

**MASARYKOVA UNIVERZITA**  
**Přírodovědecká fakulta**  
Ústav matematiky a statistiky

## **Diplomová práce**

**Brno 2019**

**Natália Gažová**



MASARYKOVA UNIVERZITA  
Přírodovědecká fakulta  
Ústav matematiky a statistiky

---



# Modelování extrémních událostí

Diplomová práce

**Natália Gažová**

Vedoucí práce: RNDr. Radim Navrátil, PhD.

Brno 2019

# Bibliografický záznam

<b>Autor:</b>	Bc. Natálie Gažová Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
<b>Název práce:</b>	Modelování extrémních událostí
<b>Studijní program:</b>	Matematika
<b>Studijní obor:</b>	Finanční matematika
<b>Vedoucí práce:</b>	RNDr. Radim Navrátil, PhD.
<b>Akademický rok:</b>	2018/2019
<b>Počet stran:</b>	$ix + 60$
<b>Klíčová slova:</b>	teorie extrémních hodnot; metoda blokových maxim; zobecněné rozdělení extrémních hodnot; sféra přitažlivosti; peaks-over-threshold; zobecněné Paretovo rozdělení

# Bibliografický záznam

<b>Autor:</b>	Bc. Natália Gažová Prírodovedecká fakulta, Masarykova univerzita Ústav matematiky a štatistiky
<b>Názov práce:</b>	Modelovanie extrémnych udalostí
<b>Študijný program:</b>	Matematika
<b>Študijný odbor:</b>	Finančná matematika
<b>Vedúci práce:</b>	RNDr. Radim Navrátil, PhD.
<b>Akademický rok:</b>	2018/2019
<b>Počet strán:</b>	$ix + 60$
<b>Kľúčové slová:</b>	teória extrémnych hodnôt; metóda blokových maxím; zovšeobecnené rozdelenie extrémnych hodnôt; sféra príťažlivosti; peaks-over-threshold; zovšeobecnené Paretovo rozdelenie

# Bibliographic Entry

**Author:** Bc. Natália Gažová  
Faculty of Science, Masaryk University  
Department of Mathematics and Statistics

**Title of Thesis:** Modelling extremal events

**Degree Programme:** Mathematics

**Field of Study:** Financial Mathematics

**Supervisor:** RNDr. Radim Navrátil, PhD.

**Academic Year:** 2018/2019

**Number of Pages:**  $ix + 60$

**Keywords:** extreme value theory; block maxima method; generalised extreme value distribution; domain of attraction; peaks-over-threshold; generalised Pareto distribution

# Abstrakt

V této diplomové práci se věnujeme analýze extrémních událostí. V modelování využíváme dvě metody, konkrétně metodu blokových maxim a metodu založenou na překročení určité meze. Úvodní část obsahuje teoretické seznámení s oběma metodami, což je důležité pro dostatečné pochopení při dalších aplikacích. Praktická část se soustředí na analýzu pojistných dat oběma způsoby s následným porovnáním výsledků.

# Abstrakt

V tejto diplomovej práci sa venujeme analýze extrémnych udalostí. V modelovaní využívame dve metódy, konkrétne metódu blokových maxim a metódu založenú na prekročení určitej medze. Úvodná časť práce obsahuje teoretické zoznámenie sa s oboma metódami, čo je dôležité pre dostatočné pochopenie pri ďalších aplikáciach. Praktická časť sa sústreďuje na analýzu poisťných dát oboma spôsobmi s následným porovnaním výsledkov.

# Abstract

This thesis deals with extreme value analysis. Two approaches to modelling extreme events are concerned in this text – block maxima method and peaks-over-threshold. In the first part, a theoretical background is given which is essential for further applications. Afterwards, modelling of insurance data is studied. Both above-mentioned methods as well as the comparison of results are included in the analysis.



## ZADÁNÍ DIPLOMOVÉ PRÁCE

Akademický rok: 2016/2017

**Ústav:** Ústav matematiky a statistiky

**Studentka:** Bc. Natália Gažová

**Program:** Matematika

**Obor:** Finanční matematika

Ředitel *Ústavu matematiky a statistiky* PŘF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje diplomovou práci s názvem:

**Název práce:** Modelování extrémních událostí

**Název práce anglicky:** Modelling extremal events

### Oficiální zadání:

Řídké (extrémní) události není vhodné modelovat pomocí klasické statistiky, která dobře popisuje chování dat okolo nějaké prostřední hodnoty. K tomuto účelu se používá teorie extrémních hodnot, která dokáže popsat i chování málo pravděpodobných (četných) jevů jako například stoletá povodeň, pojistný nárok vyšší než milion korun apod. Studentka v práci popíše dva základní přístupy modelování extrémních událostí: metodu blokových maxim a metodu peeks over threshold. Použití obou metod poté ilustruje a porovná ve vhodném statistickém softwaru, např. R.

### Literatura:

COLES, Stuart. *An introduction to statistical modeling of extreme values*. New York: Springer, 2001. xiv, 208. ISBN 1852334592.

EMBRECHTS, Paul, Claudia KLÜPPELBERG a Thomas MIKOSCH. *Modelling extremal events for insurance and finance*. Heidelberg: Springer, 1997. xv, 648. ISBN 9783540609315.

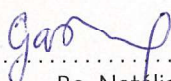
**Jazyk závěrečné práce:** slovenština

**Vedoucí práce:** RNDr. Radim Navrátil, Ph.D.


**Datum zadání práce:** 7. 3. 2017

**V Brně dne:** 24. 3. 2017

Souhlasím se zadáním (podpis, datum):

 14. 4. 2017

Bc. Natália Gažová  
studentka

 27. 3. 2017

RNDr. Radim Navrátil, Ph.D.  
vedoucí práce



prof. RNDr. Jan Slovák, DrSc.  
ředitel Ústavu matematiky a  
statistiky

# PodĎakovanie

Na tomto mieste by som chcela poĎakovať vedúcemu tejto diplomovej práce, RNDr. Radimovi Navrátilovi, PhD., za ochotu a svoje cenné pripomienky, ktoré mi počas vedenia práce poskytol a za čas, ktorý mi venoval.

# Prohlášení

Prohlašuji, že jsem svoji diplomovou práci vypracovala samostatně s využitím informačních zdrojů, které jsou v práci citovány.

Brno 20. prosince 2018

.....  
Natália Gažová



# Obsah

<b>Úvod</b> .....	<b>ix</b>
<b>Kapitola 1. Teória extrémnych hodnôt</b> .....	<b>1</b>
<b>Kapitola 2. Metóda blokových maxím</b> .....	<b>3</b>
2.1 Sféra príťažlivosti rozdelení extrémnych hodnôt .....	9
2.2 Max-stabilita rozdelení extrémnych hodnôt .....	16
2.3 Modelovanie metódou blokových maxím .....	18
2.3.1 Odhad parametrov GEV rozdelenia .....	18
2.3.2 Spracovanie dát .....	22
<b>Kapitola 3. Peaks-over-threshold</b> .....	<b>26</b>
3.1 Stabilita rozdelení pre modelovanie excesov .....	29
3.2 Metóda POT pri praktickom spracovaní dát .....	32
3.2.1 Odhad parametrov GP rozdelenia .....	32
3.2.2 Spracovanie dát .....	34
<b>Kapitola 4. Analýza dát</b> .....	<b>38</b>
4.1 Dáta .....	38
4.2 Modelovanie metódou blokových maxím .....	39
4.3 Modelovanie POT metódou .....	45
4.4 Porovnanie metódy blokových maxím a POT metódy .....	50
4.5 Parametrický model .....	53
<b>Závěr</b> .....	<b>57</b>
<b>Seznam použité literatury</b> .....	<b>59</b>

# Úvod

Extrémne udalosti, ako už ich názov napovedá, nie sú spojené s bežnými situáciami, no ich výskyt znamená veľké pohyby, vysoké straty či vysoký dopad v určitej sfére záujmu. Preto je ich štúdium zaujímavé z rôznych hľadísk a pre rôzne skupiny populácie.

Typickým príkladom extrémnej udalosti je napr. storočná povodeň, t.j. taká povodeň, ktorá sa vyskytne v dlhodobom priemere jedenkrát za sto rokov. Ďalšou oblasťou bohatou na tieto javy je meteorológia, ktorá zahŕňa pôsobenie hurikánov, rekordných zmien teploty, či vlny horúčav. Práve v oblasti meteorologických a klimatických javov sa analýza extrémnych udalostí začala rozvíjať. Extrémne javy však figurujú aj vo finančníctve, a to napríklad v podobe výraznej zmeny hodnoty aktíva za určitý čas. Veľkou sférou, kde sa metódy taktiež využívajú je poisťovníctvo, a to najmä jeho neživotná časť. V niektorých portfóliách sa môže príležitostne vyskytnúť vysoký poistný nárok, ktorý môže ohroziť jeho stabilitu. Okrem takých udalostí ako hurikány, zemetrasenia či letecké nehody sa vysoké poistné nároky vyskytujú aj v iných oblastiach. Patria medzi ne poistné nároky z automobilového poistenia či z poistenia proti požiarom (priemyselné požiare a podobne). Je preto prínosné sa proti všetkým takýmto okolnostiam brániť vopred, s čím pomáha práve analýza extrémnych udalostí.

Prvá kapitola práce poskytuje veľmi stručné zoznámenie s metódami, ktoré sa na modelovanie extrémnych situácií využívajú. Prvá z nich – metóda blokových maxim je podrobnejšie študovaná v kapitole 2. Okrem iného popisuje aj sféry príťažlivosti a overuje max-stabilitu jednotlivých rozdelení extrémnych hodnôt. Ďalej je v nej pozornosť venovaná odhadu parametrov rozdelenia extrémnych hodnôt a riešeniu praktických úloh spojených s analýzou dát, ako napr. hľadanie doby návratu extrémneho javu. Nasledujúca kapitola 3 potom popisuje podobné charakteristiky pre metódu peaks-over-threshold. V poslednej kapitole 4 je priestor venovaný analýze dát z oblasti poisťovníctva. Aplikované sú obe zmienené metódy, ktoré sú na záver porovnané.

# Kapitola 1

## Teória extrémnych hodnôt

Klasická štatistika sa, zjednodušene povedané, zameriava na „priemerné“ správanie nejakého náhodného procesu, teda na javy realizujúce sa v okolí strednej hodnoty nejakého rozdelenia.

Na druhú stranu, teória extrémnych hodnôt (ďalej len EVT, z angl. *Extreme Value Theory*) sa sústreďuje na extrémne, ojedinelé a vzácne javy, ktoré sa spravidla realizujú na chvostoch rozdelení. Z pohľadu štatistiky ide o tzv. odľahlé pozorovania. Ukazuje sa totiž, že pravdepodobnosť nastania extrémnej udalosti je v skutočnosti vyššia, ako sa predpokladá na základe teoretického rozdelenia pravdepodobnosti, t.j. že rozdelenia majú ťažšie chvosty.

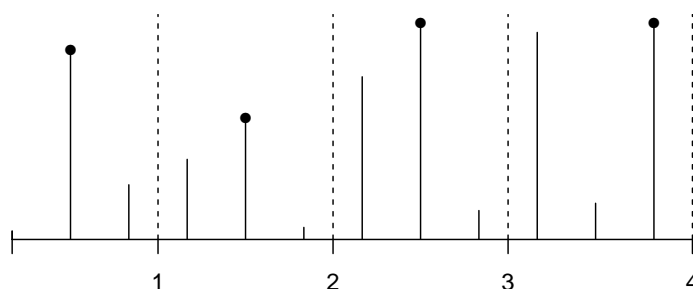
Cieľom EVT je skúmanie správania týchto odľahlých pozorovaní. Zaujímať nás napríklad môže stanovenie veľkosti extrémneho javu, ktorý nastane raz za zvolenú frekvenciu, prípadne naopak, stanovenie priemernej frekvencie výskytu extrémneho javu danej veľkosti a ďalšie. To znamená, že musíme nájsť vhodnú matematickú metódu, aby sme mohli vysvetliť okolnosti, ktoré sa vyskytujú s relatívne malou pravdepodobnosťou, no majú vysoký dopad.

Najčastejšie sa na modelovanie týchto situácií využívajú dve metódy, a to metóda blokových maxím a metóda peaks-over-threshold (POT).

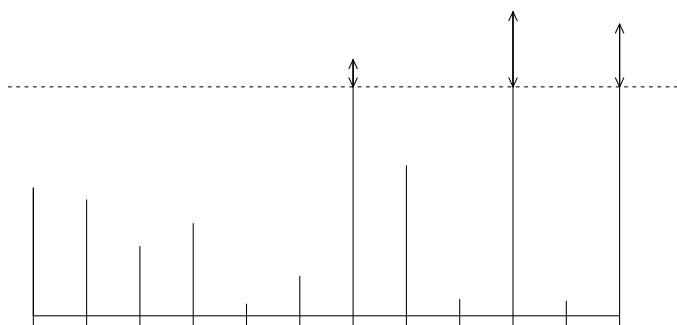
### Metóda blokových maxím

Podstatou metódy blokových maxím je rozdelenie získaných pozorovaní do určitého počtu blokov, pričom v každom bloku zistíme maximálne pozorovanie. Ideou potom je nemodelovať samotné pozorovania, ale získané blokové maximá. Obrázok 1.1 ilustruje koncept metódy blokových maxím. Pozorovania sú v tomto prípade rozdelené do štyroch blokov po troch pozorovaniach.

Prístup k modelovaniu dát pomocou blokových maxím má však jednu zjavnú nevýhodu - v rámci bloku uvažujeme len jednu extrémnu hodnotu. Vo všeobecnosti môže nastať situácia, kedy druhá najvyššia hodnota daného bloku je vyššia ako maximum nejakého iného, čím v rámci analýzy strácame cenné extrémne dáta.



Obr. 1.1: Blokové maximá.



Obr. 1.2: Peaks-over-threshold

## Peaks-over-threshold

Spomenutý nedostatok metódy blokových maxím odstraňuje metóda peaks-over-threshold (POT). Je považovaná za modernejší spôsob analýzy extrémnych udalostí a je založená na prekročení určitej pevne zvolenej medze. Opäť sa nebudeme snažiť modelovať dané pozorovania. Namiesto toho sa budeme zaoberať prekročeniami pozorovaní (excesmi) vopred danej hranice, v prípade, že k tomuto prekročeniu došlo. Tým sú dáta využité efektívnejšie ako v predchádzajúcom prípade. Na obrázku 1.2 je táto hranica vyobrazená prerušovanou čiarou.

V nasledujúcich kapitolách si obe metódy rozoberieme podrobnejšie. Zistené poznatky ďalej využijeme pri spracovaní dát.

# Kapitola 2

## Metóda blokových maxím

Majme postupnosť náhodných veličín  $X_1, X_2, \dots, X_n$ , pričom jednotlivé náhodné veličiny sú nezávislé a rovnako rozdelené s distribučnou funkciou  $F$  (ďalej ich budeme skrátene označovať ako i.i.d. náhodné veličiny, z angl. *independent identically distributed*). Označme maximum tohto náhodného výberu

$$M_n = \max(X_1, \dots, X_n),$$

pričom toto maximum chceme ďalej skúmať. Obdobne by sme mohli vyšetrovať vlastnosti minima, pričom je v takom prípade potrebné previesť transformáciu

$$\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n).$$

Zaujíma nás, ako toto maximum modelovať, aké rozdelenie je vhodné využiť. Preto s využitím i.i.d. vlastnosti náhodných veličín  $X_1, \dots, X_n$  zostrojme distribučnú funkciu pre  $M_n$ :

$$G_n(x) = P(M_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = \prod_{i=1}^n P(X_i \leq x) = F^n(x). \quad (2.1)$$

Intuitívne očakávame, že maximálne extrémne situácie sa odohrávajú na pravom konci distribučnej funkcie  $F$ , čo nám potvrdzuje aj nasledujúca veta.

**Veta 2.0.1.** *Nech  $M_n$  je maximum i.i.d. náhodných veličín  $X_1, \dots, X_n$  a  $x_F$  je pravý koncový bod rozdelenia s distribučnou funkciou  $F$  definovaný ako*

$$x_F := \sup\{x \in \mathbb{R}, F(x) < 1\}.$$

*Potom pre  $n \rightarrow \infty$  platí*

$$M_n \xrightarrow{s.i.} x_F.$$

*Dôkaz.* Musíme rozlíšiť dve situácie. Pre  $x < x_F$ ,  $x \in \mathbb{R}$  máme

$$\lim_{n \rightarrow \infty} P(M_n \leq x) = \lim_{n \rightarrow \infty} F^n(x) = 0,$$

a zároveň pre  $x \geq x_F$  a  $x_F < \infty$  platí

$$P(M_n \leq x) = F^n(x) = 1.$$

Z toho okamžite plynie konvergencia maxima  $M_n$  k pravému koncovému bodu  $x_F$  v pravdepodobnosti. Nakoľko však postupnosť maxim  $M_n$  pre  $n \rightarrow \infty$  je neklesajúca, môžeme povedať, že  $M_n$  konverguje k  $x_F$  skoro iste.  $\square$

Z predchádzajúcej vety môžeme vidieť, že limitné rozdelenie maxima v sebe nenesie žiadnu informáciu, nakoľko má degenerovanú distribučnú funkciu. Ponúka sa možnosť normovať maximá vhodnými postupnosťami konštánt. Princíp normovania je podobný tomu, ktorý využíva centrálna limitná veta.

Hľadáme preto nedegenerovanú distribučnú funkciu, označme ju  $H(x)$ , ktorá bude limitnou distribučnou funkciou normovanej náhodnej veličiny  $M_n$ , čo znamená, že nás zaujíma taká  $H(x)$ , pre ktorú pre každé  $x \in \mathbb{R}$  platí

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{n \rightarrow \infty} H(x),$$

kde  $a_n > 0$ ,  $b_n \in \mathbb{R}$  sú vhodne zvolené konštanty. Potom distribučná funkcia normovaných maxim je tvaru

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = P(M_n \leq a_n x + b_n) = G_n(a_n x + b_n) = [F(a_n x + b_n)]^n, \quad (2.2)$$

kde  $G_n(a_n x + b_n)$  je distribučná funkcia nenormovaných maxim v danom bode. Ak konštanty  $a_n > 0$ ,  $b_n \in \mathbb{R}$  existujú, potom je  $H(x)$  distribučná funkcia tzv. *rozdelenia extrémnych hodnôt*, o čom hovorí nasledujúca veta.

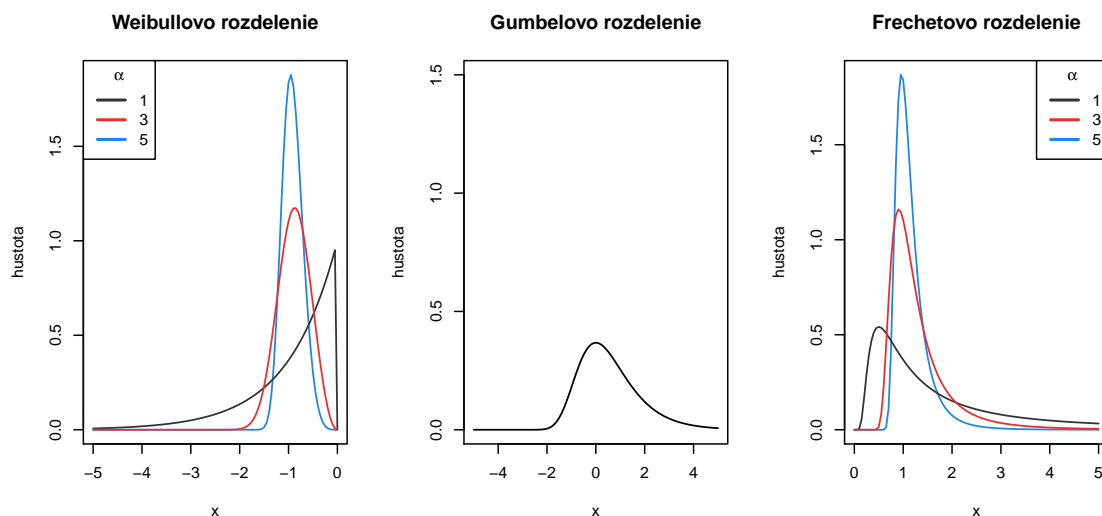
**Veta 2.0.2** (Fisherova-Tippettova). *Nech  $X_1, \dots, X_n$  sú i.i.d. náhodné veličiny,  $M_n = \max\{X_1, \dots, X_n\}$ . Ak existujú konštanty  $a_n > 0$ ,  $b_n \in \mathbb{R}$  a nedegenerovaná distribučná funkcia  $H(x)$  taká, že*

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{n \rightarrow \infty} H(x),$$

potom má  $H(x)$  jeden z nasledujúcich tvarov:

$$\begin{aligned} G_0(x) &= \exp\{-e^{-x}\}, & x \in \mathbb{R}, \\ G_1(x) &= \begin{cases} 0, & x \leq 0 \\ \exp\{-x^{-\alpha}\}, & x > 0, \end{cases} \\ G_2(x) &= \begin{cases} \exp\{-(-x)^\alpha\}, & x \leq 0, \\ 1, & x > 0, \end{cases} \end{aligned}$$

pričom  $\alpha > 0$  sa nazýva parameter tvaru.



Obr. 2.1: Hustoty rozdelení extrémnych hodnôt v štandardizovanom tvare.

*Dôkaz.* Dôkaz vety je pomerne náročný, preto ho na tomto mieste nebudeme uvádzať. Náčrt dôkazu je možné nájsť napr. v Embrechts a kol. [7] v kapitole 3. Celý dôkaz v úplnej všeobecnosti prvýkrát uviedol Gnedenko v roku 1943, pozri [10]. □

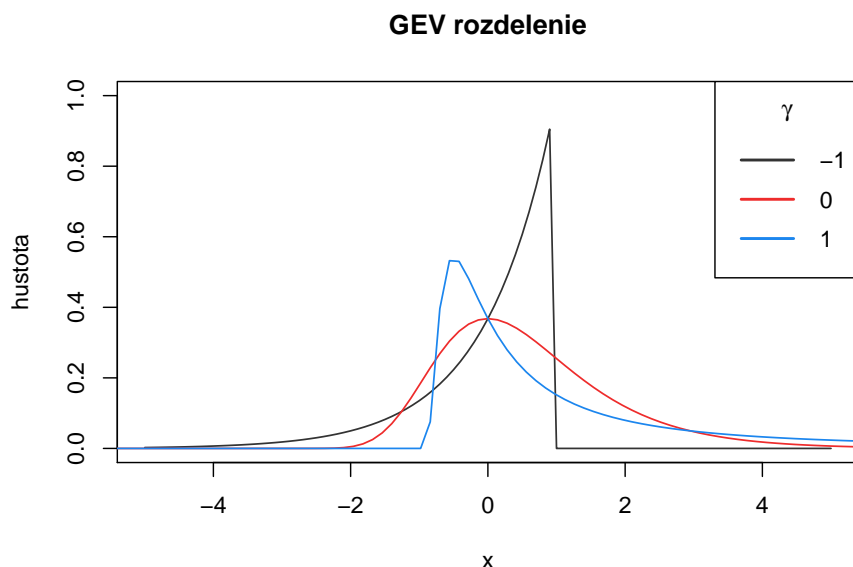
Fisherova–Tippetova veta teda inými slovami vraví, že limitné rozdelenie normovaných maxim nejakého náhodného výberu musí byť jedného z troch zmiených typov, nezávisle na pôvodnom rozdelení náhodného výberu. Rozdelenie s distribučnou funkciou  $G_0(x)$  sa nazýva Gumbelovo rozdelenie, rozdelenie s distribučnou funkciou  $G_1(x)$  je Fréchetovo rozdelenie, pričom tieto sa využívajú pri modelovaní maxim. Rozdelenie s distribučnou funkciou  $G_2(x)$  sa nazýva a Weibullovo<sup>1</sup> rozdelenie a slúži na modelovanie minim. Tieto tri rozdelenia nazývame *rozdelenia extrémnych hodnôt* alebo aj *EV rozdelenia* (*Extreme Value*). Na obrázku 2.1 sú zobrazené hustoty rozdelení extrémnych hodnôt pre rôzne hodnoty parametru  $\alpha$ .

