platí podmínka druhého druhu pro $\rho < 0$, $\gamma \neq \rho$ a $\gamma \neq 0$ a navíc platí (5.13), pak podle [9] (Corollary 3.1) pro $k_1 \to \infty$, $\frac{k_1}{n_1} \to 0$ platí

$$\frac{k^*_{0,n_1}}{k_0(n)} \left(\frac{n_1}{n}\right)^{\frac{-2\rho}{2\rho-1}} \xrightarrow{P} 1, \tag{5.28}$$

kde k^*_{0,n_1} minimalizuje bootstrapový odhad střední kvadratické chyby (5.27) tj.

$$k^*_{0,n_1} = \arg \min_{1 \le k_1 < n_1} \mathrm{MSE}^*(n_1, k_1).$$

Odtud dostaneme odhad, který je závislý na parametru druhého druhu ρ. Abychom závislost odstranili, budeme uvažovat další bootstrapový náhodný výběr rozsahu $n_2 < n$ z rozdělení o empirické distribuční funkci $\hat{F}$. Stejnými úpravami jako v sekci 5.2.2 dostaneme odhad

$$\widehat{k_0}(n) = \widehat{k_0}(n, n_1) = \frac{[k^*_{0,n_1}]^2}{k^*_{0,n_1^2/n}}, \tag{5.29}$$

který závisí na volbě n_1. Tato závislost bude blíže zkoumána na simulovaných datech v sekci 6.2.3.

Oproti odhadu (5.17) dostáváme odhad, který adaptivně odhaduje hodnotu pomocné statistiky $\hat{\gamma}_{M_{aux}}(k)$. Pokud nepředpokládáme speciální tvar funkce A z (5.13), je potřeba zavést odhad parametru druhého druhu ρ, který bývá značně obtížný a nepřesný. Tento obecnější přístup lze nalézt v [9].

5.4. Pickandsův odhad

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z rozdělení o distribuční funkcí $F \in \mathcal{D}(G_\gamma)$, $\gamma \in \mathbb{R}$. Podle [15] str. 85 pro $k \to \infty$, $\frac{k}{n} \to 0$ platí

$$\frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}} \xrightarrow{P} \frac{4^\gamma - 1}{2^\gamma - 1} - 1 = 2^\gamma.$$

Odtud dostáváme nejjednodušší a nejstarší konzistentní odhad γ z roku 1975:

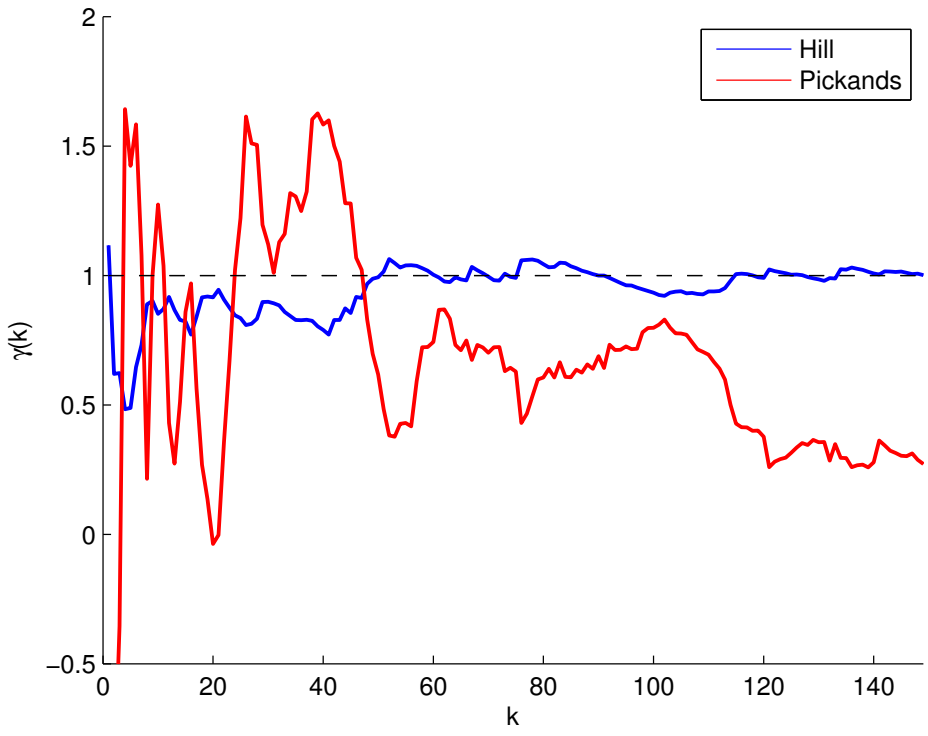
$$\hat{\gamma}_{\mathrm{P}}(k) = (\log 2)^{-1} \log \frac{X_{n-k,n} - X_{n-2k,n}}{X_{n-2k,n} - X_{n-4k,n}}, \quad 1 \le k \le \frac{n}{4},\ k \in \mathbb{N},$$

který se nazývá Pickandsův odhad viz. [26]. Jedná se o kvantilový odhad zobecněného Paretova rozdělení, pro které je medián roven $\frac{2^\gamma - 1}{\gamma}$ a 0.75 kvantil roven $\frac{4^\gamma - 1}{\gamma}$. Navíc za určitých předpokladů podle [15] (Theorem 3.3.5) platí

$$\sqrt{k}(\hat{\gamma}_P - \gamma) \xrightarrow{D} \mathrm{N}(\lambda b_{\rho,\gamma}, \mathrm{D}(\gamma)),$$

kde $\lambda b_{\rho,\gamma}$ je vychýlení závislé na parametrech ρ a γ.

Pickandsův odhad je velmi citlivý na volbu k. Dokonce malá změna k může vyvolat značnou změnu hodnoty odhadu viz obrázek 5.7.



Obrázek 5.7: Pickandsův a Hillův odhad na simulovaném náhodném výběru rozsahu $n = 1000$ z Cauchyho rozdělení.

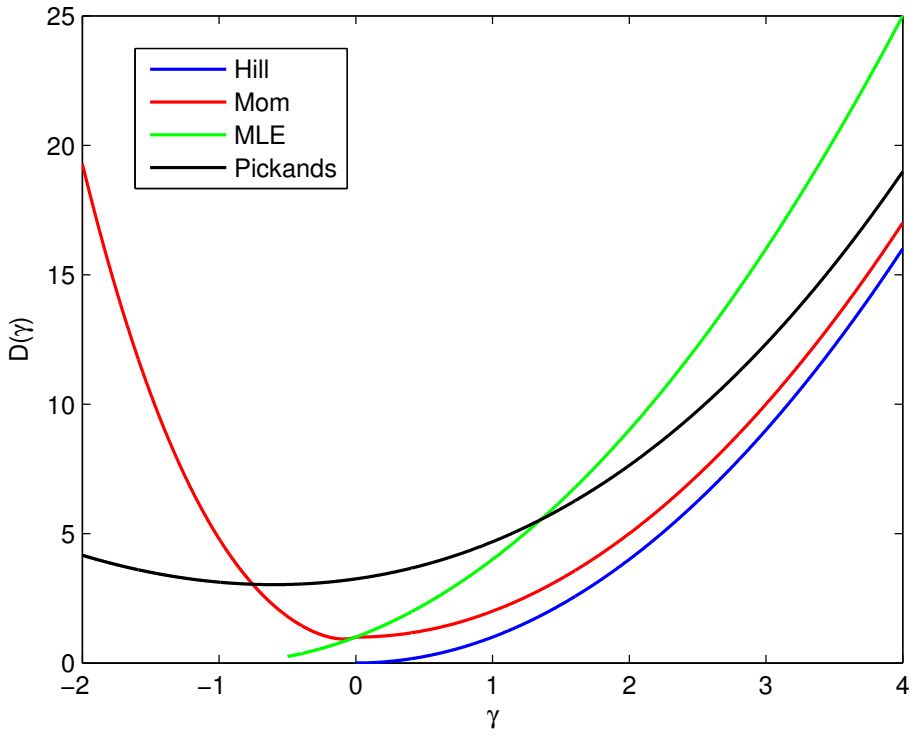
5.5. Shrnutí některých asymptotických vlastností odhadů

Maximálně věrohodný $\hat{\gamma}_{MLE}$, Hillův $\hat{\gamma}_H$, momentový $\hat{\gamma}_M$ a Pickandsův $\hat{\gamma}_P$ odhad jsou konzistentními odhady indexu extrémní hodnoty γ. Další důležitou vlastností všech jmenovaných odhadů je za splnění určitých předpokladů asymptotická normalita, tj. pro $k \to \infty$, $\frac{k}{n} \to 0$

$$\sqrt{k}(\hat{\gamma} - \gamma) \xrightarrow{D} \mathrm{N}\left(\lambda b_{\gamma,\rho}, \mathrm{D}(\gamma)\right),$$

kde $\lambda b_{\gamma,\rho}$ je nějaké vychýlení závislé na γ a parametru druhého druhu ρ.

Obrázek 5.8 představuje závislosti asymptotických rozptylů daných odhadů na parametru γ. Hillův odhad má nejmenší asymptotický rozptyl $\mathrm{D}(\gamma)$ pro kladné hodnoty γ. Maximálně věrohodný odhad má nejmenší asymptotický rozptyl pro záporné hodnoty γ. Pro srovnání odhadů je rovněž potřeba uvážit jejich vychýlení, které závisí na dvou neznámých parametrech. Poto je velmi komplikované odhady analyticky porovnat. Podrobněji viz. [15].



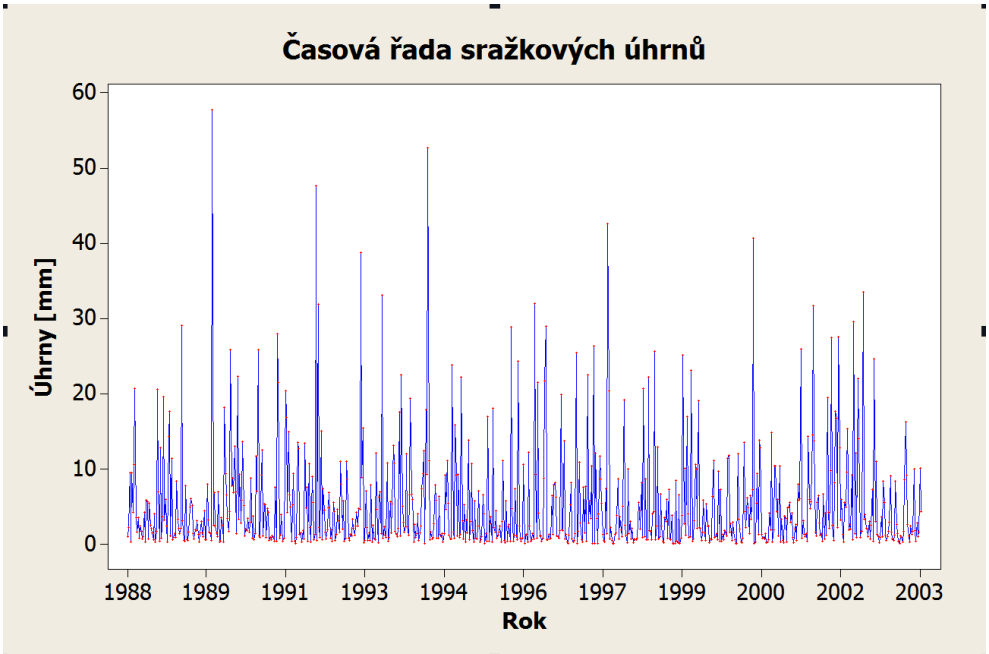
Obrázek 5.8: Asymptotický rozptyl.

