TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta přírodovědně – humanitní a pedagogická

POSTUPY ZALOŽENÉ NA BOOTSTRAPU V MODELECH EXTRÉMNÍCH UDÁLOSTÍ

DISERTAČNÍ PRÁCE

LIBEREC 2017 **RNDr. VÁCLAV KOHOUT**

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta přírodovědně-humanitní
a pedagogická

Postupy založené na bootstrapu v modelech extrémních událostí

Disertační práce

Studijní program: Aplikovaná matematika

Studijní obor: Matematické modely a jejich aplikace

Autor práce: RNDr. Václav Kohout

Vedoucí práce: Prof. RNDr. Jan Picek, CSc.

Prohlášení

Byl jsem seznámen s tím, že na mou disertační práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé disertační práce pro vnitřní potřebu TUL.

Užiji-li disertační práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Disertační práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé disertační práce.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Poděkování

Především bych chtěl poděkovat svému školiteli Prof. RNDr. Janu Pickovi, CSc. za jeho velkou podporu a péči, bez které by tato práce nikdy nevznikla. Zároveň děkuji za mnoho společných chvil, které mi v průběhu studia věnoval. Tento čas mne obohatil odborně, ale i lidsky. Za každou společně strávenou chvíli velmi děkuji.

Dále bych chtěl poděkovat své manželce Petře a všem dětem – Vendule, Lence, Soně, Petře, Vašíkovi a Josífkovi za podporu a motivaci pro další práci. Děkuji i ostatním nejbližším z rodiny.

Plzeň

Leden 2017 Václav Kohout

Anotace

Teorie extrémních hodnot je v současné době aplikována v mnoha oblastech, které se dotýkají běžného denního života. Velmi intenzivně jsou využívány například v klimatologii, kdy se snažíme odhadnout nejvyšší kvantily např. průtoku vody, teplotu ovzduší.

Proto abychom mohli takovéto jevy exaktně popsat, vytváříme statistické modely, které jsou závislé na hodnotách parametrů výsledné limitní náhodné veličiny. V takových případech se v odpovídajících modelech, které jsou vytvářeny především pomocí pravděpodobnostních rozděleních, zajímáme spíše o chvosty těchto rozdělení než o jejich centrální části.V teorii extremálních hodnot se touto problematikou zabývá tzv. věta Fisher – Tippet – Gnědenko – de Haan. Je známo, že limitní náhodná veličina je tříparametrické extremální rozdělení tzv. zobecněné rozdělení extrémních hodnot. Pro vyšetřování extrémních událostí je důležitý především „shape“ parametr, který lze popsat pomocí EVI (extremal value index). Proto jsou tak významné odhady tohoto indexu. Pomocí znalosti tohoto indexu můžeme popsat, jak jsou těžké chvosty rozdělení výsledné distribuční funkce limitní náhodné veličiny.

V práci se zabýváme především semiparametrickými odhady EVI, které jsou založeny na *k* nejvyšších pořádkových statistikách. Systematicky u každého odhadu ukazujeme jeho tvorbu a základní vlastnosti od konzistence až po asymptotickou normalitu. Práce byla zaměřena na využití metody bootstrap při tvorbě těchto odhadů, proto metodu bootstrap zavádíme a obecně studujeme. Dále uplatňujeme pro každý odhad tzv. bootstrapovou metodu „optimal sample fraction“, která nám umožní nalézt optimální volbu hodnoty k a tím i hodnotu EVI.

Pro základní odhady EVI jsme prostudovali a zjednodušili algoritmy metody „optimal smaple fraction“. Zjistili jsme, jaké jsou její optimální vstupní hodnoty, aby výsledek byl zpracován v co možná nejkratším čase a zároveň přesně. Kromě klasických odhadů EVI jsme se zabývali také odhady, které jsou souhrnně nazývány MVRB(Minimum-Variance reduced-Bias) a PORT(Peaks over Random Treshold). Pro třídu těchto odhadů jsme také připravili bootstrapové algoritmy.

Výsledky výše uvedené jsme aplikovali pro případ kvantilové regrese. Využili jsme poznatky uvedené v Dienstbier (2011) a pomocí vět 5.8 - 5.10 jsme mohli na danou situaci použít metodu „optimal sample fraction“ na odhad typu Pickands. Výsledkem bylo nalezení odhadu EVI chyby v regresním lineárním modelu (předpokládáme, že tato chyba pochází z rozdělení z Frèchetovy sféry přitažlivosti).

Pro další práci je možné vylepšovat stávající algoritmy a především aplikovat metodu „optimal sample fraction“ na další typy odhadů EVI.

Klíčová slova: Teorie extrémních hodnot, EVI, metoda optimal sample faction, kvantilová regresní analýza, bootstrap

Summary

Nowadays, the extreme value theory is applied in many fields that are related to everyday life. For example, it is very often used in climatology, where the aim is to estimate high quantiles of a water discharge, air temperate etc.

To describe such events we build statistical models which depend on unknown parameters of the limit random variable. In these probability distribution models we are interested rather in the tails of these distributions than in their central parts. In the extreme value theory this problem is treated by the Fisher – Tippet – Gnědenko – de Haan theorem. It is known that the resulting distribution is the generalized extreme value distribution which is a distribution with three parameters. For the description of extremal events the shape parameter is of the highest interest. This parameter can be described by extreme value index (EVI). That makes the estimation of this index very important. Using this index we can quantify the thickness of the tails of the limit distribution function.

This work is devoted to the semiparametric estimation of EVI which is based on the k highest order statistics. For each estimate we show its construction and its basic properties as its consistency or asymptotic normality. The work is focused on application of bootstrap methods for estimation. That is why the bootstrap methods are introduced and studied as well. We apply a bootstrap method called "optimal sample fraction" that allows us to find the optimal choice of value k as well as EVI.

We studied and simplified the algorithms of the "optimal sample fraction" method for basic EVI estimations. We found out what are its optimal initial values for the result to be processed as fast and exact as possible. Besides the EVI estimation we deal with so called MVRB (minimum-variance reduced-bias) and PORT (peaks over random treshold) estimators. Also we developed bootstrap algorithms for the class of these estimators. The above mentioned results were applied in a linear regression model.

In the future the current algorithms can be improved or the "optimal sample fraction" method can be applied on other types of EVI estimators.

Keywords: extreme value theory, extreme value index, optimal sample fraction method, quantile regression.

Obsah

Seznam grafů:

Seznam tabulek

Seznam zkratek a symbolů

$\overset{d}{=}$	rovnost v distribuci
$\overset{d}{\to}$	konvergence v distribuci
$\overset{P}{\to}$	konvergence v pravděpodobnosti
$\overset{s.j.}{\to}$	komvergence skoro jistě
$f(x) = O\big(g(x)\big), x \to \infty$	$\overset{\exists}{M > 0}\ \overset{\exists}{x_0}\ ; x > x_0, \lvert f(x)\rvert \le M\lvert g(x)\rvert$
$f(x) = o(g(x))$	$\lim_{x\to\infty} \frac{f(x)}{g)x)} = 0$
nebo	
$f(n) = o(g(n))$	$\overset{\forall}{\varepsilon > 0}\ \overset{\exists}{N > 0}\ ; n > N, \lvert f(n)\rvert < \varepsilon\ \lvert g(n)\rvert$
$X_n = O_P(a_n)$	$\overset{\forall}{\varepsilon > 0}\ \overset{\exists}{M > 0}\ ; \overset{\forall}{n}\ P\left(\left\lvert\frac{X_n}{a_n}\right\rvert > M\right) < \varepsilon$
$X_n = o_P(a_n)$	$\frac{X_n}{a_n} \overset{P}{\to} 0$
$a(t) \sim b(t)$	$\lim_{t\to\infty} \frac{a(t)}{b(t)} = 1$
$a(t) \approx b(t)$	limitně rovno, aproximace
$a(t) \cong b(t)$	přibližně rovno
α	index chvostu
γ	extreme value index
Γ	gamma funkce
$1_{(p)}$	charakteristická funkce: pro hodnotu p=pravda je rovna 1 a pro p=0 rovna 0
a_+	max(a,0)
a_-	min(a,0)
$[a]$	celá část z čísla a
s.v.	skoro všude

F_n	zprava spojitá empirická distribuční funkce
GP	zobecněné Paretovo rozdělení
$i.i.d.$	nezávislé a stejně rozdělené náhodné veličiny
U	(obvykle zleva spojité) inversní zobrazení k $\frac{1}{1-F}$
x^*	$\sup\{x; F(x) < 1\} = U(\infty)$
${}_*x$	$inf\{x; F(x) < 1\}$
f	hustota náhodné veličiny
n.v.	náhodná veličina
d.f.	distribuční funkce
EVI	extremal value index
EVT	extremal value theory

Seznam náhodných veličin užívaných v práci.

Rozdělení Burrovo. V naší práci budeme užívat rozdělení XII. třídy. Jde o tříparametrické rozdělení Burr(q, a, b).

Hustota tohoto rozdělení je rovna:

$$f(q,a,b)\colon x \mapsto \begin{cases} ab^{-a}qx^{-1+a}(1+(\frac{x}{b})^{a})^{-1-q} & , \quad x>0 \\ 0 & , \quad x\leq 0 \end{cases}$$

Rozdělení Gamma. Používáme dvouparametrické gamma rozdělení $\Gamma(\alpha,\beta)$.

Hustota tohoto rozdělení je rovna:

$$\Gamma(\alpha,\beta)\colon x \mapsto \begin{cases} \frac{e^{-\frac{x}{\beta}}x^{-1+\alpha}\beta^{-\alpha}}{\Gamma(\alpha)} & , \ x>0 \\ 0 & , \ x\leq 0 \end{cases}.$$

$\Gamma(\alpha)$ je hodnota gamma funkce v bodě α.

Paretovo rozdělení. Používáme dvouparametrické Paretovo rozdělení $P(k,\alpha)$.

Hustota tohoto rozdělení je rovna:

$$P(k,\alpha)\colon x \mapsto \begin{cases} (\frac{x}{k})^{-\alpha} & , \quad x>k \\ 0 & , \quad x\leq k \end{cases}$$

V práci se užívá také zobecněné Paretovo rozdělení. Jeho definice je uvedena v (2.18).

Beta rozdělení. Používáme klasické beta rozdělení se dvěma parametry $\mathrm{B}(\alpha,\beta)$.

Hustota tohoto rozdělení je rovna:

$$\mathrm{B}(\alpha,\beta)\colon x \mapsto \begin{cases} \frac{(1-x)^{-1+\beta}x^{-1+\alpha}}{\mathrm{Beta}(\alpha,\beta)} & , \ 0<x<1 \\ 0 & , \quad jinak \end{cases}$$

$\mathrm{Beta}(\alpha,\beta)$ je hodnota beta funkce v bodě (α,β).

F rozdělení (Fisher-Snedecorovo rozdělení) s parametry n, m.

Hustota tohoto rozdělení je rovna:

$$F(n,m)\colon x \mapsto \begin{cases} \frac{m^{m/2}n^{n/2}x^{-1+\frac{n}{2}}(m+nx)^{\frac{1}{2}(-m-n)}}{\mathrm{Beta}[\frac{n}{2},\frac{m}{2}]} & , \quad x>0 \\ 0 & , \quad x\leq 0. \end{cases}$$

Rozdělení Cauchy s parametry a, b .

Hustota tohoto rozdělení je rovna:

$$C(a,b): x \longmapsto \frac{1}{b\pi(1+\frac{(-a+x)^2}{b^2})}$$

Rovnoměrné rozdělení s parametry min, max.

Hustota tohoto rozdělení je rovna:

$$U(min,max): x \longmapsto \begin{cases} \frac{1}{\max-\min} & , \quad \min \le x \le \max \\ 0 & , \quad (x < \min) \vee (x > \max) \end{cases}$$

Normální rozdělení s parametry μ, σ.

Hustota tohoto rozdělení je rovna:

$$N(\mu,\sigma): x \longmapsto \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

Gamma rozdělení s parametry α, β.

Hustota tohoto rozdělení je rovna:

$$Gamma(\alpha,\beta): x \longmapsto \begin{cases} \frac{e^{-\frac{x}{\beta}} x^{-1+\alpha} \beta^{-\alpha}}{\Gamma[\alpha]} & , \quad x > 0 \\ 0 & , \quad x \le 0 \end{cases}$$

$\Gamma[\alpha]$ je hodnota funkce gama v bodě α.

$\boldsymbol{\chi^2}$ rozdělení s parametrem ν.

Hustota tohoto rozdělení je rovna:

$$\chi^2(\nu): x \longmapsto \begin{cases} \frac{2^{-\nu/2} e^{-x/2} x^{-1+\frac{\nu}{2}}}{\Gamma[\frac{\nu}{2}]} & , \quad x > 0 \\ 0 & , \quad x \le 0 \end{cases}$$

Studentovo t rozdělení s parametrem ν.

Hustota tohoto rozdělení je rovna:

$$t(\nu): x \longmapsto \frac{(\frac{\nu}{x^2+\nu})^{\frac{1+\nu}{2}}}{\sqrt{\nu}\mathrm{Beta}[\frac{\nu}{2},\frac{1}{2}]}$$

1. Úvod

Teorie extremálního rozdělení (dále EVT – Extremal Value Theory) má své kořeny v roce 1920, kdy Leonard Tippet nalezl způsob řešení úlohy přetržení vlákna bavlny. Toto přetržení nastávalo při jistých stavech, kdy bylo vlákno nejtenčí. Při řešení této úlohy bylo zapotřebí vytvořit pravděpodobnostní model, který by kvantifikoval chování vlákna. Jde o první modelovou situaci, která vedla k šetření extremálního stavu, extremální události.

S pomocí Ronalda Fishera vymezil novou statistickou disciplínu, která nalezla v krátké době velké praktické užití. Teoretickým rámcem byla tzv. Fisher-Tippetova věta (Fisher, Tippet (1928)). Na jejím základě bylo možno začít studovat události, které mají svou extremální povahu nebo nastávají vzácně a ojediněle. Souběžně řešil problémy minimalizace množiny náhodných veličin i Frèchet (1927). Zásadní krok k rozšíření a formalizaci výše uvedeného tvrzení provedl Boris Vladimirovič Gněděnko (Gnedenko (1943)), který syntetizoval a formalizoval všechny znalosti o extremálních jevech ve své zásadní větě z EVT.Rozšířením a úpravou předchozího tvrzení je poslední verze základní věty EVT, kterou zformuloval v roce 1972 Laurens de Haan (de Haan (1972)).

Pomocí teorie založené na předchozích tvrzeních se daří řešit mnoho úloh extremálního typu z různých oblastí vědy a života. V současné době jsme obklopováni pojmy o extrémech v počasí (orkány, přílivové vlny typu tsunami, prudké zvraty teplot, např. Clauset (2009)), v hydrologii (extremální stavy v povodí řek, Katz (2002), Bruneti (2004)), v biologii (výzkum DNA, Joyce (2008)), v ekonomii (změna hodnoty cenných papírů, pojišťovnictví Farmer (1999), Embrechts (1997)), ve fyzice (sluneční aktivita, Sibergleit (1999), kosmologie, Colombi (2011)), ale i ve společenských vědách (sociologie Kaiser (2013), psychologie Cooligan (2009)). Je tedy zřejmá snaha po popisu takových jevů, po jejich možném předvídání. Takový popis nabízí klasická i nová teorie extremálních statistik.

Všechny tyto oblasti mají přímé etické, sociální, ekonomické a jiné přesahy do běžného života celé společnosti. To je jeden z významných důvodů k mohutnému rozvoji teorie i praxe extremálním statistiky v posledních desetiletích. Ve skutečnosti mohou mít vzácné události (jevy) katastrofické následky pro lidskou společnost, pro přírodu i přírodní prostředí a to prostřednictvím jejich dopadu na reálné životní prostředí. Metodika pro odhad a predikci vzácných událostí je využívána k záchraně ohrožených přírodních zdrojů, k tvorbě klimatologických modelů, k modelování vzniku zemětřesení, vln tsunami a podobných měřitelných přírodních jevů jako jsou například teplota, srážky a povodně. Situace, kdy máme do činění s velkými riziky nebo s velmi nízkou pravděpodobností překročení resp. podkročení vysokých nebo nízkých hodnot.

Výše uvedené teoretické vlastnosti i praktické aplikace byly uvedeny a rozvíjeny ve významných monografiích Coles(2001), Beierland(2004), Castilo(2005), de Haan, Ferreira(2006), Resnick(2007) a mnoha tisíci článcích.

V kapitole 2. předkládané práce ukazujeme základní vlastnosti z teorie extremálních hodnot. Věnujeme se především budování této teorie, hledáním tzv. podmínek 1. a 2. řádu, pomocí nichž je možno nalézt různé typy semiparametrických odhadů základního parametru extremálního rozdělení. Jde o parametr EVI – extreme value index. Jsou popsány tři základní varianty tzv. sfér přitažlivosti. V této kapitole se zabýváme konstrukcí základních odhadů EVI(extremal value index). U každého odhadu se zabýváme jeho vlastnostmi včetně konzistence a asymptotické normality. V závěru kapitoly se zabýváme popisem některých

novějších odhadů EVI. Poznamenejme, že v současné době je známo několik desítek takových odhadů

Protože se v kapitole zaměřujeme především na semiparametrické přístupy, popíšeme dále v krátkosti parametrické přístupy k řešení problematiky EVT. Jedním z nich je Gumbelův přístup neboli postup, který se nazývá metoda blokových maxim. Jde o metodu, která využívá metodu maximální věrohodnosti pro odhady parametrů $(\lambda, \delta, \alpha)$ aplikované na distribuční funkci extremálního rozdělení $EV_\alpha\left(\left(\frac{x-\delta}{\lambda}\right)\right)$, kde $EV_\alpha(x)$ je následující distribuční funkce extremálního rozdělení

$$EV_\alpha(x) = \begin{cases} \exp(-(1+\alpha\, x)^{-\frac{1}{\alpha}}, 1+\ \alpha\, x > 0, \quad \alpha\ \neq 0 \\ \\ \quad exp\big(-exp(-x)\big), x \in \mathbb{R}, \qquad \alpha = 0 \end{cases}$$

Výpočetní detaily tohoto přístupu lze nalézt v článcích Hosking (1985), MacLeod (1989), Prescott, Walden (1980), (1983) a Smith (1985). Protože je tato metoda numericky velmi obtížně řešitelná, rozvinuly se další alternativní metody. Jednou z nich je metoda PWM(probability weighted moment) zavedená v článcích Landwehr at al. (1979) a Greenwood at al. (1979), jde o zajímavou alternativu k metodě maximální věrohodnosti. Hlavní myšlenka této metody je porovnání hodnot momentů

$E(X^p(F(X))^r(1-F(X))^s)$,

kde p, r , s jsou reálná čísla s jejich empirickými verzemi, podobně jak se to provádí v klasické momentové metodě.