Predchádzajúce definície distribučných funkcií rozdelení extrémnych hodnôt sú v štandardizovanom tvare, čo znamená, že parameter polohy  $\mu = 0$  a parameter merítka  $\sigma = 1$ . Úplné triedy týchto rozdelení získame zavedením týchto parametrov polohy a merítka, t.j.

$$G_{i,\mu,\sigma} = G_i\left(\frac{x - \mu}{\sigma}\right), \quad i = 0, 1, 2, \quad \mu \in \mathbb{R}, \sigma > 0.$$

Jenkinson [14] ukázal, že Gumbelovo, Fréchetovo a Weibullovo rozdelenie môžeme zapísať aj úspornejšie, pomocou jednej distribučnej funkcie *zovšeobecneneého rozdelenia extrémnych hodnôt*.

<sup>1</sup>V niektorej literatúre sa nami uvažované Weibullovo rozdelenie používané na modelovanie extrémnych situácií označuje aj ako inverzné alebo extrémálne Weibullovo rozdelenie.



Obr. 2.2: Hustota zovšeobecneného rozdelenia extrémnych hodnôt.

**Definícia 2.0.1.** Zovšeobecnené rozdelenie extrémnych hodnôt (*GEV - generalised extreme value*) s parametrom  $\gamma \in \mathbb{R}$  má distribučnú funkciu v tvare

$$G_\gamma(x) = \exp \left\{ -(1 + \gamma x)^{-\frac{1}{\gamma}} \right\}, \quad 1 + \gamma x > 0, \gamma \neq 0, \quad (2.3)$$

pričom  $\gamma$  sa nazýva parameter tvaru.

Vzťah (2.3) uvádza tvar distribučnej funkcie GEV rozdelenia v štandardizovanom tvare. Opäť môžeme pridať parametre polohy  $\mu \in \mathbb{R}$  a merítka  $\sigma > 0$ , pričom platí

$$G_{\gamma, \mu, \sigma}(x) = G_\gamma \left( \frac{x - \mu}{\sigma} \right), \quad 1 + \frac{\gamma(x - \mu)}{\sigma} > 0.$$

Hustota pravdepodobnosti GEV rozdelenia pre rôzne hodnoty parametra  $\gamma$  je zobrazená na obr. 2.2.

Lahko sa môžeme presvedčiť, že definícia GEV rozdelenia je korektná. Po spojitom dedefinovaní hodnoty parametru tvaru  $\gamma \rightarrow 0$ , a teda následným limitným prechodom získame Gumbelovo rozdelenie, konkrétne

$$\begin{aligned} \lim_{\gamma \rightarrow 0} G_\gamma(x) &= \lim_{\gamma \rightarrow 0} \exp \left\{ -(1 + \gamma x)^{-\frac{1}{\gamma}} \right\} = \\ &= \exp \left\{ - \lim_{\gamma \rightarrow 0} (1 + \gamma x)^{-\frac{1}{\gamma}} \right\} = \exp \{ -e^{-x} \}, \end{aligned}$$

čo je naozaj distribučná funkcia Gumbelovho rozdelenia.

Ďalej pre  $\gamma > 0$  a  $\gamma = 1/\alpha$  platí

$$G_\gamma(x) = G_{1/\alpha}(x) = \exp \left\{ - \left( 1 + \frac{x}{\alpha} \right)^{-\alpha} \right\},$$



čo je distribučná funkcia Fréchetovho rozdelenia v bode  $1 + x/\alpha$ , kde  $\alpha > 0$ .

Nakoniec, pre  $\gamma < 0$  a  $\gamma = -1/\alpha$  dostávame

$$G_\gamma(x) = G_{-1/\alpha}(x) = \exp \left\{ - \left( 1 - \frac{x}{\alpha} \right)^\alpha \right\},$$

čím potvrdzujeme, že sa jedná o distribučnú funkciu Weibullovoho rozdelenia v bode  $-1 + x/\alpha$ . GEV rozdelenie teda zahŕňa všetky prípady rozdelení extrémnych hodnôt.

**Príklad 2.0.1.** Majme náhodný výber  $X_1, \dots, X_n$  z exponenciálneho rozdelenia s parametrom  $\lambda = 1$ . K akému rozdeleniu konverguje maximum tohoto náhodného výberu?

Exponenciálne rozdelenie s parametrom  $\lambda = 1$  má distribučnú funkciu v tvare

$$F(x) = 1 - e^{-x}, \quad x \geq 0.$$

Konštanty zvolíme v tvare

$$a_n = 1, \quad b_n = \ln n.$$

Prečo sme zvolili práve takýto tvar normujúcich konštánt si vysvetlíme v nasledujúcej podkapitole 2.1. Dostávame

$$P \left( \frac{M_n - b_n}{a_n} \leq x \right) = P(M_n - \ln n \leq x) = P(M_n \leq x + \ln n).$$

To je distribučná funkcia normovaných maxím v bode  $x + \ln n$ . Aplikovaním vzťahu (2.2) získavame

$$P(M_n \leq x + \ln n) = [F(x + \ln n)]^n = [1 - e^{-(x + \ln n)}]^n = \left[ 1 - \frac{1}{n} e^{-x} \right]^n.$$

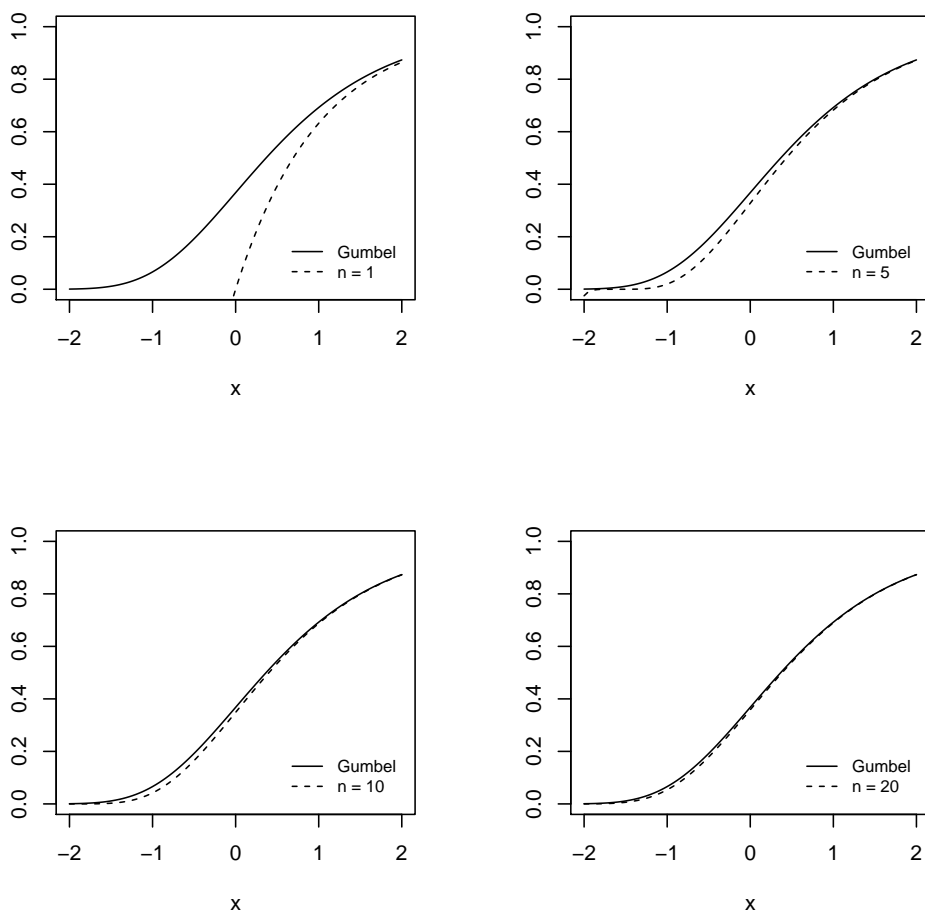
Ďalej platí, že

$$\left[ 1 - \frac{1}{n} e^{-x} \right]^n \rightarrow \exp \{ -e^{-x} \}$$

pre  $n \rightarrow \infty$ , čo je distribučná funkcia Gumbelovho rozdelenia (v štandardizovanom tvare), a teda limitným rozdelením maxím exponenciálneho rozdelenia pre vhodné zvolené normovacie konštanty je práve Gumbelovo rozdelenie. Na obrázku 2.3 je táto konvergencia znázornená pre rôzne rozsahy náhodného výberu pri zvolených konštantách  $a_n = 1$  a  $b_n = \ln n$ . Vidíme, že konvergencia rastie pomerne rýchlo, pričom už rozdelenie maxima výberu s rozsahom  $n = 20$  je veľmi blízke Gumbelovmu rozdeleniu.

Na príklade sme si ukázali platnosť Fisherovej-Tippetovej vety pre konkrétne rozdelenie, a teda, že limitné rozdelenie maxima pre vhodné konštanty  $a_n > 0$ ,  $b_n \in \mathbb{R}$  konverguje k nejakému rozdeleniu extrémnych hodnôt (v prípade, že takéto konštanty existujú).

Poznamenajme, že v prípade zvolenia nejakej inej konštantnej hodnoty pre  $a_n$  v príklade 2.0.1, by sa konvergencia ku Gumbelovmu rozdeleniu nezmenila, zmenila



Obr. 2.3: Distribučná funkcia Gumbelovho rozdelenia a normalizovaných maxím náhodného výberu z exponenciálneho rozdelenia pre rôzny rozsah výberu  $n$ .

by sa však hodnota parametrov rozdelenia. Z toho vyplýva nejednoznačnosť určenia konštánt  $a_n > 0$ ,  $b_n \in \mathbb{R}$ , pričom so zmenou výberu konštánt sa môže zmeniť aj rýchlosť konvergencie limitného rozdelenia maxim k nejakému z EV rozdelení, pozri [4]. Môžeme preto nájsť viaceré hodnoty, pre ktoré budú normované maximá konvergovať. Výsledné limitné rozdelenia sa však budú líšiť len v parametroch polohy a merítka.

## 2.1 Sféra príťažlivosti rozdelení extrémnych hodnôt

Ak vhodne štandardizované maximum i.i.d. náhodných veličín konverguje v distribúcii, potom je otázka, ako spoznať, ku ktorému z troch rozdelení extrémnych hodnôt sa tak deje. Definujme najskôr pojem sféry príťažlivosti.

**Definícia 2.1.1.** Nech  $X_1, \dots, X_n$  je náhodný výber z rozdelenia s distribučnou funkciou  $F(x)$ . Pokiaľ distribučná funkcia  $F(x)$  spĺňa podmienky Fisherovej-Tippetovej vety, t.j. ak pre vhodné konštanty  $a_n > 0$  a  $b_n \in \mathbb{R}$  a pre všetky  $x \in \mathbb{R}$  platí

$$[F(a_n x + b_n)]^n \rightarrow G_\gamma(x), \quad n \rightarrow \infty$$

potom hovoríme, že  $F(x)$  patrí do *sféry príťažlivosti* rozdelenia extrémnych hodnôt s distribučnou funkciou  $G_\gamma(x)$ , píšeme

$$F \in \text{MDA}(G_\gamma).$$

Nakoľko sa v prípade teórie extrémnych hodnôt zaujímame o rozdelenie maxima, je prirodzené sa zaoberať pravým chvostom rozdelenia náhodného výberu. Preto by sféra príťažlivosti mala závisieť len na tvare a vlastnostiach chvosta rozdelenia a nie na jeho zvyšku. Toto potvrdzuje aj nasledujúca veta. Pred jej uvedením zavedme označenie

$$\bar{F}(x) = 1 - F(x) = P(X > x), \quad x \in \mathbb{R},$$

pričom funkcia  $\bar{F}(x)$  sa bežne nazýva funkcia prežívania alebo distribučná funkcia chvostu rozdelenia.

**Veta 2.1.1.** Rozdelenie pravdepodobnosti s distribučnou funkciou  $F(x)$  patrí do sféry príťažlivosti EV rozdelenia s distribučnou funkciou  $G_i(x)$ ,  $i = 0, 1, 2$ , a normovacími konštantami  $a_n > 0$  a  $b_n \in \mathbb{R}$  práve vtedy, ak platí

$$\lim_{n \rightarrow \infty} n \bar{F}(a_n x + b_n) = -\ln G_i(x), \quad x \in \mathbb{R}.$$

*Dôkaz.* Podľa definície 2.1.1 a Fisherovej-Tippetovej vety 2.0.2 platí

$$F \in \text{MDA}(G_i) \Leftrightarrow \lim_{n \rightarrow \infty} [F(a_n x + b_n)]^n = G_i(x) \neq 0.$$

Rovnicu môžeme pomocou funkcie prežívania upraviť do tvaru

$$\lim_{n \rightarrow \infty} [1 - \bar{F}(a_n x + b_n)]^n = G_i(x).$$

Následne zlogaritmovaním rovnice a využitím rovnosti  $-\ln(1-x) \approx x$  pre  $x \rightarrow 0$  (túto aproximáciu môžeme využiť, nakoľko  $\lim_{n \rightarrow \infty} \bar{F}(a_n x + b_n) = 0$ ) získavame

$$\lim_{n \rightarrow \infty} -n \ln [1 - \bar{F}(a_n x + b_n)] \approx \lim_{n \rightarrow \infty} -n \bar{F}(a_n x + b_n) = \ln G_i(x),$$

z čoho vyplýva vzťah z tvrdenia vo vete. □

Vzhľadom na to, že predpokladáme, že sféra príťažlivosti úzko súvisí s chvostami rozdelení, definujme, že rozdelenie s distribučnou funkciou  $F_X$  a rozdelenie s distribučnou funkciou  $F_Y$  sú *chvostovo ekvivalentné* ak pre príslušné funkcie prežívania platí

$$\lim_{x \rightarrow \infty} \frac{\bar{F}_X(x)}{\bar{F}_Y(x)} = c,$$

kde  $0 < c < \infty$  je nejaká konštanta. Potom platí, že pokiaľ sú dve rozdelenia chvostovo ekvivalentné, tak patria do rovnakej sféry príťažlivosti, nakoľko konstantu  $c$  je možné zahrnúť do normujúcich konštánt  $a_n$  a  $b_n$  (pozri [15]).

Ďalej si popíšeme sféry príťažlivosti pre jednotlivé EV rozdelenia.

## Fréchetovo rozdelenie

Sféra príťažlivosti Fréchetovho rozdelenia má relatívne elegantné vyjadrenie pomocou pomaly sa meniacej funkcie v nekonečne. Preto tento pojem najskôr definujeme.

**Definícia 2.1.2.** Kladná Lebesgueovsky merateľná funkcia  $L$  na intervale  $(0, \infty)$  je

- *pomaly sa meniaci* v nekonečne, ak platí

$$\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = 1, \quad t > 0,$$

- *pravidelne sa meniaci* v nekonečne s indexom  $\rho \in \mathbb{R}$ , ak platí

$$\lim_{x \rightarrow \infty} \frac{L(tx)}{L(x)} = t^\rho, \quad t > 0.$$

Typickým príkladom pomaly sa meniacej funkcie (vynechaním triviálneho prípadu konštantnej funkcie) je napríklad logaritmus  $\ln(x)$ . Môžeme si všimnúť, že regulárne meniacu sa funkciu možno vyjadriť ako súčin mocninnej a pomaly sa meniacej funkcie.

Podľa vyššie uvedeného predpokladu, sféra príťažlivosti súvisí s chvostami rozdelení. Chvost distribučnej funkcie Fréchetovho rozdelenia má pre  $\alpha > 0$  s využitím Taylorovho rozvoja exponenciály približne tvar

$$\overline{G}_1(x) = 1 - \exp(-x^{-\alpha}) \approx x^{-\alpha}, \quad x > 0.$$

Preto môžeme približne očakávať, že rozdelenia patriace do  $MDA(G_1)$  budú mať tzv. ťažké chvosty, ktoré klesajú mocninne. Ako sa od tejto predstavy môžeme „oddialiť“ formuluje nasledovné tvrdenie.

**Veta 2.1.2.** *Rozdelenie s distribučnou funkciou  $F(x)$  patrí do sféry príťažlivosti Fréchetovho rozdelenia s distribučnou funkciou  $G_1(x)$ , práve keď platí*

$$\overline{F}(x) = x^{-\alpha}L(x), \quad \alpha > 0 \quad (2.4)$$

pre nejakú pomaly sa meniacu funkciu  $L$ .

Naviac, ak  $F \in MDA(G_1)$ , potom

$$F^n(a_n x) \rightarrow G_1(x), \quad n \rightarrow \infty$$

a normujúce konštanty môžeme voliť v tvare

$$a_n = F^{-1}(1 - n^{-1}), \quad b_n = 0. \quad (2.5)$$

*Dôkaz.* Pozri [7], kap. 3. □

Veta nám hovorí, že sféru príťažlivosti Fréchetovho rozdelenia tvoria distribučné funkcie, ktorých chvosty sú regulárne sa meniace v nekonečne s indexom  $-\alpha$ . Nakoľko sme spomenuli, že do  $MDA$  Fréchetovho rozdelenia patria rozdelenia s ťažkými chvostami, z hľadiska finančných aplikácií sa táto trieda využíva napr. pri modelovaní trhov s vysokou volatilitou.

**Príklad 2.1.1.** Ukážte, že Paretovo rozdelenie s distribučnou funkciou pre  $\alpha > 0$

$$F(x) = 1 - x^{-\alpha}, \quad x \geq 1$$

patrí do sféry príťažlivosti Fréchetovho rozdelenia.

Túto skutočnosť môžeme ukázať dvomi spôsobmi. Distribučná funkcia chvosta Paretovoho rozdelenia je tvaru

$$\overline{F}(x) = x^{-\alpha},$$

čo môžeme triviálne vyjadriť ako súčin konštantnej funkcie  $L(x) \equiv 1$  a mocninnej funkcie  $x^{-\alpha}$ , preto platí (2.4) a  $F \in MDA(G_1)$ .

Druhým spôsobom je použiť priamy výpočet konvergenzie pri zvolených konštantách podľa (2.5). Vyjadrime preto tvar konštanty  $a_n$  pomocou kvantilovej funkcie  $F^{-1}(1 - n^{-1})$  nasledovne:

$$\begin{aligned} F(x) &= 1 - n^{-1} \\ 1 - x^{-\alpha} &= 1 - n^{-1} \\ x &= n^{1/\alpha} \end{aligned}$$

Zvoľme preto  $a_n = n^{1/\alpha}$ ,  $b_n = 0$  a overme, či platí konvergencia podľa vety 2.1.1. Konkrétne

$$\begin{aligned} [F(a_n x + b_n)]^n &= [F(n^{1/\alpha} x)]^n = [1 - (n^{1/\alpha} x)^{-\alpha}]^n \\ &= \left[1 - \frac{x^{-\alpha}}{n}\right]^n \xrightarrow{n \rightarrow \infty} \exp(-x^{-\alpha}), \end{aligned}$$

čo je distribučná funkcia Fréchetovho rozdelenia, čím sme potvrdili, že  $F \in \text{MDA}(G_1)$ .

Do sféry príťažlivosti Fréchetovho rozdelenia patria mnohé ďalšie rozdelenia, pozri [2], [7] alebo [17]. Sú to napr. zovšeobecnené Paretovo,  $F$  rozdelenie, ďalej Studentovo  $t$  rozdelenie, inverzné gama alebo Cauchyho či logaritmické gama rozdelenie.

## Weibullovo rozdelenie

Fréchetovo a Weibullovo rozdelenie majú medzi sebou úzky vzťah, nakoľko pre  $\alpha > 0$  platí

$$G_2(-x^{-1}) = G_1(x), \quad x > 0. \quad (2.6)$$

Skutočne,

$$G_2(-x^{-1}) = \exp\{-(-(-x^{-1}))^\alpha\} = \exp\{-x^{-\alpha}\} = G_1(x).$$

Toto vzájomné vyjadrenie zohráva veľkú úlohu pri odvodení a popísaní sféry príťažlivosti Weibullovoho rozdelenia.

**Veta 2.1.3.** *Rozdelenie s distribučnou funkciou  $F(x)$  patrí do sféry príťažlivosti Weibullovoho rozdelenia s distribučnou funkciou  $G_2(x)$ , práve keď je pravý koncový bod rozdelenia  $x_F < \infty$  a zároveň*

$$\bar{F}(x_F - x^{-1}) = x^{-\alpha} L(x), \quad \alpha > 0 \quad (2.7)$$

pre nejakú pomaly sa meniacu funkciu  $L$ .

Naviac, ak  $F \in \text{MDA}(G_2)$ , potom

$$F^n(a_n x + x_F) \rightarrow G_2(x), \quad n \rightarrow \infty$$

a normujúce konštanty môžeme voliť v tvare

$$a_n = x_F - F^{-1}(1 - n^{-1}), \quad b_n = x_F. \quad (2.8)$$

*Dôkaz.* Dôkaz nutnej podmienky je náročný, je možné ho nájsť v [19].

Postačujúcu podmienku dokážeme nasledovne. Predpokladajme, že  $x_F < \infty$  a platí (2.7). Označme

$$F_*(x) = F(x_F - x^{-1}), \quad x > 0.$$

Podľa Vety 2.1.2 potom  $F_*(x) \in \text{MDA}(G_1)$ , pričom normujúce konštanty môžeme zvoliť ako  $a_n^* = F_*^{-1}(1 - n^{-1})$  a  $b_n = 0$ . To znamená, že keďže  $F_*$  patrí do sféry príťažlivosti Fréchetovho rozdelenia, platí

$$F_*^n(a_n^*x) \rightarrow G_1(x),$$

a preto využitím vzťahu (2.6) získavame

$$F^n(x_F - (a_n^*x)^{-1}) \rightarrow G_2(-x^{-1}).$$

Nakoniec, substitúciou  $x = -y^{-1}$  získavame

$$F^n\left(x_F + \frac{y}{a_n^*}\right) \rightarrow G_2(y), \quad y < 0, \quad (2.9)$$

čo je distribučná funkcia Weibullovoho rozdelenia v bode  $y$ . Zostáva určiť tvar normujúcich konštánt. Z (2.9) okamžite vidíme, že  $b_n = x_F$ . Ďalej  $a_n = (a_n^*)^{-1}$ , pričom podľa (2.5) je

$$a_n^* = F_*^{-1}(1 - n^{-1}) = \inf \{x \in \mathbb{R} : F_*(x) \geq 1 - n^{-1}\}.$$

Zavedením substitúcie  $u = x_F - x^{-1}$  a ďalšími úpravami získavame

$$\begin{aligned} a_n^* &= \inf \{(x_F - u)^{-1} : F(u) \geq 1 - n^{-1}\} = \\ &= (x_F - \inf \{u : F(u) \geq 1 - n^{-1}\})^{-1} = \\ &= (x_F - F^{-1}(1 - n^{-1}))^{-1}, \end{aligned}$$

z čoho vyplýva, že  $a_n = (a_n^*)^{-1} = x_F - F^{-1}(1 - n^{-1})$ , čo sme chceli dokázať. □

Z prechádzajúcej vety vidíme, že rozdelenia patriace do sféry príťažlivosti Weibullovoho rozdelenia majú obmedzený pravý koncový bod. Preto sa toto rozdelenie vo veľkej miere nevyužíva na modelovanie situácií vo finančných aplikáciách. Napriek tomu môžu byť rozdelenia tejto sféry vhodné pri nejakých špecifických stratégiách ako napr. zaistenie (hedging), kedy sa určitou kombináciou aktív alebo derivátov dokáže vytvoriť situácia, v ktorej je strata (prípadne zisk) investora limitovaný.

**Príklad 2.1.2.** Rovnomerné spojité rozdelenie na intervale  $(0, 1)$  patrí do  $\text{MDA}(G_2)$ .

Distribučná funkcia je tvaru  $F(x) = x$ , pre  $x \in (0, 1)$ . Zrejme potom  $x_F = 1 < \infty$ . Ďalej overme, či platí (2.7):

$$\bar{F}(x_F - x^{-1}) = \bar{F}(1 - x^{-1}) = 1 - (1 - x^{-1}) = x^{-1},$$

takže podmienka je splnená pre  $L(x) \equiv 1$  a  $\alpha = 1$ , z čoho vyplýva, že  $F \in \text{MDA}(G_2)$ .