6. Analýza dešťových řad

6.1. Data

Hydrologická data jsou vyhodnocována za účelem stanovení odtokového množství z malých povodí a predikci rozvodnění toků. Tím bude docílen spolehlivý provoz městského odvodnění. Tyto systémy mají zabránit vypouštění odpadních vod překračující mezní koncentrace fyzikálních, chemických a biologických parametrů, které ohrožují kvalitu přírodní vody. Navíc tyto systémy mají zajistit ochranu osob a majetku před škodlivými účinky hydrologických situací. Cílem je nalézt kompromis mezi hydrologickou a ekologickou spolehlivostí a vzniklými náklady. Odtoky z malých povodí jsou způsobeny krátkodobými přívalovými dešti. U přívalových dešťů hlavně sledujeme jejich trvání, intenzitu a množství napadlých srážek, které se měří určitým druhem srážkoměru. Konkrétně se používá měřicí přístroj ombrograf, kterým se v České republice obvykle měří od května do září. Nevýhodou ombrografu je, že nedokáže zachytit měření v době, kdy se teplota okolního vzduchu pohybuje pod bodem mrazu a srážky jsou v pevném skupenství, proto přes zimní období neprobíhají většinou žádná měření. Nicméně nejvíce srážkových úhrnů nastává právě mezi měsíci květen a září.

Srážkové úhrny byly vyvzorkovány pro prahový model metodou PDS (Partial Duration Series) podle [19], přičemž bylo získáno 12 různých řad trvající 5, 10, 15, 20, 30, 45, 60, 90, 120, 180, 240, 360 minut. Stanovená kritéria ale nejsou schopna zcela zaručit statistickou nezávislost vzorků. Z tohoto pohledu by měly být dešťové řady dále studovány. Vliv závislých pozorování na semi-parametrické metody odhadu parametru γ a pohled na řadu jako stochastický proces je nad rámec této práce. Statistická analýza těchto řad s využitím metody maximální věrohodnosti již byla provedena v [23].



Obrázek 6.1: Časová řada 16ti let vyvzorkovaných srážkových úhrnů trvajících 360 minut ze stanice Brno - Žabovřesky. Celkem 781 měření.

Statistická analýza bude dále zaměřena na kratké dešťové řady z oblasti Brno - Žabovřesky, pro kterou byla vybrána nejkratší řada obsahující 781 pozorování s dobou trvání 360 minut. Měření proběhla během 16ti let, konkrétně mezi lety 1987 a 2003 viz. obrázek 6.1. Aby bylo možné kvalitně predikovat zřídka se vyskytující jevy, je potřeba co nejkvalitněji odhadnout index extrémní hodnoty γ. Cílem je doporučit pomocí simulací některou z metod odhadu parametru γ uvedenou v této práci.

6.1.1. Testy dobré shody

V tomto odstavci je cílem nalézt vhodná rozdělení, která by co nejlépe popisovala již zmíněnou dešťovou řadu. Účelem je pouze orientačně navrhnout vhodná rozdělení k simulačním studiím pro neparametrické metody. Za pomocí statistického softwaru minitab 16 budou vykresleny probability ploty (pravděpodobnostní grafy) a proveden Aderson-Darlingův test. Testy budou provedeny na všech možných rozděleních, které knihovna v Minitabu obsahuje. Konkrétně se jedná o rozdělení lognormální, normální, exponenciální, Weibullovo, extrémní hodnoty, gamma, logistické, loglogistické.

Probability plot má stejný význam jako kvantilový graf. Odlišnost spočívá pouze v hodnotách na osách, které jsou v případě kvantilového grafu kvantily teoretického testovaného rozdělení a kvantily empirické distribuční funkce. V minitabu u probability plotu jsou funkční hodnoty teoretické distribuční funkce vynášeny v procentech na svislé ose. Ideální shoda by nastala, pokud by všechny hodnoty ležely na vyznačené úsečce. Malá hodnota testovací statistiky indukuje dobrou shodu s testovaným rozdělením.

Na obrázku 6.2 jsou uvedeny výsledky Anderson-Darlingových testů a popisné statistiky srážkové řady. Na základě p-hodnot u jednotlivých rozdělení zamítáme hypotézu, že data pochází z některého testovaného rozdělení, na hladině významnosti $\alpha = 0.05$.

Distribution ID Plot for Žabovřesky PDS 360

Descriptive Statistics

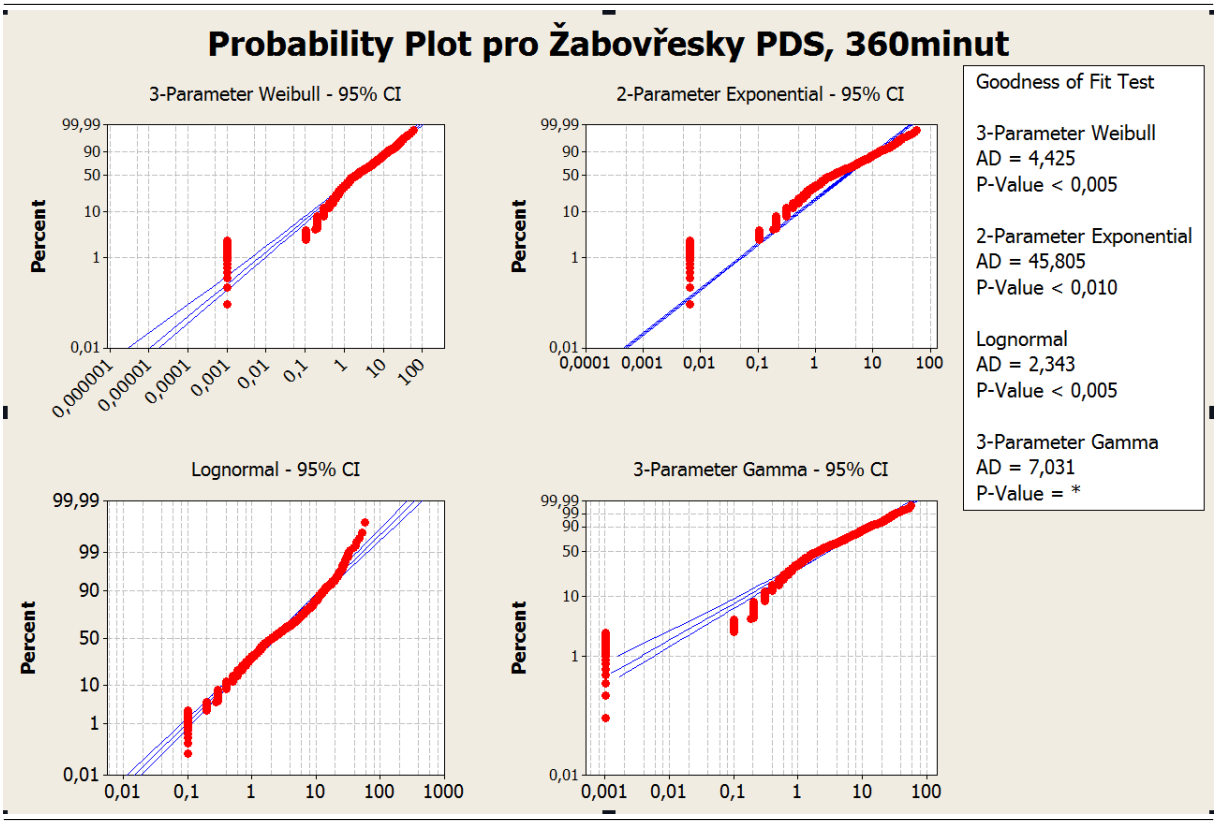
N	N*	Mean	StDev	Median	Minimum	Maximum	Skewness	Kurtosis
781	0	5,12495	7,23387	2,1	0,1	57,7	2,85243	10,8505

Box-Cox transformation: Lambda = 0
Goodness of Fit Test

Distribution	AD	P	LRT P
Normal	78,864	<0,005	
Box-Cox Transformation	2,343	<0,005	
Lognormal	2,343	<0,005	
3-Parameter Lognormal	2,022	*	0,044
Exponential	37,612	<0,003	
2-Parameter Exponential	45,805	<0,010	0,000
Weibull	7,876	<0,010	
3-Parameter Weibull	4,425	<0,005	0,000
Smallest Extreme Value	124,910	<0,010	
Largest Extreme Value	48,186	<0,010	
Gamma	12,445	<0,005	
3-Parameter Gamma	7,031	*	0,000
Logistic	55,379	<0,005	
Loglogistic	3,825	<0,005	
3-Parameter Loglogistic	3,075	*	0,000

Obrázek 6.2: Výsledky Anderson-Darlingových testů pro 360 minutovou dešťovou řadu ze stanice Brno - Žabovřesky. Celkem 781 měření.

Z probability plotů na obrázku 6.3 je patrná bimodalita. Je možné pozorovat celou skupinu odlehlých pozorování. Jednou z příčin je, že ombrograf někdy zaznamenává velmi malé intenzity. Může se například jednat o rosu. Pro lepší shodu s teoretickým rozdělením je potřeba tyto hodnoty odfiltrovat. Dále budou prováděny testy dobré shody na kratší dešťové řadě, která obsahuje srážkové úhrny vyšší než 5mm.



Obrázek 6.3: Některé probability ploty pro 360 minutovou dešťovou řadu ze stanice Brno - Žabovřesky. Celkem 781 měření.

Z výsledků na obrázku 6.4 již nezamítáme hypotézu, že data pochází z Weibullova nebo exponenciálního rozdělení, na hladině významnosti $\alpha = 0.05$. p-hodnota pro exponenciální rozdělení je hraniční. Příčinou jsou velké odchylky pro malé kvantily, které vzhledem k zájmu modelování chvostu nejsou podstatné. Vzhledem k výsledkům těchto testů dále zaměřme simulační studie na Weibullovo a exponenciální rozdělení.