Další parametrické metody lze najít v literatuře: Jde například o metodu lineárního nejlepšího nevychýleného odhadu uvedeného v článku Balakrishnan, Chan (1992), klasickou metodu momentů uvedenou v Christopeit (1994), metodu odhadu minimální vzdálenosti Dietrich, Hüsler (1996).

V třetí kapitole se věnujeme teoretickým základům metody bootstrap. Jsou zde uvedeny a studovány postupy, které zavedl ve své monografii Hall (1992). Zde popisuje principy redukce odchylky (bias) pomocí metody bootstrap, ukazuje přímou aplikací Edgeworthova rozvoje například nalezení přesného intervalového odhadu a nachází i formule pro přesnost metody bootstrap. V monografii Hall (1992) jsou uvedeny základní principy Edgeworthova a Cornish-Fisherova rozvoje a především jsou použity pro zpřesnění bootstrapu a pro ukázání rychlosti a přesnosti konvergence této metody. V knize je také načrtnuta základní myšlenka metody „sample fraction“ – nalezení vztahu mezi velikostí jednotlivých částí výběru a velikostí vlastního základního výběru. V samotné kapitole ukazujeme pomocí teoretických nástrojů – Berry-Essenovy nerovnosti, Polyovy věty a silného zákona velkých čísel konzistenci metody bootstrap. Hlavním teoretickým výsledkem jsou dále věty 3.14 a 3.15, které dokazují konvergenci metody bootstrap s.j., za jistých podmínek. Ve zbylé části kapitoly se zabýváme využitím Edgeworthova rozvoje v bootstrapu.

V kapitole 4. se zabýváme praktickým teoretickým odvozením základních postupů pro využití metody bootstrap při aplikaci na jednotlivé typy odhadů EVI. Jedná se postupně o Hillův odhad, momentový odhad, Pickandsův odhad a odhad MVRB. Na všechny tyto odhady jsme aplikovali metodu „sample fraction“, kterou myšlenkově připravil Hall ve své monografii z roku 1992. Ukazuje se, že vlastní aplikace je pro jednotlivé metody unikátní, protože používáme vždy jiné postupy jak pro důkaz konvergence metody, tak i pro jejich vlastní realizaci. Pro každý typ odhadu vždy odvodíme algoritmus, který budeme v dalším využívat. V této části také odvodíme teoreticky konvergenci metody PORT a ukážeme i algoritmus pro vlastní výpočet EVI pomocí této metody. Závěrečná část je věnována simulační studii, v níž jsme využili výše uvedené algoritmy pro jednotlivé typy odhadů a pro různé hodnoty jejich základních parametrů. Zajímá nás rychlost procedury v závislosti

na počtu prvků základního výběru, zajímá nás také hodnota AMSE(asymptotický druhý moment), optimální poměr pro jednotlivé odhady, vztah mezi jednotlivými částmi konstruovaných množin atd. Všechny tyto možnosti jsme zobrazili pro jednotlivé odhady, ale i mezi nimi navzájem. Grafy i vlastní algoritmy jsme realizovali pomocí software Mathematica.

V matematice se velmi často setkáváme s konstrukcemi, které jsou založené na metodě nejmenších čtverců. Jde o přirozenou metodu, která umožňuje měřit odchylky mezi množinami údajů. Velké využití má tato metoda například v numerické matematice a také ve statistice a to především v regresní analýze. Protože jsme schopni analyzovat stále složitější a komplexnější reálná data, ukázalo se významné najít i jiné způsoby měření přesnosti matematického modelu, který jsme pro danou situaci připravili. Protože je matematický model našeho světa stochastický, nebude nikdy dostatečně přesný. Chceme vždy navrhnout model, který bude popisovat reálnou situaci co nejlépe. Je proto nutné najít vhodné kritérium, pomocí něhož rozhodneme, zda je námi navržený model dobrý nebo dokonce nejlepší.

Intuitivním kritériem se může zdát vyjádřit chybu jako součet rozdílů mezi skutečnými a modelovanými hodnotami. Ukazuje se, že takové kritérium má mnoho záporů. hlavní spočívá v tom, že velké záporné chyby se kompenzují velkými kladnými chybami. Dalším kritériem může být součet absolutních hodnot příslušných chyb. Odtud je již jen malý krok k různým metodám nejmenších čtverců., která se užívá v širokém měřítku.

Ukazuje se ovšem, že modely založené na absolutních chybách lépe popisují například případy s dlouhodobými strukturálními změnami než modely založené na metodě nejmenších čtverců. Tato regrese je někdy známa pod zástupnými názvy – LAD (Least Absolute DeviatianValue) modely, MAD (Minimum Absolute Deviation) modely, mediánová regrese a L_1 - modely. Také v modelech s extrémními pozorováními je kritérium nejmenších absolutních chyb robustnější než metoda nejmenších čtverců. Odtud je již jen krok k tomu používat například v situacích s extrémními hodnotami nebo například v případě, kdy náhodné chyby nejsou popsány normálním rozdělením jiné kritérium přesnosti. Budeme chtít v takovýchto modelech studovat analogie postupů uvedených výše. Pro takovéto případy se v praxi používá necelých 45 let starý postup – kvantilová regrese.

V článku Koenker, Bassett (1978) se poprvé objevil popis kvantilové regrese. V druhém článku Koenker, Bassett (1982) zavedli autoři novou třídu testů heteroskedasticity, založenou na kvantilech a prokázali, že tyto testy jsou robustní, na rozdíl od podobných testů založených na metodě nejmenších čtverců reziduí.

Speciálním případem metody kvantilové regrese je zmíněná mediánová regrese. Pomocí ní odhadujeme medián závisle proměnné pomocí hodnot nezávisle proměnných. Postup je v podstatě velmi podobný metodě nejmenší čtverců s tím rozdílem, že pomocí metody nejmenších čtverců odhadujeme podmíněné průměry závisle proměnné, zatímco mediánová regrese minimalizuje součet absolutních hodnot reziduí.

Teoretické základy kvantilové regrese nalezneme například v Koenker (2005).

V poslední páté kapitole jsme se nejprve zabývali teoretickým zavedením kvantilové regrese. Poté jsme uvedli základní vlastnosti kvantilové regrese. V třetí části této kapitoly jsme uvedli několik základních vlastností kvantilové regrese. Čtvrtá část obsahuje model kvantilové regrese včetně jeho řešení. Soustředili jsme nejen na výpočty příslušných regresních funkcí, ale především na jejich grafickou realizaci. Posléze jsme využili znalosti z kapitoly 4, výsledků práce Diensbier(2001) a použili jsme metodu bootstrap na stanovení odhadu parametru tvaru rozdělení chyb v lineárním regresním. Výsledkem jsou postupně věty 5.7 – 5.10. V závěru této kapitoly jsme zrealizovali rozsáhlejší simulační studii na toto téma včetně mnoha grafických výstupů.

2. Extremální rozdělení

2.1. Úvod

Hlavním zájmem statistiky extremálních hodnot je nalézt možnou limitu výběrových maxim resp. minim posloupnosti nezávislých a stejně rozdělených náhodných veličin.

Nechť $\{X_1, X_2, \dots, X_n\}$ je náhodný výběr, tj. posloupnost n nezávislých a stejně rozdělených (i.i.d.) náhodných veličin (n.v.), a nechť F je jejich distribuční funkce (d.f.). Symbolem $X_{i:n}$, $i=1,\dots,n$ označme i-tou pořádkovou statistiku, v dalším textu budeme pracovat s uspořádaným náhodným výběrem $\{X_{1:n}, X_{2:n}, \dots, X_{n:n}\}$.. Speciálně $X_{1:n}$ a $X_{n:n}$ representují výběrové minimum resp. výběrové maximum. Vzhledem k jednoduchému vztahu mezi výběrovým minimem a výběrovým maximem, platí totiž, že $m_n = min(\{X_1, X_2, \dots, X_n\}) = -max(\{-X_1, -X_2, \dots, -X_n\})$, je zřejmé, že stačí vyšetřovat jen problematiku výběrových maxim. Označme dále symbolem $M_n = max(\{X_1, X_2, \dots, X_n\})$. Pro rozdělení M_n je možné nalézt její d.f. Pro všechna $x \in R$ zřejmě platí

$$F_{M_n}(x) = P(M_n \le x) = P(X_1 \le x, X_2 \le x, \dots, X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = F^n(x).$$

2.2. Extremální limitní věty

Zajímá nás především chování výběrových maxim pro libovolnou velikost výběru. Platí následující tvrzení

Věta 2.1

Nechť F je výše uvedená d.f. posloupnosti n.v. $\{X_1, X_2, \dots, X_n\}$ a nechť $x^* = \sup\{x; F(x) < 1\}$. Potom

$$max(X_1, X_2, \cdots, X_n) \xrightarrow{P} x^*, \quad n \to \infty ,$$

kde $\xrightarrow{P}$ označuje konvergenci v pravděpodobnosti.

Důkaz:

Máme dokázat, že pro $\forall\, \epsilon > 0$ je

$$\lim_{n\to\infty} P(|M_n - x^*| \ge \epsilon) = 0$$

Platí

$$P(|M_n - x^*| \ge \epsilon) = P\big((M_n \ge x^* + \epsilon) \vee (M_n \le x^* - \epsilon)\big) =$$
$$= P(M_n \ge x^* + \epsilon) + P(M_n \le x^* - \epsilon) = 0 + P(M_n \le x^* - \epsilon) =$$
$$= F^n(x^* - \epsilon).$$

Protože $x^* = \sup\{x; F(x) < 1\}$, je

$$\lim_{n\to\infty} P(|M_n - x^*| \ge \epsilon) = \lim_{n\to\infty} F^n(x^* - \epsilon) = 0.$$

Podle hodnoty x^* je možné určit i následující limitu

$$\lim_{n\to\infty} F^n(x) \to \begin{cases} 0, & x < x^* \\ 1, & x \geq x^* \end{cases}.$$

Odtud vyplývá, že je nutno pro získání limitní náhodné veličiny, která nebude degenerovaná, provést normalizaci náhodné veličiny (je rovna maximu uvedenému ve větě 2.1).

Předpokládejme tedy, že existují reálné posloupnosti $a_n > 0$ a b_n takové, že

$$\frac{max(X_1, X_2, \cdots, X_n) - b_n}{a_n} \tag{2.1}$$

konverguje v distribuci k nedegenerované náhodné veličině *U* s distribuční funkcí *G*

$$\frac{M_n - b_n}{a_n} \xrightarrow{D} U \tag{2.2}$$

neboli

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

v každém bodě spojitosti funkce G, která není degenerovaná.

2.3. Sféry přitažlivosti

Dříve než vyřešíme problém nalezení limitní funkce G, uvedeme dvě důležité definice.

Definice 2.2

Nechť F je d.f., pro kterou existují reálné posloupnosti $\{a_n > 0\}$, $\{b_n\}$ takové, že platí

$$\lim_{n\to\infty} P(M_n \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \tag{2.3}$$

pro každý bod spojitosti *G*. Potom řekneme, že F je ve **sféře přitažlivosti** *G* (domain of attraction) a označujeme $F \in D(G)$.

Definice 2.3

Řekneme, že dvě d.f. F_1 a F_2 jsou **stejného typu** (distribution functions of the same types), jestliže existují dvě reálná čísla $a > 0$, b taková, že

$$F_1(x) = F_2(a\,x + b) \text{ pro všechna reálná čísla } x. \tag{2.4}$$

Jsou – li tedy F_1 a F_2 stejného typu, pak vyplývá z výše uvedeného, že existují parametry polohy a měřítka takové, že dané distribuční funkce patří do stejné množiny distribučních funkcí, které jsou invariantní vzhledem ke změně měřítka a posunutí (nazývají se také location – scale family).

Problém nalezení výše uvedeného rozdělení postupně řešili Fisher a Tippet (1928), Gnedenko (1943) a zformalizoval a upřesnil de Haan (1970).

Věta 2.4

Jestliže je $F \in D(G)$, potom limitní distribuční funkce G je jedním ze tří typů:

I. $\Lambda(x) = \exp(-\exp(-x))\,,\ x \in R\ ;$

II. $\Phi_\alpha(x) = \begin{cases} 0\,, & x \le 0 \\ exp(-x^{-\alpha}), & x > 0,\ \alpha > 0; \end{cases}$

III. $\Psi_\alpha(x) = \begin{cases} exp(-(-x)^\alpha), x < 0\,, \alpha > 0 \\ 1, \quad x \ge 0 \end{cases},$

kde parametr α je označován jako tzv. **shape** parametr nebo parametr tvaru. Tento parametr popisuje chování chvostu d.f. *F* v nekonečnu.

Důkaz jednotlivých verzí věty je možno najít postupně ve výše uvedených článcích. V nejnovější verzi je založen na řešení funkcionálních rovnic.

Jednotlivé druhy výsledné d.f. *G* jsou označovány následujícím způsobem:

I. Jde o třídu rozdělení typu **Gumbel.**
II. Jde o třídu rozdělení typu **Frèchet.**
III. Jde o třídu rozdělení typu **Weibull.**

Jednotně bylo toto výsledné rozdělení popsáno v Jenkinson (1955) a je označována jako rozdělení extremálních hodnot. Takto jsou označována rozdělení, jejichž d.f. je typu

$$G_\gamma(x) = exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), 1 + \gamma x > 0, \qquad (2.5)$$

kde γ je reálná hodnota a pro γ = 0 se pravá strana interpretuje jako $exp(-e^{-x})$. Hodnotu γ nazýváme **indexem chvostu** d.f. *F* (alternativně je označován také jako **Extreme Value Index – EVI**). Mezi shape indexem α a indexem chvostu γ je následující vztah:

Třída I. Pro hodnotu $\alpha = 0$ je $\gamma = 0$

Třída II. Pro hodnotu $\gamma > 0$ je index chvostu definován jako $\alpha = \frac{1}{\gamma} > 0$

Třída III. Pro hodnotu $\gamma < 0$ je index chvostu definován jako $\alpha = -\frac{1}{\gamma} > 0$

Poznamenejme, že hodnoty parametru $\gamma = 0\,, \gamma > 0\ a\ \gamma < 0$ distribuční funkce G_γ odpovídají kvalitativně naprosto odlišným rozdělením pravděpodobnosti. Proto odhad na základě pozorování je základním problémem z teoretického hlediska, ale především z pohledu konkrétních aplikací. Velmi významné je rozlišení případu $\gamma = 0$ (exponenciální pravý chvost) a případů $\gamma > 0$ (algebraický, těžký pravý chvost).

Definice 2.5
Nechť F je d.f., pak funkce *U(t)* daná vztahem

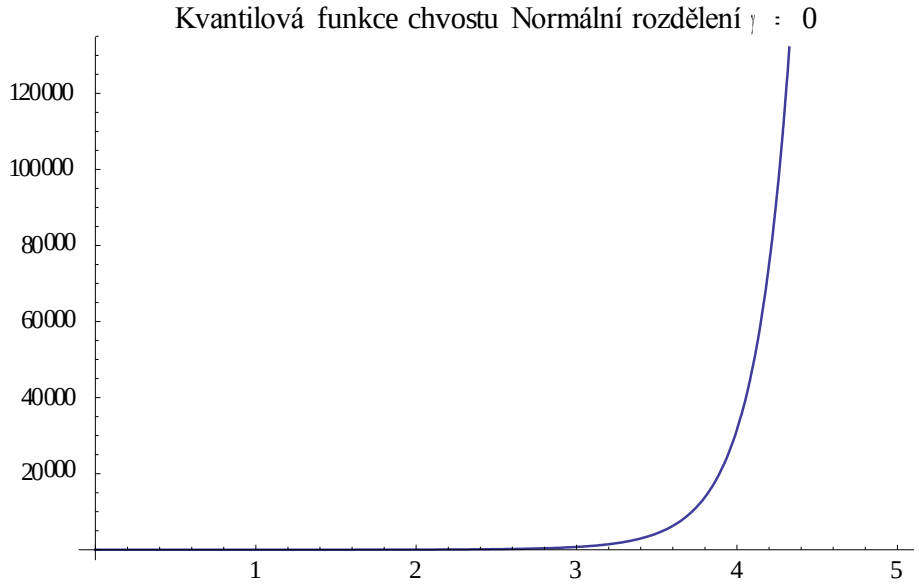
$$U(t) = inf\left\{y; F(y) \ge 1 - \frac{1}{t}\right\}, t\ \ge 1$$

a nazývá **kvantilová funkce chvostu** (tail quantile function).