Pri druhom spôsobe vypočítame normujúce konštanty podľa (2.8), teda

$$b_n = x_F = 1.$$

Ďalej

$$\begin{aligned} F(x) &= 1 - n^{-1}, \\ x &= 1 - n^{-1}, \end{aligned}$$

a preto

$$a_n = x_F - F^{-1}(1 - n^{-1}) = 1 - (1 - n^{-1}) = n^{-1}.$$

Overíme konvergenciu

$$F^n(a_n x + b_n) = \left[1 + \frac{x}{n}\right]^n \xrightarrow{n \rightarrow \infty} \exp(x).$$

Pritom  $\exp(x)$  je distribučná funkcia Weibullovoho rozdelenia pre hodnotu  $\alpha = 1$ , a teda  $F \in \text{MDA}(G_2)$ .

Okrem rovnomerného spojitého rozdelenia patrí do  $\text{MDA}(G_2)$  napríklad aj beta rozdelenie, pozri [7].

## Gumbelovo rozdelenie

Popísať sféru príťažlivosti Gumbelovho rozdelenia je zložitejšie ako v predchádzajúcich dvoch prípadoch. Ak sa pozrieme na chvost jeho distribučnej funkcie, zistíme, že (využitím aproximácie exponenciály prvými dvoma členmi príslušného Taylorovho radu)

$$\overline{G}_0(x) \approx \exp(-x), \quad x \rightarrow \infty.$$

Preto môžeme očakávať, že chvosty rozdelení, ktoré budú patriť do  $\text{MDA}(G_0)$ , budú klesať exponenciálne. Opäť musíme rozhodnúť, koľko môžeme z tejto predstavy zľaviť, aby sa neporušili podmienky sféry príťažlivosti.

Do  $\text{MDA}(G_0)$  patrí široká trieda rozdelení, ktoré nemajú rovnaké charakteristiky chvostov. Patrí sem jednak normálne rozdelenie s ľahkými chvostami, ale aj napr. lognormálne rozdelenie s chvostami relatívne ťažkými. Túto triedu popisujeme pomocou tzv. von Misesových funkcií.

**Definícia 2.1.3.** Nech  $F$  je distribučná funkcia náhodnej veličiny  $X$  s pravým koncovým bodom  $x_F \leq \infty$ . Predpokladajme, že existuje  $z \in \mathbb{R}$  také, že pre  $\overline{F}$  platí

$$\overline{F}(x) = K \exp \left\{ - \int_z^x \frac{1}{d(t)} dt \right\}, \quad z < x < x_F \quad (2.10)$$

pre nejakú konštantu  $K > 0$  a funkciu  $d(x)$ , ktorá je kladná a absolútne spojitá s deriváciou  $d'(x)$ , pre ktorú platí

$$\lim_{x \uparrow x_F} d'(x) = 0.$$

Potom funkcia  $F$  sa nazýva *von Misesova funkcia*.



Príkladom von Misesovej funkcie je napr. distribučná funkcia exponenciálneho rozdelenia  $F$  s funkciou  $d(x) = \lambda^{-1}$ , nakoľko

$$\bar{F}(x) = \exp(-\lambda x), \quad x \geq 0,$$

a tým pádom platí (2.10).

Postačujúcou podmienkou náležitosti rozdelenia do Gumbelovej sféry príťažlivosti je, že ak distribučná funkcia  $F$  je von Misesova funkcia, potom  $F \in \text{MDA}(G_0)$  (pozri [7]). To však nie je jej úplná charakterizácia, kedy je potrebné von Misesovu funkciu určitým spôsobom zovšeobecniť.

**Veta 2.1.4.** *Rozdelenie s distribučnou funkciou  $F$  patrí do sféry príťažlivosti Gumbelovho rozdelenia práve vtedy, keď existuje nejaké  $z < x_F$  tak, že  $\bar{F}(x)$  môžeme vyjadriť ako*

$$\bar{F}(x) = K(x) \exp \left\{ - \int_z^x \frac{g(t)}{d(t)} dt \right\}, \quad z < x < x_F, \quad (2.11)$$

kde  $K$  a  $g$  sú merateľné funkcie spĺňajúce

$$\lim_{x \uparrow x_F} K(x) = K > 0 \quad \text{a} \quad \lim_{x \uparrow x_F} g(x) = 1$$

a  $d(x)$  je kladná a absolútne spojitá funkcia, pre ktorú platí  $\lim_{x \uparrow x_F} d'(x) = 0$ .

Naviac, ak  $F \in \text{MDA}(G_0)$ , tak pre  $n \rightarrow \infty$

$$F^n(a_n x + b_n) \rightarrow G_0(x)$$

a normujúce konštanty môžeme voliť

$$a_n = d(b_n), \quad b_n = \bar{F}(1 - n^{-1}), \quad (2.12)$$

pričom možný tvar funkcie  $d(x)$  je

$$d(x) = \int_x^{x_F} \frac{\bar{F}(t)}{\bar{F}(x)} dt.$$

*Dôkaz.* Pozri [19]. □

**Príklad 2.1.3.** Ako sme už ukázali v Príklade 2.0.1, exponenciálne rozdelenie s parametrom  $\lambda = 1$  patrí do sféry príťažlivosti Gumbelovho rozdelenia. Táto skutočnosť sa vzťahuje na exponenciálne rozdelenie pre všetky hodnoty parametru  $\lambda > 0$ .

Zvoľme  $K(x) = 1$ ,  $g(x) = 1$ ,  $z = 0$  a  $d(x) = \lambda^{-1}$ . Následne dosadením do vzťahu (2.11) získavame

$$\bar{F}(x) = \exp \left\{ - \int_0^x \lambda dt \right\} = \exp \{-\lambda x\},$$

čo je distribučná funkcia exponenciálneho rozdelenia, a teda  $F$  spĺňa danú podmienku a patrí do sféry príťažlivosti Gumbelovho rozdelenia.

Náležitost do  $\text{MDA}(G_0)$  môžeme overiť aj priamym výpočtom pomocou normujúcich konštánt vyjadrených na základe vzťahu (2.12). Takže

$$\begin{aligned} F(x) &= 1 - n^{-1}, \\ 1 - e^{-\lambda x} &= 1 - n^{-1}, \\ x &= \frac{\ln n}{\lambda}, \end{aligned}$$

preto zvolíme  $b_n = \lambda^{-1} \ln n$  a

$$a_n = d \left( \frac{\ln n}{\lambda} \right) = \frac{1}{\lambda}.$$

Ďalej stačí overiť, že  $F^n(a_n x + b_n)$  konverguje pre  $n \rightarrow \infty$  ku  $G_0(x)$ , takže

$$F^n(a_n x + b_n) = [1 - \exp\{-(x + \ln n)\}]^n = \left[1 - \frac{e^{-x}}{n}\right]^n$$

pričom

$$\left[1 - \frac{e^{-x}}{n}\right]^n \xrightarrow{n \rightarrow \infty} \exp\{e^{-x}\},$$

a preto  $F \in \text{MDA}(G_0)$ .

Do sféry príťažlivosti Gumbelovho rozdelenia patrí okrem exponenciálneho široké spektrum často využívaných pravdepodobnostných rozdelení - normálne, gama, (klasické) Weibullovo, lognormálne, či  $\chi^2$  rozdelenie. Podobnejšie je táto skutočnosť rozoberaná napr. v [2], [4] alebo [7].

## 2.2 Max-stabilita rozdelení extrémnych hodnôt

Rozdelenia extrémnych hodnôt majú zaujímavú vlastnosť – sú tzv. max-stabilné.

**Definícia 2.2.1.** Rozdelenie náhodného výberu  $X_1, \dots, X_n$  sa nazýva *max-stabilné*, ak pre jeho nedegenerovanú distribučnú funkciu  $F(x)$  platí, že existuje vhodná postupnosť konštánt  $a_n > 0$  a  $b_n \in \mathbb{R}$  takých, že

$$[F(a_n x + b_n)]^n = F(x)$$

pre všetky  $x \in \mathbb{R}$ .

Pripomeňme, že  $[F(a_n x + b_n)]^n$  z predchádzajúcej definície je distribučná funkcia normovaného maxima náhodného výberu. Vlastnosť max-stability majú všetky rozdelenia extrémnych hodnôt, čo si postupne ukážeme.

Nech  $X_1, \dots, X_n$ ,  $n \in \mathbb{N}$  je náhodný výber z **Gumbelovho rozdelenia** s parametrami  $\mu$  a  $\sigma$ , t.j. jeho distribučná funkcia má tvar

$$G_{0,\mu,\sigma} = \exp\left\{-e^{-\frac{x-\mu}{\sigma}}\right\}.$$

Potom distribučná funkcia maxima je podľa (2.1) tvaru

$$G_n(x) = P(M_n \leq x) = [F(x)]^n = \left[ \exp \left\{ -e^{-\frac{x-\mu}{\sigma}} \right\} \right]^n$$

Postupne upravujeme

$$\begin{aligned} G_n(x) &= \exp \left\{ -n \cdot e^{-\frac{x-\mu}{\sigma}} \right\} = \exp \left\{ -e^{\ln n} e^{-\frac{x-\mu}{\sigma}} \right\} = \exp \left\{ -e^{\frac{\sigma \ln n - (x-\mu)}{\sigma}} \right\} = \\ &= \exp \left\{ -e^{-\frac{x - (\mu + \sigma \ln n)}{\sigma}} \right\} = G_{0, \mu^*, \sigma}(x), \end{aligned}$$

čo je distribučná funkcia Gumbelovho rozdelenia s parametrom polohy  $\mu^* = \mu + \sigma \ln n$  posunutým vzhľadom na pôvodné rozdelenie a s rovnakým parametrom merítka  $\sigma$ .

Podobne, pre **Fréchetovo rozdelenie** s distribučnou funkciou s parametrami  $\mu$  a  $\sigma$  definovanou ako

$$G_{1, \mu, \sigma}(x) = \exp \left\{ - \left( \frac{x - \mu}{\sigma} \right)^{-\alpha} \right\}, \quad \alpha > 0$$

počítajme

$$\begin{aligned} G_n(x) &= [F(x)]^n = \left[ \exp \left\{ - \left( \frac{x - \mu}{\sigma} \right)^{-\alpha} \right\} \right]^n = \exp \left\{ -n \left( \frac{x - \mu}{\sigma} \right)^{-\alpha} \right\} = \\ &= \exp \left\{ - \left( \frac{x - \mu}{n^{1/\alpha} \sigma} \right)^{-\alpha} \right\} = G_{1, \mu, \sigma^*}(x). \end{aligned}$$

Úpravami sme prišli opäť k distribučnej funkcii Fréchetovho rozdelenia s parametrami  $\mu$  a  $\sigma^* = \sigma \cdot n^{1/\alpha}$ .

Nakoniec uvažujme **Weibullovo rozdelenie** s distribučnou funkciou

$$G_{2, \mu, \sigma}(x) = \exp \left\{ - \left[ - \left( \frac{x - \mu}{\sigma} \right) \right]^{-\alpha} \right\}, \quad \alpha < 0,$$

s parametrami  $\mu$  a  $\sigma$ . Platí

$$\begin{aligned} G_n(x) &= [F(x)]^n = \left[ \exp \left\{ - \left[ - \left( \frac{x - \mu}{\sigma} \right) \right]^{-\alpha} \right\} \right]^n = \exp \left\{ -n \left[ - \left( \frac{x - \mu}{\sigma} \right) \right]^{-\alpha} \right\} = \\ &= \exp \left\{ - \left[ - \left( \frac{x - \mu}{n^{1/\alpha} \sigma} \right) \right]^{-\alpha} \right\} = G_{2, \mu, \sigma^*}(x). \end{aligned}$$

Pritom  $G_{2, \mu, \sigma^*}(x)$  je znova distribučnou funkciou Weibullovoho rozdelenia, tentokrát s parametrami  $\mu$  a  $\sigma^* = n^{1/\alpha} \sigma$ .

Z výpočtov plynie, že ak má náhodný výber nejaké z rozdelení extrémnych hodnôt, tak potom maximum tohto výberu má opäť príslušné rozdelenie extrémnych hodnôt, s prípadne zmenenými parametrami polohy alebo merítka. Tým sme ukázali, že EV rozdelenia sú max-stabilné.

## 2.3 Modelovanie metódou blokových maxím

V nasledujúcej časti sa zameriame na praktickejšie úlohy pri analýze blokových maxím.

Nech  $X_1, \dots, X_N$ ,  $N \in \mathbb{N}$  je náhodný výber z nejakého rozdelenia s distribučnou funkciou  $F(x)$ . Pre účely metódy rozdelíme dáta do  $m$  blokov po  $n$  pozorovaní. Celkovo teda máme  $N = n \cdot m$  pozorovaní. Ak  $n \cdot m \neq N$ , môžeme túto skutočnosť pre dostatočne veľký počet pozorovaní zanedbať, ak sa počet pozorovaní v blokoch nelíši dramaticky. V každom bloku potom nájdime maximum  $M_i$ ,  $i = 1, \dots, m$ . Blokové maximá tvoria náhodný výber, s ktorým budeme ďalej pracovať.

### 2.3.1 Odhad parametrov GEV rozdelenia

Pre dostatočne veľký počet pozorovaní v jednotlivých blokoch by postupnosť blokových maxím mala konvergovať ku GEV rozdeleniu. Budeme potom ďalej predpokladať, že  $M_1, \dots, M_m$  je náhodný výber, ktorý môžeme aproximovať GEV rozdelením s parametrami  $\gamma$ ,  $\mu$  a  $\sigma$ . Tieto parametre však nepoznáme a musíme ich odhadnúť vhodnou metódou, napríklad metódou maximálnej vierohodnosti, či metódou vážených momentov.

### Metóda maximálnej vierohodnosti

Jedným z najbežnejšie používaných odhadov je maximálne vierohodný odhad. Podstatou tejto metódy je určenie vierohodnostnej funkcie náhodného výberu z nejakého rozdelenia. Následnou maximalizáciou tejto funkcie získame takzvané maximálne vierohodné odhady parametrov rozdelenia.

Hustota GEV rozdelenia je tvaru

$$g_{\gamma, \mu, \sigma}(x) = \frac{1}{\sigma} \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma} - 1} \cdot \exp \left\{ - \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\},$$

pre  $1 + \frac{\gamma(x - \mu)}{\sigma} > 0$ ,  $\gamma \neq 0$ ,  $\gamma \in \mathbb{R}$ ,  $\mu \in \mathbb{R}$  a  $\sigma > 0$ . Potom logaritmus hustoty GEV rozdelenia je

$$\ln g_{\gamma, \mu, \sigma}(x) = -\ln \sigma - \left( \frac{1}{\gamma} + 1 \right) \ln \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right) - \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}}.$$

Teraz môžeme zhotoviť logaritmickú vierohodnostnú funkciu  $l(\gamma, \mu, \sigma; M_1, \dots, M_m)$  pre náhodný výber blokových maxím  $M_1, \dots, M_m$ , a teda

$$\begin{aligned} l(\gamma, \mu, \sigma; M_1, \dots, M_m) &= \sum_{i=1}^m \ln g_{\gamma, \mu, \sigma}(M_i) = \\ &= -m \ln \sigma - \left( \frac{1}{\gamma} + 1 \right) \sum_{i=1}^m \ln \left( 1 + \frac{\gamma(M_i - \mu)}{\sigma} \right) - \sum_{i=1}^m \left( 1 + \frac{\gamma(M_i - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}}. \end{aligned}$$

Príslušné maximálne vierohodnostné odhady  $\hat{\gamma}$ ,  $\hat{\mu}$  a  $\hat{\sigma}$  získame ako argument maxima logaritmickej vierohodnostnej funkcie. Výraz máme maximalizovať pre všetky hodnoty  $\gamma \in \mathbb{R}$ ,  $\mu \in \mathbb{R}$ ,  $\sigma > 0$  spĺňajúce podmienku

$$1 + \frac{\gamma(M_i - \mu)}{\sigma} > 0, \quad i = 1, \dots, m.$$

Explicitné riešenie tejto úlohy neexistuje, a preto sa na výpočet odhadov parametrov používajú numerické metódy.

Poznamenajme, že množina prípustných hodnôt odhadovaných parametrov závisí na pozorovaniach, čo znamená, že nie sú splnené podmienky regularity. Práve podmienkami regularity sa zaoberal Smith [21]. Dokázal, že pre hodnoty  $\gamma > -0,5$  sú odhady asymptoticky nestranné, konzistentné a majú asymptoticky normálne rozdelenie. V prípade  $-1 < \gamma \leq -0,5$  maximálne vierohodné odhady existujú, no nutne nemajú štandardné asymptotické vlastnosti. Nakoniec, pre  $\gamma < -1$  takéto odhady zvyčajne vôbec neexistujú. V praxi to však nie je prekážka, nakoľko prípad, kedy parameter  $\gamma < -0,5$  nastáva len naozaj zriedkavo.

## Metóda vážených momentov

Metóda pravdepodobnostne vážených momentov je zovšeobecnenie klasickej momentovej metódy pre odhad parametrov. Definujme najskôr všeobecné pravdepodobnostne vážené momenty.

**Definícia 2.3.1.** Nech  $X$  je náhodná veličina z nejakého rozdelenia s distribučnou funkciou  $F(x)$ . Potom veličiny

$$M_{p,r,s} = E \{ X^p [F(X)]^r [1 - F(X)]^s \},$$

sa nazývajú *pravdepodobnostne vážené momenty*, pričom  $p$ ,  $r$  a  $s$  sú reálne čísla.

Ak je možné kvantilovú funkciu  $F^{-1}(u) = \inf \{ x \in \mathbb{R}; F(x) \geq u \}$  k distribučnej funkcii  $F$  vyjadriť explicitne, potom môžeme vážené momenty ekvivalente zapísať v tvare

$$M_{p,r,s} = \int_0^1 (F^{-1})^p F^r (1 - F)^s dF. \quad (2.13)$$

Tento tvar býva často vhodnejší pre výpočet vážených momentov. Práve preto ho využijeme na odhad parametrov GEV rozdelenia, pričom Hosking a kol. [12] odporúčajú uvažovať vážené momenty typu

$$\beta_r = M_{1,r,0} = E \{ X [F(X)]^r \}, \quad r = 0, 1, 2.$$

V prípade GEV rozdelenia s distribučnou funkciou

$$G_{\gamma,\mu,\sigma}(x) = \exp \left\{ - \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\}, \quad 1 + \frac{\gamma(x - \mu)}{\sigma} > 0, \gamma \neq 0,$$

majú momenty  $\beta_r$  tvar

$$\beta_r = \frac{1}{r+1} \left\{ \mu - \frac{\sigma}{\gamma} [1 - (r+1)^\gamma \Gamma(1-\gamma)] \right\}, \quad \gamma < 1, \gamma \neq 0. \quad (2.14)$$

Vzťah (2.14) odvodíme. Podľa (2.13) platí

$$\beta_r = \int_0^1 G^{-1} G^r dG$$

Kvantilová funkcia GEV rozdelenia je tvaru

$$G^{-1} = \mu + \frac{\sigma}{\gamma} [(-\ln G)^{-\gamma} - 1],$$

a preto pokračujeme

$$\beta_r = \int_0^1 \left\{ \mu + \frac{\sigma}{\gamma} [(-\ln G)^{-\gamma} - 1] \right\} G^r dG.$$

Zavedením substitúcie  $t = -\ln G$  dostávame

$$\begin{aligned} \beta_r &= \int_0^\infty \left[ \mu + \frac{\sigma}{\gamma} (t^{-\gamma} - 1) \right] e^{-t(r+1)} dt = \\ &= \left( \mu - \frac{\sigma}{\gamma} \right) \int_0^\infty e^{-t(r+1)} dt + \frac{\sigma}{\gamma} \int_0^\infty t^{-\gamma} e^{-t(r+1)} dt = \\ &= \left( \mu - \frac{\sigma}{\gamma} \right) (r+1)^{-1} + \frac{\sigma}{\gamma} \int_0^\infty t^{-\gamma} e^{-t(r+1)} dt. \end{aligned}$$

Druhá substitúcia  $u = (r+1)t$  nás pomaly dovedie k výsledku:

$$\beta_r = \left( \mu - \frac{\sigma}{\gamma} \right) (r+1)^{-1} + \frac{\sigma}{\gamma} (r+1)^{\gamma-1} \int_0^\infty u^{-\gamma} e^{-u} du,$$

kde integrál vo výraze je gama funkcia<sup>2</sup>  $\Gamma(1-\mu)$ , a preto vzťah (2.14) platí, čím sme odvodili pravdepodobnostne vážené momenty zovšeobecneného rozdelenia extrémnych hodnôt.

Rovnice na získanie odhadov  $\hat{\gamma}$ ,  $\hat{\mu}$ ,  $\hat{\sigma}$  parametrov  $\gamma$ ,  $\mu$ ,  $\sigma$  dostaneme pomocou vážených momentov GEV rozdelenia, pričom pre  $r = 0$  podľa (2.14) platí

$$\beta_0 = \mu - \frac{\sigma}{\gamma} [1 - \Gamma(1-\gamma)]. \quad (2.15)$$

Ďalej pre  $r = 1$  priebežným využitím vyššie odvodeného vzťahu pre  $\beta_0$  dostávame

<sup>2</sup>Gama funkcia je definovaná ako  $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$  pre  $x > 0$ .

$$\begin{aligned}
\beta_1 &= \frac{1}{2} \left\{ \mu - \frac{\sigma}{\gamma} [1 - 2^\gamma \Gamma(1 - \gamma)] \right\}, \\
2\beta_1 &= \mu - \frac{\sigma}{\gamma} [1 - \Gamma(1 - \gamma) + \Gamma(1 - \gamma) - 2^\gamma \Gamma(1 - \gamma)], \\
2\beta_1 &= \mu - \frac{\sigma}{\gamma} [1 - \Gamma(1 - \gamma)] + \frac{\sigma}{\gamma} \Gamma(1 - \gamma)(2^\gamma - 1), \\
2\beta_1 - \beta_0 &= \frac{\sigma}{\gamma} \Gamma(1 - \gamma)(2^\gamma - 1).
\end{aligned}$$

Obdobne získame postupnými úpravami rovnicu pre  $r = 2$ , a teda

$$\begin{aligned}
\beta_2 &= \frac{1}{3} \left\{ \mu - \frac{\sigma}{\gamma} [1 - 3^\gamma \Gamma(1 - \gamma)] \right\}, \\
3\beta_2 &= \mu - \frac{\sigma}{\gamma} [1 - \Gamma(1 - \gamma)] + \frac{\sigma}{\gamma} \Gamma(1 - \gamma)(3^\gamma - 1), \\
3\beta_2 - \beta_0 &= \frac{\sigma}{\gamma} \Gamma(1 - \gamma)(3^\gamma - 1), \\
3\beta_2 - \beta_0 &= \frac{2\beta_1 - \beta_0}{2^\gamma - 1}, \\
\frac{3\beta_2 - \beta_0}{2\beta_1 - \beta_0} &= \frac{3^\gamma - 1}{2^\gamma - 1}.
\end{aligned}$$

Celkovo sme obdržali tri rovnice o troch neznámých

$$\begin{aligned}
\beta_0 &= \mu - \frac{\sigma}{\gamma} [1 - \Gamma(1 - \gamma)], \\
2\beta_1 - \beta_0 &= \frac{\sigma}{\gamma} \Gamma(1 - \gamma)(2^\gamma - 1), \\
\frac{3\beta_2 - \beta_0}{2\beta_1 - \beta_0} &= \frac{3^\gamma - 1}{2^\gamma - 1},
\end{aligned} \tag{2.16}$$

ktoré je potreba vyriešiť. Pritom  $\beta_r$ ,  $r = 0, 1, 2$  nahradíme nejakým vhodným odhadom. Napríklad Landwehr [16] odvodil nestranný odhad pre zoradený náhodný výber  $M_{(1)} \leq \dots \leq M_{(n)}$  v tvare

$$\hat{\beta}_r = \frac{1}{n} \sum_{j=1}^n \left( \prod_{l=1}^r \frac{j-l}{n-l} \right) M_{(j)}, \quad r = 1, 2,$$

pričom  $\hat{\beta}_0 = n^{-1} \sum_{j=1}^n M_{(j)}$ . Potom odhad  $\hat{\gamma}$  parametru  $\gamma$  získame z poslednej rovnice sústavy (2.16) numerickým riešením. Ďalej, po dosadení tohto odhadu do druhej rovnice, máme

$$\hat{\sigma} = \frac{\hat{\gamma}(2\hat{\beta}_1 - \hat{\beta}_0)}{\Gamma(1 - \hat{\gamma})(2^{\hat{\gamma}} - 1)}.$$

Nakoniec, z prvej rovnice získame odhad parametru  $\mu$

$$\hat{\mu} = \hat{\beta}_0 + \frac{\hat{\sigma}}{\hat{\gamma}} [1 - \Gamma(1 - \hat{\gamma})].$$

Odhady  $\hat{\gamma}$ ,  $\hat{\mu}$ ,  $\hat{\sigma}$  rovnako ako v prípade maximálne vierohodných odhadov nespĺňajú podmienky regularity, avšak sú asymptoticky normálne pre  $\gamma < 0,5$ . Navyiac, odhady metódou vážených momentov vykazujú lepšie správanie ako maximálne vierohodné odhady v prípade malých výberov, pozri [12].