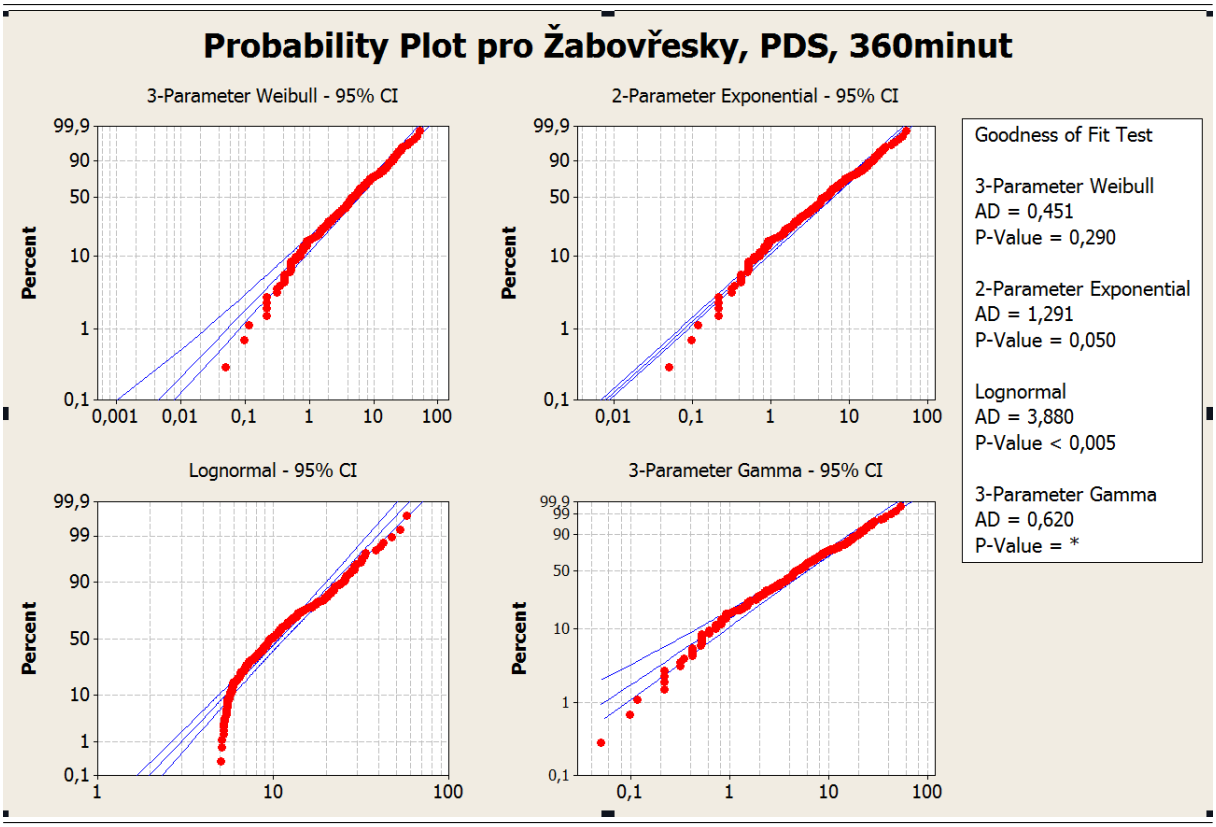
Minitab navíc doporučuje na základě maximalizace věrohodnostní funkce Weibullova rozdělení odhad parametru tvaru $\alpha = 0.91$ a měřítka $\lambda = 7.61$. Distribuční funkce pro Weibullovo rozdělení je tvaru

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^{\alpha}\right), \quad x > 0.$$

Pro exponenciální rozdělení dostaneme parameter měřítka $\frac{1}{\lambda} = 7.9$. Distribuční funkce je tvaru

$$F(x) = 1 - \exp\left(-\lambda x\right), \quad x > 0.$$

Vzhledem k tomu, že odhady jsou invariantní vůči změně měřítka a rozdělení leží v Gumblově oboru atraktivity, není nutné brát tyto parametry v potaz.



Obrázek 6.4: Některé probability ploty pro 360 minutovou dešťovou řadu ze stanice Brno - Žabovřesky. Řada obsahuje pouze hodnoty vyšší než 5.

6.2. Porovnání odhadů na základě simulací

Obvykle se ve statistice hodnotí kvalita odhadu podle střední kvadratické chyby. Ideální odhad je takový, který má nulové vychýlení a co nejmenší rozptyl. Tyto charakteristiky jsou ale analyticky velmi obtížně odvoditelné. Prostor dostávají numerické simulační metody, především metoda Monte Carlo.

6.2.1. Popis simulací

Simulační studie bude zaměřena na maximálně věrohodý odhad $\hat{\gamma}_{MLE}$, Hillův odhad $\hat{\gamma}_H$, momentový odhad $\hat{\gamma}_M$ a Pickandsův odhad $\hat{\gamma}_P$. Jako výchozí testovací model bude použito Gumblovo a exponenciální rozdělení z sekce 6.1.1.

Nejprve bylo z daného rozdělení vygenerováno $MC = 500$ náhodných výběrů o rozsahu $n = 781$ a provedeny výše zmiňované odhady. Za účelem ještě vyššího snížení variability byla simulace opakována $R = 5$ krát. Simulaci velmi zpomaluje maximálně věrohodný odhad, u kterého je potřeba řešit $R \times MC \times (n-1)$ nelineárních rovnic. Pro tento odhad byla využita matlabovská funkce gpfit. Pro zbylé odhady byla navíc zvlášť provedena simulace rozsahu $MC = 5000$, $R = 100$ za účelem zvýšení přesnosti.

Optimální index prahové statistiky je počítán jako

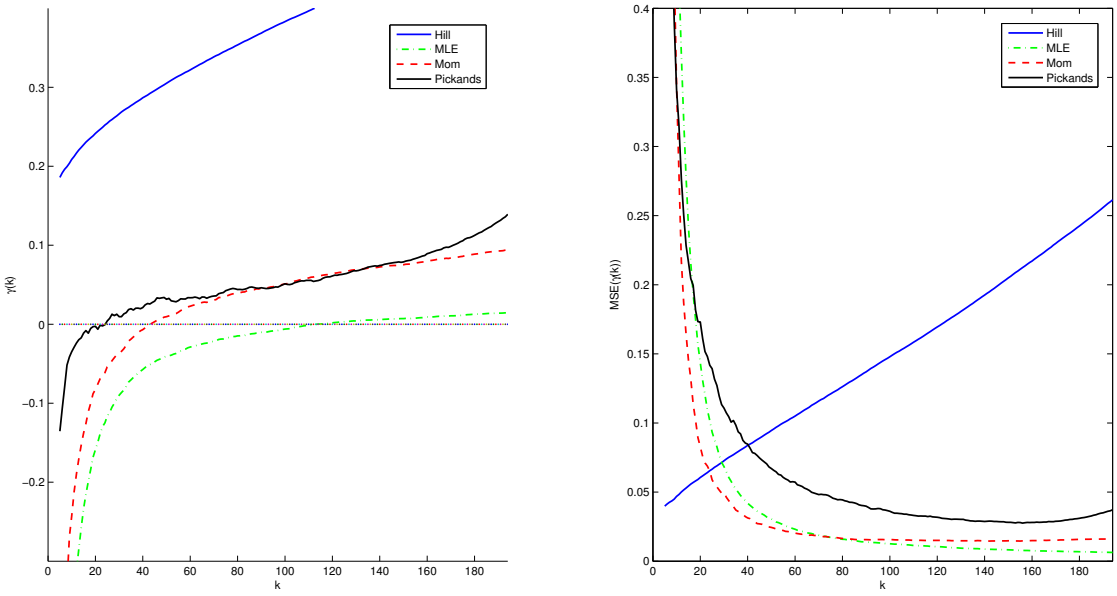
$$\overline{\overline{k_0}} = \frac{1}{R}\sum_{r=1}^{R}\left(\arg\min_{1\le k<n}\widehat{MSE}_r\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\arg\min_{1\le k<n}\sum_{j=1}^{MC}(\hat{\gamma}_{r,j}(k)-\gamma)^2\right),$$

kde $\hat{\gamma}_{r,j}$ značí momentový, Hillův nebo maximálně věrohodný odhad. V případě Pickandsova odhadu je potřeba dbát zřetel na to, pro které hodnoty k je definovaný. Mez pro optimální index prahové statistiky je potřeba upravit. Konkrétně se pro Pickansův odhad spočte

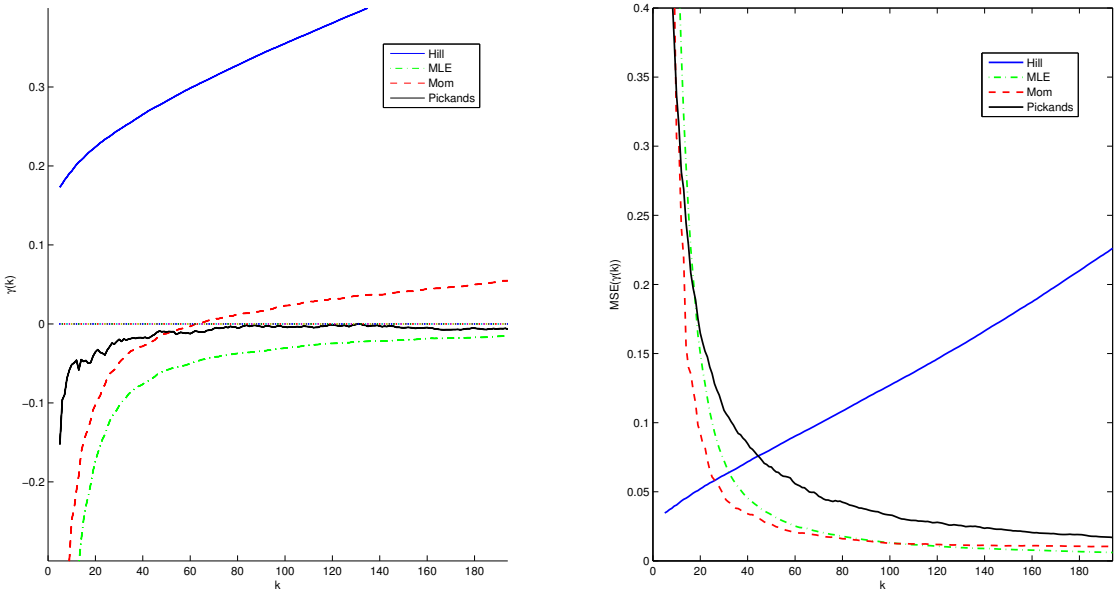
$$\overline{\overline{k_0}} = \frac{1}{R}\sum_{r=1}^{R}\left(\arg\min_{1\le k\le 194}\sum_{j=1}^{MC}(\hat{\gamma}_{P_{r,j}}(k)-\gamma)^2\right).$$

6.2.2. Výsledky simulací

Na obrázku 6.5, případně 6.6, jsou porovnány všechny zmíněné odhady na Weibullově, případně exponenciálním, rozdělení. Oba obrázky dávají přibližně podobné výsledky. Pickandsův odhad je v tomto případě definován pouze pro $k < 195$. Vzhledem k počtu opakování dochází k vyhlazování průběhů odhadů. Není překvapující, že Hillův odhad je s rostoucím k nejvíce vychýlený, přestože je stále konzistentním odhadem parametru $\gamma = 0$, pro $k \to \infty$ $\frac{k}{n} \to 0$. Tento odhad se hodí především pro vyšší hodnoty γ, pro které je dokonce odhadem s nejmenším rozptylem vůči ostatním zmiňovaným odhadům. Jestliže je γ blízká nule, hrozí, že Hillův odhad přesáhne skutečnou hodnotu γ. Tato situace může mít za následek příliš pesimistické predikce. Pickandsův odhad dává rozumnější výsledky, ale obecně není tolik doporučován kvůli jeho vysoké citlivosti na volbu k a navíc je počítán z velmi malého počtu pozorování. Nejlepší výsledky dávají momentový a maximálně věrohodný odhady. Z grafů plyne podezření, že střední kvadratická chyba momentového nebo věrohodného odhadu může dále klesat pro $k > 194$. Tyto dva odhady budou porovnány ještě jednou.