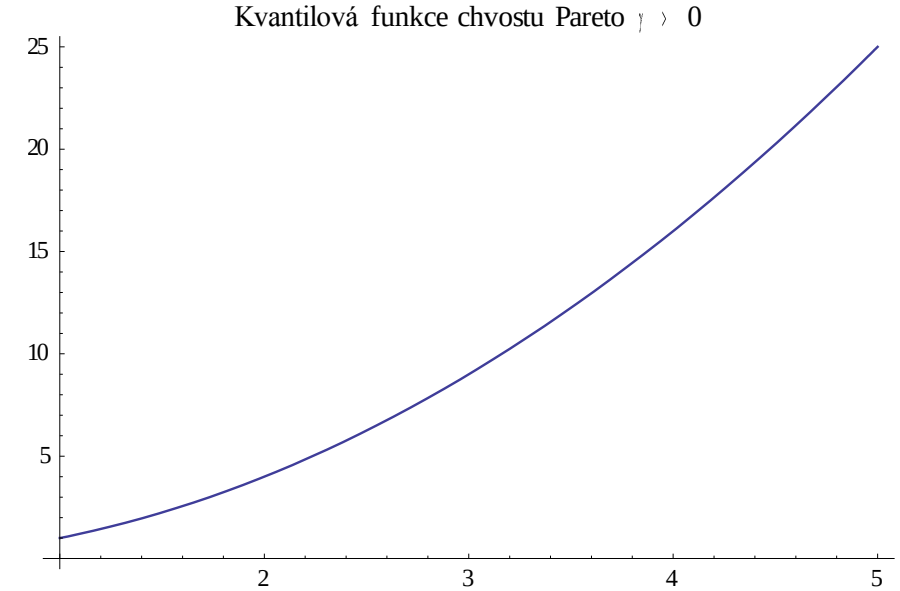
Uveďme některé vlastnosti kvantilové funkce chvostu:

1. Funkce *U(t)* je neklesající na intervalu $< 1 ; \infty)$.
2. $U(1) = inf\{x; F(x) \geq 0\} = x_*$, kde x_* je levý koncový bod *F*.
3. $U(\infty) = \lim_{t \to \infty} U(t) = inf\{x; F(x) \geq 1\} = sup\{x; F(x) < 1\} = x^*$

Na následujících obrázcích jsou ilustrovány dvě různé kvantilové funkce chvostu.



Graf 1 - Kvantilová funkce chvostu normálního rozdělení



Graf 2 - Kvantilová funkce chvostu rozdělení Pareto s parametry 1,1

Kvantilová funkce chvostu normálního rozdělení nabývá v odpovídajících bodech mnohem větších hodnot než u Paretova rozdělení.

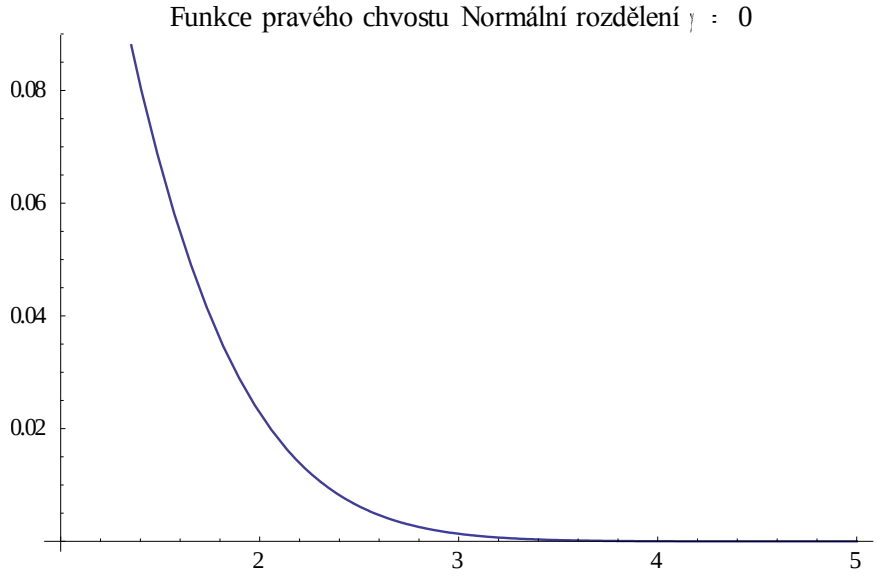
Definice 2.6

Nechť je F d. f., pak funkce $\bar{F}(t)$ daná vztahem

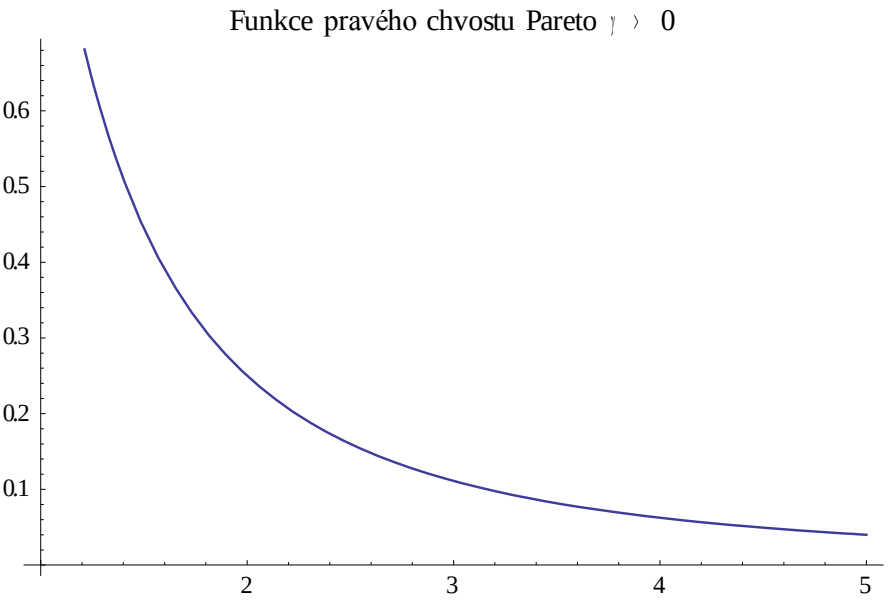
$$\bar{F}(t) = 1 - F(t)$$

se nazývá **funkce pravého chvostu** (right tail of a distribution function).

Na následujících obrázcích jsou podobně ilustrovány dvě funkce pravého chvostu:



Graf 3 - Funkce pravého chvostu normální rozdělení



Graf 4 - Funkce pravého chvostu Pareto s parametry 1,1

Na první pohled je zřejmé, že hodnota funkce pravého chvostu normálního rozdělení je podstatně menší než u rozdělení Paretova.

Z výše uvedených obrázků můžeme usoudit, že je významný rozdíl mezi případy tříd I., II., III. Popíšeme dále jednoduché rozdíly mezi jednotlivými třídami, které budeme dále upřesňovat.

Třída I. Pro případ $\gamma = 0$ jsou rozdělení (sféra přitažlivosti typu Gumbel), která patří do této třídy, označována za rozdělení s exponenciálním pravým chvostem, která mají konečný nebo nekonečný pravý bod x^*. Pro rozdělení z této třídy existují všechny momenty. Příkladem takových rozdělení jsou například normální rozdělení, exponenciální rozdělení, lognormální rozdělení.

Třída II. Pro případ $\gamma > 0$ jsou rozdělení (sféra přitažlivosti typu Frèchet), která patří do této třídy, označována jako rozdělení s pravým těžkým (algebraickým) chvostem a nekonečným bodem x^*. Momenty řádu vyššího než $\frac{1}{\gamma}$ neexistují. Tato třída obsahuje rozdělení například jako jsou Paretovo rozdělení, Cauchyho rozdělení, Studentovo rozdělení.

Třída III. Pro případ $\gamma < 0$ jsou rozdělení (sféra přitažlivosti typu Weibull), která patří do této třídy, označována jako rozdělení s lehkým pravým chvostem, která mají konečný pravý bod x^*. Příkladem takových rozdělení jsou např. rovnoměrné rozdělení, Beta rozdělení.

Ve svém článku Fisher a Tippet (1928) nalezli výše uvedené tři třídy rozdělení a zároveň popsali jejich chování:

Výběrové maximum o velikosti k n může být nalezeno jako největší prvek k náhodných výběrů o rozsahu n. Jestliže existuje pro každý takový výběr limitní distribuční funkce, musí platit $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$. Pro k výběrových maxim musí tedy platit $\lim_{n\to\infty} F^{kn}(a_n x + b_n) = G^k(x)$. Pokud bychom ovšem vyšetřovali toto maximum jako výběr o rozsahu k n bude platit $\lim_{n\to\infty} F^{kn}(a_{kn} x + b_{kn}) = G(x)$, kde $\{a_{kn} > 0\}, \{b_{kn}\}$ jsou dvě reálné normalizační posloupnosti. Ovšem Fisher a Tippet (1928) ve svém článku ukázali, že oba výsledky jsou ekvivalentní. Existují proto dvě reálné posloupnosti $\{A_k > 0\}, \{B_k\}$ takové, že platí

$$G(x) = G^k(A_k x + B_k) \text{ pro všechna reálná čísla x a celá čísla k > 0.} \qquad (2.6)$$

Tato funkcionální rovnice je známá jako rovnice stability (**stability equation**). Distribuční funkce *G*, která splňuje výše uvedenou funkcionální rovnici se nazývá **max-stable** distribuční funkce.

Autoři výše uvedeného článku určili všechna možná řešení předchozí funkcionální rovnice. Jsou jimi třídy funkcí I., II. a III. typu.

Když řešíme problém nalezení d. f. *G*, která je distribuční funkcí n.v. odpovídající limitě posloupnosti $\frac{M_n - b_n}{a_n}$, je přirozené se ptát: Jaké jsou nutné a postačující podmínky kladené na d. f. *F*, aby tato distribuční funkce patřila k sféře přitažlivosti *G*?

Takovou otázkou se zabýval mimo jiné von Mieses (1936), který nalezl sadu postačujících podmínek pro to, aby d. f. *F* patřila dané sféře přitažlivosti *G*.

Věta 2.7 (von Miesesovy postačující podmínky pro $F \in D(G_\gamma)$, von Mieses (1936))

Nechť je *F* absolutně spojitá d. f. náhodné veličiny *X*. Nechť dále existuje hustota $f(x) = F'(x)$ této náhodné veličiny *X* a $F''(x)$. Položme dále $h(x) = \frac{f(x)}{1-F(x)}$.

1) Předpokládejme, že $h(x) \neq 0$ a diferencovatelné vzhledem k *x* v okolí bodu x^*. Jestliže je

$$\lim_{x\to x^*}\left(\frac{1}{h(x)}\right)' = 0\,,$$

potom $F \in D(G_0)$, kde G_0 odpovídá třídě rozdělení typu Gumbel.

2) Předpokládejme, že $x^* = \infty$ a F' existuje. Nechť existuje $\gamma > 0$ takové, že

$$\lim_{x\to\infty} x\,h(x) = \frac{1}{\gamma} = \alpha\,,$$

potom je $F \in D(G_\gamma)$, kde G_γ odpovídá třídě rozdělení typu Frèchet.

3) Předpokládejme, že $x^* < \infty$ a F' existuje pro všechna $x < x^*$. Nechť existuje $\gamma < 0$ takové, že platí

$$\lim_{x\to x^*}(x^* - x)\,h(x) = -\frac{1}{\gamma} = \alpha\,,$$

potom je $F \in D(G_\gamma)$, kde G_γ odpovídá třídě rozdělení typu Weibull.

Důkaz dané věty je uveden například v de Haan (1976) nebo de Haan, Fereira (2006).

Předchozí větu lze přeformulovat takto:

Věta 2.8

Nechť platí podmínky předchozí věty. Nechť dále je

$$\lim_{x\to x^*}\left(\frac{1}{h(x)}\right)' = \gamma\,,$$

potom je $F \in D(G_\gamma)$.

Uvedené podmínky von Miesese jsou jen postačující, navíc požadujeme existenci první a druhé derivace d. f. *F*. Nejsou proto použitelné na rozdělení, která nesplňují předpoklad existence těchto derivací.

Gněděnko (1943) nalezl postačující i nutné podmínky pro to, aby daná distribuční funkce *F* patřila do sféry přitažlivosti G_γ .

Věta 2.9

1) $F \in D(G_0)$ právě když

$$x^* \le \infty \text{ a } \lim_{x\to x^*}\frac{\bar{F}(t+xg(t))}{\bar{F}(t)} = \exp(-x), \text{ pro všechna reálná x,}$$

kde $g(t)$ je spojitá a monotónní kladná reálná funkce.

2) $F \in D(G_\gamma)$, $\gamma > 0$ právě když $\gamma > 0$ a $\gamma = \frac{1}{\alpha}$,

$$x^* = \infty\,,\ \lim_{t\to x^*}\frac{\bar{F}(tx)}{\bar{F}(x)} = t^{-\frac{1}{\gamma}} = t^{-\alpha}\,,\ \text{pro všechna } t > 0.$$

3) $F \in D(G_\gamma)$, $\gamma < 0$ právě když $\gamma < 0$ a $\gamma = -\frac{1}{\alpha}$,

$$x^* < \infty\,,\ \lim_{t\to x^*}\frac{\bar{F}(x^*-tx)}{\bar{F}(x^*-x)} = t^{-\frac{1}{\gamma}} = t^{\alpha}\,,\ \text{pro všechna } t > 0.$$

Výše uvedené podmínky nejsou konstruktivní v tom smyslu, že neumožňují nalézt dříve uvedené posloupnosti $\{a_n > 0\}, \{b_n\}$ nutné pro normalizaci náhodné veličiny M_n.

Ovšem pomocí těchto podmínek zavedl de Haan (1970) důležité pojmy, které umožňují důkladnější analýzu základních postupů při práci s extremálními statistikami.

Definice 2.10 (pravidelně se měnící funkce)

Řekneme, že kladná a měřitelná funkce $f: R^+ \to R$ se nazývá **pravidelně se měnící funkce** (v nekonečnu) (regular variation) s indexem α, právě když

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

Takovouto funkci značíme $f \in R_\alpha$. V případě hodnoty $\alpha = 0$, nazýváme funkci $f \in R_0$ **pomalu se měnící funkcí** (v nekonečnu).

Pro další komplexnější analýzy jsou zavedeny zobecněné pojmy pravidelně se měnících funkcí.

Definice 2.11 (rozšíření pojmu pravidelně se měnících funkcí a třída funkcí Π)

Předpokládejme, že $f: R^+ \to R$ je měřitelná funkce, pro kterou existuje kladná reálná funkce *a* taková, že pro všechna kladná reálná čísla *x* platí

$$\lim_{t\to\infty} \frac{f(tx)-f(t)}{a(t)} = \tau(x), \qquad (2.7)$$

kde $\tau(.)$ je nekonstantní funkce definovaná takto:

$$\tau(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0, \\ \log(x), & \gamma = 0, \end{cases} \text{ pro všechny } x \in R.$$

Navíc je $a \in R_\gamma$.

Potom funkci f nazveme rozšířeně pravidelně se měnící funkcí. Píšeme $f \in ER_\gamma$ a funkci *a* nazveme pomocnou funkcí pro *f*.

Pro případ $\gamma = 0$ říkáme, že *f* náleží třídě Π a píšeme $f \in \Pi$ nebo $f \in \Pi(a)$.

2.4. Volba normalizované posloupnosti a_n a b_n

Uveďme nyní dvě základní věty s komentářem. Jde o tvrzení formulovaná a dokázaná Laurensem de Haanem.

Věta 2.12 (de Haan, Ferreira (2006), Věta 1.1.2)

Buďte $\{a_n > 0\}$ a $\{b_n\}$ dvě reálné posloupnosti a G nedegenerovaná d. f. Následující výroky jsou ekvivalentní:

i. $$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \qquad (2.8),$$

ve všech bodech spojitosti *G*.

ii. $$\operatorname{Lim}_{t\to\infty} t\{1 - F(a(t)x + b(t))\} = -\log G(x) \qquad (2.9)$$

ve všech bodech spojitosti *G*, pro které $0 < G(x) < 1$, $a(t) = a_{[t]}$, $b(t) = b_{[t]}$ ($[t]$ označuje celou část *t*).

iii. $$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D(x) \qquad (2.10)$$

pro všechny kladné body spojitosti funkce $D(x) = G^{-1}\left(\exp\left(-\frac{1}{x}\right)\right)$, $a(t) = a_{[t]}$, $b(t) = b_{[t]}$.

Z předchozích tvrzení již víme, že $G(x) = G_\gamma(x)$ pro jisté γ. Protože je tvar této funkce znám, můžeme popsat i její inversní funkci ${G_\gamma}^{-1}(x)$ jako:

$${G_\gamma}^{-1}(x) = \begin{cases} \frac{1}{\gamma(-\log x)^\gamma} - \frac{1}{\gamma}, & \gamma \neq 0 \\ -log(-logx), & \gamma = 0 \end{cases},$$

pro 0 < x < 1. Dále můžeme určit i funkci $D(x)$:

$$D(x) = G^{-1}\left(\exp\left(-\frac{1}{x}\right)\right) = D_\gamma(x) = \begin{cases} \dfrac{1}{\gamma\left(-log\left(exp\left(-\frac{1}{x}\right)\right)\right)^\gamma} - \dfrac{1}{\gamma} = \dfrac{x^\gamma - 1}{\gamma}, & \gamma \neq 0 \\ -log\left(-log\left(exp\left(-\frac{1}{x}\right)\right)\right) = logx, & \gamma = 0 \end{cases} \qquad (2.11)$$

pro všechna $x > 0$.

Na základě těchto výpočtů je možno přeformulovat předchozí větu na následující tvrzení.

Věta 2.13 (de Haan, Ferreira (2006), věta 1.1.6)

Nechť $\gamma \in R$ a uvažujme d. f. G_γ jako v předchozích tvrzeních. Potom jsou následující výroky ekvivalentní:

i. Existují dvě reálné posloupnosti $\{a_n > 0\}$, $\{b_n\}$ takové, že

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \qquad (2.12)$$

ve všech bodech spojitosti funkce G_γ.

ii. Existuje kladná funkce a taková, že pro $x > 0$ platí

$$\lim_{t\to\infty} \frac{U(tx)-b(t)}{a(t)} = D_\gamma(x) = \frac{x^\gamma-1}{\gamma}, \tag{2.13}$$

kde pro $\gamma = 0$ je pravá strana interpretována jako $\log x$.

iii. Existuje kladná funkce a taková, že

$$\lim_{t\to\infty} t\{1-F(a(t)x+U(t))\} = -\log G_\gamma(x) = (1+\gamma x)^{-\frac{1}{\gamma}},$$

ve všech bodech x spojitosti funkce G_γ, kde navíc $1+\gamma x > 0$.

iv. Existuje kladná funkce $g(x)$ taková, že

$$\lim_{x\to x^*} \frac{\bar{F}(t+xg(t))}{\bar{F}(t)} = (1+\gamma x)^{-\frac{1}{\gamma}}, \tag{2.14}$$

pro každé x, pro které je $1+\gamma x > 0$. Navíc posloupnosti $\{a_n > 0\}$, $\{b_n\}$ uvedené v části i) této věty můžeme určit jako

$$a_n = a(n), \qquad b_n = U(n).$$

Funkce g z předchozího vztahu je určena takto: $g(t) = a\left(\frac{1}{\bar{F}(t)}\right)$.