Existuje mnoho ďalších typov odhadov pre parametre EV a GEV rozdelení. Často sa využíva Hillov odhad ([2], [7], [17]), ďalej Pickandsov či Dekkers–Einmahl–de Haanov odhad ([7]) alebo klasická metóda momentov ([3]), prípadne metóda vážených najmenších štvorcov ([2]).

### 2.3.2 Spracovanie dát

V nasledujúcej časti sa zameriame na vybrané postupy pre štatistickú analýzu extrémnych hodnôt. Postupne rozoberieme niekoľko otázok – aká je pravdepodobnosť, že náhodný jav nadobudne nejakej extrémnej hodnoty a aká je úroveň a doba návratu extrémneho javu.

### Pravdepodobnosť nastania extrémnej udalosti

Uvažujme príklad, kedy máme k dispozícii dáta o množstve napadnutých zrážok na určitom území. Aká je pravdepodobnosť, že nasledujúci deň napadne napr. 100 mm zrážok?

Matematicky formulujme úlohu nasledovne. Nech  $X_1, \dots, X_N$  je náhodný výber z rozdelenia s distribučnou funkciou  $F(x)$ , pričom pozorovania sú rozdelené do  $m$  blokov s počtom pozorovaní  $n$  v každom bloku. Odhadnime teraz pravdepodobnosť  $P(X_i > x)$  pre nejaké  $x$  veľké. Ak je teda dĺžka bloku  $n$  pozorovaní, tak podľa (2.1) platí, že distribučná funkcia blokového maxima  $M_n$  je

$$P(M_n \leq x) = [P(X_i \leq x)]^n = [1 - P(X_i > x)]^n,$$

pričom v poslednom vyjadrení sa nachádza člen, ktorý je predmetom nášho záujmu a ktorý chceme v určitej rozumnej forme vyjadriť. Zároveň poznáme tvar distribučnej funkcie  $G_{\gamma, \mu, \sigma}(x)$  zovšeobecneného rozdelenia extrémnych hodnôt (2.3), pomocou ktorého modelujeme blokové maximá, a preto platí

$$G_{\gamma, \mu, \sigma}(x) = [1 - P(X_i > x)]^n,$$

z čoho vyplýva, že

$$P(X_i > x) = 1 - (G_{\gamma, \mu, \sigma}(x))^{\frac{1}{n}}.$$

Po dosadení distribučnej funkcie GEV rozdelenia do predchádzajúceho vzťahu, pričom  $\gamma \neq 0$  a  $1 + \gamma(x - \mu)/\sigma > 0$ , výraz upravíme do tvaru



$$\begin{aligned} P(X_i > x) &= 1 - \left[ \exp \left\{ - \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\} \right]^{\frac{1}{n}} \\ &= 1 - \exp \left\{ -\frac{1}{n} \left( 1 + \frac{\gamma(x - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\}. \end{aligned}$$

Pre veľký rozsah výberu môžeme výraz aproximovať pomocou prvých dvoch členov Taylorovho polynómu pre rozvoj funkcie  $e^x$ , kde približne platí, že  $e^x = 1 + x$ . Parametre GEV rozdelenia nahradíme príslušnými odhadmi. Celkovo teda môžeme vysoké (extrémne) pravdepodobnosti odhadnúť vzťahom

$$P(\widehat{X}_i > x) = \frac{1}{n} \left( 1 + \frac{\widehat{\gamma}(x - \widehat{\mu})}{\widehat{\sigma}} \right)^{-\frac{1}{\widehat{\gamma}}}, \quad (2.17)$$

kde  $\widehat{\mu}$ ,  $\widehat{\sigma}$  a  $\widehat{\gamma}$  sú vhodné odhady príslušných parametrov.

## Odhad extrémnych kvantilov – úroveň a doba návratu

Opäť na príklade dát o množstve napadnutých zrážok uvažujme situáciu, kedy nás zaujíma, aké extrémne množstvo zrážok sa priemerne vyskytne napr. raz za 10 rokov. Tomuto pojmu sa hovorí **úroveň návratu** (*return level*), pričom teda chceme určiť veľkosť extrémneho javu, ktorý bude v priemere nastávať so známou zvolenou frekvenciou. V našom kontexte sa jedná o odhad vysokých (extrémnych) kvantilov GEV rozdelenia.

Hľadáme úroveň, ktorá bude  $i$ -tým blokovým maximom  $M_i$ ,  $i = 1, \dots, m$ , prekročená raz za  $k$  časových okamihov, čo sa dá pomocou definície  $(1 - 1/k)$ -kvantilu vyjadriť ako

$$P(M_i > q_{1-1/k}) = \frac{1}{k}, \quad k \in \mathbb{N}, \quad (2.18)$$

pričom  $q_{1-1/k}$  je hľadaná úroveň. Pre väčšiu prehľadnosť ďalšieho výpočtu označme  $q := q_{1-1/k}$ . Pre dostatočný rozsah náhodného výberu aproximujme výraz (2.18) pomocou distribučnej funkcie GEV rozdelenia (2.3):

$$1 - G_{\gamma, \mu, \sigma}(q) = 1 - \exp \left\{ - \left( 1 + \frac{\gamma(q - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\} = \frac{1}{k},$$

z ktorého chceme vyjadriť úroveň návratu  $q$ , pričom  $\gamma \neq 0$  a  $1 + \gamma(x - \mu)/\sigma > 0$ . Preto pokračujme vo výpočte

$$\begin{aligned} \exp \left\{ - \left( 1 + \frac{\gamma(q - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} \right\} &= 1 - \frac{1}{k}, \\ - \left( 1 + \frac{\gamma(q - \mu)}{\sigma} \right)^{-\frac{1}{\gamma}} &= \ln \left( 1 - \frac{1}{k} \right), \\ 1 + \frac{\gamma(q - \mu)}{\sigma} &= \left[ -\ln \left( 1 - \frac{1}{k} \right) \right]^{-\gamma}, \end{aligned}$$

odkiaľ

$$q = \mu + \frac{\sigma}{\gamma} \left[ \left( -\ln \left( 1 - \frac{1}{k} \right) \right)^{-\gamma} - 1 \right].$$

Po nahradení  $\mu, \sigma$  a  $\gamma$  prislúchajúcimi odhadmi  $\hat{\mu}, \hat{\sigma}, \hat{\gamma}$  získame

$$\hat{q} = \hat{q}_{1-\frac{1}{k}} = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\gamma}} \left[ \left( -\ln \left( 1 - \frac{1}{k} \right) \right)^{-\hat{\gamma}} - 1 \right], \quad (2.19)$$

čo je tzv. odhad úrovne návratu.

S extrémnymi kvantilmi súvisí aj ďalší pojem – **doba návratu** (*return period*) extrémnej udalosti. V tomto prípade je situácia opačná, snažíme sa napr. určiť, priemerne za aké obdobie denné zrážky prekročia 150 mm. To znamená, že k dispozícii máme údaj o úrovni návratu  $q = q_{1-1/k}$ , no priemerná frekvencia, s ktorou bude jav nastávať, nie je známa a snažíme sa ju určitým spôsobom odhadnúť.

Zo vzťahu (2.18)

$$P \left( M_i > q_{1-\frac{1}{k}} \right) = \frac{1}{k}$$

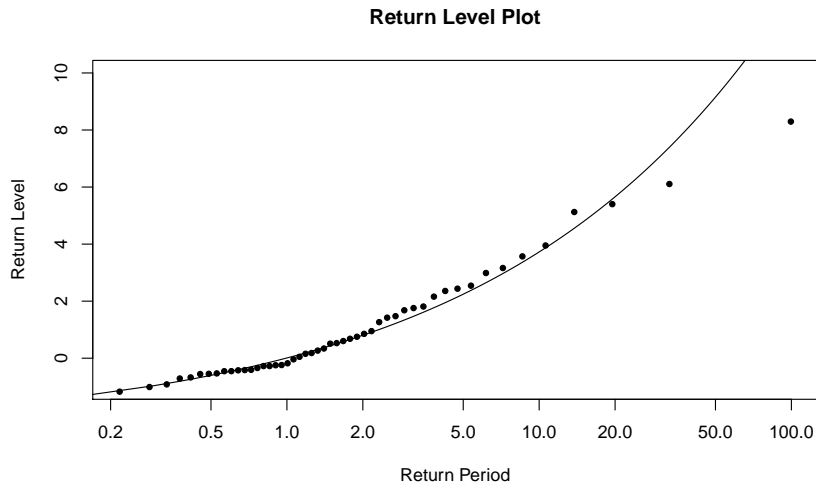
preto vyjadríme dobu návratu  $k$

$$k = \frac{1}{P \left( M_i > q_{1-\frac{1}{k}} \right)} = \frac{1}{1 - \exp \left\{ - \left( 1 + \frac{\gamma(q-\mu)}{\sigma} \right)^{-1/\gamma} \right\}}$$

Prislúchajúci odhad je potom tvaru

$$\hat{k} = \frac{1}{1 - \exp \left\{ - \left( 1 + \frac{\hat{\gamma}(q-\hat{\mu})}{\hat{\sigma}} \right)^{-1/\hat{\gamma}} \right\}}.$$

Vzťah medzi dobou a úrovňou návratu graficky znázorňuje tzv. *return-level plot*. Tvar grafu závisí na hodnote parametru tvaru  $\gamma$  zovšeobecneného rozdelenia extrémnych hodnôt. Ak  $\gamma = 0$ , graf je lineárny. Ak  $\gamma > 0$ , graf má konvexný tvar a ak  $\gamma < 0$ , závislosť je konkávneho tvaru. Na obr. 2.4 môžeme vidieť príklad takéhoto grafu, pričom na vodorovnej osi je znázornená doba návratu a na zvislej úrovň návratu.

Obr. 2.4: Return–level plot pre  $\gamma = 0,5$ .

Nakoniec tejto kapitoly zostáva otázkou, ako je vhodné zvoliť počet pozorovaní v jednotlivých blokoch, čo patrí medzi najkomplikovanejšie úlohy v metóde blokových maxím. Tu proti sebe vystupujú dve skutočnosti. Pokiaľ určíme bloky, ktoré obsahujú malý počet pozorovaní, získame tak mnoho blokových maxím a to má za následok zlé asymptotické vlastnosti modelu a jeho veľké vychýlenie. No na druhú stranu, veľký počet pozorovaní v blokoch vedie k malému počtu blokových maxím, čo vytvára model s veľkým rozptylom.

V praxi sa doposiaľ nezažíval žiadny jednotný postup na určenie veľkosti blokov a počtu pozorovaní v nich. Niektorí autori navrhujú využiť Hillov odhad parametru  $\gamma$ , ktorý sa vynesie do grafu v závislosti na počte pozorovaní v jednotlivých blokoch, a ďalej odhadom určiť bod, od ktorého sa graf stabilizuje (pozri napr. [22]). Nevýhodou tohto postupu je akási subjektivita každého jedinca pri určovaní bodu stability. Poznamenajme však, že bloky sa v praxi mnohokrát volia tak, aby korešpondovali s časovým obdobím, ku ktorému sa analyzované dáta viažu. Blokové maximá sú potom často štvrťročné, ročné, desaťročné maximá a podobne.

# Kapitola 3

## Peaks-over-threshold

V nasledujúcej metóde uvažujeme postupnosť i.i.d. náhodných veličín  $X_1, \dots, X_n$  s distribučnou funkciou  $F$ . Viac, ako veľkosť a správanie sa týchto veličín nás však budú zaujímať náhodné veličiny  $Y_1, \dots, Y_{N_u}$ ,  $N_u \leq n$ , ktoré udávajú veľkosť prekročenia nejakej vopred danej medze  $u$ . Prekročenia sú definované ako

$$Y_i = X_i - u, \quad X_i \geq u,$$

pričom  $N_u$  udáva počet prekročení hranice  $u$  náhodnými veličinami  $X_1, \dots, X_n$ , matematicky zapísané ako

$$N_u = \text{card} \{i : i = 1, \dots, n; X_i \geq u\}.$$

Náhodné veličiny  $Y_i$ , niekedy nazývané aj excesy, má preto zmysel uvažovať, len pokiaľ k prekročeniu danej medze došlo. Zostrojme teraz distribučnú funkciu excesov  $F_u(x)$  pre náhodné veličiny  $Y_i$ ,  $i = 1, \dots, N_u$ . Zrejme sa preto jedná o podmienenú distribučnú funkciu

$$F_u(x) = P(Y_i \leq x | X_i \geq u) = P(X_i - u \leq x | X_i \geq u). \quad (3.1)$$

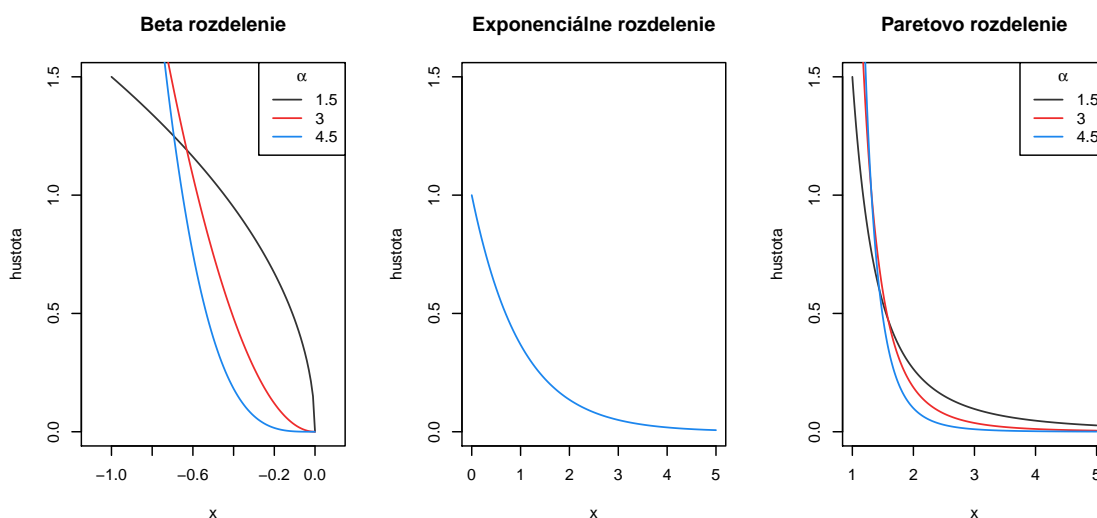
Využitím vzťahu pre podmienenú pravdepodobnosť získavame

$$F_u(x) = \frac{P(u \leq X_i \leq u + x)}{P(X_i \geq u)} = \frac{F(u + x) - F(u)}{1 - F(u)}, \quad x \geq 0. \quad (3.2)$$

Môžeme si všimnúť, že podmienená distribučná funkcia  $F_u(x)$  závisí na výške zvolenej medze  $u$ . V predchádzajúcom odvodení sme uvažovali prekročenia nad určitú medzu, t.j. pre maximálne extrémne udalosti. Analogicky môžeme uvažovať aj extrémne udalosti opačným smerom, čiže udalosti, ktoré (extrémne) klesli pod určitú hranicu. Ďalej sa v celej práci budeme zaoberať situáciami, kedy pozorovanie prekročí vopred danú hranicu smerom nahor.

Podobne ako pri metóde blokových maxím v predchádzajúcej kapitole, vhodné normované excesy budú konvergovať k určitej triede rozdelení, čo ukážeme neskôr. Uvedme preto teraz tieto rozdelenia spolu s ich distribučnými funkciami.

**Definícia 3.0.1.** Pre modelovanie excesov používame rozdelenia definované pomocou nasledujúcich distribučných funkcií:



Obr. 3.1: Hustoty beta, exponenciálneho a Paretovho rozdelenia pre rôzne hodnoty parametru  $\alpha$ .

$$\begin{array}{lll}
 \text{Exponenciálne:} & W_0(x) = 1 - e^{-x}, & x \geq 0, \\
 \text{Paretovo:} & W_1(x) = 1 - x^{-\alpha}, & x \geq 1, \\
 \text{Beta:} & W_2(x) = 1 - (-x)^\alpha, & -1 \leq x \leq 0.
 \end{array}$$

kde  $\alpha > 0$  sa nazýva parameter tvaru.

Hustoty rozdelení z definície 3.0.1 môžeme vidieť na obrázku 3.1. Exponenciálne a Paretovo rozdelenie sa používajú pri modelovaní maxim a beta rozdelenie pri modelovaní miním.

Všeobecný tvar rozdelení dostaneme zavedením parametru polohy  $\mu$  a parametru merítka  $\sigma$ , t.j.

$$W_{i,\mu,\sigma}(x) = W_i\left(\frac{x - \mu}{\sigma}\right), \quad \mu \in \mathbb{R}, \sigma > 0, i = 0, 1, 2.$$

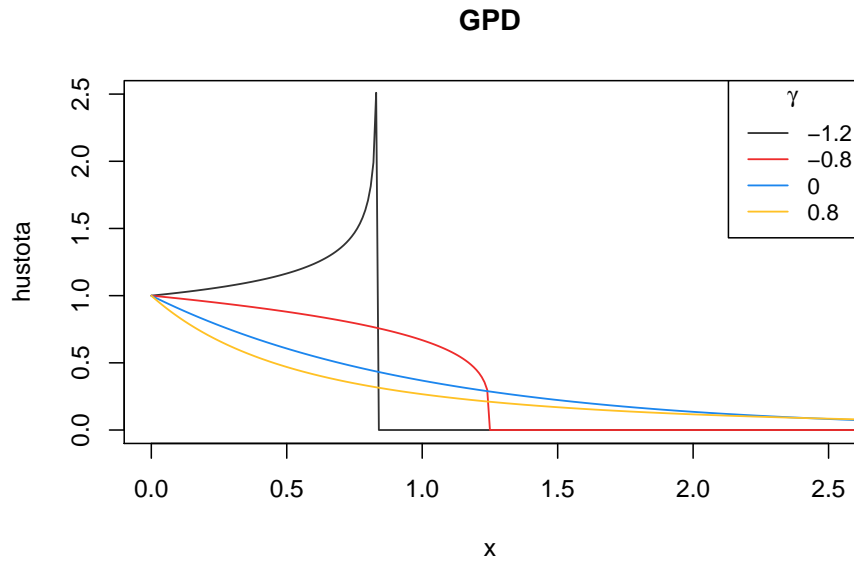
Rozdelenia z definície 3.0.1 môžeme definovať aj jednotným kompaktnjším spôsobom, pomocou *zovšeobecneného Paretovho rozdelenia*.

**Definícia 3.0.2.** Zovšeobecnené Paretovo (*GP - Generalised Pareto*) rozdelenie má distribučnú funkciu v tvare

$$W_\gamma(x) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}}, \quad \gamma \in \mathbb{R}, \gamma \neq 0,$$

pričom  $\gamma$  sa nazýva parameter tvaru.

Hustota zovšeobecneného Paretovho rozdelenia pre rôzne hodnoty parametru  $\gamma$  je zobrazená na obr. 3.2. Nakoľko GP rozdelenie má ľavý koncový bod rovný nule,



Obr. 3.2: Hustota zovšeobecneného Paretovho rozdelenia s rôznymi hodnotami parametru  $\gamma$ .

parameter polohy  $\mu$  v našom prípade nepotrebuje uvažovať. Úplnú triedu GP rozdelenia potom získame pridaním parametru merítka  $\sigma$ , t.j.

$$W_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}, \quad \gamma \in \mathbb{R}, \sigma > 0.$$

Poznamenajme, že GP a GEV rozdelenia majú medzi sebou úzky vzťah, ktorý sa dá jednoducho vyjadriť ako

$$W_{\gamma}(x) = 1 + \ln G_{\gamma}(x).$$

Ďalej sa môžeme presvedčiť, že pre rôzne hodnoty parametru tvaru  $\gamma$  je GP rozdelenie len pozmeneným zápisom rozdelení z definície 3.0.1.

Pre  $\gamma = 0$  sa jedná o exponenciálne rozdelenie, distribučnú funkciu GP rozdelenia však musíme spojitivo dodefinovať v nule, a teda previesť limitný prechod. Preto

$$\lim_{\gamma \rightarrow 0} W_{\gamma}(x) = \lim_{\gamma \rightarrow 0} 1 - (1 + \gamma x)^{-\frac{1}{\gamma}} = 1 - \lim_{\gamma \rightarrow 0} (1 + \gamma x)^{-\frac{1}{\gamma}} = 1 - e^{-x},$$

čo je distribučná funkcia exponenciálneho rozdelenia.

V prípade  $\gamma > 0$  a  $\gamma = 1/\alpha$ ,  $\alpha > 0$  pre distribučnú funkciu GP rozdelenia získavame

$$W_{\gamma}(x) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}} = 1 - \left(1 + \frac{x}{\alpha}\right)^{-\alpha} = W_1\left(1 + \frac{x}{\alpha}\right),$$

čím sme prišli k distribučnej funkcii Paretovho rozdelenia v bode  $1 + x/\alpha$ .

Nakoniec, pre  $\gamma < 0$  a zároveň  $\gamma = -1/\alpha$ ,  $\alpha > 0$  máme

$$W_{\gamma}(x) = 1 - (1 + \gamma x)^{-\frac{1}{\gamma}} = 1 - \left(1 - \frac{x}{\alpha}\right)^{\alpha} = W_2\left(-1 + \frac{x}{\alpha}\right),$$

čiže sme získali beta rozdelenie v bode  $-1 + x/\alpha$ , a tým aj overili korektnosť tvrdenia o zovšeobecnenom Paretovom rozdelení.

Zaoberajme sa teraz limitným rozdelením excesov. Špeciálne nás zaujíma, aké rozdelenie budú mať so zvyšujúcou sa hranicou  $u$ . Preto teraz uvedieme akúsi analógiu Fisherovej-Tippetovej vety 2.0.2 pre excesy.

**Veta 3.0.1** (Balkemova-de Haanova, Pickandsova). *Nech  $X_1, \dots, X_n$  je náhodný výber z nejakého rozdelenia. Ak existujú postupnosti konštant  $a_n > 0$ ,  $b_n \in \mathbb{R}$  tak, že*

$$F_u(a_n x + b_n) \xrightarrow{u \uparrow x_F} H(x),$$

pre nejakú spojitú distribučnú funkciu  $H(x)$ , potom platí

$$\lim_{u \uparrow x_F} |F_u(a_n x + b_n) - W_{\gamma, \sigma(u)}(x)| = 0, \quad x \in \mathbb{R},$$

kde  $W_{\gamma, \sigma(u)}(x)$  je distribučná funkcia GP rozdelenia.

*Dôkaz.* Dôkaz je opäť náročný, je inšpirovaný Gnedenkovým dôkazom Fisherovej-Tippetovej vety a ako prví ho nezávisle na sebe sformulovali Balkema a de Haan ([1]) a Pickands ([18]).  $\square$

Prechádzajúca veta nám vraví, že jediné možné limitné rozdelenie excesov náhodného výberu pre  $u \uparrow x_F$ <sup>1</sup> je zovšeobecnené Paretovo rozdelenie. Budeme teda predpokladať, že excesy  $Y_i = X_i - u$  (za podmienky  $X_i \geq u$ ) majú asymptoticky GP rozdelenie

$$Y_i \approx W_{\gamma, \sigma}$$

a budeme ich modelovať pomocou tohto rozdelenia, pričom parametre  $\gamma$  a  $\sigma$  odhadneme z dát. Ukázať konvergenciu na konkrétnom príklade, ako sme to urobili v predchádzajúcej kapitole, už nie je také jednoduché, nakoľko sa v limite má zvyšovať nielen počet pozorovaní náhodného výberu, ale aj výška thresholdu  $u$ . Tak isto z predchádzajúcej vety vidíme, že variabilita  $\sigma(u)$  závisí na voľbe hranice  $u$ . Čím nižšie threshold zvolíme, tým bude variabilita väčšia a opačne. S tým súvisí problém, ako vhodne zvoliť hranicu  $u$ , ktorému sa budeme venovať nižšie.