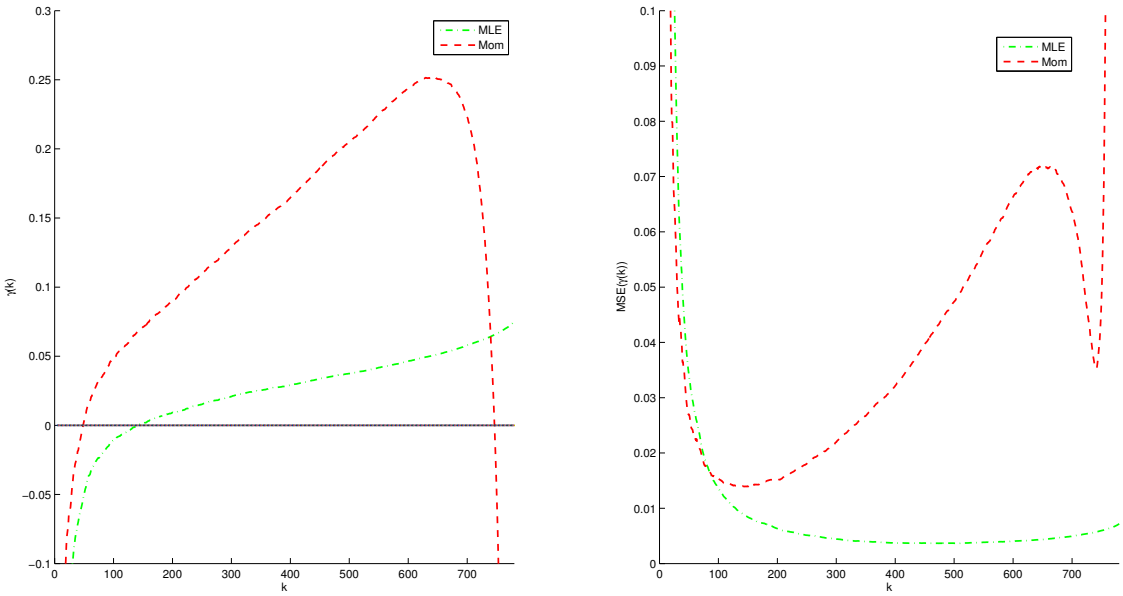


Obrázek 6.5: Simulace z Weibullova rozdělení o rozsahu 781 s počtem opakování $R = 5$ krát $MC = 500$. Průměrné odhady parametru γ v závislosti na k (vlevo). Střední kvadratická chyba v závislosti na k (vpravo).



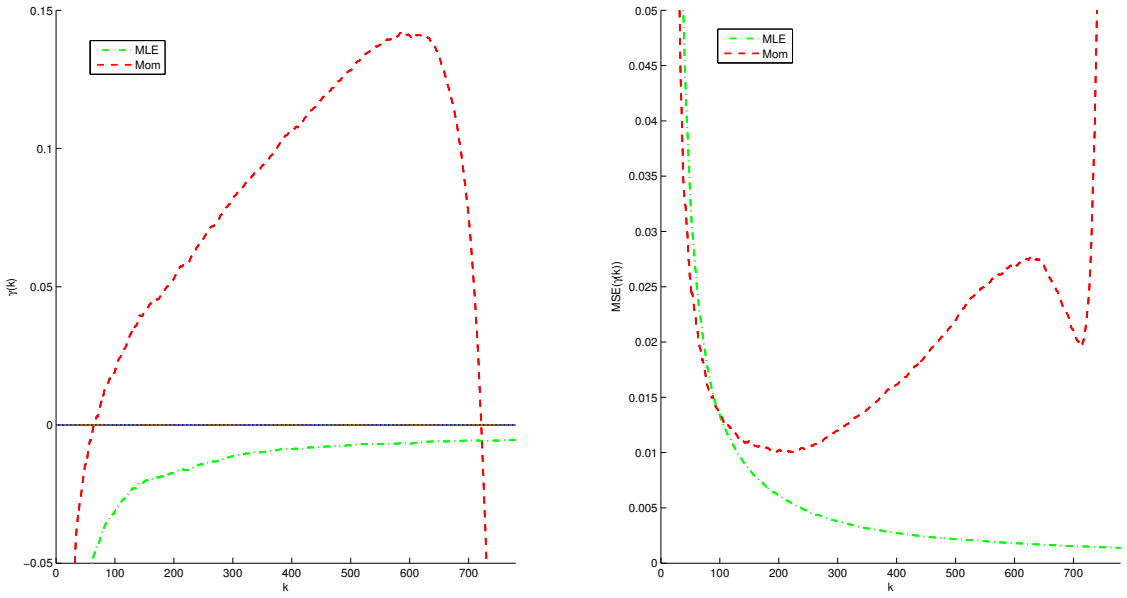
Obrázek 6.6: Simulace z exponenciálního rozdělení o rozsahu 781 s počtem opakování $R = 5$ krát $MC = 500$. Průměrné odhady parametru γ v závislosti na k (vlevo). Střední kvadratická chyba v závislosti na k (vpravo).

Srovnání momentového a maximálně věrohodného odhadu na Weibullově rozdělení je znázorněno na obrázku 6.7. Pro hodnoty k blízké n může docházet k numerické nestabilitě. Momentový odhad je totiž počítán pomocí zlogaritmovaných hodnot. Malá pozorování mohou být blízká nule, ale nejsou až tak zajímavá. Větší pozornost si zaslouží hodnoty momentového odhadu pro $k = 100, \ldots, 250$, pro které vychází střední kvadratická chyba tohoto odhadu nejmenší a dává velmi podobné hodnoty. Oproti maximálně věrohodnému odhadu je momentový odhad více vychýlen, a tím dává o něco horší výsledky.



Obrázek 6.7: Simulace z Weibullova rozdělení o rozsahu 781 s počtem opakování $R = 5$ krát $MC = 500$. Průměrné odhady parametru γ v závislosti na k (vlevo). Střední kvadratická chyba v závislosti na k (vpravo).

Srovnání momentového a maximálně věrohodného odhadu na exponenciálním rozdělení je znázorněno na obrázku 6.7. Dostáváme obdobné průběhy jako u Weibullova rozdělení. Opět dochází k numerické nestabilitě pro hodnoty k blízké n. Střední kvadratická chyba je ale u momentového odhadu nízka pro $k = 150, \ldots, 300$. Není překvapující, že maximálně věrohodný odhad dosahuje nejmenší chyby pro $k = 780$. Tento odhad vychází z rozdělení ,,velkých pozorování", která mají přibližně zobecněné Paretovo rozdělení. Pro $\gamma = 0$ je exponenciální rozdělení speciálním případem zobecněného Paretova rozdělení.



Obrázek 6.8: Simulace z exponenciálního rozdělení o rozsahu 781 s počtem opakování $R = 5$ krát $MC = 500$. Průměrné odhady parametru γ v závislosti na k (vlevo). Střední kvadratická chyba v závislosti na k (vpravo).

	Hill	MLE	Mom	Pickands
Weibull$(7.6, 0.92)$, $MC = 5000$, $R = 100$, $\gamma = 0$				
$\overline{\overline{k_0}}$	3	-	136	152
$\overline{\overline{\hat{\gamma}}}(\overline{\overline{k_0}})$	0.17254	-	0.06821	0.08546
$\overline{RMSE}$	0.19420	-	0.11938	0.17099
Weibull$(7.6, 0.92)$, $MC = 500$, $R = 5$, $\gamma = 0$				
$\overline{\overline{k_0}}$	2	474	142	156
$\overline{\overline{\hat{\gamma}}}(\overline{\overline{k_0}})$	0.16332	0.01460	0.07356	0.08413
$\overline{RMSE}$	0.19252	0.06053	0.12067	0.16604
exponenciální(0.13), $MC = 5000$, $R = 100$, $\gamma = 0$				
$\overline{\overline{k_0}}$	2	-	184	194
$\overline{\overline{\hat{\gamma}}}(\overline{\overline{k_0}})$	0.10154	-	0.05233	-0.00337
$\overline{RMSE}$	0.18054	-	0.10154	0.12962
exponenciální(0.13), $MC = 500$, $R = 5$, $\gamma = 0$				
$\overline{\overline{k_0}}$	2	780	213	194
$\overline{\overline{\hat{\gamma}}}(\overline{\overline{k_0}})$	0.15224	-0.00551	0.05732	-0.00648
$\overline{RMSE}$	0.17984	0.03731	0.10019	0.12998

Tabulka 6.1: Výsledky simulací

Souhrn výsledků simulací je uveden v tabulce 6.1. Statistika $\overline{RMSE}$ představuje odmocninu ze střední kvadratické chyby pro $\overline{\overline{\gamma}}(\overline{\overline{k_0}})$. Dále je uvedena průměrná hodnota odhadu parametru γ opět pro $\overline{\overline{k_0}}$. Pickandsův odhad je kvantilovým odhadem zobecněného

Paretova rozdělení. Proto pro tento odhad v případě exponenciálního rozdělení vychází nejmenší kvadratická chyba pro maximální možné k. Na základě těchto výsledků vychází vítězně maximálně věrohodný odhad. Ovšem momentový odhad také nedává špatné výsledky.

6.2.3. Momentový odhad pomocí metody bootstrap

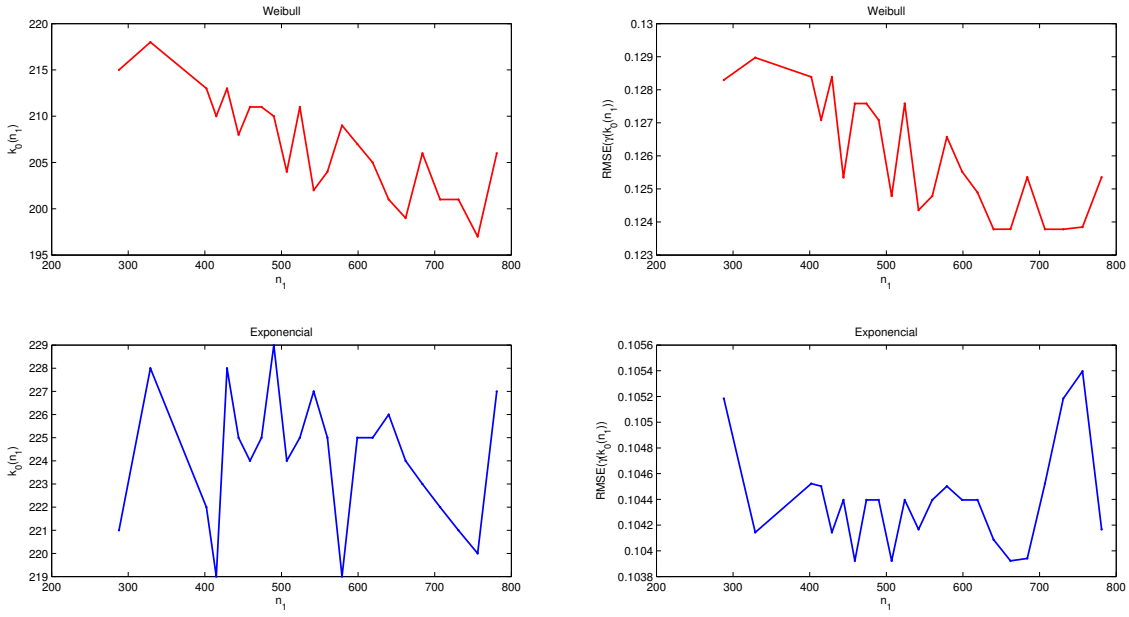
V praxi bývá velmi obtížné odhadnout prahový index k_0. V této části bude blíže zkoumána volba tohoto prahu pro momentový odhad pomocí metody bootstrap na simulovaných datech z rozdělení, která co nejlépe popisují zmíněnou dešťovou řadu. Jako model tedy uvažujme opět Weibullovo a exponenciální rozdělení.