Podmínka v části ii) byla poprvé uvedena v de Haan (1984) a nazývá se **podmínkou prvního řádu** (first order condition). Z části ii) dále plyne, že $F \in D(G_\gamma) \Leftrightarrow U \in ER_\gamma$.

Laurens de Haan (1984) upravil také postačující podmínky, které formuloval von Mieses, s použitím kvantilové funkce chvostu U.

Věta 2.14

Uvažujme platnost podmínek věty 2.4. Jestliže

$$\lim_{x\to x^*} \left(\frac{1}{h(x)}\right)' = \gamma,$$

nebo ekvivalentně

$$\lim_{x\to x^*} \frac{1}{h(x)} \frac{f'(x)}{f(x)} = -\gamma-1 \tag{2.15}$$

nebo užitím kvantilové funkce chvostu

$$\lim_{t\to\infty} \frac{t\,U''(t)}{U'(t)} = \gamma - 1,$$

potom je $F \in D(G_\gamma)$.

Podobně je možné pomocí kvantilové funkce chvostu *U* přepsat i nutnou a postačující podmínku formulovanou Gnědenkem.

Věta 2.15

Nechť $\gamma \neq 0$. Potom $F \in D(G_\gamma) \Leftrightarrow$

1. Pro $\gamma > 0$ a pro $x > 0$

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma \Leftrightarrow U \in R_\gamma;$$

2. Pro $\gamma < 0$ a pro $x > 0$

$$\lim_{t\to\infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = x^\gamma \Leftrightarrow U(\infty) - U \in R_\gamma$$

Důkazy obou předchozích vět lze nalézt v de Haan, Ferreira (2006). Obě věty ukazují vztah funkce *U* a třídy pravidelně se měnících funkcí.

Posledním problémem, který zbývá vyřešit, je nalezení vhodných posloupností $\{a_n > 0\}$, $\{b_n\}$, které potřebujeme pro normalizaci řešeného problému. Jak už víme, volba těchto posloupností není jednoznačná de Haan, Ferreira (2006).

Věta 2.16 (Volba vhodných posloupností, Gnedenko (1943), de Haan, Ferreira (2006))

Nechť $F \in D(G_\gamma)$, potom

Třída I. $\gamma = 0$,

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

pro libovolné reálné *x* zvolíme

$$a_n = F^{-1}\left(1 - \frac{1}{ne}\right) - F^{-1}\left(1 - \frac{1}{n}\right) = U(ne) - U(n)$$
$$b_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n)$$

Třída II. $\gamma > 0$,

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

pro všechna x > 0 zvolíme

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n)$$
$$b_n = 0$$

Třída III. $\gamma < 0$

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

pro všechna x < 0 zvolíme

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

Tuto konstrukci posloupností $\{a_n > 0\}$, $\{b_n\}$ používáme v případě, jestliže můžeme předpokládat, že d. f. *F* náleží do jedné konkrétní třídy.

Pro většinu důležitých a známých náhodných veličin uvedené posloupnosti umíme zkonstruovat.

Konstrukcí posloupností $\{a_n > 0\}$, $\{b_n\}$ se zabýval také Laurens de Haan. Využil postačující podmínky von Mises uvedené ve větě 2.7 a získal následující tvrzení.

Věta 2.17

Nechť $F \in D(G_\gamma)$ a nechť jsou splněny předpoklady věty 2.7.

Potom

$$b_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n)$$

$a_n = \frac{1}{h(b_n)} = \frac{1}{nf(b_n)} = n\,U'(n).$

Musíme ovšem podotknout, že takto vytvořené posloupnosti jsou jiné než posloupnosti v předchozí větě. Podstatné je však to, že v této větě nepředpokládáme žádné informace o tom, do jaké třídy patří d. f. *F*.

2.5. Odhady parametrů

Odhady parametrů výsledného extremálního rozdělení jsou počátečním bodem v statistickém vyvozování závěrů z hodnot dané populace. Existují dva základní přístupy.

Klasický přístup je **parametrický**, kdy se snažíme za pomoci předpokladu znalosti daného rozdělení a pomocí standardních metod najít vztah mezi daty a parametry očekávané limity výběrových maxim. Základní úlohy pak spadají do oblasti teorie odhadu a do oblasti testování hypotéz. V této práci se nebudeme těmito přístupy zabývat.

Druhou možností je **semiparametrický** přístup. Parametrický přístup vzbuzuje v některých situacích otázky a pochyby o správnosti předpokladů. Při výpočtech používáme empirickou d. f., která aproximuje skutečnou d. f. F a parametrický model je potom nutno upravit na asymptoticko parametrický. Výsledky se potom zdají být nepřesvědčivé a nerealistické.

Navíc v mnohých aplikacích EVT není hlavním problémem popis dat pomocí teoretického a nerealistického modelu, ale podstatou je popis „chování“ extrémních hodnot.

V sedmdesátých letech minulého století počínaje pionýrskými pracemi Pickandse (1975) a Hilla (1975) se objevil nový přístup pro statistickou inferenci v EVT. Tento nový přístup se nazývá semiparametrický přístup. Tento termín odráží skutečnost, že hlavním zájmem zůstává odhad parametrů extrémních událostí speciálně EVI, zároveň však užíváme částečné předpoklady o neznámé d. f. F. Počátek vývoje těchto metod je svázán především se jménem Laurens de Haan.

Při používání semiparametrického přístupu nemáme k dispozici žádný parametrický model, který by závisel na parametru tvaru γ. Nepředpokládáme nic ani o tvaru d. f. F. Podstatné jsou předpoklady o chování chvostu, z nichž chceme získat podstatné závěry. Pomocí tohoto přístupu získáme informace o tom, do jaké třídy patří d. f. F.

Z podmínek věty 2.12 vyplývá, že

$$F \in D(G_\gamma) \Leftrightarrow \lim_{x \to x^*} \frac{\bar{F}(t+xg(t))}{\bar{F}(t)} = (1+\gamma x)^{-\frac{1}{\gamma}},$$

pro všechna x, kde je $1+\gamma x > 0$.

Dále platí

$$\frac{\bar{F}(t+xg(t))}{\bar{F}(t)} = \bar{F}_{X|X>t}(t+g(t)x)$$

Tato rovnost inspirovala autory článků Balkema, de Haan (1974) a Pickands (1975) k zavedení zobecněného Paretova rozdělení.

Definice 2.18 (Zobecněné Paretovo rozdělení)

Zobecněné Paretovo rozdělení (GP) je popsáno pomocí následující distribuční funkce

$$H_\gamma(x) = \begin{cases} 1-(1+\gamma x)^{-\frac{1}{\gamma}}, 1+\gamma x > 0, x \geq 0, & \gamma \neq 0 \\ 1-\exp(-x), x \geq 0, & ,\gamma = 0 \end{cases} \quad (2.16)$$

Pro $\gamma < 0, \gamma = 0, \gamma > 0$ se GP redukuje postupně na beta rozdělení, exponenciální rozdělení a Paretovo rozdělení.

Z definice 2.18 vyplývá, že

$$\lim_{x\to x^*} \frac{\bar{F}(t+xg(t))}{\bar{F}(t)} = \lim_{x\to x^*} \bar{F}_{X|X>t}(t+g(t)x) = \overline{H_\gamma}(x) =$$
$$= \begin{cases} (1+\gamma x)^{-\frac{1}{\gamma}}, 1+\gamma x > 0, x \geq 0 & ,\gamma \neq 0 \\ \exp(-x), x \geq 0 & ,\gamma = 0 \end{cases}$$

nebo ekvivalentně

$$\lim_{x\to x^*} F_{X|X>t}(t+g(t)x) = H_\gamma\left(\frac{x-t}{g(t)}\right),$$

kde $H_\gamma(.)$ je GP a $g(t)$ je funkce z Věty 2.12 .

Z poslední rovnosti vyplývá, že pro dostatečně velké $x > t$, může být chvost d. f. F aproximován výrazem

$$\bar{F}(x) \cong \bar{F}(t)\left(1 - H_\gamma\left(\frac{x-t}{g(t)}\right)\right), \quad (2.17)$$

Z tohoto vztahu je zřejmé, že hodnotu $\bar{F}(x)$ lze pro dostatečně velká *x* aproximovat vhodnou volbou dostatečně velké hodnoty *t*. Většinou ji volíme jako $X_{n-k:n}$. Tato hodnota se obecně nazývá práh (threshold), inference je potom založena na nejvyšších k+1 pořádkových statistikách a ne jen na výběrovém maximu M_n. Jde o správnou myšlenku, protože je nerealistické předpokládat, že jen výběrové maximum (jedno pozorování) obsahuje podstatnou informaci o chvostu distribuční funkce. Je proto vhodnější inferenci založit na množině nejvyšších pořadových statistik.

Intuitivně, jestliže se bude zvětšovat *n*, bude se zvětšovat i *k*. Volba hodnoty *k* bude potom zřejmě záviset na změně *n* takto:

1. $k = k_n$, kde $k = k_n \to \infty$, jestliže $n \to \infty$, (2.18)
2. $\frac{k_n}{n} \to 0$, jestliže $n \to \infty$. (2.19)

Posloupnost k_n nazveme prostřední (intermediate), jestliže splňuje předchozí dvě podmínky. Podobně pořádkovou statistiku nazveme prostřední pořádkovou statistikou, jestliže hodnota k splňuje obě předchozí podmínky.

Určení hodnoty k je důležitá otázka v semiparametrickém přístupu. Volba správné hodnoty k není jednoduchá a mnoho autorů nabízí různá řešení, žádné z nich ale nemůžeme přijmout obecně.

Nechť $\{X_1, X_2, \dots, X_n\}$ jsou i.i.d. náhodné veličiny s d. f. *F*. Nevychýleným odhadem *F* je empirická distribuční funkce

$$F_n(x) = \frac{\sum_{i=1}^{n} I_{(-\infty,x>}(x_i)}{n}, \tag{2.20}$$

kde $I_A(.)$ je charakteristická funkce množiny *A*. Užijeme–li aproximaci chvostu d. f. *F* pro hodnotu $t = X_{n-k:n}$ a zvolíme – li $k = k_n \to \infty$, $\frac{k}{n} \to 0, n \to \infty$ získáme následující vztah

$$\bar{F}(x) \cong \bar{F}(X_{n-k:n})\left(1 - H_\gamma\left(\frac{x - X_{n-k:n}}{g(X_{n-k:n})}\right)\right), x > X_{n-k:n},$$

využijeme – li nyní dále vztah $\bar{F}(X_{n-k:n}) \cong \bar{F}_n(X_{n-k:n}) = \frac{k}{n}$ a fakt, že

$$g(X_{n-k:n}) = a\left(\frac{1}{\bar{F}(X_{n-k:n})}\right),$$

získáme po úpravě vztah

$$\bar{F}(x) \cong \frac{k}{n}\left(1 - H_\gamma\left(\frac{x - X_{n-k:n}}{a\left(\frac{n}{k}\right)}\right)\right), x > X_{n-k:n}.$$

Tedy abychom zjistili hodnotu poslední aproximace, potřebujeme odhadnout „shape“ parametr γ a konstantu $a\left(\frac{n}{k}\right)$. Tato aproximace je správná pro hodnoty *x*, které jsou větší než $X_{n-k:n}$ a dokonce i pro hodnoty větší než $X_{n:n}$.

V následující části se budeme zabývat jen nejznámějšími semiparametrickými odhady parametru γ. Půjde tedy o omezený výčet těchto odhadů, množina všech možných odhadů je příliš široká, některé jsou vytvořeny jen unikátně pro určitý typ dat, jiné jsou především úpravou právě těch nejznámějších, práce proto nemůže popsat zdaleka všechny. Důležité je popsat ty odhady, které jsou významné i pro jejich vlastnosti.

Dosud jsme vycházeli z požadavků zformulovaných ve větě 2.11, část iii), které se nazývají podmínky prvního řádu. Pro další vlastnosti například asymptotickou normalitu je nutné tyto podmínky rozšířit. Systém těchto podmínek se nazývá **podmínky druhého řádu**.

Tyto podmínky nám garantují vhodné vlastnosti odhadů EVI. Podmínky prvního řádu znamenají, že

$$F \in D(G_\gamma) \Leftrightarrow \lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D_\gamma(x) = \begin{cases} \dfrac{x^\gamma - 1}{\gamma}, & \gamma \neq 0, \\ \log x\,, & \gamma = 0, \end{cases}$$

pro každé $x > 0$ a pro kladnou měřitelnou pomocnou funkci a, pro niž platí $a \in R_\gamma$.

Jak jsme uvedli výše, s rostoucí velikostí výběru n se také zvětšuje hodnota k - počet prostředních pořadových statistik použitých pro odhad γ. Ovšem příliš velké k vede ke zvýšení rozptylu v odhadech EVI. K tomu, abychom tento rozptyl řídili, musíme získat dodatečnou informaci o chvostu d. f. *F*. Tato informace je známá podmínky jako podmínky druhého řádu. Pomocí těchto podmínek můžeme nalézt resp. odhadnout rychlost konvergence podmínek prvního řádu. Při volbě k použijeme tedy právě tyto podmínky, které budou rozhodovat o přiměřené rychlosti konvergence.

Definice 2.19 (Podmínky druhého řádu – Second order conditions)

Řekneme že funkce *U* (nebo s ním asociovaná d. f. *F*) splňuje podmínky druhého řádu, jestliže existuje kladná funkce *a(t)* a jestliže existuje kladná nebo záporná funkce *A(t)* neměnící znamení v okolí nekonečna, pro níž platí navíc $\lim_{t\to\infty} A(t) = 0$ tak, že existuje funkce

$$H(x) = \lim_{t\to\infty} \frac{\frac{U(xt)-U(t)}{a(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)}, x > 0 \qquad (2.21)$$

Funkci A vystupující ve výše uvedeném výrazu nazveme pomocnou funkcí podmínky druhého řádu.

Pro další přístup je nutné určit funkci $H(x)$ a její vlastnosti. Úplnou odpověď nám dává následující věta.

Věta 2.20 (de Haan, Ferreira (2006), Věta 2.3.3 a doplněk 2.3.4)

Předpokládejme, že platí (2.21) a funkce *H* není násobkem funkce D_γ a není identicky rovná nule. Potom existuje kladná funkce a a funkce A kladná nebo záporná (neměnící znaménko v okolí nekonečna) a parametr $\rho \leq 0$ takový, že

$$\lim_{t\to\infty} \frac{\frac{U(xt)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = H_{\gamma,\rho}(x) = \begin{cases} \frac{1}{\gamma}\left(x^\gamma log x - \frac{x^\gamma - 1}{\gamma}\right) & , \rho = 0 \neq \gamma \\ \frac{1}{\rho}\left(\frac{x^\rho - 1}{\rho} - log x\right) & , \rho \neq 0 = \gamma, \\ \frac{1}{2}(log\ x)^2 & \rho = 0 = \gamma. \end{cases} \qquad (2.22)$$

pro x > 0.

Parametr ρ se označuje jako parametr druhého řádu a popisuje rychlost konvergence podmínek prvního řádu.

Navíc funkce $A(t)$ má následující vlastnost

$$\lim_{t\to\infty} \frac{A(tx)}{A(t)} = x^{\rho} ,$$

pro $x > 0$. Tedy $|A| \in R_{\rho}$.

Funkce A popisuje rychlost konvergence podmínek prvního řádu. Jestliže je $\rho < 0$ je konvergence polynomiálního typu a jestliže je $\rho = 0$ je konvergence pomalejší, například logaritmická. Rychlost konvergence posloupnosti k_n musí být ve shodě s rychlostí konvergence podmínek prvního řádu. Poznamenejme, že platnost podmínek druhého řádu implikuje platnost podmínek prvního řádu.

Dekkers, de Haan (1989) ukázali, že podmínky druhého řádu platí pro většinu známých n. v. (normální rozdělení, gama rozdělení, exponenciální rozdělení, rovnoměrné rozdělení, rozdělení Paretova typu atd.).

Pro konstrukci některých odhadů, pro stanovení intervalových odhadů EVI a stanovení rychlosti konvergence jednotlivých odhadů ke skutečným hodnotám EVI byly dále zavedeny i podmínky vyšších řádů. Detaily je možné nalézt v Gomes, de Haan, Peng (2002), Fraga Alves at all. (2003), Fraga Alves at all. (2006).

V následující části přistoupíme k popisu jednotlivých významných typů odhadů EVI. Před tímto krokem ale znovu sjednotíme označení, abychom se v dalším vyhnuli zmatkům. Nechť $\{X_1, X_2, \dots, X_n\}$ je náhodný výběr o rozsahu n a nechť F je distribuční funkce taková, že $F \in D(G_\gamma)$, pro nějaké $\gamma \in R$, kde G_γ je definovaná v (2.5). Označme dále $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$, které odpovídají pořádkovým statistikám v neklesajícím uspořádání. Naším cílem je nalézt odhad parametru γ odpovídající prostředním statistikám $X_{n-k:n}, k = 1, \dots, n-1$ (odpovídající *k + 1* nejvyšším pořadovým statistikám).

2.5.1. Hillův odhad EVI $\gamma > 0$

K zavedení Hillova odhadu je možné využít část 2. z věty 2.9. V ní je uvedena nutná a postačující podmínka pro případ, že $F \in D(G_\gamma)$ pro $\gamma > 0$:

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

Tento výraz je ekvivalentní s

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

dále použijeme integraci per partes, platí

$$\int_t^{\infty} (1-F(s))\frac{ds}{s} = \int_t^{\infty} (\log u - logt) dF(u) ,$$

nyní dosadíme do původního výrazu a je

$$\lim_{t\to\infty} \frac{\int_t^{\infty}(\log u - log t)dF(u)}{1-F(t)} = \gamma \qquad (2.23)$$

Pro odvození odhadu nyní využijeme asymptotických výsledků, nahradíme v (2.23) parametr t hodnotou prostřední pořádkové statistiky $X_{n-k:n}$ a d. f. F nahradíme empirickou distribuční funkcí F_n definovanou v (2.20). Tedy

$$\frac{\int_{X_{n-k:n}}^{\infty}(\log u - log X_{n-k:n})dF_n(u)}{1-F_n(X_{n-k:n})} = \gamma$$

Definice 2.21 Hillův odhad

Nechť že $F \in D(G_\gamma)$ a $\gamma > 0$, potom Hillovým odhadem parametru γ nazveme

$$\hat{\gamma}_H(n,k) = \frac{1}{k}\sum_{i=0}^{k-1} \log X_{n-i:n} - \log X_{n-k:n} \qquad (2.24)$$

Tento odhad byl formulován v článku Hilla (1975). V dalším uvedeme některé podstatné vlastnosti tohoto odhadu. Tento odhad je nejznámějším a nejvíce užívaným odhadem EVI, o jeho vlastnostech byla napsána celá řada článků. Je implementován v mnohých základních statistických programech.