Medzi GP a GEV rozdelením je, ako už bolo povedané, blízky vzťah. Dokonca platí, že rozdelenia, pre ktoré normované maximá konvergujú k GEV rozdeleniu, tvoria množinu, pre ktorú príslušné rozdelenie excesov konverguje s rastúcou medzou k zovšeobecnenému Paretovmu rozdeleniu. Zaujímavé pritom je, že parameter tvaru  $\gamma$  nie je závislý na voľbe thresholdu  $u$  a je v oboch prípadoch rovnaký (pozri [7]).

### 3.1 Stabilita rozdelení pre modelovanie excesov

V predchádzajúcej kapitole sme ukázali, že EV rozdelenia sú max-stabilné. Podobnú vlastnosť majú aj rozdelenia excesov, je však nutné pozmeniť definíciu max-stability, nakoľko nás nezaujíma stabilita blokových maxím, ale stabilita prekročení nad hranicu  $u$ .

<sup>1</sup> $u \uparrow x_F$  značí, že  $u$  sa blíži k  $x_F$  zdola.

**Definícia 3.1.1.** Rozdelenie náhodného výberu  $X_1, \dots, X_n$  sa nazýva *stabilné voči excesom*, ak pre jeho nedegenerovanú distribučnú funkciu  $F(x)$  platí, že existuje vhodná postupnosť konštant  $a_n > 0$  a  $b_n \in \mathbb{R}$  takých, že

$$F_u(a_n x + b_n) = F(x)$$

pre všetky  $x \in \mathbb{R}$ , pričom  $F_u(x)$  distribučná funkcia excesov definovaná podľa (3.1).

Túto vlastnosť overíme najskôr na samotnom GP rozdelení. Nech preto náhodný výber  $X_1, \dots, X_n$  je zo **zovšeobecneného Paretoho rozdelenia** s distribučnou funkciou

$$W_{\gamma, \sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}.$$

Pre pevne zvolené  $u$  a excesy  $Y_i = X_i - u$  (za podmienky  $X_i \geq u$ ),  $i = 1, \dots, n$  podľa (3.2) platí

$$F_u(x) = \frac{P(u \leq X_i \leq u + x)}{P(X_i \geq u)} = \frac{W_{\gamma, \sigma}(u + x) - W_{\gamma, \sigma}(u)}{1 - W_{\gamma, \sigma}(u)}.$$

Dosadením a malou úpravou získavame

$$F_u(x) = \frac{\left(1 + \frac{\gamma}{\sigma}u\right)^{-\frac{1}{\gamma}} - \left(1 + \frac{\gamma}{\sigma}(u + x)\right)^{-\frac{1}{\gamma}}}{\left(1 + \frac{\gamma}{\sigma}u\right)^{-\frac{1}{\gamma}}} = 1 - \left[\frac{1 + \frac{\gamma}{\sigma}(u + x)}{1 + \frac{\gamma}{\sigma}u}\right]^{-\frac{1}{\gamma}},$$

pričom úpravou na spoločného menovateľa dostaneme

$$F_u(x) = 1 - \left[\frac{\sigma + \gamma(u + x)}{\sigma + \gamma u}\right]^{-\frac{1}{\gamma}} = 1 - \left[1 + \frac{\gamma x}{\sigma + \gamma u}\right]^{-\frac{1}{\gamma}} = W_{\gamma, \sigma^*}(x),$$

čo je opäť distribučná funkcia GP rozdelenia s rovnakým parametrom tvaru  $\gamma$  a zmeneným parametrom  $\sigma^* = \sigma + \gamma u$ . Tým sme ukázali, že excesy majú znova GP rozdelenie.

Už sme ukázali, že GP rozdelenie je len iným zápisom rozdelení na modelovanie excesov z definície (3.0.1), čo automaticky znamená, že tieto rozdelenia sú taktiež stabilné voči excesom. Môže byť však zaujímavé pozrieť sa, ako sa zmenia jednotlivé parametre polohy a merítka v týchto konkrétnych prípadoch. Prejdime preto postupne ku všetkým trom rozdeleniam. Nech náhodný výber  $X_1, \dots, X_n$  je z **exponenciálneho rozdelenia** s distribučnou funkciou

$$W_{0, \mu, \sigma}(x) = 1 - \exp\left\{-\frac{x - \mu}{\sigma}\right\}, \quad \mu \in \mathbb{R}, \sigma > 0.$$

Potom podmienená distribučná funkcia excesov má tvar



$$F_u = \frac{W_{0,\mu,\sigma}(u+x) - W_{0,\mu,\sigma}(u)}{1 - W_{0,\mu,\sigma}(u)} = \frac{\exp\left\{-\frac{u-\mu}{\sigma}\right\} - \exp\left\{-\frac{u+x-\mu}{\sigma}\right\}}{\exp\left\{-\frac{u-\mu}{\sigma}\right\}},$$

a teda postupnými úpravami prídeme až k

$$\begin{aligned} F_u(x) &= 1 - \frac{\exp\left\{-\frac{u+x-\mu}{\sigma}\right\}}{\exp\left\{-\frac{u-\mu}{\sigma}\right\}} = 1 - \exp\left\{-\frac{u+x-\mu}{\sigma} + \frac{u-\mu}{\sigma}\right\} = \\ &= 1 - \exp\left\{-\frac{x}{\sigma}\right\} = W_{0,\mu^*,\sigma}. \end{aligned}$$

Tým sme prišli k distribučnej funkcii exponenciálneho rozdelenia s parametrom polohy  $\mu^* = 0$  a rovnakým parametrom merítka  $\sigma$ . Táto vlastnosť exponenciálneho rozdelenia je známa aj v iných oblastiach a aplikáciach matematiky, kedy hovoríme, že exponenciálne rozdelenie je tzv. bez pamäti.

Podobne, pre náhodný výber z **Pareto** rozdelenia s distribučnou funkciou

$$W_{1,\mu,\sigma}(x) = 1 - \left(\frac{x-\mu}{\sigma}\right)^{-\alpha}, \quad \alpha > 0$$

môžeme odvodiť

$$F_u(x) = \frac{W_{1,\mu,\sigma}(u+x) - W_{1,\mu,\sigma}(u)}{1 - W_{1,\mu,\sigma}(u)} = \frac{\left(\frac{u-\mu}{\sigma}\right)^{-\alpha} - \left(\frac{u+x-\mu}{\sigma}\right)^{-\alpha}}{\left(\frac{u-\mu}{\sigma}\right)^{-\alpha}},$$

čo po úpravách vedie ku konečnému výsledku

$$F_u(x) = 1 - \left[\frac{x - (\mu - u)}{u - \mu}\right]^{-\alpha} = W_{1,\mu^*,\sigma^*}(x).$$

Toto ukazuje, že excesy nad určitú medzu z náhodného výberu z Pareto rozdelenia majú tak isto Pareto rozdelenie, no so zmenou parametrov, kde parameter polohy je  $\mu^* = \mu - u$  a parameter merítka  $\sigma^* = u - \mu$ .

Nakoniec, táto vlastnosť platí aj pre náhodný výber z **beta** rozdelenia s distribučnou funkciou

$$W_{2,\mu,\sigma} = 1 - \left(-\frac{x-\mu}{\sigma}\right)^{\alpha}, \quad \alpha > 0.$$

Obdobným postupom získavame

$$F_u(x) = \frac{W_{2,\mu,\sigma}(u+x) - W_{2,\mu,\sigma}(u)}{1 - W_{2,\mu,\sigma}(u)} = \frac{\left(-\frac{u-\mu}{\sigma}\right)^{\alpha} - \left(-\frac{u+x-\mu}{\sigma}\right)^{\alpha}}{\left(-\frac{u-\mu}{\sigma}\right)^{\alpha}}.$$

Úpravou tohto výrazu sa získame

$$F_u(x) = 1 - \left(-\frac{x - (\mu - u)}{\mu + u}\right)^{\alpha} = W_{2,\mu^*,\sigma^*}(x).$$

Tým overujeme platnosť stability voči excesom aj pre beta rozdelenie. V tomto prípade sa zmenil parameter polohy na  $\mu^* = \mu - u$  a parameter merítka na  $\sigma^* = \mu + u$ . Všetky rozdelenia sú preto stabilné voči excesom.

## 3.2 Metóda POT pri praktickom spracovaní dát

Zopakujme, že uvažujeme náhodný výber  $X_1, \dots, X_n$  s distribučnou funkciou  $F$ . Pri použití POT metódy pri modelovaní dát je našou úlohou najskôr odhadnúť parametre GP rozdelenia, ďalej si vhodne zvolíť prah  $u$  a definovať excesy. Potom môžeme skúmať rôzne vlastnosti v závislosti na požadovaných výsledkoch, čomu sú venované nasledujúce časti práce.

### 3.2.1 Odhad parametrov GP rozdelenia

Podľa vety 3.0.1 vieme, že pre dostatočne veľké  $u$  môžeme excesy  $Y_i = X_i - u$ ,  $X_i \geq u$ ,  $i = 1, \dots, N_u$  aproximovať zovšeobecneným Paretoovým rozdelením. Predpokladajme teda, že náhodný výber  $Y_1, \dots, Y_{N_u}$  je náhodný výber z GP rozdelenia s distribučnou funkciou  $W_{\gamma, \sigma}(x)$ . Parametre  $\gamma$  a  $\sigma$  však nepoznáme, a teda ich musíme odhadnúť z dát. V nasledujúcej časti sú popísané dve metódy zhodné s metódami použitými na odhad parametrov GEV rozdelenia vyššie, a to metóda maximálnej vierohodnosti a metóda vážených momentov.

### Metóda maximálnej vierohodnosti

Podstata tejto metódy je obdobná s prípadom z predchádzajúcej kapitoly, postupujeme preto analogicky. Hustotu GP rozdelenia získame deriváciou jeho distribučnej funkcie, a preto

$$w_{\gamma, \sigma}(x) = \frac{1}{\sigma} \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma} - 1}, \quad \gamma \in \mathbb{R}, \sigma > 0.$$

Logaritmus hustoty teda vyzerá nasledovne

$$\ln w_{\gamma, \sigma}(x) = -\ln \sigma - \left(\frac{1}{\gamma} + 1\right) \ln \left(1 + \frac{\gamma x}{\sigma}\right).$$

Potom logaritmická vierohodnostná funkcia bude v tvare

$$l(\gamma, \sigma; Y_1, \dots, Y_{N_u}) = -N_u \ln \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{N_u} \ln \left(1 + \frac{\gamma Y_i}{\sigma}\right), \quad (3.3)$$

pričom tento výraz maximalizujeme pre  $\sigma > 0$ ,  $\gamma \in \mathbb{R}$  a za podmienky

$$1 + \frac{\gamma Y_i}{\sigma} > 0, \quad i = 1, \dots, N_u.$$

Maximalizáciu je potrebné previesť priamo numericky, a príslušné maximálne vierohodné odhady  $\hat{\gamma}$  a  $\hat{\sigma}$  získame ako argument maxima logaritmickkej vierohodnostnej funkcie (3.3).

Poznamenajme, že pri odhadovaní parametrov nie sú splnené podmienky regularity. Avšak pre hodnoty  $\gamma > -0,5$  táto metóda funguje klasicky a odhady sú asymptoticky nestranné, konzistentné a majú asymptoticky normálne rozdelenie (pozri [8]).

## Metóda vážených momentov

Odhad parametrov GP rozdelenia pomocou metódy pravdepodobnostne vážených momentov prevedieme obdobne ako pri GEV rozdelení. Pripomeňme, že pravdepodobnostne vážené momenty sú definované ako

$$M_{p,r,s} = E \{ X^p [F(X)]^r [1 - F(X)]^s \},$$

kde  $F$  je distribučná funkcia náhodnej veličiny  $X$ . V prípade odhadu parametrov zovšeobecneného Paretovh rozdelenia odporúčajú Hosking a Wallis (pozri [13]) uvažovať vážené momenty v tvare

$$\alpha_s = M_{1,0,s} = E \{ X [1 - F(X)]^s \}, \quad s = 0, 1.$$

Toto môžeme podľa (2.13) vyjadriť ako

$$\alpha_s = \int_0^1 F^{-1}(1 - F)^s dF, \quad (3.4)$$

kde  $F^{-1}$  je kvantilová funkcia k distribučnej funkcii  $F$ . Pripomeňme, že GP rozdelenie má distribučnú funkciu

$$W_{\gamma,\sigma}(x) = 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}, \quad \gamma \in \mathbb{R}, \sigma > 0.$$

Potom kvantilová funkcia bude

$$W^{-1} = \frac{\sigma}{\gamma} [(1 - W)^{-\gamma} - 1].$$

Dosadením tejto funkcie do (3.4) získame

$$\alpha_s = \int_0^1 W^{-1}(1 - W)^s dW = \int_0^1 \frac{\sigma}{\gamma} [(1 - W)^{-\gamma} - 1] (1 - W)^s dW$$

V integráli zavedieme substitúciu  $t = 1 - W$ , čím získame riešenie nasledovne

$$\alpha_s = \frac{\sigma}{\gamma} \int_0^1 (t^{-\gamma} - 1) t^s dt = \frac{\sigma}{\gamma} \left( \frac{1}{s - \gamma + 1} - \frac{1}{s + 1} \right)$$

a teda

$$\alpha_s = \frac{\sigma}{(s - \gamma + 1)(s + 1)}.$$

Nakoľko chceme odhadnúť dva parametre, potrebujeme vyjadriť  $\alpha_0$  a  $\alpha_1$ , čiže

$$\alpha_0 = \frac{\sigma}{1 - \gamma},$$

$$\alpha_1 = \frac{\sigma}{2(2 - \gamma)}.$$

Úpravou týchto rovníc získame odhady  $\hat{\gamma}$  a  $\hat{\sigma}$  parametrov  $\gamma$  a  $\sigma$  v tvaroch

$$\hat{\gamma} = \frac{4\alpha_1 - \alpha_0}{2\alpha_1 - \alpha_0},$$

$$\hat{\sigma} = \frac{2\alpha_0\alpha_1}{\alpha_0 - 2\alpha_1},$$

pričom  $\alpha_s$  sa nahradí nejakým odhadom. Napríklad Landwehr navrhol v [16] ne-stranný odhad pre zoradený náhodný výber  $Y_{(1)}, \dots, Y_{(N_u)}$  v tvare

$$\hat{\alpha}_s = \frac{1}{n} \sum_{j=1}^{N_u} \left( \prod_{k=1}^s \frac{N_u - j - k + 1}{N_u - k} \right) Y_{(j)}.$$

Tieto odhady opäť nespĺňajú podmienky regularity. Avšak pre hodnotu parametru  $\gamma > -0,5$  je v [13] ukázané, že v takomto prípade majú odhady štandardné vlastnosti neutrannosti, konzistentnosti a majú taktiež asymptoticky normálne rozdelenie.

### 3.2.2 Spracovanie dát

Analogicky ako pri metóde blokových maxim opísanej v predchádzajúcej kapitole, sa teraz zamerajme na vybrané štatistické úlohy, ktoré nás môžu pri analýze dát zaujímať. Odhadneme pravdepodobnosť, že nastane extrémna udalosť a pozrieme sa na úroveň a dobu návratu extrémnej udalosti v prípade uvažovania POT metódy.

#### Pravdepodobnosť nastania extrémnej udalosti

Predstavme si, že je našou úlohou odhadnúť pravdepodobnosť, že si klient poisťovne bude nárokovať nejaké veľmi vysoké poistné plnenie. Túto otázku môžeme riešiť nasledovne. Nech  $X_1, \dots, X_n$  je náhodný výber z rozdelenia s distribučnou funkciou  $F$ . Zaujímá nás pravdepodobnosť

$$P(X_i > x), \quad \text{pre nejaké veľké } x.$$

Pravdepodobnosť v prípade POT metódy uvažujeme ako podmienenú a pre prípad, že došlo k prekročeniu medze  $u$ , t.j. ak  $x \geq u$ . Preto podľa definície podmienenej pravdepodobnosti máme

$$P(X_i > x | X_i \geq u) = \frac{P(X_i > x)}{P(X_i \geq u)}.$$

Odtiaľ vyjadríme a postupne upravujeme

$$\begin{aligned} P(X_i > x) &= P(X_i \geq u) \cdot P(X_i > x | X_i \geq u) = \\ &= P(X_i \geq u) \cdot (1 - P(X_i \leq x | X_i \geq u)), \end{aligned}$$

pričom zavedením excesu  $Y_i = X_i - u$ ,  $i = 1, \dots, N_u$ ,  $X_i \geq u$  získame

$$\begin{aligned} P(X_i > x) &= P(X_i \geq u) \cdot (1 - P(Y_i \leq x - u | X_i \geq u)) \\ &= P(X_i \geq u) \cdot (1 - F_u(x - u)), \end{aligned}$$

kde  $F_u(x - u)$  je podmienená distribučná funkcia excesov. Podľa vety 3.0.1 môžeme pre dostatočne veľkú hranicu  $u$  aproximovať  $F_u(x - u)$  pomocou GP rozdelenia s distribučnou funkciou  $W_{\gamma, \sigma}(x - u)$ . Preto

$$P(X_i > x) = P(X_i \geq u) \cdot \left(1 + \frac{\gamma(x - u)}{\sigma}\right)^{-\frac{1}{\gamma}}. \quad (3.5)$$

Potom parametre  $\gamma$  a  $\sigma$  nahradíme ich príslušnými odhadmi  $\hat{\gamma}$  a  $\hat{\sigma}$  získanými nejakou vhodnou metódou. Ďalej odhadnime  $P(X_i \leq u)$  výrazom  $N_u/n$ , kde  $N_u$  je počet pozorovaní, ktoré prekročili danú medzu  $u$  a  $n$  je počet všetkých pozorovaní. Celkovo získavame vzťah

$$P(\widehat{X_i \geq x}) = \frac{N_u}{n} \left(1 + \frac{\hat{\gamma}(x - u)}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\gamma}}}, \quad (3.6)$$

ktorý môžeme použiť na odhad extrémnych pravdepodobností.

## Odhad extrémnych kvantilov – úroveň a doba návratu

Obdobne ako v predchádzajúcej metóde, **úroveň návratu** je vlastne veľkosť extrémneho javu, ktorý bude priemerne nastávať s vopred zvolenou frekvenciou. Zmena však nastáva v chápaní plynutia frekvencie. Kým pri metóde blokových maxim sme ju merali v blokoch (t.j. typicky to je ročná, desaťročná frekvencia a pod.), pri POT metóde sa frekvencia meria v pozorovaniach, čo už častokrát nie je také intuitívne. V nasledujúcom texte budeme uvažovať úroveň návratu, ktorá bude prekročená extrémnym javom raz za  $k$  pozorovaní. Poznamenajme však, že ak by nás zaujímala úroveň návratu za  $K$  rokov, môžeme zaviesť substitúciu

$$k = K \cdot n_y,$$

kde  $n_y$  je počet pozorovaní za rok. Toto môžeme využiť v prípade, že  $n_y$  je známe a v jednotlivých rokoch sa príliš nelíši (pri veľkom počte pozorovaní môžeme malé rozdiely zanedbať).

Hľadáme teraz úroveň návratu  $q := q_{1-1/k}$ , ktorá bude  $i$ -tým pozorovaním prekročená raz za  $k$  pozorovaní ako  $(1 - 1/k)$ -kvantil, t.j. podľa definície

$$P(X_i > q) = \frac{1}{k}, \quad k \in \mathbb{N}.$$

Podľa vzťahu (3.5), ktorý sme odvodili vyššie máme

$$P(X_i \geq u) \cdot \left(1 + \frac{\gamma(q - u)}{\sigma}\right)^{-\frac{1}{\gamma}} = \frac{1}{k},$$

z čoho úpravami vyjadríme požadovaný kvantil

$$q = u + \frac{\sigma}{\gamma} [(k \cdot P(X_i \geq u))^\gamma - 1].$$

Potom príslušný odhad bude

$$\hat{q} = \hat{q}_{1-\frac{1}{k}} = u + \frac{\hat{\sigma}}{\hat{\gamma}} \left[ \left( \frac{k N_u}{n} \right)^\gamma - 1 \right], \quad (3.7)$$

kde  $\hat{\gamma}$ ,  $\hat{\sigma}$  sú vhodné odhady parametrov GP rozdelenia  $\gamma$ ,  $\sigma$  a  $N_u/n$  je odhad pre  $P(X_i \geq u)$ , pričom  $N_u$  je počet pozorovaní, ktoré prekročili medzu  $u$  a  $n$  je celkový počet pozorovaní.

Ďalšou charakteristikou, ktorá nás zaujíma, je **doba návratu** extrémnej udalosti. Je to vlastne priemerná frekvencia výskytu extrémneho javu pri známej úrovni návratu  $q = q_{1-1/k}$ , t.j. snažíme sa nejakým spôsobom odhadnúť  $k$ . Vychádzajme zo vzťahu pre  $(1 - 1/k)$ -kvantil rozdelenia

$$P(X_i > q) = \frac{1}{k}.$$

Potom

$$k = \frac{1}{P(X_i > q)} \stackrel{(3.5)}{=} \frac{1}{P(X_i \geq u) \left( 1 + \frac{\gamma(q-u)}{\sigma} \right)}.$$

Príslušný odhad  $\hat{k}$  je teda tvaru

$$\hat{k} = \left[ \frac{N_u}{n} \left( 1 + \frac{\hat{\gamma}(q-u)}{\hat{\sigma}} \right) \right]^{-1}, \quad (3.8)$$

pričom  $\hat{\gamma}$ ,  $\hat{\sigma}$  sú vhodné odhady  $\gamma$ ,  $\sigma$ ,  $N_u$  je počet prekročení nad medzu  $u$  a  $n$  je celkový počet pozorovaní.

## Voľba thresholdu $u$

Voľba thresholdu  $u$  je jednou z najkritickejších úloh POT analýzy. Ak zvolíme  $u$  príliš vysoko, získame model s málo excesmi a tým pádom s veľkým rozptylom. Naopak, pri nízko zvolenej hranici  $u$  vzniká vychýlený model. Je preto potrebné nájsť nejaký kompromis.

K nájdeniu hranice  $u$  sa používajú viaceré metódy. Jednou z nich je grafická metóda založená na linearite strednej hodnoty excesov pre GP rozdelenie, ktorú si najskôr definujeme.

**Definícia 3.2.1.** Stredná hodnota excesov (*mean excess function*) je funkcia definovaná ako

$$e(u) = E(X - u | X > u), \quad u \in \mathbb{R},$$

kde  $u$  je zvolená hranica prekročenia.

Poznamenajme, že ak má náhodná veličina  $Z$  zovšeobecnené Paretovo rozdelenie s parametrami  $\gamma$  a  $\sigma$ , potom stredná hodnota tejto veličiny je

$$EZ = \frac{\sigma}{1 - \gamma} \quad \gamma < 1.$$

Ďalej sme v kapitole 3.1 odvodili, že excesy  $Y_i = X_i - u$  za podmienky  $X_i \geq u$ , majú zovšeobecnené Paretovo rozdelenie s parametrami  $\gamma$  a  $\sigma^* = \sigma + \gamma u$ . Skombinovaním týchto dvoch informácií dostaneme pre strednú hodnotu excesov GP rozdelenia vzťah

$$e(u) = \frac{\sigma + \gamma u}{1 - \gamma},$$

čo je lineárna funkcia premennej  $u$ . Toto je charakteristická vlastnosť GP rozdelenia, v prípade uvažovania iného rozdelenia by sme získali zložitejšie vyjadrenie.