Uvažujme metodu popsanou v sekci 5.3.2. Odhad $\widehat{k_0}$ definovaný vztahem (5.29) závisí na volbě velikosti bootstrapového výběrů $n_1 < n$. Nejlépe bychom volili n_1 takové, že hodnota $\widehat{k_0}(n_1)$ je rovna hodnotám z tabulky 6.1. Generujeme proto $MC = 500$ náhodných výběrů, pro které vypočítáme

$$\overline{\widehat{k_0}}(n_1) = \frac{1}{MC} \sum_{i=1}^{MC} \widehat{k_{0i}}(n_1),$$

kde volíme $n_1 = n^{1-\epsilon}$, $\epsilon \in (0, 1)$. Počet bootstrapových výběrů n^* je zvolen podle zvyklostí 250.

Na obrázku 6.9 je znázorněn průběh $\overline{\widehat{k_0}}(n_1)$ pro různě zvolené hodnoty ϵ. V případě Weibullova rozdělení lze pozorovat mírný pokles střední kvadratické chyby s rostoucím n_1.



Obrázek 6.9: Závislost boostrapového odhadu indexu k_0 na volbě rozsahu boostrapového náhodného výběru n_1 na simulovaných datech z exponenciálního (modře) a Weibullova(červeně) rozdělení s rozsahem $n = 781$ a počtem opakování $MC = 500$, $n^* = 250$. Průměrné odhady $\overline{\widehat{k_0}}(n_1)$ (vlevo). Odmocnina ze střední kvadratické chyby odhadu $\hat{\gamma}_M(k_0(n_1))$ (vpravo).

Souhrn výsledků této simulace je uveden v tabulce 6.2. Nejlepší výsledky jsou dosaženy pro $\epsilon = 0,005$, respektive $n_1 = 756$. V případě exponenciálního rozdělení se odhad indexu k_0 příliš neliší od výsledků z tabulky 6.1. V případě Weibullova rozdělení dostáváme stále rozumné výsledky vzhledem k střední kvadratické chybě z obrázku 6.7.

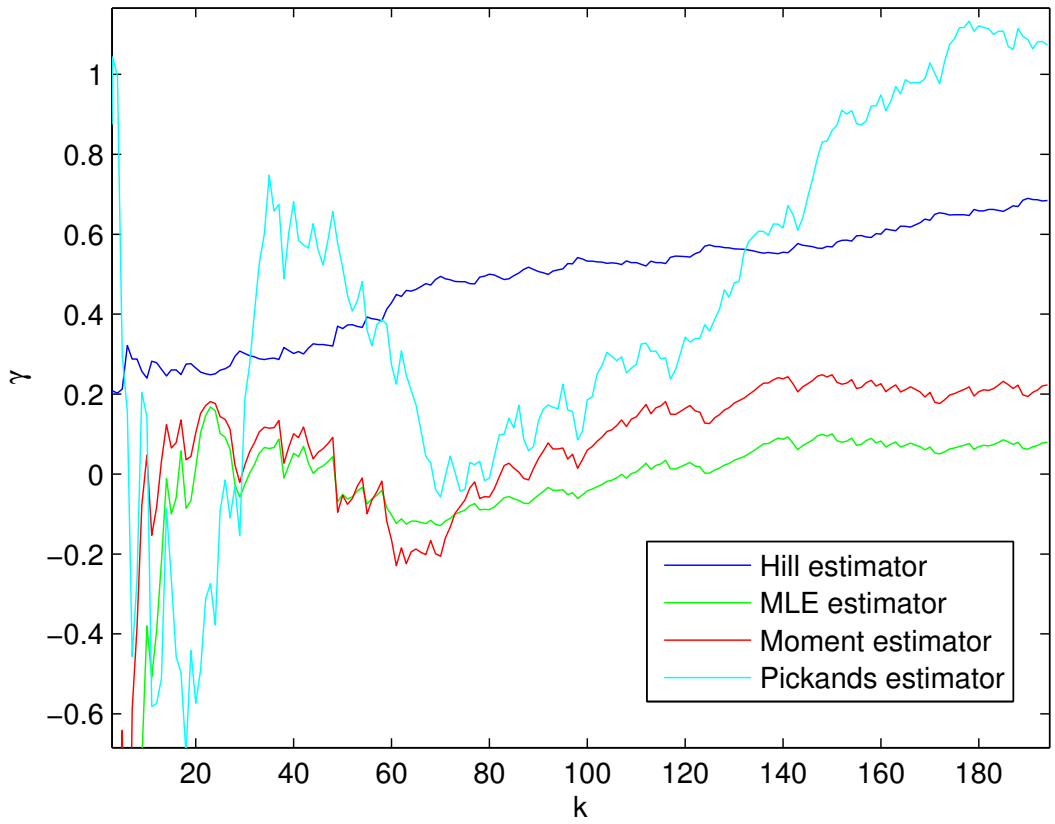
Vstup			Weibull$(7.6, 0.92)$, $\gamma = 0$			exponenciální(0.13), $\gamma = 0$		
ϵ	n_1	n_2	$\widehat{\widehat{k_0}}$	$\overline{\hat{\gamma}}_M(\widehat{k_0})$	RMSE$(\overline{\hat{\gamma}}_M(\widehat{k_0}))$	$\widehat{\widehat{k_0}}$	$\overline{\hat{\gamma}}_M(\widehat{k_0})$	RMSE$(\overline{\hat{\gamma}}_M(\widehat{k_0}))$
0,000	781	781	206	0,09560	0,12535	227	0,06328	0,10417
0,005	756	732	197	0,09270	0,12385	220	0,06199	0,10540
0,010	731	684	201	0,09313	0,12378	221	0,06219	0,10518
0,015	707	640	201	0,09313	0,12378	222	0,06222	0,10452
0,020	684	599	206	0,09560	0,12535	223	0,06204	0,10394
0,025	662	561	199	0,09316	0,12378	224	0,06254	0,10392
0,030	640	524	201	0,09313	0,12378	226	0,06282	0,10409
0,035	619	491	205	0,09453	0,12489	225	0,06250	0,10440
0,040	599	459	207	0,09605	0,12552	225	0,06250	0,10440
0,045	579	429	209	0,09658	0,12657	219	0,06100	0,10450
0,050	560	402	204	0,09415	0,12479	225	0,06250	0,10440
0,150	288	106	215	0,09897	0,12830	221	0,06219	0,10518

Tabulka 6.2: Výsledky simulací

6.3. Odhady na reálných datech

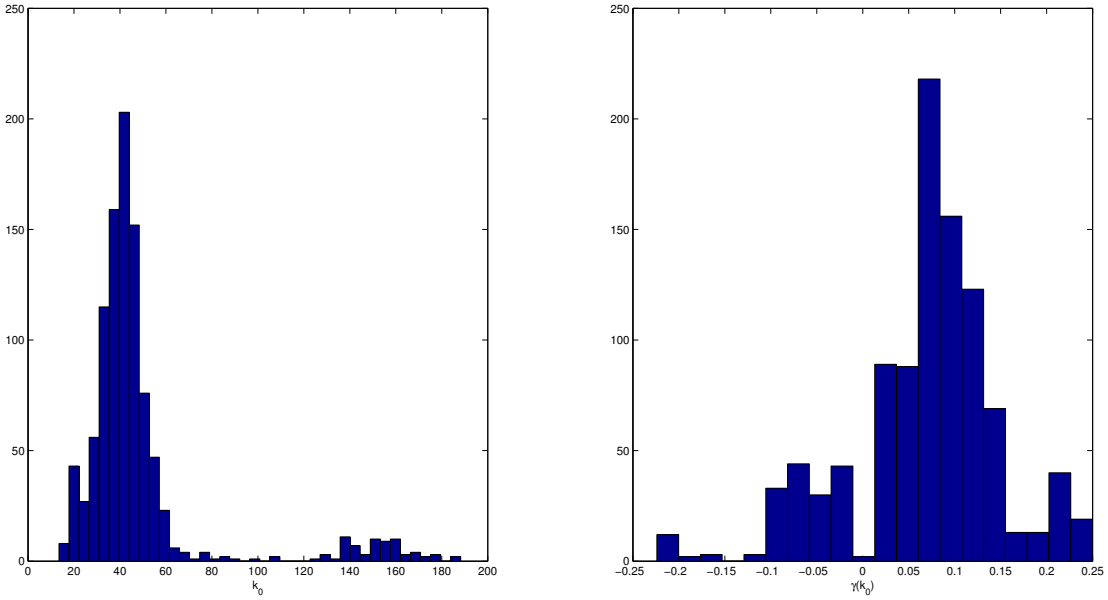
Vraťme se k srážkám ze stanice Brno - Žabovřesky s dobou trvání 360 minut, které jsou udávány v milimetrech na 3 desetinná místa. Řady srážkových úhrnů mohou obsahovat stejné hodnoty. Nejčastější příčinou může být zaokrouhlování a nepřesná měření. Odhady parametru γ byly odvozeny s úvahou různých hodnot náhodného výběru. Proto je do těchto dat vnesen šum, který má rovnoměrné rozdělení R(-0.0005,0.0005).

Na obrázku 6.10 je znázorněn průběh všech odhadů zavedených v této práci na vybrané dešťové řadě. Vidíme, že Pickandsův odhad má vysokou variabilitu. Hillův odhad je značně vychýlen kvůli hodnotám $\hat{\gamma}$, které jsou blízké nule. Pro $k = 1, 2, \ldots, 80$ dávají momentové a maximálně věrohodné odhady velmi podobné výsledky.



Obrázek 6.10: Odhady parametru γ na dešťové řadě vyvzorkovaných srážkových úhrnů trvajících 360 minut ze stanice Brno - Žabovřesky. Celkem 781 měření.

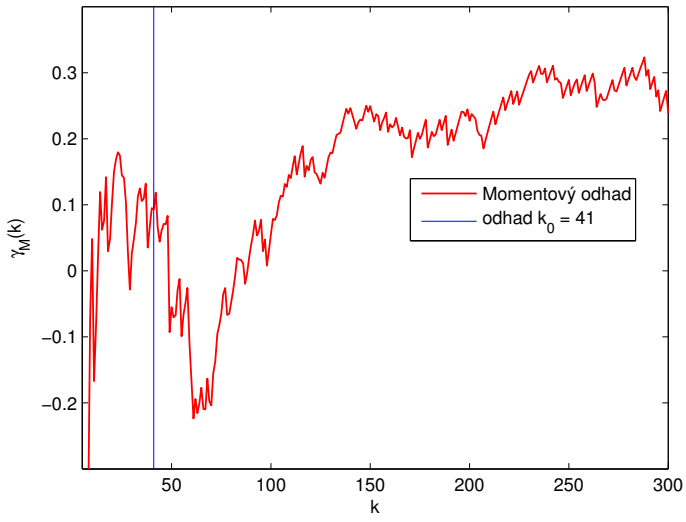
Pro momentový odhad je dále proveden odhad k_0 pomocí metody bootstrap. Byl zvolen rozsah bootstrapového náhodného výběru $n_1 = 756$, $n^* = 250$. Odhad byl opakován $MC = 1000$ krát. Odhady jsou znázorněny pomocí histogramů 6.11. Vzhledem k některým odlehlým odhadům $\widehat{k_0}$ je vhodnější místo výběrového průměru použít robustní mediánový odhad střední hodnoty. Výsledný odhad prahového indexu je zaznamenán v tabulce 6.3 včetně směrodatné odchylky odhadu $\hat{\gamma}_M(\widehat{k_0})$, která může posloužit k sestavení intervalu spolehlivosti.