Uveďme nejprve konzistenci odhadu.

Věta 2.22 (Slabá konzistence Hillova odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F. Nechť dále $F \in D(G_\gamma)$, kde $\gamma > 0$ a nechť $\{k_n\}$ je posloupnost reálných čísel splňujících (2.18) a (2.19), potom

$$\hat{\gamma}_H(n,k) \xrightarrow{P} \gamma$$

Důkaz daného tvrzení je uveden například v de Haan, Ferreira (2006), věta 3.2.2.

Věta 2.23 (Silná konzistence Hillova odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F. Nechť dále $F \in D(G_\gamma)$, kde $\gamma > 0$ a nechť $\{k_n\}$ je posloupnost čísel splňujících (2.18) a (2.19) a nechť dále platí $\frac{k_n}{log(logn)} \to \infty$, potom

$$\hat{\gamma}_H(n,k) \xrightarrow{s.j.} \gamma$$

Důkaz tohoto tvrzení je uveden např. v monografii Embrechts et all. (1997).

Důležitou teoretickou ale i praktickou vlastností Hillova odhadu je jeho asymptotická normalita.

Věta 2.24 (Asymptotická normalita Hillova odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí *F*. Nechť tato distribuční funkce splňuje podmínky druhého řádu z definice 2.19, tj. pro *x > 0*

$$\lim_{t \to \infty} \frac{\frac{U(xt)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{2.24}$$

kde $\gamma > 0$, $\rho \leq 0$, *A* je kladná nebo záporná funkce, pro kterou $\lim_{t\to\infty} A(t) = 0$.

Potom

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

kde *N* je normální rozdělení a jestliže $\{k = k_n\}$ je posloupnost čísel splňujících (2.18) a (2.19), navíc platí

$$\lim_{n\to\infty} \sqrt{k}\; A\left(\frac{n}{k}\right) = \lambda \tag{2.25}$$

Věta 2.24 je zformulována a dokázána například v de Haan, Ferreira (2006) jako věta 3.2.5.

Poznámka 2.25

V následujícím textu určíme řád konvergence posloupnosti $\{k_n\}$ pro jeden speciální případ. Nechť tedy dále je

$$1 - F(x) = c_1 x^{-\frac{1}{\gamma}} + c_2 x^{-\frac{1}{\gamma}+\frac{\rho}{\gamma}}\left(1 + o(1)\right),\ x \to \infty \tag{2.26}$$

s konstantami $c_1 > 0$, $c_2 \neq 0$, $\gamma > 0$, $\rho < 0$. Lze ověřit, že potom platí podmínky druhého řádu.

Parametr ρ řídí obecně rychlost konvergence asymptotické normality Hillova odhadu $\hat{\gamma}_H$. V případě distribuční funkce, pro kterou platí (2.26), jsou podmínky druhého řádu splněny s $A\left(\frac{1}{1-F(t)}\right) = \rho\,\gamma^{-1} c_2 {c_1}^{-1} t^{\frac{\rho}{\gamma}}$, odtud $A(t) = \rho\,\gamma^{-1} c_2\, {c_1}^{-1} t^{\frac{\rho}{\gamma}}$, navíc jestliže platí (2.25) s $\lambda > 0$, potom je možné ověřit, že

$$k_n \sim \left(\frac{\lambda\gamma}{\rho c_2} {c_1}^{1-\rho}\right)^{\frac{2}{2-\rho}} n^{-\frac{2\rho}{1-2\rho}}. \tag{2.27}$$

Rychlost konvergence ve větě 2.24 je proto řádu $n^{\frac{\rho}{1-2\rho}}$.

Hillův odhad je užíván v teoretických, ale i praktických situacích. Má však i některé negativní vlastnosti. V první řadě je zapotřebí připomenout, že není invariantní vzhledem k posunutí, naopak je invariantní vzhledem k změně měřítka. Podobně jako i u jiných odhadů

je velmi citlivý na správné určení hodnoty k. V případě příliš malých nebo příliš velkých hodnot k dochází v tomto odhadu k velkým zkreslením výsledku.

Odstranit nedostatky invariance tohoto odhadu vůči posunutí se pokusila řada autorů. Například Fraga Alves (2001) upravila Hillův odhad takto:

$$\hat{\gamma}_{H,F}(k_0,k) = \frac{1}{k_0}\sum_{i=0}^{k_0-1}\log\left(\frac{X_{n-i:n}-X_{n-k:n}}{X_{n-k_0:n}-X_{n-k:n}}\right), \tag{2.28}$$

kde $k \to \infty, k_0 \to \infty, \frac{k}{n} \to 0, \frac{k_0}{k} \to 0$. Ve výše uvedeném článku je ukázána slabá i silná konzistence i asymptotická normalita odhadu. Zároveň je zde studována optimální volba parametru k_0 .

V práci Caeiro, Gomes (2002) autoři zavedli a studovali vlastnosti upraveného Hillova odhadu

$$\hat{\gamma}_{H,CG}(k,\alpha) = \frac{\Gamma(\alpha)}{M_n^{(\alpha-1)}(k)}\left(\frac{M_n^{(2\alpha)}(k)}{\Gamma(2\alpha+1)}\right)^{\frac{1}{2}}, \tag{2.29}$$

kde $\Gamma(.)$ je gama funkce a

$$M_n^{(\alpha)}(k) = \frac{1}{k}\sum_{i=1}^{k-1}(logX_{n-i:n} - logX_{n-k:n})^{\alpha},\ \alpha > 0 \tag{2.30}$$

Odhad $\hat{\gamma}_{H,CG}$ je Hillova typu, navíc kromě vlastnosti invariance vzhledem k měřítku je invariantní vzhledem k posunutí. Odhad je slabě i silně konzistentní a asymptoticky normální.

Na tuto práci navázali Li, Peng, Nadarajah (2008) a zkonstruovali odhad následujícím způsobem

$$\hat{\gamma}_{H,LPN}(k_0,k) = \frac{\Gamma(\alpha)}{M_n^{(\alpha-1)}(k_0,k)}\left(\frac{M_n^{(2\alpha)}(k_0,k)}{\Gamma(2\alpha+1)}\right)^{\frac{1}{2}}, \alpha \geq 1 \tag{2.31}$$

kde $\Gamma(.)$ je gama funkce a

$$M_n^{(\alpha)}(k_0,k) = \frac{1}{k_0}\sum_{i=1}^{k_0-1}\left(log\frac{X_{n-i:n}-X_{n-k:n}}{X_{n-k_0:n}-X_{n-k:n}}\right)^{\alpha},\ \alpha > 0 \tag{2.32}$$

Autoři prokázali platnost slabé konzistence a asymptotické normality. Navíc se jim podařilo odvodit asymptotický rozvoj odhadu (2.31), pomocí něhož dále konstruovali intervalové odhady pro EVI.

2.5.2. Pickandsův odhad EVI $\gamma \in R$

Připojme opět krátkou motivaci k odvození tohoto odhadu, který je jedním z prvních odhadů EVI γ. Byl uveřejněn v článku Pickandse (1975) prakticky souběžně s článkem Hilla (1975).

Podle věty 2.13 část iii) je pro libovolné $\gamma \in R$ a $x > 0$

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

protože je funkce U monotónní, platí tento vztah lokálně stejnoměrně. Proto můžeme pro $0 < x, y < \infty$ psát

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{U(ty)-U(t)} = \frac{x^{\gamma}-1}{y^{\gamma}-1}, \tag{2.33}$$

Nejdříve upravíme následující zlomek

$$\frac{X_{n-k:n}-X_{n-2k:n}}{X_{n-2k:n}-X_{n-4k:n}} = \frac{X_{n-k:n}-X_{n-4k:n}}{X_{n-2k:n}-X_{n-4k:n}} - 1 \tag{2.34}$$

Zavedeme dále pomocné označení $X_{n-i:n} = U(Y_{n-i:n})$, potom mají náhodné veličiny $Y_{n-i:n}$ společnou distribuční funkci rovnou 1 – 1/x. Využijeme – li Lemma 3.3.2 v de Haan, Ferreira (2006), můžeme výraz (2.34) upravit následujícím způsobem

$$\frac{X_{n-k:n} - X_{n-2k:n}}{X_{n-2k:n} - X_{n-4k:n}} = \frac{X_{n-k:n} - X_{n-4k:n}}{X_{n-2k:n} - X_{n-4k:n}} - 1 = \frac{U\left(\frac{Y_{n-k:n}}{Y_{n-4k:n}}\, Y_{n-4k:n}\right) - U(Y_{n-4k:n})}{U\left(\frac{Y_{n-2k:n}}{Y_{n-4k:n}}\, Y_{n-4k:n}\right) - U(\, Y_{n-4k:n})} - 1,$$

Podle výše uvedeného lemma platí

$$\frac{Y_{n-k:n}}{Y_{n-4k:n}} \xrightarrow{P} 4 \text{ a } \frac{Y_{n-2k:n}}{Y_{n-4k:n}} \xrightarrow{P} 2. \tag{2.35}$$

Použijeme – li nyní (2.28) a (2.30) získáme

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

Z tohoto výsledku již můžeme odvodit Pickandsův odhad.

Poznámka 2.26

Při vyšetřování zobecněného Paretova rozdělení (definice 2.18) bychom zjistili, že hodnota $\frac{2^{\gamma}-1}{\gamma}$ je jeho mediánem a hodnota $\frac{4^{\gamma}-1}{\gamma}$ je jeho horní kvartil. Pickandsův odhad tedy vytváříme pomocí kvantilů výsledného limitního rozdělení, zatímco Hillův odhad vytváříme pomocí momentu tohoto limitního rozdělení.

Definice 2.27 Pickandsův odhad

Nechť $F \in D(G_{\gamma})$, $\gamma \in R$. Pickandsovým odhadem parametru γ označíme

$$\hat{\gamma}_P(n,k) = \frac{1}{log2}\, log\left(\frac{X_{n-k:n}-X_{n-2k:n}}{X_{n-2k:n}-X_{n-4k:n}}\right) \tag{2.36}$$

Vlastnosti tohoto odhadu zavedl a vyšetřoval nejdříve Pickands (1975) a výsledky rozšířili Dekkers, de Haan (1989).

Věta 2.27 (Slabá konzistence, Pickands (1975))

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí *F*. Nechť dále $F \in D(G_\gamma)$ a $\{k\}$ je posloupnost reálných čísel splňující podmínky (2.18) a (2.19), potom

$$\hat{\gamma}_P(n,k) \xrightarrow{P} \gamma.$$

Věta 2.28 (Silná konzistence, Dekkers, de Haan (1989))

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí *F*. Nechť dále $F \in D(G_\gamma)$ a nechť $\{k_n\}$ je posloupnost čísel splňujících (2.18) a (2.19) a nechť dále platí $\frac{k_n}{log(logn)} \to \infty$, potom

$$\hat{\gamma}_P(n,k) \xrightarrow{a.s.} \gamma.$$

Podobně jako u Hillova odhadu je Pickandsův odhad asymptoticky normální.

Věta 2.29 (Asymptotická normalita Pickandsova odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí *F*. Nechť tato distribuční funkce splňuje podmínky druhého řádu z definice 2.19, tj. pro $x > 0$

$$\lim_{t \to \infty} \frac{\frac{U(xt)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho},$$

kde $\gamma > 0$, $\rho \leq 0$, A je kladná nebo záporná funkce, pro kterou $\lim_{t\to\infty} A(t) = 0$.

Potom pro posloupnost reálných čísel $\{k = k_n\}$ splňujících (2.18) a (2.19) a

$$\lim_{n\to\infty} \sqrt{k}\ A\left(\frac{n}{k}\right) = \lambda, \tag{2.37}$$

kde λ je reálné číslo, platí

$$\sqrt{k}\,(\hat{\gamma}_P(n,k) - \gamma) \xrightarrow{d} N(\lambda\, b_{\gamma,\rho}; VAR_\gamma),$$

kde N je normální rozdělení s parametry:

$$b_{\gamma,\rho} = \begin{cases} \dfrac{4^{-\rho}\gamma\left((4^{\gamma+\rho}-1)-(2^{\gamma}+1)(2^{\gamma+\rho}-1)\right)}{\rho\, 2^{\gamma}\,(\rho+\gamma)\,(2^{\gamma}-1)\,log2}, & \rho<0, \gamma \neq 0, \\ \dfrac{1-2^{-\rho+1}+4^{-\rho}}{\rho^2\,(log2)^2}, & \rho<0, \gamma = 0, \\ 1\,, & \rho = 0, \end{cases}$$

a

$$VAR_{\gamma} = \begin{cases} \dfrac{\gamma^2\,(2^{2\gamma+1}+1)}{4\,(log2)^2\,(2^{\gamma}-1)^2}, & \gamma \neq 0, \\ \dfrac{3}{4\,(log2)^4}, & \gamma = 0. \end{cases}$$

Důkaz tohoto tvrzení je uveden například v de Haan, Ferreira (2006), věta 3.3.5.

Pickandsův odhad je snadno aplikovatelný, je relativně jednoduchý a je na rozdíl od mnohých dalších odhadů invariantní vůči posunutí i změně měřítka (scale/location invariant) a je aplikovatelný pro libovolnou hodnotu $\gamma \in R$. Zároveň ovšem má tento odhad i velmi negativní vlastnosti, je charakterizován velkým asymptotickým rozptylem a velkými změnami optimální hodnoty k, takže je velmi obtížné provést skutečnou optimální volbu prostřední pořadové statistiky vhodné k určení hodnoty odhadu. Všechny pozitivní i negativní vlastnosti inspirovali velké množství autorů k úpravám tohoto odhadu, které by některé negativní vlastnosti odstranily. Mezi mnohými uveďme například Dreese (1995) nebo Segers (2005). Drees definoval upravený odhad založený na kombinaci Pickandsových odhadů s různými volbami počátečních indexů k. Tento typ posléze zobecnil ve své práci Segers:

$$\hat{\gamma}_{P,S}(c,v) = \frac{1}{logv}\, log\left(\frac{X_{n-[ck]:n}-X_{n-k:n}}{X_{n-[cvk]:n}-X_{n-[vk]:n}}\right). \qquad (2.38)$$

Ve své práci prokázal slabou i silnou konzistenci i asymptotickou normalitu tohoto odhadu. Zároveň zjišťoval optimální volbu parametrů $c\,, v$ ve smyslu MSE.

2.5.3. Momentový odhad EVI pro $\gamma \in R$

Tento odhad je založen na takové modifikaci Hillova odhadu, aby byl platný pro libovolné reálné γ. Jestliže totiž aplikujeme Hillův odhad na případy, kdy $\gamma \leq 0$, zjistíme snadno, že pro případ, že $U(\infty) < 0$, nejsou výrazy, které se nachází v Hillově odhadu vůbec definovány. Abychom upravili hodnotu $U(\infty) > 0$, musíme posunout data. Ovšem toto posunutí má vliv na chování Hillova odhadu, který není obecně invariantní vzhledem k posunutí. Pro další potřeby zavedeme ještě následující definici.

Definice 2.30

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F. Nechť dále $F \in D(G_\gamma)$, $x^* = U(\infty) > 0$. Potom

$$M_{n,k}^{(j)} = \frac{1}{k} \sum_{i=0}^{k-1} (logX_{n-i:n} - logX_{n-k:n})^j \tag{2.39}$$

Dá se dokázat (například lemma 3.5.1 v de Haan. Ferreira (2006)), že Hillův odhad pro $\gamma \leq 0$ konverguje k nule. Nepřináší tedy pro nekladné hodnoty γ žádnou informaci. Jestliže použijeme výše zmíněné lemma 3.5.1, platí pro $\gamma < 0$ a $\{k = k_n\}$ splňujících (2.18) a (2.19)

$$\frac{\left(M_{n,k_n}^{(1)}\right)^2}{M_{n,k_n}^2} \xrightarrow{P} \frac{1-2\gamma_-}{2(1-\gamma_-)}, \tag{2.40}$$

kde $\gamma_- = \min(\gamma, 0)$ a $\gamma_+ = max(\gamma, 0)$, protože dále platí

$$\hat{\gamma}_H(n,k) \xrightarrow{P} \gamma_+, \tag{2.41}$$

můžeme vhodnou kombinací již získat odhad obecného γ. Zvolíme následující tvar

$$\hat{\gamma}_M(n,k) = M_{n,k_n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_{n,k_n}^{(1)}\right)^2}{M_{n,k_n}^2}\right)^{-1} \tag{2.42}$$

Skutečně, jestliže nyní je:

1. $\gamma > 0$. Potom je výraz $\hat{\gamma}_M(n,k) \xrightarrow{P} \gamma + 1 - \frac{1}{2}(1 - \frac{1}{2})^{-1} = \gamma + 1 - 1 = \gamma$
2. $\gamma = 0$. Potom je výraz $\hat{\gamma}_M(n,k) \xrightarrow{P} 0 + 1 - \frac{1}{2}(1 - \frac{1}{2})^{-1} = 0 + 1 - 1 = 0 = \gamma$
3. $\gamma < 0$. Potom je výraz $\hat{\gamma}_M(n,k) \xrightarrow{P} 0 + 1 - \frac{1}{2}\left(1 - \frac{1-2\gamma}{2(1-\gamma)}\right)^{-1} = 1 - \frac{1}{2}\left(\frac{1}{2(1-\gamma)}\right)^{-1} = \gamma$

Nyní tedy již můžeme zavést definici momentového odhadu parametru γ.