V praxi, pre náhodný výber  $X_1, \dots, X_n$  sa funkcia  $e(u)$  môže odhadnúť výberovým priemerom excesov, t.j.

$$\hat{e}(u) = \frac{1}{N_u} \sum_{i=1}^{N_u} Y_i,$$

kde  $Y_i$  je exces, ak k nemu došlo a  $N_u$  je počet prekročení hranice  $u$ . Následne vynesieme do grafu dvojice bodov

$$\{ [X_{(i)}, e(X_{(i)})], i = 1, \dots, n \},$$

čím vznikne tzv. *mean excess plot* a  $X_{(1)}, \dots, X_{(n)}$  je pritom usporiadaný náhodný výber. Pokiaľ pozorovania pochádzajú z GP rozdelenia, potom by sa mal graf po určitom čase stabilizovať a správať sa približne lineárne. Potom hodnotu thresholdu  $u$  vyberieme na základe toho, od ktorého pozorovania sa graf javí lineárny. Nevýhodou je subjektivita jednotlivca pri výbere hranice touto optickou metódou, nakoľko interpretácia tohto grafu je častokrát náročná. Ďalšou prekážkou môže byť fakt, že stabilizácia grafu v skutočnosti častokrát vôbec nenastane, nakoľko sa rozdelenie excesov nepodarí dostatočne aproximovať GP rozdelením, čo robí úlohu určenia hranice  $u$  ešte komplikovanejšou.

Niektoré ďalšie metódy určenia thresholdu  $u$  boli navrhnuté, pozri [20]. Jednou z možností je určiť hranicu prekročenia  $u$  ako nejaký vysoký kvantil, často sa používa 90 % alebo 96 % kvantil. Ďalšou alternatívou je označiť ako  $u$  hodnotu pozorovania  $X_{(n-j+1)}$ , zoradeného náhodného výberu  $X_{(1)} \leq \dots \leq X_{(n)}$ , pričom  $j = \sqrt{n}$  alebo  $j = n^{2/3} \ln(\ln(n))$ . Pri praktickom spracovaní dát je vhodné zvoliť viacero hodnôt pre threshold  $u$  a výsledné modely vzájomne porovnať.

# Kapitola 4

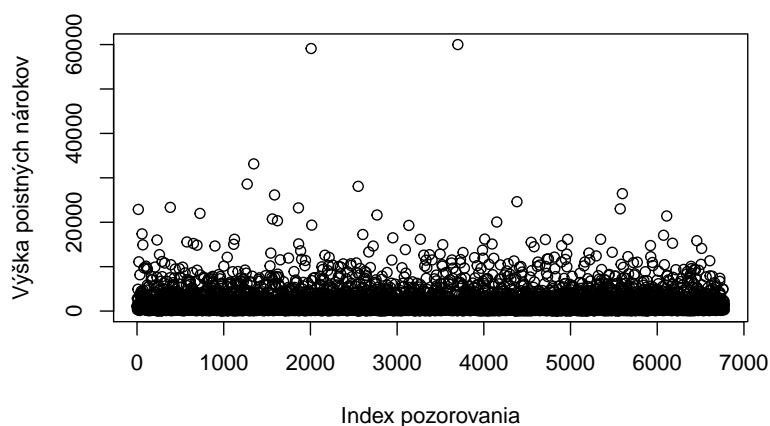
## Analýza dát

Analýza extrémnych udalostí sa uplatňuje v mnohých oblastiach. V nasledujúcej kapitole ju aplikujeme na dáta z oblasti poisťovníctva. Naším cieľom bude na základe známych historických dát poskytnúť odhad extrémnej udalosti a iné vyššie zmieňované charakteristiky. Odpovede na tieto otázky sú v tejto oblasti dôležité, nakoľko pomáhajú poisťným spoločnostiam vykonať rozhodnutie napríklad o alokácii kapitálu. Na záver budú výsledky rôznych metód porovnané.

Dáta boli modelované v jazyku R , pričom zdrojové kódy je možné nájsť v elektronickom archíve tejto práce.

### 4.1 Dáta

Dáta sú tvorené informáciami o poisťných nárokoch americkej poisťnej spoločnosti z oblasti súkromného poistenia automobilov [9]. Premenná, ktorá je ďalej v práci využitá obsahuje údaje o veľkosti čiastky, ktorá bola spoločnosťou poistenému vyplatená. Priebeh tejto premennej obsahujúcej 6 773 pozorovaní je zobrazený na obr. 4.1.



Obr. 4.1: Výška poisťných nárokov.



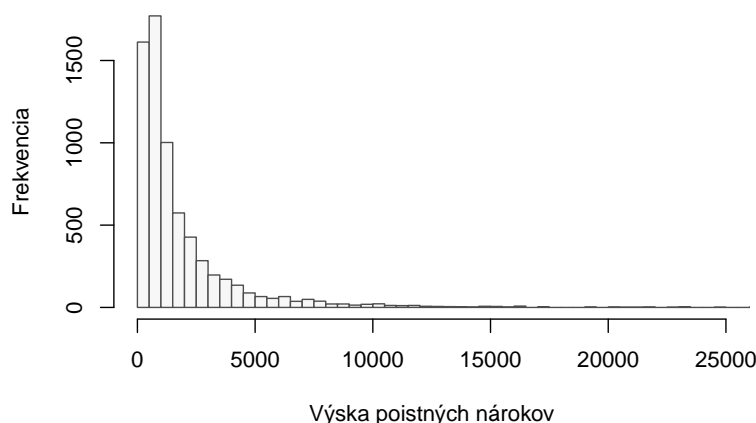
Minimálna vyplatená čiastka je 9,5 \$, maximálna 60 000 \$. Ďalšie základné charakteristiky sú zhrnuté v tabuľke 4.1.

Priemer	1 853	Rozptyl	7 006 129
Medián	1 002	Šmerodajná odchýlka	2 649
1. kvartil	524	Šikmosť	6,23
3. kvartil	2 137	Špicatosť	84,25

Tabuľka 4.1: Základné charakteristiky dátového súboru.

Šikmosť je kladná, preto očakávame dáta zošikmené doprava a hodnota špicatosti naznačuje ťažké chvosty rozdelenia dát. Tieto vlastnosti vizuálne potvrdzuje aj histogram dát na obr. 4.2 (s hodnotami vyššími ako 25 000 odseknutými).

**Histogram poistných nárokov**

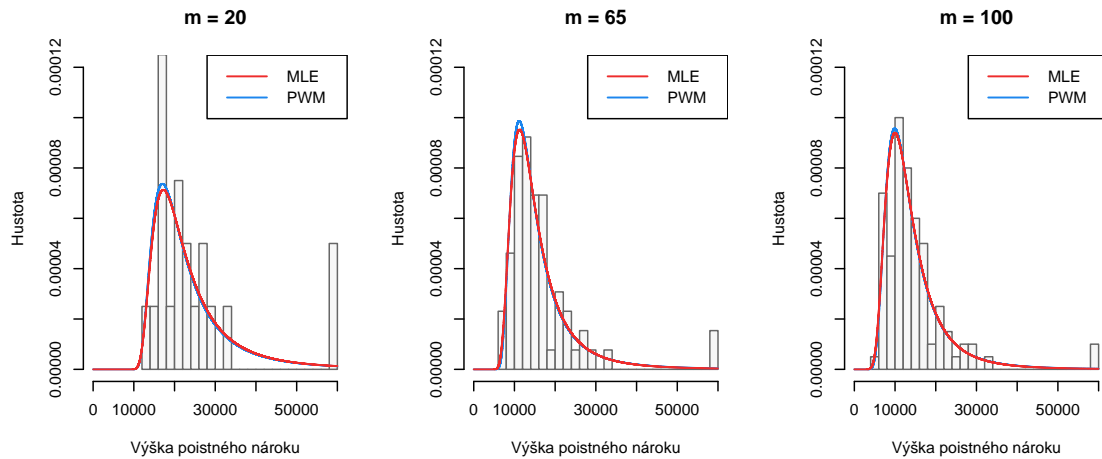


Obr. 4.2: Histogram výšky poistných nárokov.

Vidíme, že dataset je výrazne zošikmený doprava s ťažkými chvostami a tým pádom by modelovanie pomocou normálneho rozdelenia nebolo vhodné. Využijeme preto metódy na modelovanie extrémnych situácií popísané v predchádzajúcich kapitolách práce.

## 4.2 Modelovanie metódou blokových maxím

Kľúčovým začiatočným rozhodnutím je výber počtu blokov, ktorý bude pre náš model použitý. Vzhľadom na to, že neexistuje nejaká zaužívaná metóda pre túto voľbu, prevedieme ju porovnaním vlastností modelov s rôznymi počtami blokov. Predovšetkým sa upriamime na to, ako dobre je výber blokových maxím aproximovaný GEV rozdelením na základe grafických nástrojov. Následne sa môžeme zamerať na odhad vybraných charakteristík.



Obr. 4.3: Histogram blokových maxim pre počet blokov  $m = 20, 65$  a  $100$ . Červená krivka je hustota odhadnutá na základe maximálne vierohodných odhadov a modrá krivka je odhadnutá hustota na základe pravdepodobnostne vážených momentov.

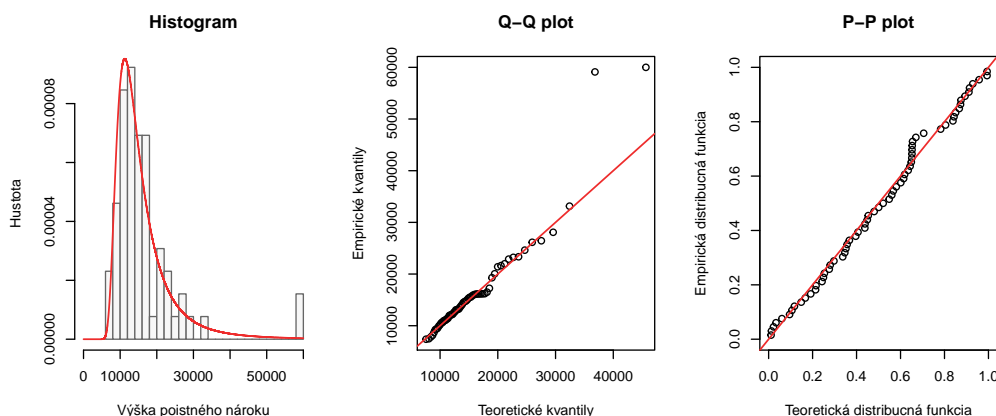
## Výber počtu blokov

Modely boli porovnávané okrem iného podľa histogramov pre rôzne počty blokových maxim. Aj na základe tejto metódy sme dospeli k rozhodnutiu, že najvhodnejšie bude zvoliť počet blokov  $m = 65$ , čím sa vytvárajú bloky s veľkosťou  $n = 104$  (resp.  $105$ ) pozorovaní. Pre ilustráciu môžeme na obr. 4.3 pozorovať histogramy pre dáta rozdelené do 20, 65 a 100 blokov. Zároveň sú v grafoch znázornené odhadnuté hustoty zovšeobecneného rozdelenia extrémnych hodnôt - červenou krivkou vzhľadom na parametre odhadnuté maximálnou vierohodnosťou (MLE) a modrou pravdepodobnostne váženými momentami (PWM).

Tak isto boli porovnané  $p$ -hodnoty testu zhody empirického rozdelenia blokových maxim s GEV rozdelením s neznámymi parametrami. Využitá bola technika založená na výberovom korelačnom koeficiente, ktorá bola navrhnutá pre rozdelenia stabilné voči maximám (kam náš prípad spadá) v publikácii [11]. Výsledky sú zhrnuté v tabuľke 4.2. Podľa tohto testu na hladine významnosti 5% nezamietame nulovú hypotézu, že rozdelenie maxim pre počet blokov 20 a 65 pochádza z GEV rozdelenia. Naopak, pre 100 blokových maxim túto hypotézu zamietame.

$m$	$p$ -hodnota
20	0,156
65	0,188
100	0,029

Tabuľka 4.2:  $P$ -hodnoty prislúchajúce testu zhody s GEV rozdelením pre rôzny počet blokov  $m$ .



Obr. 4.4: Histogram, Q-Q plot a P-P plot pre maximálne vierohodné odhady parametrov GEV rozdelenia.

## Odhad parametrov GEV rozdelenia

Odhady parametrov pre výber blokových maxím  $M_i$ ,  $i = 1, \dots, 65$  boli prevedené metódou maximálnej vierohodnosti a metódou pravdepodobnostne vážených momentov. Pre nami vybraný počet blokov sú ich hodnoty zhrnuté v tabuľke 4.3, pričom sa oba prípady od seba vzájomne markantne nelíšia.

	$\hat{\gamma}$	$\hat{\mu}$	$\hat{\sigma}$
MLE	0,30	12 269	4 023
PWM	0,33	12 239	3 919

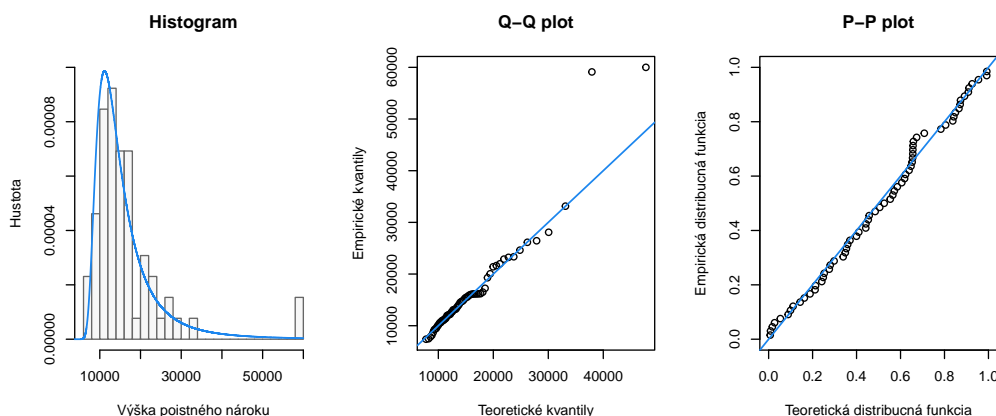
Tabuľka 4.3: Odhady parametrov GEV rozdelenia metódou maximálnej vierohodnosti (MLE) a metódou pravdepodobnostne vážených momentov (PWM).

Pozrime sa aj na grafické metódy pre posúdenie vhodnosti vybraného modelu, ktoré sú založené na porovnávaní teoretického a empirického rozdelenia. Q-Q plot porovnáva teoretické a empirické kvantily rozdelenia a P-P plot teoretickú a empirickú distribučnú funkciu. Pretože sú tieto diagnostické grafy približne lineárne, môžeme povedať, že GEV rozdelenie aproximuje naše dáta relatívne dobre. Na obr. 4.4 vidíme tieto grafy pre maximálne vierohodné odhady a na obr. 4.5 pre pravdepodobnostne vážené momentové odhady.

V oboch prípadoch je približná linearita relatívne dobre zachovaná, menšiu komplikáciu ale predstavujú dve pozorovania, ktoré sú odľahlejšie ako ostatné, čo je badateľné napr. z Q-Q plotu. To môže budúce odhady pravdepodobností do istej miery skresliť.

## Extrémne pravdepodobnosti

V predchádzajúcej časti práce sme odvodili, aký je odhad pravdepodobnosti, že nasledujúce pozorovanie bude mať vyššiu hodnotu ako nejaké veľké  $x$ , čo odhadujeme pomocou vzťahu (2.17). V našom prípade to znamená pravdepodobnosť, že nasledujúci zákazník, ktorý do poisťovne príde, bude požadovať poistný nárok vyšší ako



Obr. 4.5: Histogram, Q-Q plot a P-P plot pre parametrov GEV rozdelenia pomocou pravdepodobnostne vážených momentov.

$x$  dolárov. Tieto pravdepodobnosti sú v prípade skúmaných dát relatívne nízke pre vysoké čiastky poistného plnenia, napr. pravdepodobnosť, že nasledujúci zákazník si bude nárokovať viac ako 35 000 \$ odhadujeme približne na 0,0349 %. Tieto charakteristiky teda nepatria medzi najužitočnejšie.

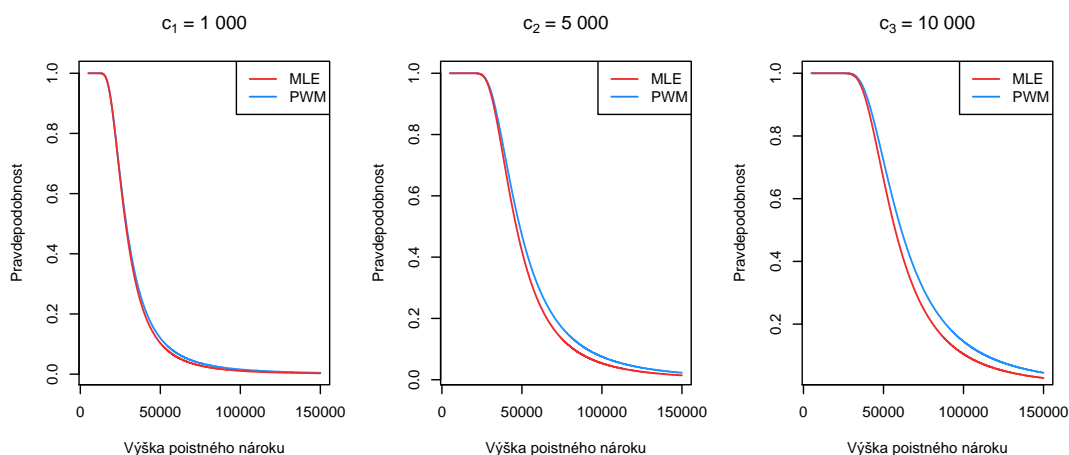
Preto je napríklad zaujímavejšie pozorovať, aká je pravdepodobnosť, že medzi nasledujúcimi  $c$  zákazníkmi príde aspoň jeden zákazník, ktorý požaduje poistný nárok vyšší ako  $x$  \$. To odhadneme nasledovnou úvahou: pravdepodobnosť, že aspoň jeden zákazník bude mať nárok vyšší ako  $x$  je doplnok k pravdepodobnosti, že žiadny z  $c$  zákazníkov nebude mať nárok vyšší ako  $x$ , čo sa dá vyjadriť ako

$$1 - P(X_1 \leq x) \cdot \dots \cdot P(X_c \leq x) = 1 - [1 - P(X_i > x)]^c,$$

nakoľko v našom modeli predpokladáme, že  $X_1, \dots, X_c$  je postupnosť i.i.d. náhodných veličín. Pravdepodobnosť  $P(X_i > x)$  už potom odhadneme známym vzťahom (2.17) odvodeným v teoretickej časti práce.

Na obrázku 4.6 vidíme odhady týchto pravdepodobností pre celú škálu výšky poistných nárokov a rôzne počty uvažovaných novoprichádzajúcich klientov  $c_i$ , ktorí majú tieto poistné nároky. Dve krivky reprezentujú výsledok výpočtu za použitia maximálne vierohodných odhadov GEV rozdelenia a odhadov získaných pomocou pravdepodobnostne vážených momentov. V prvom prípade pre  $c_1 = 1000$  klientov sú pritom takmer na nerozoznanie a relatívne presne sa kopírujú. V ostatných príkladoch pre počty klientov  $c_2 = 5000$  a  $c_3 = 10000$  rozdiely mierne rastú. Na grafoch taktiež pozorujeme, že čím vyšší počet klientov  $c$  uvažujeme, tým je krivka reprezentujúca odhad pravdepodobnosti menej „strmá“ a viac posunutá doprava, teda odhadovaná pravdepodobnosť sa s vyšším počtom klientov taktiež zvyšuje. Je potrebné poznamenať, že pre nízke poistné nároky treba brať dané grafy s rezervou, je dobré však vidieť aspoň približný priebeh. Metódy modelovania extrémnych udalostí sú navrhnuté pre vysoké  $x$ , t.j. vysoké hodnoty požadovaných poistných nárokov a tým pádom v opačnom prípade dobre nefungujú. Na otázku, čo už sú extrémne poistné plnenia nie je jednoznačná odpoveď, no vzhľadom na povahu dát by mohli byť

odhady relevantné približne od čiastky 40 000 dolárov.



Obr. 4.6: Odhad extrémnych pravdepodobností - pravdepodobnosť, že medzi nasledujúcimi  $c$  klientmi bude aspoň jeden s poistným nárokom vyšším ako  $x$  dolárov. MLE - maximálne vierohodné odhady, PWM - odhady metódou pravdepodobnostne vážených momentov pre parametre GEV rozdelenia.

V tabuľke 4.4 vidíme konkrétny príklad odhadnutých pravdepodobností. Označme pritom  $\hat{p}_{c_i}$ ,  $i = 1, 2, 3$  ako odhad pravdepodobnosti, že medzi nasledujúcimi  $c_i$  zákazníkmi sa objaví aspoň jeden s nárokom vyšším ako  $x$  \$. Počty zákazníkov  $c_i$  sú pritom opäť  $c_1 = 1\,000$ ,  $c_2 = 5\,000$  a  $c_3 = 10\,000$ . Aj číselne sa potvrdili zvyšujúce sa rozdiely odhadov pravdepodobností použitím MLE a PWM metódy pre rastúci počet uvažovaných prichádzajúcich zákazníkov. Interpretácia napríklad druhého riadku tabuľky (a) je, že pravdepodobnosť, že medzi nasledujúcimi 1 000 zákazníkmi sa objaví aspoň jeden klient s nárokom vyšším ako 60 000 \$ je 6 %, že sa objaví medzi nasledujúcimi 5 000 zákazníkmi je 26 % a medzi nasledujúcimi 10 000 zákazníkmi je 45 %, všetko pri použití maximálne vierohodných odhadov GEV rozdelenia. Teda potvrdilo sa zrejme, čím vyšší počet prichádzajúcich zákazníkov budeme uvažovať, tým vyššia je pravdepodobnosť, že sa medzi nimi objaví aspoň nejaký s požiadavkou na vysoké poistné plnenie.

$x$	$\hat{p}_{c_1}$	$\hat{p}_{c_2}$	$\hat{p}_{c_3}$
40 000	0,20	0,68	0,90
60 000	0,06	0,26	0,45
80 000	0,02	0,11	0,21
100 000	0,01	0,05	0,10

(a)

$x$	$\hat{p}_{c_1}$	$\hat{p}_{c_2}$	$\hat{p}_{c_3}$
40 000	0,22	0,72	0,92
60 000	0,07	0,31	0,52
80 000	0,03	0,14	0,26
100 000	0,02	0,08	0,14

(b)

Tabuľka 4.4: Odhad extrémnych pravdepodobností (a) maximálne vierohodnými odhadmi (b) odhadmi pomocou pravdepodobnostne vážených momentov GEV rozdelenia.

## Úroveň a doba návratu

V prípade dát o poistných nárokoch je zaujímavé preskúmať, aká veľkosť poistného nároku bude prekonaná v priemere jedenkrát za  $c$  zákazníkov. Inak povedané, hľadáme úroveň návratu, ktorá bude priemerne nastávať so zvolenou frekvenciou. Matematicky teda odhadujeme extrémny kvantil GEV rozdelenia podľa vzťahu (2.19). Jediná odlišnosť tkvie v tom, že kým sme v predchádzajúcom prípade počítali s periódou  $k$  ekvivalentnou dĺžke jedného bloku, tentokrát sa zameriame na frekvenciu v podobe počtu zákazníkov. Preto  $k$  vyjadríme ako

$$k = \frac{c}{n},$$

kde  $c$  je počet zákazníkov a  $n$  je počet pozorovaní v jednom bloku.

V tabuľke 4.5 sú odhadnuté úrovne návratu  $q$  pre rôzny počet zákazníkov pre obe metódy odhadov parametrov rozdelenia, ktoré sú popísané vyššie. Vidíme napríklad, že priemerne raz za 10 000 klientov sa objaví taký, ktorý si bude od poisťovne nárokovat poistné plnenie vyššie ako 51 239, resp. 54 094 dolárov. Ako je zrejmé, úrovne návratu s rastúcim počtom zákazníkov taktiež stúpajú. Odhad úrovne návratu pomocou pravdepodobnostne vážených momentov je pritom vyšší ako maximálne vierohodný, pričom ich rozdiel navzájom sa prehľbuje. Kým pre 5 000 zákazníkov sa odhadnuté úrovne líšia o približne 4%, pre 75 000 zákazníkov sú odlišné o 10%.