Obrázek 6.11: Histogramy odhadů $\widehat{k_0}$(vlevo) a $\hat{\gamma}_M(\widehat{k_0})$ (vpravo).

Stanice Brno - Žabovřesky. Srážky s dobou trvání 360 minut.								
Vstup						Výstup		
n	ϵ	n_1	n_2	MC	n^*	$\overline{\hat{\gamma}}_M(\widehat{k_0})$	$\widehat{k_0}$	směr. odchylka
781	0.005	756	732	1000	250	0.0701	41	0.0814

Tabulka 6.3: Výsledky momentového odhadu na reálných datech



Obrázek 6.12: Momentový odhad parametru γ a odhad optimálního prahu k_0 pomocí metody bootstrap na datech ze Stanice Brno - Žabovřesky. Srážky s dobou trvání 360 minut.

6.4. Shrnutí

Srážkové úhrny vybrané deštové řady lze přibližně popsat pomocí Weibullova nebo exponenciálního rozdělení. Tato rozdělení leží v oboru atraktivity rozdělení Gumbelova typu. Při simulačních studiích na těchto rozděleních vychází vítězně parametrický odhad maximální věrohodnoti založený na věrohodnostní funkci zobecněného Paretova rozdělení. Rozumné výsledky dává i semi-parametrický momentový odhad, který může být mírně vychýlený. Pro tento odhad lze dále rozumně odhadnout index prahové statistiky na základě metody bootstrap i přesto, že index extrémní hodnoty $\gamma = 0$. S touto ,,singularitou" má většina semi-parametrických odhadů problémy a bývá předmětem mnoha dnešních studií. Rovněž lze dále zkoumat závislost odhadu na volbě počtu bootstrapových výběrů n^* nebo uvažovat rozdělení z jiné než Gumbelovy sféry.

Na konkrétní dešťové řadě byl proveden zmíněný momentový odhad parametru γ za pomocí metody bootstrap, který vychází podle očekávání blízko nule. V modelu se předpokládá mimo jiné nezávislost a stejné rozdělení pozorování, splnění podmínky druhého druhu pro parametr $\rho < 0$, $\rho \neq \gamma$ a speciální tvar funkce A podle vztahu 5.13. Dešťovou řadu nejspíš netvoří nezávislá stejně rozdělená pozorování. Dále je vhodné zkoumat vliv stochastických procesů na momentový odhad. Rovněž lze zkoumat odhad parametru ρ, pomocí kterého lze získat odhad optimální prahové statistiky pro obecný tvar funkce A.

7. Závěr

První část diplomové práce je věnována teorii rozdělení extrémních hodnot. Je zde zformulována a dokázána limitní věta pro rozdělení maxim a uvedeny základní vlastnosti rozdělení extrémního typu. Tato část také obsahuje spoustu teoretických poznatků, které jsou dále využity při zavádění odhadů parametru rozdělení extrémního typu.

Stěžejní část diplomové práce tvoří semi-parametrické metody odhadu indexu extrémní hodnoty. Tyto metody včetně parametrické metody maximální věrohodnosti jsou teoreticky popsány a implementovány s ohledem na výpočetní rychlost. Dále jsou srovnány na základě simulací z vhodně vybraných rozdělení, která jsou velmi blízká rozdělení srážkových úhrnů z vybrané dešťové řady. Z těchto simulačních studií vychází vítězně parametrický maximálně věrohodný odhad. Pro danou řadu je dále navržen nový odhad, který kombinuje momentový odhad s metodou bootstrap. Tento odhad vznikl zjednodušením odhadu uvedeného v článku [9] a jeho hlavní výhodou je nepotřeba odhadu parametru druhého druhu. Navržený odhad nalezne uplatnění i pro jiná data než srážkové řady.

Součástí práce je simulační program s propracovaným uživatelským prostředím pro odhad optimálního prahového indexu pro vybraná rozdělení patřící do oboru atraktivity rozdělení Fréchetova typu. V programu lze mimo jiné odhadovat index extrémní hodnoty z reálných dat.

Literatura

[1] Anděl, J.: *Základy matematické statistiky.* MATFYZPRESS, Praha, ISBN 80-7378-001-1, 2007.

[2] Baxevani A.: *Computer Intensive Statistical Methods (přednášky).* Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, SE-412 96 Göteborg, Sweden, 2011.

[3] Brito, M., Freitas, M. C. A.: *Consistent estimation of the tail index for dependent data.* Statistics & Probability Letters, Volume 80, Issues 23–24, 1–15, Pages 1835-1843, December 2010.

[4] Brockwell, J. P., Davis, A. R.: *Time series: Theory and methods.* 2-nd ed., Springer-Verlag, New York, 1991 (corrected 2-nd printing 1993).

[5] Beirlant, J.; Goegeburg, Y.; SEGERS, J.; TEUGELS, J. *Statistics of Extremes: Theory and Applications.* Wiley, 522 s. ISBN 978-0-471-97647-9, 2004.

[6] Coles, S.,: *An Introduction to Statistical Modeling of Extreme Values* Springer-Verlag London Berlin Heidelberg, ISBN 1-85233-459-2, 2001.

[7] Danielsson, J., de Haan, L., Peng L., de Vries C. G.: *Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation.* Journal of Multivariate Analysis 76, 226-248, 2001.

[8] Dienstbier, J.: *Problémy s extrémy.* [online]. 2008, [cit. 2013-06-04]. Dostupné na: http://www.karlin.mff.cuni.cz/ antoch/robust08/postery/dienstbier.pdf.

[9] Draisma, G., De Haan, L., Peng L.: *A Bootstrap-based Method to Achive Optimality in Estimating the Extreme-value Index.* Econometric Institute Report EI 2000-18/A.

[10] Drees, H., Kaufmann, E.:*Selecting the optimal sample fraction in univariate extreme value estimation.* Stochastic Processes and their Applications 75 149-172, 1998.

[11] Efron B.: *Bootstrap methods: another look at the jackknife.* Ann. Statist. 7, 1 – 26, 1979.

[12] Efron B., and Tibshirani R.: *An Introduction to the Bootstrap.* Chapman & Hall/CRC. ISBN 0-412-04231-2, 1994.

[13] Fraga Alves, M.I., *A location invariant Hill-type estimator.*In Extremes, Vol.4, No.3, pp. 199–217, 2001. Kluwer Academic Publishers, 2002.

[14] Gomes, M. I., Oliveira, O.: *The Bootstrap Methodology in Statistics of Extremes - Choice of the Optimal Sample Fraction.* In Extremes, Vol.4, No.4, pp. 331-358. Kluwer Academic Publishers, 2002

[15] de Haan, L.,Ferreira, A.: *Extreme Value Theory. An Interduction.* Springer, 421 p. ISBN 0387239464, 2006.

[16] Hill, B.M.: *A simple general approach to inference about the tail of a distribution.* Ann. Statist. 3, 1163-1174. 1975.

[17] Kotz, S.; Nadarajah, S.: *Extreme Value Distributions: Theory and Applications.* World Scientific, 2001. 185 s. ISBN 1860942245.

[18] Ling, S., Peng, L.: *Hill's estimator for the tail index of an ARMA.* Journal of Statistical Planning and Inference, Volume 123, Issue 2, Pages 279-293, 1 July 2004.

[19] Madsen, H., Mikkelsen, P. S., Rosbjerg, D., Harremoes, P.,:*Regional estimation of rainfall intensity–duration–frequency curves using generalized least squares regressions of partial duration series.* Water Resources Research 38 (11), 1239. doi:10.1029/2001WR001125. 2002.

[20] Matlab. Programový systém pro provádění matematických výpočtů. Komerční software, verze R2012b. [online]. http://www.mathworks.com. The MathWorks, květen 2013.

[21] Michálek, J.: *Pravděpodobnost a statistika I (přednášky).* Vysoké učení technické v Brně, Fakulta strojního inženýrství, Matematické inženýrství. 2010.

[22] Michálek, J.: *Pravděpodobnost a statistika II (přednášky).* Vysoké učení technické v Brně, Fakulta strojního inženýrství, Matematické inženýrství. 2011.

[23] Hellebrand R., MichálekJ., Fusek M., Prax P.,:*Modelling precipitation extremes for the purpose of urban drainage in the Czech Republic.* Brno University of Technology, Faculty of Civil Engineering, Institute of Municipal Water, 2012.

[24] Morton, P. D., Mak W., Wood K. R.,; *Monte Carlo bounding techniques for determining solution quality in stochastic programs.* Operations Research Letters 24, 47-56, 1999.

[25] Picek J.: *Testy a odhady Paretova indexu.* ROBUST 2004: Sborník prací 13. letní školy JČMF ROBUST 2004 uspořádané Jednotou českých matematiků a fyziků., Praha, JČMF 57-552-04, ISBN 80-7015-972-3, 2004.

[26] Pickands, J.: *Statistical inference using extreme order statistics.* Ann. Statist. 3, 119-131, 1975.

[27] Protter, H. M.: *Basic Elements of Real Analysis.* Springer, ISBN 0-387-98479-8, 1998.

[28] Radford M. N.: *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Technical Report, Department of Computer Science, University of Toronto, 25, CRG-TR-93-1, 1993.

[29] Rao, R. C.:*Lineární metody statistické indukce a jejich aplikace.* 1. vyd. Praha: Academia nakladatelství ČSAV, 1978.

[30] Robert, C. P., Casella, G.,: *Monte Carlo Statistical Methods.* (2nd ed.). New York: Springer, ISBN 0-387-21239-6, 2004.

[31] van der Vaart A. W.: *Asyptotic Statistic.* Cambridge University Press, ISBN 0 521 49603 9, 1998.