Definice 2.31 Momentový odhad

Nechť $F \in D(G_\gamma)$, $\gamma \in R$. Momentovým odhadem nazveme výraz

$$\hat{\gamma}_M(n,k) = M_{n,k_n}^{(1)} + 1 - \frac{1}{2}\left(1 - \frac{\left(M_{n,k_n}^{(1)}\right)^2}{M_{n,k_n}^2}\right)^{-1}.$$

Podobně jako oba předchozí odhady i momentový odhad je slabě a silně konzistentní.

Věta 2.32 (Slabá konzistence momentového odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F. Nechť dále $F \in D(G_\gamma)$ a $\{k\}$ je posloupnost reálných čísel splňující podmínky (2.18) a (2.19), potom

$$\hat{\gamma}_M(n,k) \xrightarrow{P} \gamma.$$

Důkaz tohoto tvrzení je uveden například v de Haan, Ferreira (2006) věta 3.5.2.

Věta 2.33 (Silná konzistence momentového odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F. Nechť dále $F \in D(G_\gamma)$ a nechť $\{k_n\}$ je posloupnost čísel splňujících (2.18) a (2.19) a nechť dále platí $\frac{k_n}{(log(n))^\tau} \to \infty$, pro nějaké $\tau > 0$, potom

$$\hat{\gamma}_M(n,k) \xrightarrow{a.s.} \gamma.$$

Poznámka 2.34

Odhad má svůj název proto, že je odvozen z momentů prvního a druhého řádu zobecněného Paretova rozdělení.

Věta 2.35 (Asymptotická normalita momentového odhadu)

Nechť $X_1, \dots, X_n$ je posloupnost nezávislých stejně rozdělených náhodných veličin se společnou distribuční funkcí F a $x^* > 0$. Nechť tato distribuční funkce splňuje podmínky druhého řádu z Definice 2.19 s $\gamma \neq \rho$, nechť dále posloupnost čísel $\{k = k_n\}$ splňuje (2.18) a (2.19), nechť

$$\lim_{n\to\infty} \sqrt{k}\, Q\left(\frac{n}{k}\right) = \lambda, \tag{2.43}$$

kde $\lambda > 0$ a Q je funkce definovaná pomocí podmínek druhého řádu pro $\log U(t)$ tj.

$$Q(t) = \begin{cases} A(t), & \gamma < \rho \leq 0, \\ \gamma_+ - \frac{a(t)}{U(t)}, & (\rho < \gamma \leq 0) \vee \big((0 < \gamma < -\rho) \wedge (l \neq 0)\big) \vee (\gamma = -\rho), \\ \frac{\rho}{\gamma+\rho} A(t), & \big((0 < \gamma < -\rho) \wedge (l = 0)\big) \vee (\gamma > -\rho > 0), \\ A(t), & \gamma > \rho = 0, \end{cases} \tag{2.44}$$

a hodnota $l = \lim_{t\to\infty} \left(A(t) - \frac{a(t)}{\gamma}\right)$, potom

$$\sqrt{k}(\hat{\gamma}_M(n,k) - \gamma) \xrightarrow{d} N\big(\lambda b_{\gamma,\rho}; var_\gamma\big), \tag{2.45}$$

kde N je normální rozdělení a konstanty $b_{\gamma,\rho}$ a var_γ jsou určeny takto

$$b_{\gamma,\rho} = \begin{cases} \frac{(1-\gamma)(1-2\gamma)}{(1-\gamma-\rho)(1-2\gamma-\rho)}, & \gamma < \rho \leq 0, \\ \frac{\gamma(1+\gamma)}{(1+\gamma)(1-3\gamma)}, & \rho < \gamma \leq 0, \\ \frac{-\gamma}{(1+\gamma)^2}, & (0 < \gamma < -\rho) \wedge (l \neq 0), \\ \frac{\gamma-\gamma\rho+\rho}{\rho(1-\rho)^2}, & \big((0 < \gamma < -\rho) \wedge (l \neq 0)\big) \vee (\gamma \geq -\rho > 0), \\ 1, & \gamma > \rho = 0. \end{cases}, \tag{2.46}$$

a

$$var_\gamma = \begin{cases} \gamma^2 + 1, & \gamma \geq 0, \\ \frac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}, & \gamma < 0\,. \end{cases} \tag{2.47}$$

Podobně jako u Hillova odhadu se objevila snaha upravit momentový odhad tak, aby daná úprava vedla k odhadu, který by zachoval invarianci vzhledem k měřítku a který by navíc získal invarianci vzhledem k posunutí.

Jedním ze zajímavých výsledků jsou práce Ling, Peng, Nadarajah (2007a) a Ling, Peng, Nadarajah (2007b). Autoři v nich studují modifikovaný momentový odhad, využívají základní práci Fragi Alves (2001), v níž je definován nový typ Hillova odhadu. Ten používají k odvození „nového“ momentového odhadu.

$$\hat{\gamma}_{M,LPN}(k_0,k) = M_n^{(1)}(k_0,k) + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}(k_0,k)\right)^2}{M_n^2(k_0,k)}\right)^{-1}, \tag{2.48}$$

kde

$$M_n^{(\alpha)}(k_0,k) = \frac{1}{k_0}\sum_{i=1}^{k_0-1}\left(log\frac{X_{n-i:n} - X_{n-k:n}}{X_{n-k_0:n} - X_{n-k:n}}\right)^\alpha,\ \alpha > 0$$

pro $j = 1,2$ a $\{k = k(n)\}$, $\{k_0 = k_0(n)\}$ jsou posloupnosti, pro které platí $0 < k_0 < k$ a platí pro ně podmínky (2.18) a (2.19).

V uvedených článcích byla dokázána slabá konzistence odhadu a navíc i asymptotická normalita a odvozeny vztahy, z nichž bylo možno konstruovat intervalový odhad EVI.

2.5.4. Odhady EVI založené na technice PORT (The Peak Over Random Threshold)

Většina dat, na která bychom rádi aplikovali odhady EVI, má tu vlastnost, že mají stanovenou hranici, od které probíhá měření (kolísání hladiny moře, měření teploty), proto je velmi důležité mít k dispozici odhad, který je invariantní vzhledem k posunutí a měřítku.

V předchozím textu jsme nalezli upravené odhady typu Hillova resp. momentového odhadu, které mají vlastnost invariance vzhledem k měřítku i posunutí (scale/location invariant). V této části se zaměříme na jistou techniku, která umožňuje vytvářet odhady stejných vlastností.

Tato technika a metodologie, jak s ní pracovat, byla popsána v článku Araújo, Fraga Alves, Gomes (2006). Je založena na práci s upraveným výběrem, kdy původní hodnoty zmenšíme o náhodně stanovenou mez $X_{n_q:n}$:

$$X^q = (X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \dots, X_{n_q+1:n} - X_{n_q:n}) \qquad (2.49)$$

kde $n_q = [nq] + 1$ a

1. $0 < q < 1$, pro d.f. F s konečným nebo nekonečným levým krajním bodem $x_* = inf\{x; F(x) > 0\}$, a proto je stanovená mez rovna empirickému kvantilu,
2. $q = 0$ s konečným levým krajním bodem, potom je náhodná mez volena jako výběrové minimum.

Parametr q nazýváme ladícím (tunning) parametrem.

Použitá metodologie pro tvorbu takovýchto odhadů je uvedena ve výše zmíněném článku. Zde jsou také odvozeny odhady založené na uvedené transformaci dat, které jsou nazvány PORT – Hillův odhad

$$\hat{\gamma}_{H(q)}(n,k) = \hat{\gamma}_{H(q)}(n,k)(X^q)$$

a PORT – Momentový odhad

$\hat{\gamma}_{M(q)}(n,k) = \hat{\gamma}_{M(q)}(n,k)(X^q)$. Podobně jako u klasických odhadů, kdy je podstatná volba parametru k, je v tomto případě podstatná správná volba q. Nabízí se volba $q = 0$, ale ve většině případů pomocí ní vytvoříme odhady, které nemusí být konzistentními odhady γ. Například v případě, že vytvoříme PORT – Hillův odhad pro data pocházející z modelu s nekonečným krajním bodem, vytvořený odhad není konzistentním odhadem parametru γ, viz Gomes (2007).

3. Metoda bootstrap

3.1. Úvod

Metoda bootstrap patří mezi tzv. intenzivní počítačové metody pro statistickou analýzu dat. Základní myšlenku této metody uveřejnil ve svém článku Efron (1979).

Samotná metoda bootstrap je založena na principu převzorkování výběru (resampling), tento princip nám poskytuje informace o výběrové náhodné veličině a statistice $T_n = T_n(X_1, \dots, X_n, F)$, kde $X_1, \dots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny, jejichž distribuční funkce F není nijak blíže určena. Jednoduchý princip použití vlastní metody vedl k vytvoření velkého množství teoretických i aplikačních prací. Vždyť jen citací základního článku je k dnešnímu dni více než 10 000.

Vlastní použití metody převzorkování poprvé využil Hubback (1923), zabýval se možnostmi odhadu sklizně rýže. Přímo touto metodou byla stanovena výnosnost plodin ve Velké Británii (Fisher, Yates (1945)). Práci Hubbacka rozšířil a teoreticky podpořil Mahalanobis (1931), který pracoval s korelovanými daty. Jeho metoda je vlastně základem tzv. blokového bootstrapu pro závislé údaje.

V 50-tých letech se rozvinula metoda tzv. „half – sampling“. Tato metoda byla používána při sčítání lidu v USA pro zlepšení odhadů. Ideové myšlenky jsou uvedeny v článku Gurney (1962), metodu teoreticky i prakticky rozšířil Mc Carthy (1969a, 1969b). Při vlastní aplikaci metody „half – sampling“ se postupně budovali základy využití. Především chtěli autoři základní myšlenky metody získávat relevantní informace jako z censu. V šedesátých letech se již běžně užívala. Později metodu citoval ve své práci i Efron(1979).

Vlastní myšlenku bootstrapu uveřejnil Simon (1969), vytvořil i proceduru pro použití metody Monte Carlo pro tvorbu náhodných výběrů z dané množiny dat Simon ((1974)). Bohužel se soustředil od uveřejnění článku Efrona na prokazování svého prvenství v tvorbě metody bootstrap.

Teoretickou základnou pro bootstrap je tzv. Edgeworthův rozvoj. Na přelomu 19. a 20. století na ní nezávisle pracovali Čebyšev (1890) a Edgeworth (1896, 1905). Pomocí základních principů můžeme prokázat konzistenci metody bootstrap v mnoha konkrétních obecných situacích.

3.2. Zavedení metody bootstrap

Mějme nezávislé stejně rozdělené náhodné veličiny (i.i.d) $X_1, \dots, X_n$, jejichž distribuční funkce F není nijak blíže určena. Úkolem je odhadnout neznámý parametr $\theta = \theta(F)$ náhodných veličin $X_i, i = 1, \dots, n$. Podobně bychom ovšem postupovali i v případě, že chceme nalézt konfidenční interval pro tento parametr nebo dokonce i pro stanovení testu pro tento parametr.

Nechť je $T_n = T_n(X_1, \dots, X_n, F)$ je statistika pro odhad parametru $\theta(F)$, označme dále $U_n = \sqrt{n}\,(T_n - \theta)$ jako standardizovanou verzi této statistiky. Nechť dále je

$$G_n(x) = P(U_n(X_1, \dots, X_n, F) \leq x)\,, \tag{3.1}$$

distribuční funkce statistiky U_n. Určit tuto distribuční funkci je v mnoha případech velmi obtížné nebo dokonce neproveditelné, a to dokonce, i když je distribuční funkce F známá.

V případě známé distribuční funkce F je možno využít metodu Monte Carlo a generovat velké množství nezávislých náhodných výběrů z rozdělení daného danou distribuční funkcí F, při každém opakování odhadnout hodnotu daného parametru a skutečné rozdělení tohoto parametru aproximovat pomocí empirické distribuční funkce získané z daných vypočtených parametrů.

V případě neznámé distribuční funkce F je možno postupovat metodou, kterou navrhl Efron (1979, 1982):

Nejprve zkonstruujeme klasickým způsobem empirickou distribuční funkci F_n, definovanou takto:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)\,, \tag{3.2}$$

kde $I(M)$ je charakteristická funkce (indikátor) množiny M.

Je známo, že funkce F_n jsou maximálně věrohodnými odhady distribuční funkce F, v případě, že neznáme žádný parametr neznámého rozdělení. Základním principem metody bootstrap je proto aproximace skutečné distribuční funkce F její empirickou verzí F_n. Jestliže zvolíme pevné $x \in R$, pak jsou náhodné veličiny $I(X_i \leq x)$ nezávislé stejně rozdělené náhodné veličiny s alternativním rozdělením s parametrem $F(x)$. Proto je $n\,F_n(x)$, jakožto součet n nezávislých alternativních rozdělení se stejným parametrem $F(x)$, rovno binomickému rozdělení $Bi(n, F(x))$. Odtud

$$E\,F_n(x) = F(x), \tag{3.3}$$

$$VAR\,F_n(x) = \frac{1}{n}\,F(x)(1 - F(x)), \tag{3.4}$$

Ze vztahů (3.3) a (3.4) se dá odvodit, že $F_n(x)$ je pro pevné x konzistentním odhadem distribuční funkce $F(x)$. Toto tvrzení vyplývá z Čebyševovy nerovnosti, viz např. Štěpán (1987), tvrzení IV. 1.10. Tedy pro libovolné $\epsilon > 0$ platí

$$P(|F_n(x) - F(x)| > \epsilon) \leq \frac{F(x)(1-F)x))}{n\,\epsilon^2}.$$

Proto pro pevné $x \in R$ je

$$\lim_{n\to\infty} P(|F_n(x) - F(x)| > \epsilon) = 0.$$

Aplikujeme – li na $F_n(x)$ silný zákon velkých čísel, viz např. Štěpán (1987), věta IV.2.1, zjistíme snadno, že

$$F_n(x) \xrightarrow{s.j.} F(x), \tag{3.5}$$

Naše úvahy se doposud omezovali jen na pevné *x*. Pomocí následující věty lze ukázat platnost vztahu (3.5) stejnoměrně vzhledem k *x*.

Věta 3.1 (Glivenko - Cantelli)

Nechť $X_1, \dots, X_n$ jsou nezávislé náhodné veličiny se společnou distribuční funkcí F. Označme dále F_n empirickou distribuční funkci. Potom platí

$$P\left(\lim_{n\to\infty} \sup_{x\in R} |F_n(x) - F(x)| \to 0\right) = 1.$$

Důkaz:

Štěpán (1987) věta IV. 2.3.

Nechť je dále $X_1^*, \dots, X_n^*$ je nezávislý náhodný výběr z F_n , tj. při daných pozorováních $X_1, \dots, X_n$ jsou $X_1^*, \dots, X_n^*$ nezávislé stejně rozdělené náhodné veličiny, z nichž každá nabývá hodnot $X_1, \dots, X_n$ s pravděpodobností $\frac{1}{n}$.

Definice 3.2

Nechť $X_1, \dots, X_n$ jsou nezávislé náhodné veličiny popsané pomocí d. f. F a nechť $T(X_1, \dots, X_n, F)$ je funkcionál. Bootstrapovou náhodnou veličinu odvozenou z *T* definujeme

$$H_{Boot}(x) = P_{F_n}(T(X_1^*, \dots, X_n^*, F_n) \leq x). \tag{3.6}$$

Poznámka 3.3

Pokud bychom chtěli zjistit distribuční funkci (3.6) přesně, museli bychom nalézt všech n^n možností (výběr se provádí s opakováním). To je však již pro relativně malé hodnoty n v reálném čase nemožné, např. pro hodnotu *n=50* bychom museli vyšetřovat zhruba $8{,}88 \cdot 10^{84}$ možností, dokonce i kdybychom v tomto případě označili za shodné výběry ty, které se shodují až na pořadí, vyšetřovali bychom potom „jen“ $\binom{2n-1}{n}$ možností, což pro náš případ je přibližně 5.04457×10^{28}. Proto hodnotu distribuční funkce $H_{Boot}(x)$ odhadujeme pomocí metody Monte Carlo, která nám umožní použít mnohem menší počet výběrů ke stanovení odhadu výrazu (3.6). Chyby takto určeného výsledku pochází ze dvou

zdrojů: 1. Náhrada F empirickou distribuční funkcí; 2. Odhadem skutečné hodnoty (3.6) pomocí odhadu vytvořeného metodou Monte Carlo. Volbou dostatečně velké hodnoty $\boldsymbol{B}$ počtu opakování v metodě můžeme chybu 2 eliminovat. Ovšem správná volba $\boldsymbol{B}$ je těžkým matematickým problémem. V článcích Hall (1986, 1989) jsou takovéto problémy řešeny a na základě těchto výsledků je zřejmé, že správná volba hodnoty $\boldsymbol{B}$ je v mnoha případech individuální v závislosti na podmínkách případu. Obecně se doporučuje nejméně hodnota 250.

V této části textu popíšeme tvorbu bootstrapových odhadů. Při následujícím popisu postupu využijeme článek Prášková (2004). Samozřejmým cílem je nalezení odhadů charakteristik náhodné veličiny $T_n = T_n(X_1, \dots, X_n, F)$, které závisí na skutečné distribuční funkci F. Jde především o vychýlení $Bias_n = ET_n - \theta$ a dále o rozptyl $VAR\, T_n$. Pro stanovení statistických závěrů je zapotřebí určit rozdělení standardizovaných statistik $U_n(X_1, \dots, X_n, F)$. Tyto charakteristiky budeme dále odhadovat aplikací metody bootstrap. V dalším nahradíme skutečnou distribuční funkci F empirickou distribuční funkci F_n. Budeme potom odhadovat výše uvedené charakteristiky pomocí bootstrapového výběru $X^* = \{X_1^*, \dots, X_n^*\}$, který je vytvořen na základě známého rozdělení určeného empirickou distribuční funkci F_n. Tedy

$$P^*(X_i^* = X_i) = \frac{1}{n}, \quad i = 1, \dots, n.$$

Pro náhodné veličiny $X_1^*, \dots, X_n^*$ odtud dále platí,

$$E^* X_j^* = \sum_{i=1}^n X_i \, P^*\left(X_j^* = X_i\right) = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}, \tag{3.7}$$

$$VAR^* X_j^* = \sum_{i=1}^n (X_i - E^* X_i^*)^2 P^*\left(X_j^* = X_i\right). \tag{3.8}$$

$E^*(.)$, $VAR^*(.)$ označují podmíněnou střední hodnotu, respektive podmíněný rozptyl, tzn., platí

$$E^*(.) = E(.|X), VAR^*(.) = VAR(.|X),$$

kde $X = (X_1, \dots, X_n)$ je původní náhodný výběr.