$c$	$\hat{q}_{MLE}$	$\hat{q}_{PWM}$
5 000	41 396	42 989
10 000	51 239	54 094
20 000	63 304	68 038
35 000	75 016	81 871
50 000	83 560	92 122
75 000	94 436	105 345

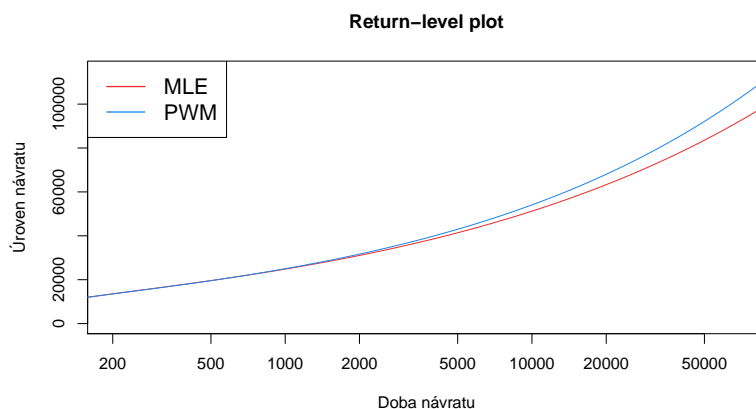
Tabuľka 4.5: Odhad úrovne návratu pre rôzny počet zákazníkov  $c$ .

Možnosťou takisto je, že poisťovňa bude naopak viac sústredená na priemernú frekvenciu nejakej hodnoty poistného plnenia. V tomto prípade sa preto zaujímame o dobu návratu extrémnej udalosti, opäť vyjadrenú pomocou počtu zákazníkov. Doby návratu pre rôzne výšky poistného plnenia  $q$  a pre obe metódy odhadov parametrov GEV rozdelenia sú vypísané v tabuľke 4.6. Môžeme napríklad pozorovať, že poistné plnenie vyššie ako 60 000 \$ sa vyskytne priemerne raz za 43 292, resp. 32 637 zákazníkov.

V odhadoch oboch charakteristík – úrovne aj doby návratu – môžeme sledovať, že GEV rozdelenie odhadnuté metódou pravdepodobnostne vážených momentov priraduje extrémnym udalostiam vyššiu váhu v porovnaní s maximálne vierohodne odhadnutým GEV rozdelením, pričom rozdiely sa stávajú pomerne vysoké s rastúcou hodnotou požadovaného poistného nároku. Inak povedané, prvé z menovaných má ťažšie chvosty voči druhému. Tento vzťah môžeme vidieť aj na obr. 4.7, pričom tento graf sa nazýva return-level plot. Zobrazuje závislosť medzi úrovňou a dobou návratu, ktorá je uvažovaná pomocou zákazníkov.

$q$	$\hat{c}_{MLE}$	$\hat{c}_{PWM}$
40 000	4 476	4 025
60 000	16 770	13 676
80 000	43 292	32 637
100 000	90 690	64 078

Tabuľka 4.6: Odhad doby návratu pre rôzne veľkosti poistného plnenia  $q$ .



Obr. 4.7: Return-level plot, kde MLE značí maximálne vierohodné odhady a PWM odhady pomocou pravdepodobnostne vážených momentov GEV rozdelenia.

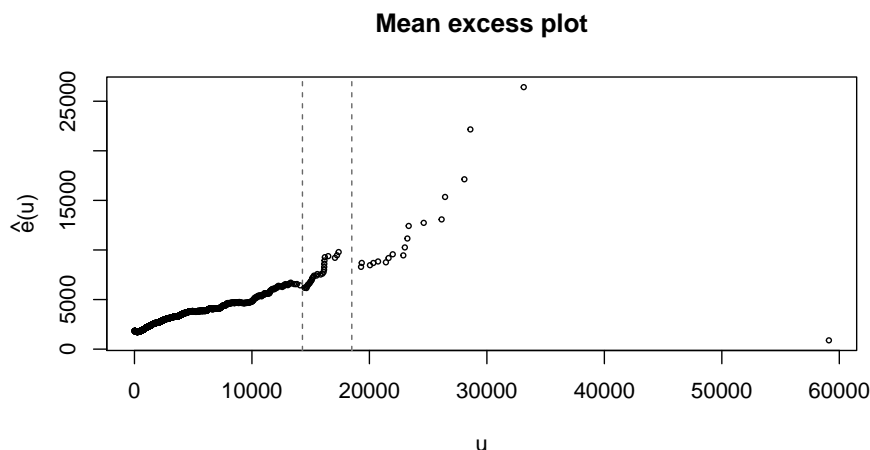
### 4.3 Modelovanie POT metódou

Druhým prístupom je modelovanie dát pomocou metódy prekročenia nad určitú medzu. Podobne, ako bolo v predchádzajúcom prípade kritickou úlohou vybrať vhodný počet blokov, do ktorých pozorovania rozdelíme, teraz je našim problémom nájsť vhodnú medzu. Po vyriešení tohto problému budú odhadnuté parametre GP rozdelenia a prevedená podobná analýza ako v prípade modelovania pomocou blokových maxim.

#### Voľba thresholdu

V kapitole 3 boli navrhnuté viaceré prístupy, ako sa k dopracovať k vhodnej voľbe hranice prekročenia  $u$ . Jedným z nich bolo vybrať takú hodnotu, od ktorej sa graf – tzv. mean excess plot – stabilizuje. Tento graf znázorňujúci závislosť strednej hodnoty excesov na výške thresholdu vidíme na obr. 4.8. Nejaví však očividné známky stabilizácie, ako sa teoreticky predpokladá. Napriek tomu sme zvolili dve veľkosti tejto medze, ktoré boli ďalej preskúmané, pričom vybrané hodnoty sa nachádzajú na horizontálnej osi a sú znázornené prerušovanou čiarou. Výška thresholdu odpovedajúca zvoleným hodnotám je 14 300, resp. 18 500 dolárov.

Takisto skúmané boli hranice zvolené na základe spomenutých heuristických odhadov. Modely boli opäť vyhodnocované najmä na základne grafických metód. Porovnanie histogramov s vykreslenou hustotou GP rozdelenia v troch prípadoch môžeme



Obr. 4.8: Mean excess plot.

pozorovať na obr. 4.9. Vidíme, že MLE a PWM odhady sú takmer identické a od seba nerozoznateľné vo všetkých vybraných možnostiach. V tabuľke 4.7 sa potom nachádzajú počty excesov prislúchajúce výškam hranice  $u$ , ktoré sú na obrázku zobrazené. Vybrané thresholdy odpovedajú (zľava) hodnote zvolenej na základe mean excess plotu, 96 % kvantilu a pozorovaniu  $X_{(n-j+1)}$ , kde  $j = n^{2/3} \ln(\ln(n))$ , pričom predpokladáme, že vstupné dáta tvoria náhodný výber  $X_1, \dots, X_n$ .

$u$	14 300	7 210	3 875
$N_u$	47	271	780

Tabuľka 4.7: Počet excesov  $N_u$  pre rôzne výšky hranice prekročenia  $u$ .

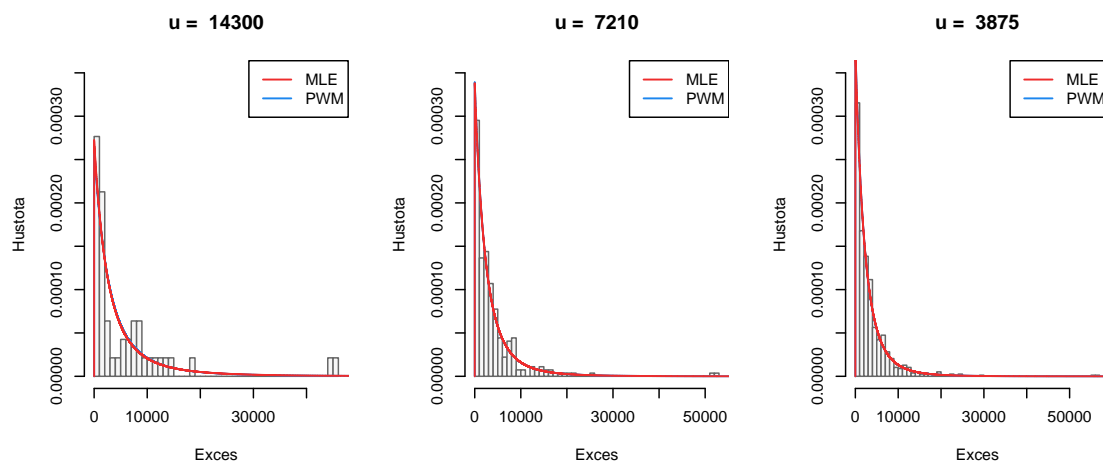
Kombináciou všetkých faktorov bola zvolená výška thresholdu ako  $u = 7\,210$ , čím získavame 271 excesov. Nakoniec bol prevedený test zhody empirického rozdelenia dát a zovšeobecneného Paretoho rozdelenia. Vybraný test bol navrhnutý v [23], pričom testovanie hypotézy prebieha metódou bootstrap. Výsledná  $p$ -hodnota je rovná 0,25, a preto na hladine významnosti 5 % nezamietame hypotézu, že výber excesov pochádza z GP rozdelenia.

## Odhad parametrov GP rozdelenia

Predpokladajme, že prekročenia  $Y_i$ ,  $i = 1, \dots, N_u$ , získané voľbou hranice  $u = 7\,210$  tvoria náhodný výber zo zovšeobecneného Paretoho rozdelenia. Odhady jeho parametrov  $\hat{\gamma}$  a  $\hat{\sigma}$  metódou maximálnej vierohodnosti a pravdepodobnostne váženými momentami sú zhrnuté v tabuľke 4.8. Obe metódy produkujú veľmi blízke odhady, dokonca odhady parametru  $\gamma$  sa v tomto prípade líšia až na vzdialenejších desiatiných pozíciách.

Vizuálne metódy posúdenia vhodnosti modelu pre maximálne vierohodné odhady GEV rozdelenia sú zobrazené na obrázku 4.10. Vidíme, že GP rozdelenie relatívne





Obr. 4.9: Histogram excesov v porovnaní s hustotou GP rozdelenia. Parametre GP rozdelenia boli odhadnuté maximálne vierohodnostnou (MLE) metódou a pravdepodobnostne váženými momentami (PWM).

	$\hat{\gamma}$	$\hat{\sigma}$
MLE	0,279	2 960
PWM	0,281	2 950

Tabuľka 4.8: Odhad parametrov GP rozdelenia pomocou metódy maximálnej vierohodnosti a pomocou pravdepodobnostne vážených momentov.

dobré aproximuje výber excesov, slabinou môžu byť dve odľahlé pozorovanie, ktoré sú zreteľné na Q-Q plote. Z dôvodu veľkej podobnosti odhadov parametrov pri použití oboch metód sú obdobné diagnostické grafy pre metódu pravdepodobnostne vážených momentov vynechané.

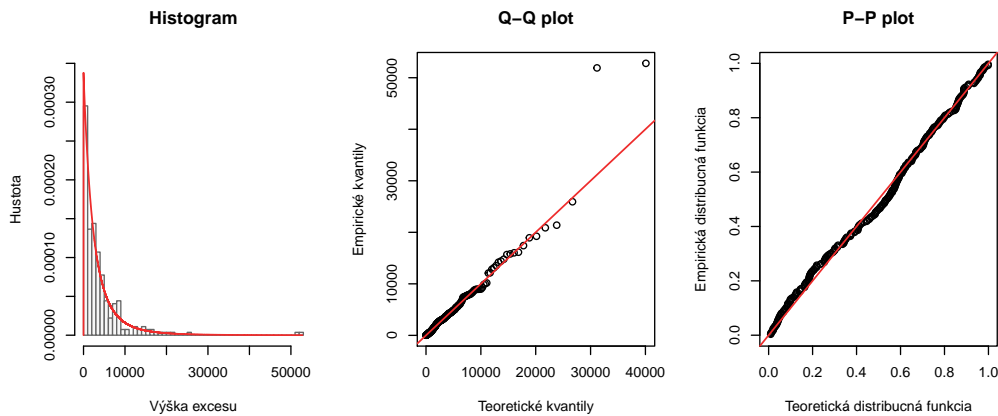
## Extrémne pravdepodobnosti

Ak by poisťovňa zaujímala pravdepodobnosť, že nasledujúci zákazník, ktorý bude mať voči nej poisťný nárok, bude požadovať čiastku vyššiu ako  $x$  dolárov, mohla by ju odhadnúť pomocou vzťahu (3.6). Konkrétnejšie, pravdepodobnosť, že nasledujúci klient bude požadovať poisťné plnenie vyššie ako 35 000 \$ by sa podľa neho odhadla na 0,0398%. Pri metóde blokových maxim sme však odhadovali pravdepodobnosti, že medzi nasledujúcimi prichádzajúcimi  $c$  zákazníkmi sa objaví aspoň jeden klient s poisťným nárokom vyšším ako nejaké vysoké  $x$ . Analogicky riešime túto úlohu metódou POT.

Obdobne, túto pravdepodobnosť vyjadríme ako

$$1 - (1 - P(X_i > x))^c$$

kde  $P(X_i > x)$  sa odhadne pomocou vzťahu (3.6). Vývoj tejto veličiny v závislosti na výške extrémneho poisťného nároku je zobrazený na obr. 4.11, pričom uvažujeme



Obr. 4.10: Histogram, Q-Q plot a P-P pre maximálne vierohodné odhady parametrov GP rozdelenia.

rôzny počet prichádzajúcich zákazníkov  $c_i$ ,  $i = 1, 2, 3$ . Zároveň znova upozorňujeme, že odhady pravdepodobností pre malé poistné čiastky sú skôr orientačné, nakoľko metóda na tieto situácie nie je stavaná, pretože uvažuje len extrémne udalosti. Tabuľka 4.9 potom ukazuje niektoré vybrané odhady pravdepodobností pre rôzne poistné nároky. Napr. pravdepodobnosť, že medzi nasledujúcimi  $c_2 = 5\,000$  zákazníkmi príde aspoň jeden zákazník s nárokom vyšším než 60 000 dolárov je odhadnutá ako 0,28 pre maximálne vierohodné odhady GP rozdelenia. Odhady pravdepodobností za použitia MLE a PWM metódy sa líšia úplne minimálne.

$x$	$\hat{p}_{c_1}$	$\hat{p}_{c_2}$	$\hat{p}_{c_3}$
40 000	0,23	0,72	0,92
60 000	0,06	0,28	0,48
80 000	0,02	0,12	0,22
100 000	0,01	0,06	0,11

(a)

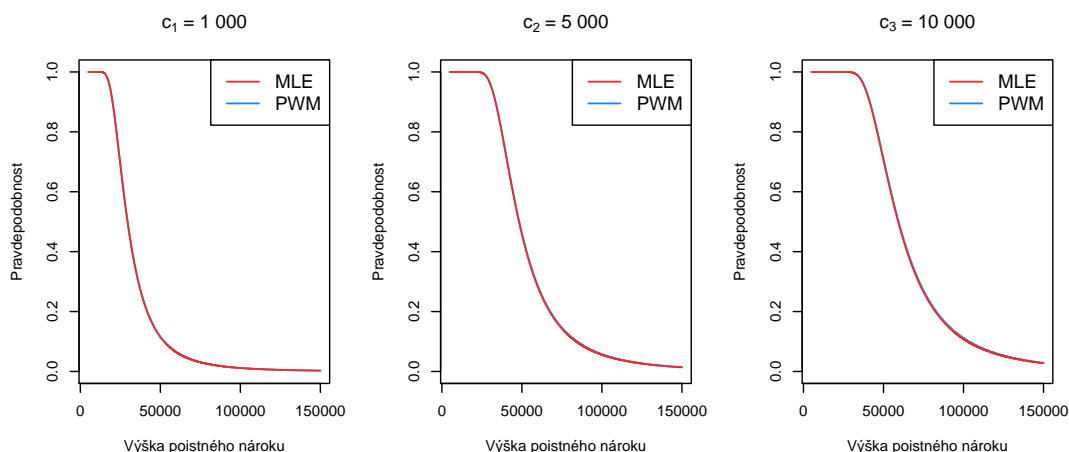
$x$	$\hat{p}_{c_1}$	$\hat{p}_{c_2}$	$\hat{p}_{c_3}$
40 000	0,23	0,73	0,93
60 000	0,07	0,29	0,49
80 000	0,03	0,12	0,22
100 000	0,01	0,06	0,11

(b)

Tabuľka 4.9: Odhad extrémnych pravdepodobností (a) maximálne vierohodnými odhadmi (b) odhadmi metódou pravdepodobnostne vážených momentov GP rozdelenia.

## Úroveň a doba návratu

Zaoberajme sa teraz opäť rovnakými úlohami ako pri predchádzajúcej metóde blokových maxim. Hľadáme preto najskôr úroveň návratu poistného nároku, ktorý bude v priemere nastávať s danou frekvenciou. Výhodou oproti prvej metóde je fakt, že kým pri blokových maximách bola chápaná frekvencia v rámci blokov a bolo treba zaviesť substitúciu, aby sme mohli úroveň návratu vyjadriť pomocou zákazníkov, teraz nič podobné nie je potrebné. Frekvencia je chápaná rovno v pozorovaniach, t.j. v zákazníkoch, a môžeme preto na odhad priamo využiť vzťah (3.7). Výsledné



Obr. 4.11: Odhad extrémnych pravdepodobností - pravdepodobnosť, že medzi nasledujúcimi  $c$  klientmi bude aspoň jeden s poistným nárokom vyšším ako  $x$  dolárov. MLE - maximálne vierohodné odhady, PWM - pravdepodobnostne vážené odhady GP rozdelenia.

odhady úrovne návratu  $q$  pre rôzny počet zákazníkov  $c$  sú viditeľné v tabuľke 4.10. Rozdiely medzi použitím MLE a PWM odhadov nie sú markantné.

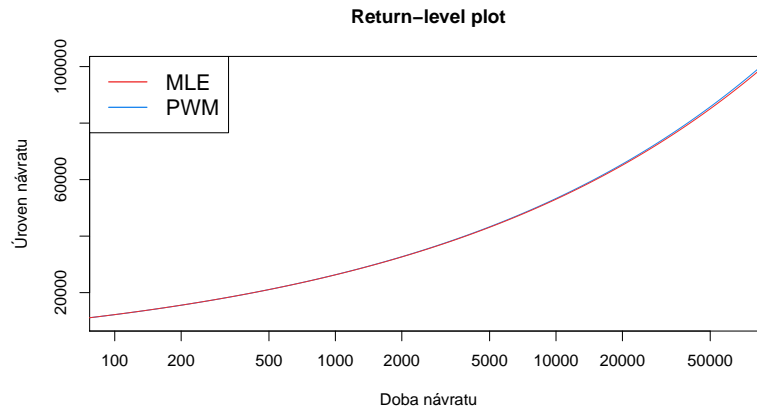
$c$	$\hat{q}_{MLE}$	$\hat{q}_{PWM}$
5 000	43 133	43 294
10 000	53 062	53 325
20 000	65 110	65 518
35 000	76 688	77 253
50 000	85 068	85 756
75 000	95 665	96 519

Tabuľka 4.10: Odhad úrovne návratu pre rôzny počet zákazníkov  $c$ .

Druhým pohľadom je snaha odhadnúť dobu návratu  $k$ . Využijeme na to opäť vzťah (3.8) odvodený v predchádzajúcich kapitolách. Pýtame sa, priemerne raz za koľko zákazníkov sa vyskytne poistné plnenie vyššie ako  $q$  dolárov? Súhrnné odhady pozorujeme v tabuľke 4.11. To znamená, že do poisťovne by mal prísť zákazník s poistným nárokom vyšším ako napr. 80 000 dolárov v priemere jedenkrát za približne 40 000 zákazníkov.

Vzťah medzi úrovňou a dobou návratu opäť pekne ilustruje return-level plot na obr. 4.12. Rozdiel v týchto odhadoch použitím MLE a PWM odhadov GP rozdelenia je nepatrný.

$q$	$\hat{c}_{MLE}$	$\hat{c}_{PWM}$
40 000	3 895	3 853
60 000	15 149	14 860
80 000	40 471	39 431
100 000	87 442	84 718

Tabuľka 4.11: Odhad doby návratu pre rôzne veľkosti poistného plnenia  $q$ .

Obr. 4.12: Return-level plot, kde MLE značí maximálne vierohodné odhady a PWM odhady GP rozdelenia pomocou pravdepodobnostne vážených momentov.

## 4.4 Porovnanie metódy blokových maxim a POT metódy

Pre lepšiu prehľadnosť sú v ďalšom texte zhrnuté výsledky z analýzy dát metódou blokových maxim aj POT metódou.

Pozrime sa najskôr na samotný odhad parametru tvaru pre obe metódy. Podľa kapitoly 3 by za splnenia všetkých teoretických predpokladov mal byť odhad parametru  $\gamma$  rovnaký pre GEV aj GP rozdelenie. Jeho odhady vo všetkých prípadoch, t.j. maximálne vierohodný odhad aj odhad pomocou pravdepodobnostne vážených momentov, môžeme vidieť v tabuľke 4.12.

	MLE	PWM
GEV	0,297	0,332
GP	0,279	0,281

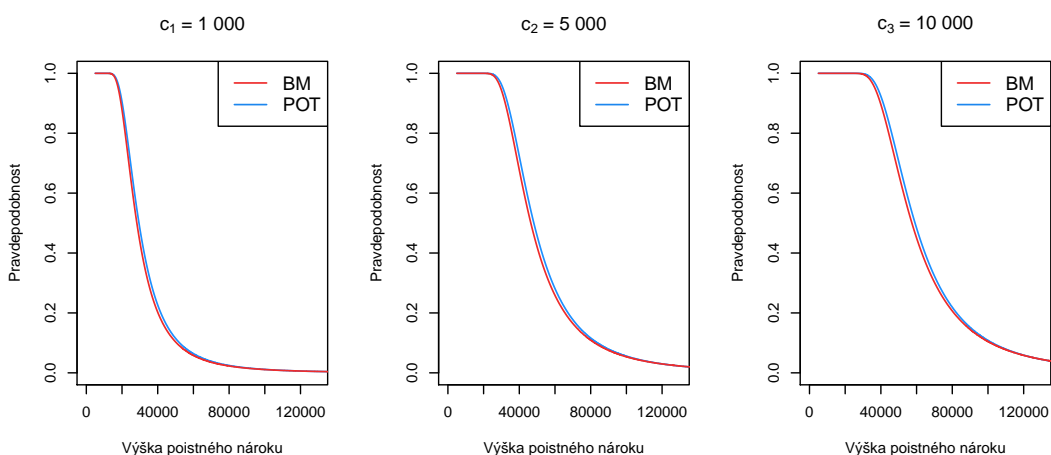
Tabuľka 4.12: Odhad parametru  $\gamma$ .

Poznamenajme, že odhady nie sú príliš rozdielne, avšak tieto výsledky by mali byť brané orientačne, nakoľko práve splnenie teoretických predpokladov uvažujeme, no nemôžeme si byť istý, že sa tak aj naozaj udialo.

## Extrémne pravdepodobnosti

Čo sa odhadu extrémnych pravdepodobností týka, obe metódy vykazujú podobné výsledky. Ako príklad uveďme odhad pravdepodobnosti, že do poisťovne príde klient s poistným nárokom vyšším ako 35 000 dolárov. Pri metóde blokových maxím ju odhadujeme na 0,0349 % a pri využití POT metódy na 0,0398 %, takže vidíme, že sa líšia až na treťom desatinnom mieste.

Ďalej sme skúmali pravdepodobnosť, že medzi nasledujúcimi  $c$  zákazníkmi sa objaví aspoň jeden s poistným nárokom vyšším ako  $x$  dolárov. Výsledky týchto odhadov pri použití maximálne vierohodných odhadov rozdelení sa pre obe metódy líšia len mierne, ako môžeme pozorovať na obrázku 4.13. Rozdiely sa zrejme prehľbujú tým, čím vyšší počet prichádzajúcich klientov  $c$  uvažujeme. Pritom metóda blokových maxím je v tomto prípade o niečo konzervatívnejšia ako POT metóda a priraduje vysokým čiastkam o čosi nižšie pravdepodobnosti.