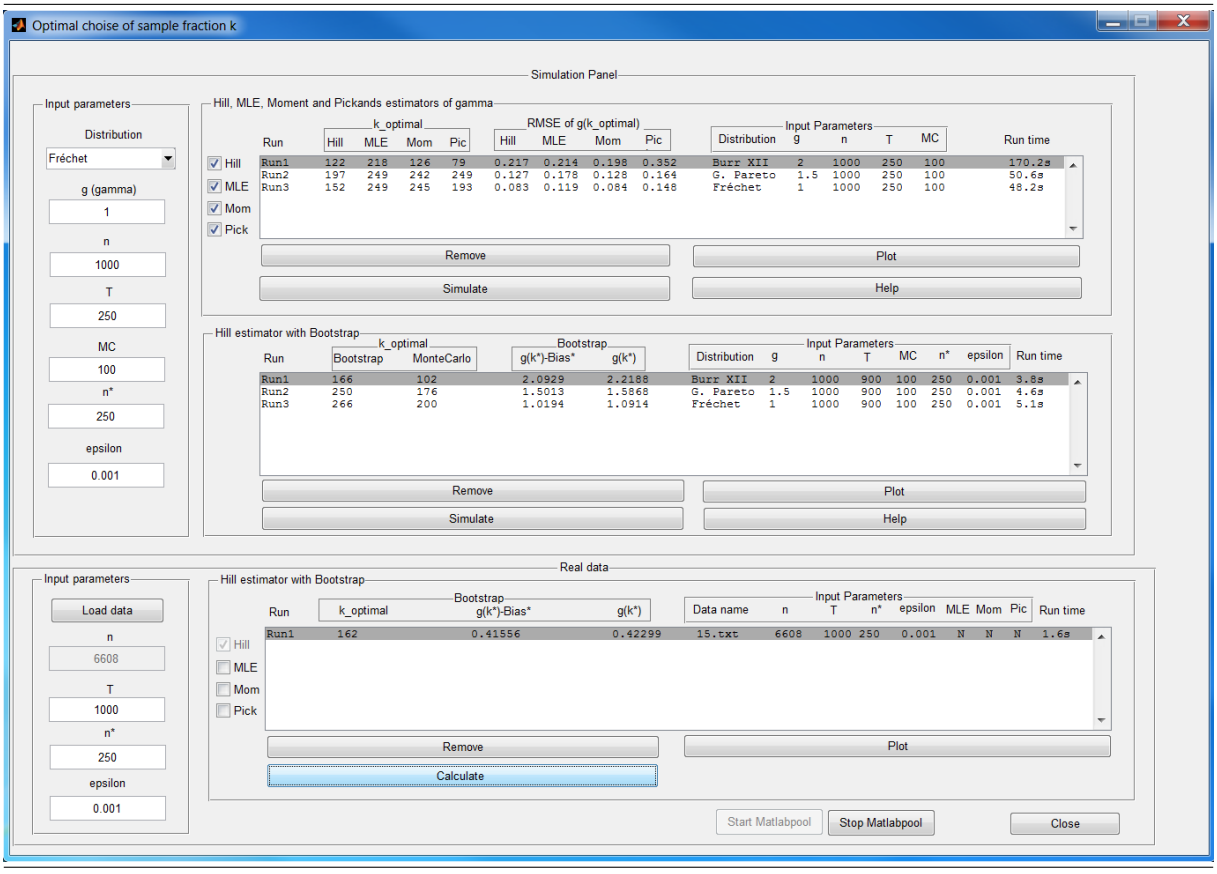
A. Simulační program

Program obsažený v příloze slouží k výpočtům odhadů indexu extrémní hodnoty a k volbě optimálního prahového indexu pro rozdělení z Fréchetova oboru atraktivity. K jeho spuštění je nutný Matlab se statistickým toolboxem nejlépe verze R2012b a novější. Program se spouští souborem My_guide2 (je potřeba se nacházet ve správném adresáři s tímto souborem).

Program se dělí na tři části. První z nich je zaměřena na srovnání odhadů na simulovaných náhodných výběrech z vybraných rozdělení. Druhou část tvoří Hillův odhad s metodou bootstrap pro simulované náhodné výběry s těžkými chvosty. Předpokládají se uživatelem korektně zadané vstupní parametry. Vstupy a výstupy pro tyto dvě části jsou detailně popsány níže.

Hillův odhad s bootstrapem lze provést na reálných datech v třetí části. Zde lze rovněž vypočítat a vykreslit maximálně věrohodný, Pickandsův a momentový odhad v závislosti na k. Momentový odhad s bootstrapem na reálných datech popsaných v sekci 6.1 lze provést samostatně mimo tento program v souboru Data_Mom_boot.

Odhady jsou naprogramovány s ohledem na výpočetní čas. Například Hillův a momentový odhad jsou počítány iterativně pro každý index prahové statistiky k. Ještě rychlejších výpočtů lze dosáhnout pomocí matlabovské funkce matlabpool, která dovede využít k výpočtu více procesorů najednou. S touto funkcí mohou nastat problémy u starších verzí Matlabu a může být potřeba přepsat cykly parfor na for.



Obrázek A.1: Simulační program.

Hillův, maximálně věrohodný, momentový a Pickandsův odhad parametru γ na simulovaných náhodných výběrech

Vstupní parametry

Vstupní údaje, které uživatel zadává jsou:

- Distribution - typ rozdělení většinou ve standardizovaném tvaru. Ze zvoleného rozdělení je pak pomocí kvantilové funkce $F^{\leftarrow}(u) = U\left(\frac{1}{1-u}\right)$ generováno MC náhodných výběrů $X_1^{(j)}, X_2^{(j)} \ldots, X_n^{(j)}$, $j = 1, 2, \ldots, MC$. Podrobnosti o rozděleních jsou uvedeny níže v tabulce. Sloupec s názvem *Vstupní parametry* obsahuje parametry, které nejsou možné zadat uživatelem. Uživatel tak pro zvolené rozdělení volí pouze index extrémní hodnoty γ a rozsah náhodného výběru n.

Tabulka A.1: Vybraná rozdělení patřící do Fréchetova oboru atraktivity ($\gamma > 0$)

Rozdělení	Distribuční funkce $F(x)$ a kvantilová funkce chvostu $U(t)$;	Vstupní parametry
Burr(η, τ, λ) (Typ XII)	$F(x) = 1 - \left(\frac{\eta}{\eta + x^{\tau}}\right)^{\lambda}, \quad \eta, \tau, \lambda > 0,\ x > 0$ $U(t) = \left(\eta(t^{\frac{1}{\lambda}} - 1)\right)^{\frac{1}{\tau}}, \quad t > 1$	$\tau = -\rho = 0.5$ $\eta = 1,\ \lambda = \frac{1}{\gamma\tau}$
F(m, n) (Fisher-Snedecor)	$F(x) = \int_0^x \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2} \mathrm{d}w, m, n > 0;\ x > 0$ Předpis pro $U(t)$ není znám. $F^{\leftarrow}(u)$ v matlabu pomocí funkce $(1, 2/\gamma)$	$m = 1,\ n = \frac{2}{\gamma}$
G. Pareto(σ, γ) (Zobecněné Paretovo)	$F(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}, \quad \sigma > 0,\ x > 0$ $U(t) = \frac{\sigma}{\gamma}(t^{\gamma} - 1), \quad t > 1$	$\sigma = 1$
Paret Pa(γ)	$F(x) = 1 - x^{-\frac{1}{\gamma}}, \quad x \geq 1$ $U(t) = t^{\gamma}, \quad t > 1$	
Frechet(γ)	$F(x) = \exp\left(-x^{-\frac{1}{\gamma}}\right), x > 0$ $U(t) = \left(\ln \frac{t}{t-1}\right)^{-\gamma}, \quad t > 1$	

- Tlačítka Hill, MLE, Mom a Pick určují, zda bude proveden Hillův, maximálně věrohodný, momentový a Pickandsův odhad.
- g - index extrémní hodnoty $\gamma > 0$.
- n - rozsah náhodného výběru.
- T - prahový index. Je možné zadat:
  - $T \in \{2, 3, \ldots, n\}$ pro Hillův a maximálně věrohodný odhad.
  - $T \in \{3, 4 \ldots, n\}$ pro momentový odhad.
  - $\frac{n}{4} \leq T \leq n$, $T \in \mathbb{N}$ pro Pickandsův odhad.
- MC - počet náhodných výběrů.

Výstup

- k_optimal - volba pořadového indexu k_0 taková, že

$$k_0(n) = \arg \min_{1 \le k \le T-1} \sum_{j=1}^{MC} (\widehat{\gamma}^{(j)}(k) - \gamma)^2,$$

kde pro $j = 1, \ldots, MC$

$$\widehat{\gamma}^{(j)}(k) = \begin{cases} \hat{\gamma}^{(j)}_{\mathrm{H}}(k), & \text{v případě Hillova odhadu,} \\ \hat{\gamma}^{(j)}_{\mathrm{MLE}}(k), & \text{v případě maximálně věrohodného odhadu,} \\ \hat{\gamma}^{(j)}_{\mathrm{M}}(k), & \text{v případě momentového odhadu,} \\ \hat{\gamma}^{(j)}_{\mathrm{P}}(k), & \text{v případě Pickandsova odhadu.} \end{cases}$$

Přitom

$$\begin{aligned} \hat{\gamma}^{(j)}_{\mathrm{H}}(k) &= \frac{1}{k} \sum_{i=1}^{k} \log X^{(j)}_{n-i+1,n} - \log X^{(j)}_{n-k,n}, \quad j = 1, \ldots, MC, \quad 1 \le k \le T-1, \\ \hat{\gamma}^{(j)}_{\mathrm{P}}(k) &= (\log 2)^{-1} \log \frac{X^{(j)}_{n-k,n} - X^{(j)}_{n-2k,n}}{X^{(j)}_{n-2k,n} - X^{(j)}_{n-4k,n}}, \quad j = 1, \ldots, MC, \quad 1 \le k \le T-1, \\ \hat{\gamma}^{(j)}_{\mathrm{M}}(k) &= \hat{\gamma}^{(j)}_{\mathrm{H}}(k) + \hat{\gamma}^{(j)}(k)_{-}, \quad j = 1, \ldots, MC, \quad 2 \le k \le T-1, \end{aligned}$$

kde pro $\hat{\gamma}^{(j)}_{\mathrm{M}}(k)$, $j = 1, \ldots, MC$, $2 \le k \le T-1$ jsou

$$\hat{\gamma}^{(j)}(k)_{-} = 1 - \frac{1}{2} \left(1 - \frac{\left(\hat{\gamma}^{(j)}_{\mathrm{H}}(k)\right)^2}{M^{(2)}_n} \right)^{-1}, \quad M^{(2)}_n = \frac{1}{k} \sum_{i=1}^{k} \left(\log X^{(j)}_{n-i+1,n} - \log X^{(j)}_{n-k,n} \right)^2.$$

Odhad $\hat{\gamma}^{(j)}(k)_{\mathrm{MLE}}$, $j = 1, \ldots, MC$, $2 \le k \le T-1$ je prováděn na základě maximalizace věrohodnostní funkce G. Pareto rozdělení pomocí matlabovské funkce gpfit($\boldsymbol{z}^{(j)}(k)$), kde

$$\boldsymbol{z}^{(j)}(k) = (X^{(j)}_{n-k+1,n} - X^{(j)}_{n-k,n}, X^{(j)}_{n-k+2,n} - X^{(j)}_{n-k,n}, \ldots, X^{(j)}_{n,n} - X^{(j)}_{n-k,n}),$$

pro $j = 1, \ldots, MC$, $1 \le k \le T-1$.

- RMSE of g(k_optimal) - odmocnina ze střední kvadratické chyby pro $k = k_0$, tedy

$$\widehat{RMSE}(k_0) = \sqrt{\widehat{MSE}(k_0)} = \sqrt{\frac{1}{MC} \sum_{j=1}^{MC} (\widehat{\gamma}^{(j)}(k_0) - \gamma)^2}.$$

- tlačítko Simulate - nasimuluje náhodné výběry a provede odhady zadané uživatelem, které jsou uvedeny výše. Výsledky dané simulace i se vstupními parametry se vypíši v příslušné liště.