Dále tedy můžeme stejným způsobem definovat $T_n^* = T_n(X_1^*, \dots, X_n^*, F_n)$ jako bootstrapový odhad parametru $\theta(F_n)$, podobně bychom mohli vytvořit bootstrapovou verzi standardizované statistiky $U_n^* = U_n(X_1^*, \dots, X_n^*, F_n)$ a její distribuční funkci H_{Boot}.

Výrazy (3.7) a (3.8) jsou definovány pro všechny hodnoty $j = 1, \dots, n$.

Jak jsme již uvedli výše, je stanovení skutečné hodnoty H_{Boot} i pro malý rozsah výběru $X^* = \{X_1^*, \dots, X_n^*\}$ téměř nemožné, protože jde o velké množství operací. Proto se využívá v dalším metoda Monte - Carlo.

Principem metody Monte – Carlo je mnohokrát (B – krát) vygenerovat bootstrapový výběr $X^* = \{X_1^*, \dots, X_n^*\}$ z původního náhodného výběru $X^* = \{X_1^*, \dots, X_n^*\}$. Tímto způsobem získáme hodnoty $T_{n,1}^*, \dots, T_{n,B}^*$. Odhad vychýlení je potom dán vztahem

$$\widehat{Bias_n^*} = \frac{\sum_{i=1}^{n} T_{n,i}^*}{B} - \theta(F_n), \tag{3.9}$$

Podobně pro výpočet odhadu rozptylu použijeme velmi podobný vztah

$$\widehat{VAR}T_n^* = \frac{1}{B}\sum_{i=1}^{n}\left(T_{n,i}^* - \frac{1}{B}\sum_{k=1}^{n} T_{n,k}^*\right)^2, \tag{3.10}$$

Při odhadu distribuční funkce H_{Boot} statistiky U_n můžeme postupovat buď pomocí verze centrální limitní věty pro stejně rozdělené náhodné veličiny, nebo využijeme opět metodu bootstrap. Bootstrapovým odhadem této distribuční funkce nazveme funkci

$$\widehat{H}^*_{Boot,n} = \frac{1}{B}\sum_{i=1}^{B} I_{[U_{n,i}^* \le x]}. \tag{3.11}$$

Příklad 3.4

Ukážeme jakým způsobem se prakticky provádí výpočet bootstrapových odhadů nebo statistik. Předpokládejme, že funkcionál $T(X_1, \dots, X_n, F) = \frac{\sqrt{n}\,(\bar{X}-\mu(F))}{\sigma(F)}$. Z (3.7) a (3.8) vyplývá, že střední hodnota a rozptyl náhodné veličiny popsané pomocí empirické distribuční funkce F_n jsou rovny $\bar{X}$ a $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}^2)$. Tedy bootstrapovým odhadem $P\left(\frac{\sqrt{n}\,(\bar{X}-\mu(F))}{\sigma(F)} \le x, F\right)$ je $P\left(\frac{\sqrt{n}\,(\bar{X_n^*}-\bar{X})}{s} \le x, F_n\right)$. Bootstrapový odhad získáme tak, že budeme B – krát provádět výběr z původního náhodného výběru $X_1, \dots, X_n$. Bootstrapový odhad $P\left(\frac{\sqrt{n}\,(\bar{X}-\mu(F))}{\sigma(F)} \le x, F\right)$ pro dané pevné x získáme jako poměr mezi počtem $\#\left\{j; \frac{\sqrt{n}\,(\bar{X_{n,J}^*}-\bar{X})}{s} \le x\right\}$ a číslem B. Tento postup se užívá při počítačovém zpracování úloh, které vedou na problematiku bootstrapu.

Je zřejmé, že celá metoda je závislá mimo jiné na správné volbě B počtu opakování v metodě Monte – Carlo. Mnoho aplikací metody bootstrap je přímo závislé na správné volbě této hodnoty. V případě využití metody bootstrap v odhadech parametrů extremálního rozdělení v části 3.3 budeme tuto volbu hodnoty B studovat důkladněji.

Pokud jsme získali tyto bootstrapové odhady parametrů či jiných charakteristik původní neznámé náhodné veličiny, z níž byl vytvořen náhodný výběr $X = (X_1, \dots, X_n)$, bude nás samozřejmě zajímat přesnost bootstrapových odhadů.

Definice 3.5

Nechť F a G jsou distribuční funkce definované na výběrovém prostoru $X = (X_1, \dots, X_n)$. Nechť dále je $\rho(F, G)$ metrika definovaná na prostoru všech distribučních funkcí na výběrovém prostoru $X = (X_1, \dots, X_n)$. Nechť jsou dále náhodné veličiny $X_1, \dots, X_n$ nezávislé, stejně rozdělené se společnou distribuční funkcí F a nechť dále je $T(X_1, \dots, X_n, F)$

je funkcionál. Nechť dále $H_n(x)$ a $H_{Boot}(x)$ jsou postupně skutečná distribuční funkce a bootstrapová distribuční funkce.

Řekneme, metoda bootstrap je slabě konzistentní vzhledem k ρ a pro T, jestliže $\rho(H_n, H_{Boot}) \xrightarrow{P} 0$, pro $n \to \infty$. Řekneme, že metoda bootstrap je silně konzistentní vzhledem k ρ a pro T, jestliže $\rho(H_n, H_{Boot}) \xrightarrow{s.j.} 0$, pro $n \to \infty$.

Nejčastěji užívanými metrikami, kterými ověřujeme konzistenci bootstrapových odhadů jsou Kolmogorovova metrika a Mallows – Wassersteinova metrika.

Kolmogorovova metrika

$$K(F, G) = \sup_{x \in R} |F(x) - G(x)|, \tag{3.12}$$

Mallows – Wassersteinova (Vasershteinova) metrika

$$\ell_2(F, G) = \inf_{\Gamma_2, F, G} (E|Y - X|^2)^{\frac{1}{2}}, \tag{3.13}$$

kde $X \sim F$, $Y \sim G$ a Γ_2, F, G je množina všech možných sdružených rozdělení vektoru (X, Y), jejichž marginální rozdělení mají postupně distribuční funkce F a G a mají konečné druhé momenty. Metrika ℓ_2 je speciálním případem metriky

$$\ell_p(F, G) = \inf_{\Gamma_p, F, G} (E|Y - X|^p)^{\frac{1}{p}}, \tag{3.14}$$

kde $X \sim F$, $Y \sim G$ a Γ_p, F, G je množina všech možných sdružených rozdělení vektoru (X, Y), jejichž marginální rozdělení mají postupně distribuční funkce F a G a mají konečné momenty p – tého řádu.

Metrika Kolmogorovova je klasickou metrikou ve smyslu matematické analýzy. ℓ_2 je metrikou, která umožňuje řešit některé statistické problémy. Platí následující tvrzení:

$\ell_2(F_n, F) \to 0 \Leftrightarrow F_n \xrightarrow{D} F$, $E_{F_n}(X^i) \to E_F(X^i)$, pro hodnoty $i = 1, 2$.

Tuto metriku můžeme použít, jestliže chceme zároveň odhadnout distribuční funkci, střední hodnotu a rozptyl.

Věta 3.6

Nechť $X_1, \dots, X_n$ jsou nezávislé, stejně rozdělené náhodné veličiny se společnou distribuční funkcí F a nechť $E_F(X_1^2) < \infty$. Nechť $T(X_1, \dots, X_n, F) = \sqrt{n}\,(\bar{X} - \mu(F))$, potom $K(H_n, H_{Boot}) \xrightarrow{s.j.} 0$ a také $\ell_2(H_n, H_{Boot}) \xrightarrow{s.j.} 0$, jestliže $n \to \infty$.

Poznámka 3.7

Předpoklad $E_F(X_1^2) < \infty$ zajišťuje, že statistika $\sqrt{n}\,(\bar{X} - \mu(F))$ splňuje předpoklady centrální limitní věty pro stejně rozdělené náhodné veličiny. Z důkazů ve výše uvedených článcích vyplývá dále, že jestliže funkcionál $T(X_1, \dots, X_n, F)$ bude splňovat předpoklad centrální limitní věty, bude bootstrap v obou metrikách aspoň slabě konzistentní.

Důkaz:

Silná konzistence pro metriku *K* je dokázána v Singh (1981). Pro metriku ℓ_2 je důkaz uveden v Bickel, Friedmann (1981). V prvním případě je důkaz založen na několika tvrzeních:

Věta 3.8 (Berry – Essenova nerovnost)

Nechť $n \in N$ a nechť jsou dány centrované nezávislé náhodné veličiny $X_1, \dots, X_n$ splňující

$$0 < S_n^2 = \sum_{i=1}^{n} VAR(X_i) < \infty$$

Potom platí

$$\sup_{x \in R} \left| P\left(\frac{1}{S_n}\sum_{i=1}^{n} X_i < x\right) - \Phi(x)\right| \le \frac{C}{S_n^3}\sum_{i=1}^{n} E|X_i|^3. \qquad (3.15)$$

(Funkce $\Phi(x)$ v tvrzení (3.15) je d.f. standardního normálního rozdělení).

Důkaz : Berry (1941), Essen (1956), Štěpán (1987), věta IV.5.1. Hodnota *C* byla postupně zpřesňována, její dolní odhad je 0,4097 a nejlepší horní 0,5600.

Jestliže v tomto tvrzení položíme $\mu = E\,X_1$ a $\sigma^2 = VAR(X_1)$, potom podle této věty platí

$$\sup_{x \in R}|H_n(x) - \Phi(x)| \le \frac{C}{\left(\sqrt{n}\sigma\right)^3}\sum_{i=1}^{n} E|X_i - \mu|^3 = \frac{C}{\sqrt{n}\sigma^3}\,E|X_1 - \mu|^3\,.$$

Jestliže $E|X_1 - \mu|^3 < \infty$, potom je $H_n(x) - \Phi(x) = O\left(n^{-\frac{1}{2}}\right)$. Toto tvrzení lze použít i na bootstrapový výběr (jedná se o nezávislé stejně rozdělené náhodné veličiny, s odlišným rozdělením vzhledem k původnímu náhodnému výběru *X*). Získáme tento odhad

$$\sup_{x \in R}|H_n(x) - H_{Boot}(x)| = \sup_{x \in R}|H_n(x) - \Phi(x) + \Phi(x) - H_{Boot}(x)| \le$$
$$\frac{C}{\sqrt{n}\sigma^3}\,E|X_1 - \mu|^3 + \frac{C}{\sqrt{n}\sigma^{*3}}\,E|X_1^* - \bar{X}|^3,$$

Z tohoto vyplývá, že konvergence bootstrapového odhadu je stejného řádu jako při aproximaci pomocí normálního rozdělení. Vylepšení řádu této aproximace se pokusíme v části věnující se **Edgeworthovu rozvoji**.

Věta 3.9. (Věta Polya)

Nechť F_n a F jsou distribuční funkce. Nechť dále $F_n \xrightarrow{D} F$, F je spojitá distribuční funkce. Potom

$$\sup_{x \in R} |F_n(x) - F(x)| \to 0 \text{ , jestliže } n \to \infty. \tag{3.16}$$

Důkaz:

Billingsley (1995), Example 8.8.

Věta 3.10 (Zygmund – Marcinkiewicz, silný zákon velkých čísel)

Nechť $Z_1, Z_2, \ldots$ jsou nezávislé stejně rozdělené náhodné veličiny se společnou distribuční funkcí F a nechť dále pro nějaké $0 < \delta < 1$ je $E_F |Z_1|^\delta < \infty$. Potom

$$n^{-\frac{1}{\delta}} \sum_{i=0}^{n} Z_i \xrightarrow{s.j.} 0. \tag{3.17}$$

Důkaz:

Marcinkiewicz, Zygmund (1939).

Nyní můžeme přistoupit k důkazu silné konzistence pro metriku K. Podle definice je

$$K(H_n, H_{Boot}) = \sup_{x \in R} \left| P_F(T_n \le x) - P_{F_n}(T_n^* \le x) \right| =$$

$$= \sup_{x \in R} \left| P_F\left(\frac{T_n}{\sigma} \le \frac{x}{\sigma}\right) - P_{F_n}\left(\frac{T_n^*}{s} \le \frac{x}{s}\right) \right| =$$

$$= \sup_{x \in R} \left| P_F\left(\frac{T_n}{\sigma} \le \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) + \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) + \Phi\left(\frac{x}{s}\right) - P_{F_n}\left(\frac{T_n^*}{s} \le \frac{x}{s}\right) \right| \le$$

$$\le \sup_{x \in R} \left| P_F\left(\frac{T_n}{\sigma} \le \frac{x}{\sigma}\right) - \Phi\left(\frac{x}{\sigma}\right) \right| + \sup_{x \in R} \left| \Phi\left(\frac{x}{\sigma}\right) - \Phi\left(\frac{x}{s}\right) \right| + \sup_{x \in R} \left| \Phi\left(\frac{x}{s}\right) - P_{F_n}\left(\frac{T_n^*}{s} \le \frac{x}{s}\right) \right| \le$$

$= I. + II. + III.$, v dalším se budeme zabývat jednotlivými částmi odhadu.

I.: Tato část konverguje k nule podle věty 3.9.

II.: Protože $s^2 \xrightarrow{s.j.} \sigma^2$, musí $s \xrightarrow{s.j.} \sigma$ (zobrazení $f: x \mapsto x^2$ je spojité). Tato část konverguje k nule, protože je funkce $\Phi(x)$ stejnoměrně spojitá.

III.: Na tuto část můžeme aplikovat větu 3.8. Tedy

$$III. \le \frac{4}{5\sqrt{n}} \cdot \frac{E_{F_n}|X_1^* - \bar{X}|^3}{\left(VAR_{F_n}(X_1^*)\right)^{\frac{3}{2}}} = \frac{4}{5\sqrt{n}} \frac{\sum_{i=1}^{n} |X_i - \bar{X}|^3}{n\, s^3} \le$$

$$\leq \frac{4}{5\, n^{3/2}\, s^3}\, 2^3 \left(\sum_{i=1}^{n} |X_i - \mu|^3 + n|\mu - \bar{X}|^3 \right) =$$

$$= \frac{32}{5\, s^3} \left(\frac{\sum_{i=1}^{n} |X_i - \mu|^3}{n^{3/2}} + \frac{|\bar{X} - \mu|^3}{n^{1/2}} \right).$$

Protože $s \to \sigma > 0$ a $\bar{X} \to \mu$, musí druhý sčítanec v závorce konvergovat s.j. k nule. Na ověření konvergence prvního sčítance použijeme větu 3.10. Položíme $Z_i = |X_i - \mu|^3$ a $\delta = \frac{2}{3}$. Potom je

$$E|Z_i|^{\delta} = E_F|X_i - \mu|^{3.2/3} = VAR_F(X_1) < \infty.$$

Tím jsme ověřili předpoklady věty 3.10. Z této věty vyplývá

$$\frac{\sum_{i=1}^{n} |X_i - \mu|^3}{n^{3/2}} = n^{-1/\delta} \sum_{i=1}^{n} Z_i \overset{s.j.}{\to} 0.$$

Tím jsme ověřili silnou konzistenci pro Kolmogorovovu metriku, protože všechny tři sčítance I. , II. a III. konvergují s.j. k nule.

Pro důkaz silné konzistence metriky ℓ_2 uvedeme skupinu tvrzení, která jsou uvedena a dokázána v Bickel, Freedman (1981).

Věta 3.11

Nechť $G_n , G \in \Gamma_2$. Potom $\ell_2(G_n, G) \to 0$ právě když

$G_n \overset{D}{\to} G$ a $\lim_{n\to\infty} \int x^i dG_n(x) = \int x^i\, dG(x)\,, \qquad i = 1,2\,.$

Věta 3.12

Nechť $G , H \in \Gamma_2$ a předpokládejme, že $Y_1 , Y_2 , \ldots , Y_n$ jsou nezávislé stejně rozdělené náhodné veličiny se společnou distribuční funkcí G a $Z_1 , Z_2, \ldots , Z_n$ jsou nezávislé stejně rozdělené náhodné veličiny se společnou distribuční funkcí H. Jestliže je $G^{(n)}$ je distribuční funkce náhodné veličiny $\sqrt{n}(\bar{Y} - \mu_G)$ a $H^{(n)}$ je distribuční funkcí náhodné veličiny $\sqrt{n}(\bar{Z} - \mu_H)$. Potom $\ell_2\left(G^{(n)}, H^{(n)}\right) \leq \ell_2(G, H)$, (pro všechna přirozená čísla n).

Věta 3.13

Nechť $X_1 , X_2 , \ldots , X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F a nechť F_n je empirická distribuční funkce. Potom $\ell_2(F_n , F) \overset{s.j.}{\to} 0$.

Vlastní důkaz silné konzistence metriky ℓ_2 vyplývá z vět 3.11, 3.12 , 3.1 a 3.13.

Ukazuje se, že předpoklad $E_F(X_1^2) < \infty$ je podstatný pro silnou konzistenci. Jde o to, že v případě, kdy není splněn výše uvedený předpoklad, může bootstrapová posloupnost

distribučních funkcí konvergovat k distribuční funkci, která není pevnou distribuční funkcí, ale pravděpodobnostní mírou. Níže uvedený výsledek ukazuje nutnou a postačující podmínku pro daný předpoklad.

Věta 3.14

Nechť $X_1, X_2, \dots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny. Potom existuje $\mu_n(X_1, X_2, \dots, X_n)$ a neklesající posloupnost reálných čísel c_n a pevná distribuční funkce $G(x)$ taková, že

$$\mathrm{P}\left(\frac{\sum_{i=1}^{n}\left(X_i^* - \mu_n(X_1, X_2, \dots, X_n)\right)}{c_n} \le x\right) \xrightarrow{s.j.} G(x) \tag{3.18}$$

právě když $E_F(X_1^2) < \infty$ a zároveň $\frac{c_n}{\sqrt{n}} \to 1$.