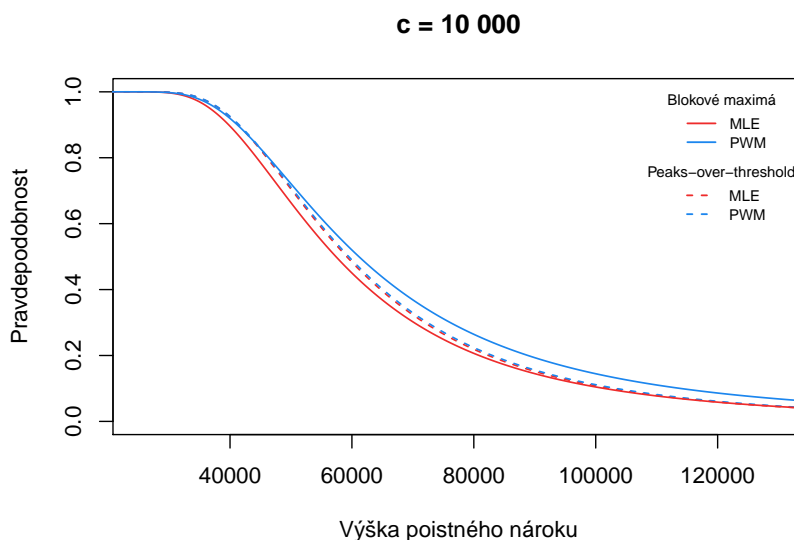


Obr. 4.13: Odhad extrémnych pravdepodobností, BM – metódou blokových maxím, POT – metódou peaks-over-threshold.

Lepšiu ilustráciu poskytuje obrázok 4.14, na ktorom sú zobrazené odhady extrémnych pravdepodobností oboma spôsobmi modelovania a taktiež pre obe metódy odhadu parametrov rozdelení. Tu vidíme, že najvyššie odhady pravdepodobností získavame použitím metódy blokových maxím a odhadov GEV rozdelenia získaných pomocou pravdepodobnostne vážených momentov. Naopak, ako najnižšie sú odhadované pri použití maximálne vierohodných odhadov pri metóde blokových maxím a niekde v strede sú odhady týchto pravdepodobností POT metódou v oboch prípadoch. Pritom posledné dve spomenuté sa asi od čiastky 100 000 dolárov ustálujú na približne rovnakých hodnotách.

## Úroveň a doba návratu

Porovnanie úrovne a doby návratu pre metódy blokových maxím a peaks-over-threshold je najjednoduchšie pomocou return-level plotu, ktorý vidíme na obr. 4.15.



Obr. 4.14: Odhad extrémnych pravdepodobností.

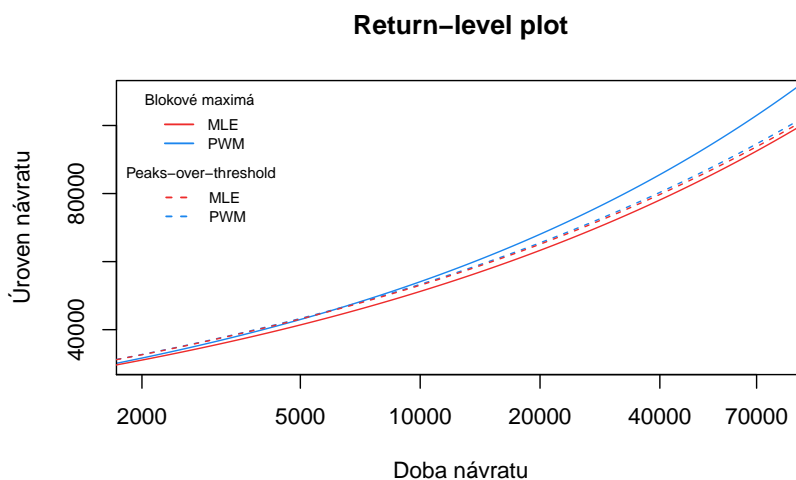
Znova sú v ňom vykreslené obe metódy s odhadmi príslušných rozdelení nadobudnutými metódou maximálnej vierohodnosti aj pomocou pravdepodobnostne vážených momentov. Získavame pritom podobné výsledky ako v prechádzajúcom prípade. Najmenšiu priemernú dobu návratu extrémneho poistného plnenia odhaduje model blokových maxim pri použití GEV rozdelenia s odhadmi pomocou pravdepodobnostne vážených momentov, pričom rozdiel oproti ostatným trom spôsobom narastá so zvyšujúcim sa uvažovaným poistným plnením.

Napríklad, ako vidíme v tabuľke 4.13, pre konkrétnu hodnotu poistného nároku vo výške 100 000 dolárov získavame odhad priemernej doby návratu o približne 26 000 zákazníkov nižší pri použití PWM odhadov zovšeobecneného rozdelenia extrémnych hodnôt ako pri použití maximálne vierohodných odhadov, čo už vytvára značný rozdiel. Odhady pomocou POT metódy sa opäť držia medzi spomínanými odhadmi metódou blokových maxim.

	MLE	PWM
GEV	90 690	64 078
GP	87 442	84 718

Tabuľka 4.13: Odhad doby návratu pre hodnotu poistného nároku  $q = 100\,000$ .

V prípade analyzovaných dát sa ako praktickejšia metóda na modelovanie javí peaks-over-threshold. Metóda blokových maxim môže totižto z analýzy veľmi ľahko vynechať dôležité dáta. Ak sa v jednom bloku aj nachádza viac udalostí s extrémnym dopadom, ďalej uvažované pri modelovaní bude len blokové maximum s najvyššou hodnotou. Naopak, do modelu sa môžu zaniest blokové maximá, ktoré nie sú až také významné. Vzhľadom na povahu dát o poistných nárokoch, kedy v rámci nich neuvažujeme žiadne časové periódy, by sme preto zvolili metódu POT, ktorá dostupné



Obr. 4.15: Return-level plot.

dáta využíva lepšie. Naviac, odhad zovšeobecneného Paretovho rozdelenia pri použití MLE a PWM odhadov je v oboch prípadoch veľmi blízky, preto ani nevzniká dilema, ktorý z týchto dvoch odhadov uprednostniť. Avšak použitie metódy blokových maximí môže byť v určitých situáciach taktiež preferované. Či už ide o stav, kedy sú k dispozícii len maximá nejakej premennej (ako napr. ročné maximálne teploty) alebo práve v prípade, kedy sa voľba veľkosti bloku vyskytuje prirodzene (dáta sú mesačné, ročné, desaťročné a pod.)

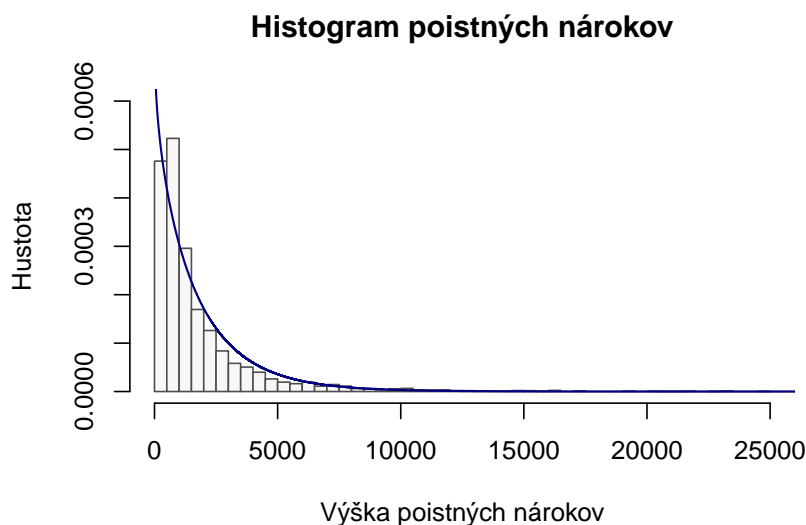
## 4.5 Parametrický model

Na záver práce sa pokúsme v stručnosti modelovať dáta klasickým parametrickým modelom a odhadnúť pritom pravdepodobnosť, že do poisťovne príde klient s vysokým poisťným nárokom ako aj úroveň a dobu návratu vysokého poisťného plnenia. Rozdiel oproti vyššie analyzovaným metódam je v tom, že tentokrát sa pokúsime aproximovať celý dátový súbor jedným pravdepodobnostným rozdelením.

Analyzovaná dáta sú, ako je uvedené na začiatku tejto kapitoly, silno zošikmené doprava a teda s ťažkými chvostami. Preto sa pri ich ďalšom spracovaní pokúsime využiť niektoré z rozdelení s podobnými vlastnosťami. Do úvahy napríklad pripadajú exponenciálne, Pareto, lognormálne, či Weibullovo<sup>1</sup>. Po preskúmaní a porovnaní možností sme sa rozhodli na modelovanie využiť Weibullovo rozdelenie. Histogram dát s hustotou Weibullovoho rozdelenia odhadnutou na základe maximálne vierohodných odhadov vidíme na obrázku 4.16, pričom hodnoty, ktoré presahujú 25 000 dolárov sú odrezané. Ani toto rozdelenie nie je na modelovanie ideálne, no spomedzi ostatných variant sa javilo ako najpriateľnejšia alternatíva.

Samotné odhady pravdepodobností objavenia sa zákazníka s vysokým poisťným

<sup>1</sup>Weibullovom rozdelením v tomto prípade myslíme „klasické“ Weibullovo rozdelenie, odlišné od toho používaného v kapitole 2.



Obr. 4.16: Histogram výšky poistných nárokov s odhadnutou hustotou Weibullovhovho rozdelenia.

nárokom sú veľmi nízke až nulové. Napr. pravdepodobnosť, že do poisťovne príde zákazník, ktorý od nej požaduje viac ako 35 000 dolárov je  $8,7 \cdot 10^{-6} \%$ . Pripomeňme, že pri použití teórie extrémnych hodnôt sme v rovnakom prípade získali odhady rádovo štyrikrát vyššie.

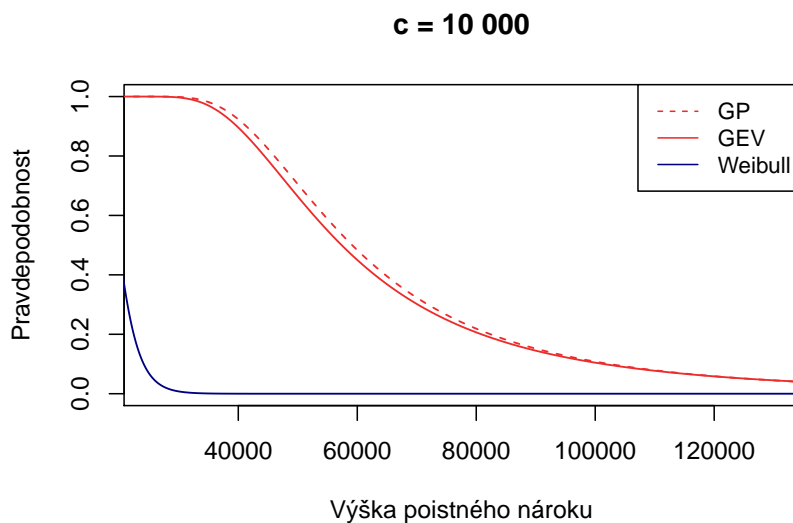
Okrem toho sa taktiež pozrime na odhad pravdepodobnosti  $\hat{p}_c$ , že medzi nasledujúcimi  $c = 10\,000$  zákazníkmi sa objaví aspoň jeden s poistným nárokom vyšším ako  $x$ . Tieto pravdepodobnosti tak isto nie sú vysoké a rýchlo klesajú. Pre konkrétne hodnoty sú tieto odhady zhrnuté v tabuľke 4.14.

$x$	$\hat{p}_c$
20 000	0,48453
25 000	0,06814
30 000	0,00770
35 000	0,00087
40 000	0,00010
45 000	0,00001

Tabuľka 4.14: Odhad pravdepodobnosti  $\hat{p}_c$  pre  $c = 10\,000$  klientov.

Dobrá vizuálnu predstavu môžeme získať taktiež z grafu 4.17. Vidíme, že klasický parametrický model priraduje pravdepodobnostiam, že do poisťovne príde medzi nasledujúcimi  $c$  zákazníkmi klient s vysokým poistným nárokom výrazne nižšie odhady v porovnaní s tými, ktoré sme získali pomocou teórie extrémnych hodnôt. Tam, kde ešte len odhady pomocou teórie extrémnych hodnôt ešte len začínajú byť zaujímavé, t.j. okolo hodnoty poistného nároku vo výške približne 40 000, tak sú už odhady pomocou parametrického modelu takmer, či úplne, nulové.





Obr. 4.17

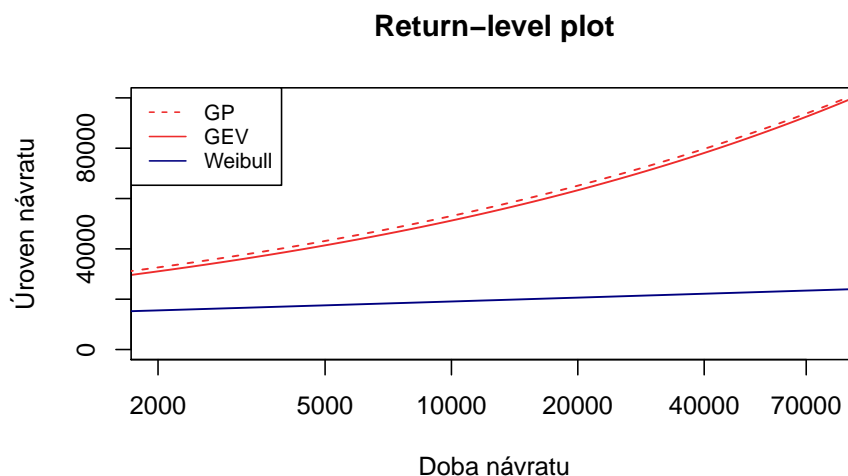
## Úroveň a doba návratu

Porovnajme teraz úroveň a dobu návratu extrémnej udalosti medzi parametrickým modelom a výsledkami získanými pomocou teórie extrémnych hodnôt. Úroveň návratu, t.j. veľkosť poistného nároku  $q$ , ktorá bude priemerne prekročená raz za  $c$  zákazníkov odhadneme ako  $(1 - 1/c)$ -kvantil Weibulloвого rozdelenia. Pre konkrétne hodnoty a maximálne vierohodné odhady Weibulloвого rozdelenia sú výsledky uvedené v tabuľke 4.15. Úroveň návratu stúpa s rastúcim počtom zákazníkov len pomaly. Zatiaľ čo výška poistného nároku, ktorá bude priemerne prekročená raz za 5 000 zákazníkov je odhadnutá na 17 561 dolárov, pre pätnásťkrát viac zákazníkov je to len o zhruba 6 000 dolárov vyššia čiastka.

$c$	$\hat{q}$
5 000	17 561
10 000	19 089
20 000	20 625
35 000	21 870
50 000	22 666
75 000	23 573

Tabuľka 4.15: Odhad úrovne návratu.

Na druhú stranu a v súlade s očakávaniami, odhad doby návratu  $c$  rapídne stúpa. Dobu návratu  $c$  chápeme ako priemerný počet novoprichádzajúcich zákazníkov, za ktorý sa vyskytne poistný nárok vo výške  $q$ . Odhady pre určité zvolené hodnoty sú zhrnuté v tabuľke 4.16. Už len rozdiel medzi odhadom doby návratu pre čiastku 25 000 a 30 000 je celkom priepastný. Vidíme, že tento model prakticky s vysokými poistnými nárokmi nepočíta, resp. uvažuje ich výskyt len s minimálnou pravdepo-



Obr. 4.18: Return-level plot.

dobnosťou.

Na obr. 4.18 vidíme return-level plot, kde je porovnaný vzťah medzi úrovňou a dobou návratu pre parametrický model a oba modely teórie extrémnych hodnôt. GEV pritom označuje model pomocou zovšeobecneného rozdelenia extrémnych hodnôt a teda metódou blokových maxím a GP značí zovšeobecnené Paretovo rozdelenie, ktoré reprezentuje modelovanie metódou peaks-over-threshold. Z grafu môžeme pozorovať, že krivka zobrazujúca závislosť odhadnutej doby a úrovne návratu parametrickým modelom síce stúpa, no výrazne pomalšie ako ostatné spomenuté.

$q$	$\hat{c}$
20 000	15 091
25 000	141 702
30 000	1 293 999
40 000	101 087 031
45 000	870 037 866

Tabuľka 4.16: Odhad doby návratu.

Parametrický model, ktorý aproximuje vstupné dáta pomocou Weibullovoho rozdelenia, výrazne podhodnocuje odhady vysokých pravdepodobností, úrovni a dôb návratu v porovnaní s metódami využívanými v analýze extrémnych udalostí. Analýzou dát sme potvrdili, že modelovaním pomocou metód študovaných v priebehu práce získavame zaujímavé výsledky, ktoré sa z dlhodobého hľadiska môžu javiť ako užitočné. Po zvážení záverov by mohli v našom konkrétnom prípade poslúžiť danej poisťovni napríklad pri posúdení, či rezervy, ktoré si vytvára, sú dostatočné na pokrytie mimoriadne vysokých poistných nárokov, ktoré sú síce vzácne, no s o to väčším dopadom.

# Záver

V tejto diplomovej práci sme sa venovali teórii extrémnych hodnôt a možnostiach jej využitia pri modelovaní dát. V prvých kapitolách sme v ucelenej podobe položili teoretický základ pre túto problematiku. Popísané sú dve široko využívané štatistické metódy pre analýzu extrémnych udalostí.

Jedným prístupom je využitie metódy blokových maxím. Podstata metódy tkvie v určení maxím pevne určeného počtu neprekrývajúcich sa blokov a následnom modelovaní tejto série získaných pozorovaní. Predpokladáme pritom, že výber blokových maxím patrí do sféry príťažlivosti zovšeobecneného rozdelenia extrémnych hodnôt, a teda dá sa týmto rozdelením aproximovať. No spomínaný postup je náchylný k neefektívnemu využitiu dostupných informácií, nakoľko z každého bloku sa vyberá len jedna hodnota. Metóda peaks-over-threshold túto nevýhodu odstraňuje. Myšlienkou je zvoliť si hranicu a ďalej modelovať tzv. excesy, t.j. výšky prekročenia nad túto hranicu, ak k tejto udalosti došlo. Potom predpokladáme, že limitné rozdelenie excesov sa dá aproximovať zovšeobecneným Paretovym rozdelením.

Metódy boli aplikované na dátach, ktoré sú tvorené informáciami o výške poisťných nárokov z automobilového priemyslu. Skúmané boli odhady extrémnych pravdepodobností, úrovne a doby návratu extrémnej udalosti a využité pritom boli maximálne vierohodné a pravdepodobnostne vážené momentové odhady príslušných rozdelení. Vo všetkých prípadoch neboli výsledky analýz markantne rozdielne. Avšak prípad modelovania pomocou blokových maxím s odhadmi GEV rozdelenia pomocou pravdepodobnostne vážených momentov sa zdá byť ako najmenej konzervatívny, pričom priraduje extrémnym udalostiam najvyššie odhady spomínaných charakteristík spomedzi všetkých možností. Problematickými úlohami pri modelovaní dát boli najmä voľba počtu blokov a výšky thresholdu. V oboch prípadoch bolo toto rozhodnutie zaťažené subjektívnym úsudkom.

Porovnaním výsledkov získaných pomocou teórie extrémnych hodnôt a klasického parametrického modelu sme dospeli k záveru, že parametrický model značne podhodnocuje odhady extrémnych pravdepodobností. Je preto dobré uvažovať nad tým, že vezmeme do úvahy práve výsledky analýzy extrémnych situácií, najmä ak by sme sa v danej oblasti chceli pripraviť aj na nečakané udalosti v rámci dlhšieho obdobia.

Zaujímavým rozšírením by mohlo byť skúmanie viacrozmerných metód na modelovanie extrémnych udalostí. Príkladom použitia môže byť situácia, kedy okrem výšky klientských poisťných nárokov uvažujeme ako druhú zložku ostatné náklady spojené s poisťným nárokom, ako napr. náklady spojené s overením pravosti nároku, prípadne právne poplatky. Tento prístup je študovaný napríklad v publikáciách [5]

a [6]. Další možností je skúmanie extrémov stacionárnych časových radov, na čo je poukázané napríklad v [2] alebo [4]. Nakoľko pozorovania v týchto sekvenciách častokrát vykazujú závislosť či sezónnosť, je potrebné základné metódy modifikovať, prípadne vyvinúť ďalšie nástroje pre ich lepšiu analýzu. Tieto a mnohé ďalšie metódy sa v súčasnosti stále rozvíjajú.

# Seznam použité literatury

- [1] BALKEMA A.A. a de HAAN L. Residual Life Time of Great Age. *The Annals of Probability*, 1974, 2(5), pp. 792–804.
- [2] BEIRLANT, J. a kol. *Statistics of Extremes: Theory and Applications*. Chichester: John Wiley & Sons, c2004. ISBN 0-471-97647-4.
- [3] CHRISTOPEIT, N. Estimating parameters of an extreme value distribution by the method of moments. *J. Statist. Plann. Inference*. 1994, (41), pp. 173–186.
- [4] COLES, S. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer, 2001. ISBN 1-85233-459-2.
- [5] COLES, S. a J. TAWN. Modelling Extreme Multivariate Events. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1991, 53(2), pp. 377–392.
- [6] DUPUIS, D. a B. JONES. Multivariate Extreme Value Theory And Its Usefulness In Understanding Risk. *North American Actuarial Journal*. 2006, 10(4).
- [7] EMBRECHTS, P., C. KLÜPPELBERG a T. MIKOSCH. *Modelling extremal events for insurance and finance*. New York: Springer, 1997. ISBN 3540609318.
- [8] FRANKE, J., W.K. Härdle a C.M. HAFNER. *Statistics of Financial Markets*. Berlin: Springer, 2011. ISBN 978-3-642-16520-7.
- [9] FREES, E.W. *Regression Modeling with Actuarial and Financial Applications*. Cambridge: Cambridge University Press, 2010.
- [10] GNEDENKO, B. V. Sur la distribution limite du terme maximum d' une série aléatoire. *Ann. Math.* 1943, (44), pp. 423 --453.
- [11] GONZÁLES-ESTRADA, E., J.A. VILLASENOR-ALVA. A Goodness-of-Fit Test for Location-Scale Max-Stable Distributions. *Communications in Statistics - Simulation and Computation*. 2010, 39(3), pp. 557–562.
- [12] HOSKING, J., J. WALLIS a E. WOOD. Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, 1985, 27(3), pp. 251-261.
- [13] HOSKING, J., J. WALLIS. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 1987, 29(3), pp. 339–349.

- [14] JENKINSON, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q.J.R. Meteorol. Soc.* 1955, (81), pp. 158–171.
- [15] KLUGMAN, S.A., H.H. PANJER a G.E. WILLMOT. *Loss models: from data to decisions*. 3rd ed. New Jersey: John Wiley & Sons, c2008. ISBN 978-0-470-18781-4.
- [16] LANDWEHR, J., N. MATALAS a J. WALLIS. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.*, 1979, (15), pp. 1055 –1064.
- [17] MCNEIL, A. J., R. FREY a P. EMBRECHTS. *Quantitative Risk Management: Concepts, Techniques and Tools*. New Jersey: Princeton University Press, 2005. ISBN 978-0-691-12255-7.
- [18] PICKANDS J. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*. 1975, 3(1), pp. 119–131.
- [19] RESNICK, S.I. *Extreme Values, Regular Variation and Point Processes*. New York: Springer, 1987.
- [20] SCARROTT C., A. MACDONALD. A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT - Statistical Journal*. 2012, 10(1), pp. 33–60.
- [21] SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*. 1985, (72), pp. 67–90.
- [22] de SOUSA, B. a G. MICHAILIDS. A Diagnostic Plot for Estimating the Tail Index of a Distribution. *J. Comput. Graph. Statist.* 2004, 13(4), pp. 974–995.
- [23] VILLASENOR-ALVA, J. A. a E. G. GONZÁLEZ. A bootstrap goodness of fit test for the generalized Pareto distribution. *Computational Statistics and Data Analysis*. 2009, 53, pp. 3835–3841.