- tlačítko Remove - smazání označeného řádku z lišty.

- tlačítko Plot - vykreslení $\overline{\widehat{\gamma}(k)} = \frac{1}{MC} \sum_{j=1}^{MC} \widehat{\gamma}^{(j)}(k)$, $\widehat{MSE}(k)$ a charakteristiky rozdělení extrémního Fréchetova typu pro parametry $\overline{\widehat{\gamma}(k_0)}$ a γ.

Hillův odhad parametru $\gamma > 0$, volba prahového indexu k a oprava vychýlení pomocí metody bootstrap na simulovaných náhodných výběrech (pro $\rho < 0$)

Vstupní parametry

Vstupní údaje, které uživatel zadává jsou:

- Distribution - typ rozdělení většinou ve standardizovaném tvaru. Ze zvoleného rozdělení je pak pomocí kvantilové funkce generováno MC náhodných výběrů $X_1^{(j)}, X_2^{(j)} \ldots, X_n^{(j)}$, $j = 1, 2, \ldots, MC$. Podrobnosti o rozděleních jsou uvedeny výše v tabulce A.1. Sloupec Vstupní parametry obsahuje parametry, které nejsou možné zadat uživatelem. Uživatel tak pro zvolené rozdělení volí pouze index extrémní hodnoty γ a rozsah náhodného výběru n.
- g - index extrémní hodnoty γ, $\gamma > 0$.
- n - rozsah náhodného výběru.
- T - prahový index. Je možné zadat $T \in \{2, 3, \ldots, n_2\}$, přičemž n_2 a n_1 jsou zvolena tak, že $n_2 = \text{round}\,(n_1^2/n)$, $n_1 = \text{round}(0.975 \cdot n)$.
- MC - počet náhodných výběrů.
- n* - počet bootstrapových náhodných výběrů z rovnoměrného diskrétního rozdělení na množině $\{x_i^{(j)}\}_{i=1}^n$, $j = 1, 2, \ldots, MC$.
- epsilon - prahová hodnota ϵ, která zabraňuje dělení nulou. Doporučeno neměnit přednastavenou hodnotu.

Výstup

- k_optimal, MonteCarlo - volba pořadového indexu k_0 taková, že

$$k_0(n) = \arg \min_{1 \le k \le T-1} \sum_{j=1}^{MC} (\widehat{\gamma}^{(j)}(k) - \gamma)^2,$$

  kde

$$\widehat{\gamma}^{(j)}(k) = \widehat{\gamma}_{\mathrm{H}}^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} \log X_{n-i+1,n}^{(j)} - \log X_{n-k,n}^{(j)}, \quad j = 1, 2, \ldots, MC, \quad 1 \le k \le T-1,$$

- k_optimal, Bootstrap - volba pořadového indexu round$(\overline{k_0^*})$ taková, že

$$\overline{k_0^*} = \overline{k_0^*}(n; k_{aux}, n_1) = \frac{1}{MC} \sum_{j=1}^{MC} k_0^{*(j)} = \frac{1}{MC} \sum_{j=1}^{MC} \frac{[{k_0}^{**j}(n_1)]^2}{{k_0}^{**j}(n_2)},$$

  přitom n_2 a n_1 jsou zvolena tak, že $n_2 = \text{round}\,(n_1^2/n)$ a $n_1 = \text{round}(0.975n)$. Dále volíme referenční index $k_{aux} = \text{round}(2\sqrt{n})$, pro kterou počítáme $\widehat{\gamma}^{(j)}(k) = \frac{1}{k_{aux}} \sum_{i=1}^{k_{aux}} \log X_{n-i+1,n}^{(j)} - \log X_{n-k_{aux},n}^{(j)}$. Odhad pořadové statistiky

$${k_0}^{**(j)}(n_i) = \arg \min_{1 \le k \le T-1} \sum_{b^*=1}^{n^*} (\widehat{\gamma^*}_{n_i}^{(j),b^*}(k) - \widehat{\gamma}^{(j)}(k_{aux}))^2, \quad j = 1, 2, \ldots, MC \quad i = 1, 2,$$

kde pro bootstrapový náhodný výběr $X_1^{*(j),b^*}, X_2^{*(j),b^*}, \ldots, X_{n_i}^{*(j),b^*}$, $b^* = 1, \ldots, n^*$, $j = 1, 2, \ldots, MC$, $i = 1, 2$ z rovnoměrného diskrétního rozdělení na množině $\{x_i^{(j)}\}_{i=1}^n$, $j = 1, 2, \ldots, MC$, $i = 1, 2$ je

$$\widehat{\gamma^*}_{n_i}^{(j),b^*}(k) = \widehat{\gamma^*}_{n_i,\mathrm{H}}^{(j),b^*}(k) = \frac{1}{k}\sum_{i=1}^{k} \log X_{n_i-i+1,n_i}^{*(j),b^*} - \log X_{n_i-k,n_i}^{*(j),b^*},$$

pro $\quad b^* = 1, \ldots, n^*,\ j = 1, 2, \ldots, MC,\ i = 1, 2.$

- g(k*) - průměrný Hillův odhad $\overline{\widehat{\gamma}}(k)$ parametru γ pro $k = \overline{k_0^*}$, kde $\overline{\widehat{\gamma}}(\overline{k_0^*}) = \frac{1}{MC}\sum_{j=1}^{MC} \widehat{\gamma}^{(j)}(\overline{k_0^*})$

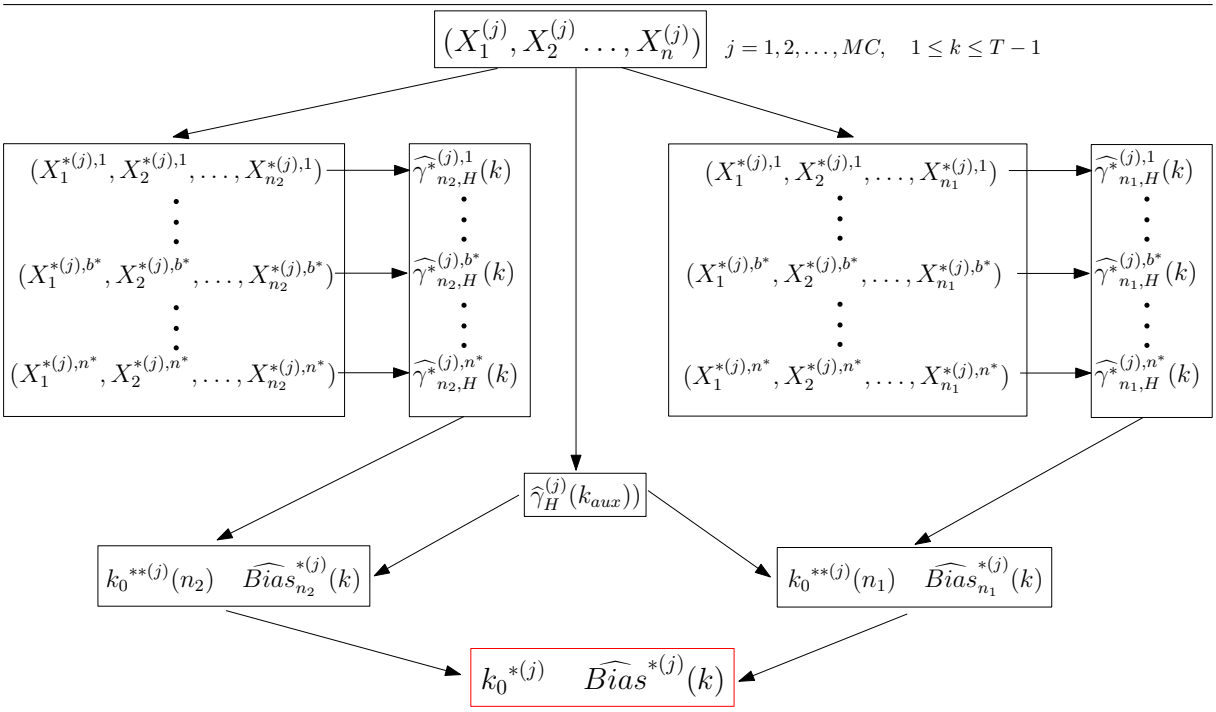
- Bias* - průměrný bootstrapový odhad vychýlení $\overline{\widehat{Bias}}^*(\overline{k_0^*}) = \frac{1}{MC}\sum_{j=1}^{MC} \widehat{Bias}^{*(j)}(\overline{k_0^*})$ takový, že pro $n_2 = \mathrm{round}\,(n_1^2/n)$ a $n_1 = \mathrm{round}(0.975n)$ je

$$\widehat{Bias}^{*(j)}(k) = \frac{[\widehat{Bias}_{n_1}^{*(j)}(k)]^2}{\widehat{Bias}_{n_2}^{*(j)}(k)}, \quad 1 \le k \le T-1, \quad j = 1, 2 \ldots, MC,$$

kde

$$\widehat{Bias}_{n_i}^{*(j)}(k) = \frac{1}{n^*}\sum_{b^*=1}^{n^*} (\widehat{\gamma^*}_{n_i}^{(j),b^*}(k) - \widehat{\gamma}^{(j)}(k_{aux})), \quad 1 \le k \le T-1,\ j = 1, 2 \ldots, MC,\ i =, 1, 2.$$

V případě, že některá z hodnot $\widehat{Bias}_{n_2}^{*(j)}(k)$, $j = 1, 2 \ldots, MC$, $1 \le k \le T-1$ je blízká nule, dochází k numerickým chybám. Tento případ je ošetřen podmínkou $\widehat{Bias}_{n_2}^{*(j)}(k) \ge \epsilon$, $j = 1, 2 \ldots, MC$, $1 \le k \le T-1$. Hodnotu ϵ volí uživatel. V případě, že podmínka není splněna pro některou dvojici (k, j), je generován nový bootstrapový náhodný výběr $X_1^{*(j),b^*}, X_2^{*(j),b^*}, \ldots, X_{n_2}^{*(j),b^*}$, $b^* = 1, \ldots, n^*$. Při špatné volbě vstupního parametru ϵ tedy hrozí zacyklení simulace. Počet vygenerovaných bootstrapových náhodných výběrů, které jsou blízké nule, lze během simulace sledovat v Command Window u proměnné num_cl_to_0.



Obrázek A.2: Algoritmus pro volbu k-té pořadové statistiky a opravu vychýlení

- tlačítko Simulate - nasimuluje náhodné výběry a provede odhady, které jsou uvedeny výše. Výsledky dané simulace i se vstupními parametry se vypíši v příslušné liště.
- tlačítko Remove - smazání označeného řádku z lišty.
- tlačítko Plot - vykreslení průměrného Hillova odhadu $\overline{\widehat{\gamma}}(k)$ (modře) a jeho korekci vychýlení metodou bootstrap $\overline{\widehat{\gamma}}(k) - \overline{\widehat{Bias}}^*(k)$ (červeně). Na dalším obrázku se vykreslí charakteristiky rozdělení Fréchetova typu pro parametry $\overline{\widehat{\gamma}}(\overline{k_0^*})$ a γ.