Důkaz věty je uveden v Athreya (1987).

Postačujícími podmínkami silné konzistence se zabýval Singh (1981). Navíc se také zabýval i rychlostí konvergence bootstrapu. Pro následující tvrzení zavedeme několik pomocných označení.

$$H_n(x) = P\left(\sqrt{n}(\bar{X}_n - \mu)\right), \qquad H_n^*(x) = P^*\left(\sqrt{n}\,(\overline{X^*}_n - \mu^*)\right),$$

$$\widetilde{H_n}(x) = P\left(\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \le x\right), \qquad \widetilde{H}_n^*(x) = P^*\left(\sqrt{n}\frac{\overline{X^*}_n - \mu^*}{\sigma^*} \le x\right),$$

kde μ, σ^2 jsou střední hodnota a rozptyl a $\mu^* = \bar{X}_n$ a $(\sigma^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Věta 3.15

Nechť $X_1, X_2, \dots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F.

Jestliže $EX_1^4 < \infty$, potom

$$\limsup_{n\to\infty} \frac{\sqrt{n}\,K(H_n^*, H_n)}{\sqrt{log(log(n))}} = \frac{\sqrt{var(X_1-\mu)^2}}{2\sigma^2\sqrt{2\pi e}}\ s.j. \tag{3.19}$$

Jestliže $E|X_1|^3 < \infty$ a F je řešetovitá, tj. existují konstanty c, h takové, že $\sum_{j=-\infty}^{\infty} P(X_1 = c + jh) = 1$, potom

$$\lim_{n\to\infty} sup\,\sqrt{n}\,K\left(\widetilde{H}_n^*, \widetilde{H}_n\right) = \frac{h}{\sqrt{2\pi}\sigma}\ s.j. \tag{3.20}$$

Jestliže $E|X_1|^3 < \infty$ a F není řešetovitá, potom

$$\sqrt{n}\,K\left(\widetilde{H}_n^*, \widetilde{H}_n\right) \xrightarrow{s.j.} 0. \tag{3.21}$$

Důkaz:

Je proveden v Singh (1981). Pomocí výrazu (3.19) můžeme odhadnout rychlost konvergence bootstrapové aproximace. Tvrzení (3.19) a (3.20) se zabývají rychlostí

bootstrapové aproximace standardizovaného výběrového průměru. Věta 3.8 udává rychlost aproximace normálním rozdělením, která je hodnoty $O\left(n^{-\frac{1}{2}}\right)$. Podle (3.21) je tedy bootstrapová aproximace pro neřešetovitá rozdělení lepší. Protože je normální rozdělení symetrické, nezachycuje v náhodném výběru informaci o šikmosti. Bootstrapová aproximace ji ovšem obsahuje. Dokonce můžeme provést korekci šikmosti pomocí prostředků Edgeworthova rozvoje. Takové vlastnosti jsou nazývány přesností druhého řádu metody bootstrap.

3.3. Bootstrap a Edgeworthův rozvoj

Ve svém článku Hall (1988) ukázal, že nutná a postačující podmínka pro vyšší přesnost bootstrapové aproximace, která je uvedená ve větě 3.15, je existence konečných absolutních momentů až do řádu 3.

V svém článku Hall (1990) uvádí příklad bootstrapu v extremálním rozdělení, který nekonverguje v distribuci k očekávanému výsledku. Podle článku Athreya (1987) autor provádí konstrukci tzv. „subsamples“, které jsou základem nové teorie „sample fraction“ v užití bootstrapu v extremálním rozdělení. Způsobem užití výše uvedené metody se budeme zabývat v části 3.3 této kapitoly.

Ve větách 3.8 a 3.15 jsme nalezli odhad řádu konvergence bootstrapové posloupnosti. Dále jsme uvedli, že v některých případech je tato konvergence lepší než konvergence pomocí centrální limitní věty. Základem takových tvrzení je tzv. Edgeworthův rozvoj.

V dalším zavedeme tento rozvoj a využijeme jej pro účely bootstrapové konvergence a načrtneme i jeho další využití. Základní informace jsou uvedeny v monografii Hall (1992).

Nechť jsou $X_1, X_2, \ldots$nezávislé, stejně rozdělené náhodné veličiny se střední hodnotou $\theta_0 = \mu$ a konečným rozptylem σ^2. Bodovým odhadem θ_0 je výběrový průměr

$$\hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n},$$

s rozptylem $\frac{\sigma^2}{n}$. Podle centrální limitní věty je tedy náhodná veličina $S_n = n^{1/2}\frac{(\hat{\theta} - \theta_0)}{\sigma}$ asymptoticky normálně rozdělena se střední hodnotou nula a rozptylem rovným jedna. Pro další potřeby označíme tuto náhodnou veličinu $N{\sim}N(0,1)$. K přesnějšímu vyjádření například intervalových odhadů střední hodnoty μ potřebujeme nalézt přesnost vyjádření $P(S_n \leq x)$ pomocí hodnot distribuční funkce N, tj. $\Phi(x)$. Jestliže označíme $F_n(x) = P(S_n \leq x)$, platí podle věty 3.1, že $F_n(x) \to \Phi(x)$, pro každé x. Odtud vyplývá, že v tomto případě k sobě konvergují příslušné charakteristické funkce. Navíc pro charakteristickou funkci $\chi_n(t)$ náhodné veličiny S_n platí $\chi_n(t) = \left(\chi\left({}^{t}/_{\sqrt{n}}\right)\right)^n$, kde $\chi(t)$ je charakteristická funkce náhodné veličiny $Y = \frac{X-\mu}{\sigma}$, konverguje k charakteristické

funkci náhodné veličiny *N(0;1)*, která je rovna $e^{-\frac{t^2}{2}}$. V následujícím vztahu je N je $N(0,1)$ potom podle konvergence charakteristických funkcí platí

$$\chi_n(t) = E\big(exp(i\,t\,S_n)\big) \to \chi(t) = E\big(exp(i\,t\,N)\big) = e^{-\frac{t^2}{2}} \tag{3.22}$$

V dalším textu zavedeme tzv. kumulanty dostatečně hladké náhodné veličiny *X*.

Definice 3.16

Nechť X je náhodná veličina s distribuční funkcí F a nechť má konečnou momentovou vytvořující funkci $\psi(t)$ v nějakém okolí nuly a $K(t) = \log \psi(t)$. Potom r-tý kumulant náhodné veličiny X je definovaný jako $\kappa_r = \frac{\mathrm{d}^{\mathrm{r}}}{\mathrm{dt}^{\mathrm{r}}} \mathrm{K(t)}|_{\mathrm{t=0}}$. Ekvivalentně jsou kumulanty hodnoty koeficientů v Taylorově řadě funkce $K(t) = \sum_{i=1}^{\infty} \kappa_i \frac{t^i}{i!}$.

Z této definice vyplývá, že

$$\chi(t) = e^{\kappa_1 it + \kappa_2 \frac{(it)^2}{2!} + \cdots + \kappa_j \frac{(it)^j}{j!} + \cdots} . \tag{3.23}$$

Protože je ale $\chi(t)$ charakteristická funkce náhodné veličiny *X*, platí pro ni také

$$\chi(t) = 1 + E(X)(it) + E(X^2)\frac{(it)^2}{2!} + \cdots + E\big(X^j\big)\frac{(it)^j}{j!} + \cdots$$

Z těchto dvou rovností lze odvodit vztah mezi kumulanty κ_j a obecnými momenty $\mu_k$náhodné veličiny X. Platí tedy

$$\sum_{j=1}^{\infty} \kappa_j \frac{(i\,t)^j}{j!} = log\left(1 + \sum_{j=1}^{\infty} E\big(X^j\big)\frac{(it)^j}{j!}\right) = \sum_{k=1}^{\infty}(-1)^{k+1}\,\frac{1}{k}\left(\sum_{j=1}^{\infty} E\big(X^j\big)\frac{(it)^j}{j!}\right)^k \tag{3.24}$$

Porovnáme – li koeficienty u mocnin $(i\,t)^j$ u první a třetí řady, můžeme odvodit vztahy u prvních kumulantů:

$$\kappa_1 = E(Y)$$

$$\kappa_2 = E(Y^2) - \big(E(Y)\big)^2 = VAR(Y)$$

$$\kappa_3 = E(Y^3) - 3\,E(Y^2)\,E(Y) + 2(E(Y))^3 = E(Y - E(Y))^3$$

$$\kappa_4 = E(Y^4) - 4\,E(Y^3)\,E(Y) - 3\big(E(Y^2)\big)^2 + 12E(Y^2)(E(Y))^2 - 6(E(Y))^4 =$$
$$= \mathrm{E(Y - E(Y))^4} - 3\,(\mathrm{E(Y - E(Y))^2})^2$$

První dva kumulanty jsou totožné se střední hodnotou a rozptylem. Třetí kumulant je až na multiplikativní konstantu roven šikmosti a podobně čtvrtý je až na multiplikativní konstantu roven špičatosti. Ze vztahu (3.24) lze i odvodit obecný vztah mezi kumulanty a obecnými momenty náhodné veličiny X. Následující rekurzivní vztah je odvozen například v Smith (1995).

Poznámka 3.17

Obecně je vztah mezi kumulanty κ_n a obecnými momenty $\mu_k = E(X^k)$ následující

$$\kappa_n = \mu_n - \sum_{k=1}^{n-1} \binom{n-1}{k-1} \mu_{n-k} \kappa_k \text{ a } \kappa_1 = E(X). \tag{3.25}$$

Jestliže budeme předpokládat, že náhodná veličina Y je normovaná tj. $E(Y) = 0 \; a \; VAR(Y) = 1$, potom můžeme její charakteristickou funkci $\chi_n(t)$ upravit takto:

$$\chi_n(t) = e^{\left(-\frac{1}{2}t^2 + n^{-\frac{1}{2}} \frac{1}{3!} \kappa_3 (it)^3 + \cdots + n^{-\frac{j-2}{2}} \frac{1}{j!} \kappa_j (it)^j + \cdots\right)} =$$

$$= e^{-\frac{t^2}{2}} \left(1 + n^{-\frac{1}{2}} r_1(it) + n^{-1} r_2(it) + \cdots + n^{-\frac{j}{2}} r_j(it) + \cdots\right). \tag{3.26}$$

Ve výrazu (3.26) jsou r_j polynomy stupně 3j s reálnými koeficienty, které závisí na hodnotách $\kappa_3, \cdots, \kappa_{j+2}$, ale jsou nezávislé na hodnotě *n*. Odtud je možné odvodit vztahy pro jednotlivé polynomy r_j. Například první dva polynomy jsou rovny

$$r_1(u) = \frac{1}{6} \kappa_3 u^3,$$

$$r_2(u) = \frac{1}{24} \kappa_4 u^4 + \frac{1}{72} \kappa_3^2 u^6 .$$

Předchozí vztahy pro hodnotu charakteristické funkce $\chi_n(t)$ můžeme dále přepsat do následujícího tvaru

$$\chi_n(t) = e^{-\frac{t^2}{2}} + n^{-\frac{1}{2}} r_1(it)\, e^{-\frac{t^2}{2}} + n^{-1} r_2(it)\, e^{-\frac{t^2}{2}} + \cdots + n^{-\frac{j}{2}} r_j(it)\, e^{-\frac{t^2}{2}} + \cdots$$

Z klasické definice charakteristické funkce plyne

$$\chi_n(t) = \int_{-\infty}^{\infty} e^{i\,t\,x}\, dP(S_n \le x) \quad \text{a}$$

$$e^{-\frac{t^2}{2}} = \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\Phi(x). \tag{3.27}$$

Na charakteristickou funkci $\chi_n(t)$ ve tvaru řady můžeme použít inversní Fourierovu transformaci a získáme tvar

$$\mathrm{P}(S_n \le x) = \Phi(x) + n^{-\frac{1}{2}} R_1(x) + n^{-1} R_2(x) + \ldots + n^{-\frac{j}{2}} R_j(x) + \cdots, \quad \text{kde} \quad R_j(x)$$

jsou funkce, jejichž Fourierova transformace je rovna $r_j(it)\, e^{-\frac{t^2}{2}}$

$$\int_{-\infty}^{\infty} e^{i\,t\,x}\, dR_j(x) = r_j(it)\, e^{-\frac{t^2}{2}}. \tag{3.28}$$

Dále odvodíme hodnotu neznámých funkcí $R_j(x)$. Nejdříve nalezneme pomocí charakteristické funkce normovaného normálního rozdělení a pomocí integrace per partes následující vztahy:

$$e^{-\frac{t^2}{2}} = \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\Phi(x) = -\frac{1}{i\,t} \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\Phi^{(1)}(x) = \left(-\frac{1}{i\,t}\right)^2 \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\Phi^{(2)}(x) =$$

$$= \cdots = \left(-\tfrac{1}{\mathrm{i\,t}}\right)^j \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\Phi^{(\mathrm{j})}(x)\,, \text{ kde } \Phi^{(\mathrm{j})}(x) = \left(\frac{d}{dx}\right)^j \Phi(x)\,.$$

Uvedené rovnosti můžeme přepsat do tvaru:

$$(i\,t)^j\, e^{-\frac{t^2}{2}} = \int_{-\infty}^{\infty} e^{i\,t\,x}\, d\{(-\mathrm{D})^{\mathrm{j}}\,\Phi\,(x)\}, \tag{3.29}$$

kde D je diferenciální operátor $\frac{d}{dx}$. Jestliže interpretujeme $r_j(-D)$ jako polynom v D, je potom $(-D)^{\mathrm{j}}\,\Phi$ diferenciální operátor. Odtud již můžeme získat

$$\int_{-\infty}^{\infty} e^{i\,t\,x}\, d\{r_j(-D)\Phi\,(x)\} = r_j(i\,t)\, e^{-\frac{t^2}{2}}\,.$$

Porovnáním předchozích vztahů získáváme základní vztah

$$R_j(x) = r_j\left(-\frac{d}{dx}\right)\,\Phi\,(x). \tag{3.30}$$

Dále lze odvodit, viz Ralston (1973, str. 119, 120) a Petrov (1975, str. 136, 137) pro $j \geq 1$

$$(-D)^j\,\Phi\,(x) = -H_{j-1}(x)\,\varphi(x), \tag{3.31}$$

kde $H_j(x)$ je Hermitův ortogonální polynom:

$$H_j(x) = j!\sum_{k=0}^{\left[\frac{j}{2}\right]} \frac{(-1)^j\, x^{j-2\,k}}{k!\;(j-2\,k)!\;2^k}\;,$$

viz Ralston (1973, str. 120) nebo pomocí rekurentní formule

$$H_0(x) = 1, H_1(x) = x,$$

$$\text{pro } j \geq 2:\; H_{j+1}(x) = 2\,x\,H_j(x) - 2\,j\,H_{j-1}(x),$$

viz Ralston (1975, str. 171).

Pro nejnižší hodnoty j jsou tyto polynomy uvedeny níže:

$H_0(x) = 1,$

$H_1(x) = x,$

$H_2(x) = x^2 - 1,$

$H_3(x) = x\,(x^2 - 3)$, …

Z předchozích rovností můžeme proto získat

$R_1(x) = -\frac{1}{6}\,\kappa_3\,(x^2 - 1)\,\varphi(x),$

a

$R_2(x) = -x\left\{\frac{1}{24}\,\kappa_4(x^2 - 3) + \frac{1}{72}\,\kappa_3^{\,2}\,(x^4 - 10\,x^2 + 15)\right\}\varphi(x).$

Pro obecné $j \geq 1$ je

$$R_j(x) = p_j(x)\,\varphi(x), \tag{3.32}$$

kde p_j je polynom stupně *3 j – 1*, který je lichý pro liché *j* a sudý pro sudé *j*.

Potom Edgeworthovým rozvojem distribuční funkce $P(S_n \leq x)$ nazveme

$$\mathrm{P}(S_n \leq x) = \Phi(x) + n^{-\frac{1}{2}}p_1(x)\varphi(x) + n^{-1}p_2(x)\varphi(x) + \ldots + n^{-\frac{j}{2}}p_j(x)\varphi(x) + \cdots, \tag{3.33}$$

kde

$$p_1(x) = -\frac{1}{6}\,\kappa_3(x^2 - 1),$$

$$p_2(x) = -x\left(\frac{1}{24}\kappa_4(x^2 - 3) + \frac{1}{72}\kappa_3^2(x^4 - 10x^2 + 15)\right).$$

Funkce $p_1(x)$ se nazývá redukce šikmosti a funkce $p_2(x)$ redukce špičatosti.

Definice 3.18

Nechť *F* je distribuční funkce náhodné veličiny *X* a $\chi(t)$ je charakteristická funkce *X*. Řekneme, že náhodná veličina *X* splňuje Cramérovu podmínku, jestliže $\limsup_{t \to \infty} |\chi(t)| < 1$.

Věta 3.19

Nechť ***X*** je náhodný výběr z absolutně spojitého rozdělení s distribuční funkcí *F* a charakteristickou funkcí $\chi(t)$, který splňuje Cramérovu podmínku a $E\left(e^{\left(\frac{X^2}{4}\right)}\right) < \infty$, potom platí

$$P(S_n \leq x) = \Phi(x) + n^{-\frac{1}{2}}p_1(x)\varphi(x) + n^{-1}p_2(x)\varphi(x) + \ldots + n^{-\frac{j}{2}}p_j(x)\varphi(x) + \cdots .$$

Z této věty můžeme dále odvodit následující tvrzení:

Věta 3.20

Nechť $\boldsymbol{X}$ je náhodný výběr z absolutně spojitého rozdělení s distribuční funkcí F a charakteristickou funkcí $\chi(t)$, který splňuje Cramérovu podmínku a $E\left(|X|^{j+2}\right) < \infty$, potom platí

$$P(S_n \leq x) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\varphi(x) + n^{-1} p_2(x)\varphi(x) + \ldots + n^{-\frac{j}{2}} p_j(x)\varphi(x) + o\left(n^{-\frac{j}{2}}\right).$$

Výše uvedené postupy lze zobecnit i na případ, kdy střední hodnotu μ a směrodatnou odchylku σ neznáme a musíme je odhadnout. Použijeme postupy uvedené v Hall (1992; věta 2.1, 2.2, 2.3 str. 52 – 67).

Uvedeme dále větu o statistikách typu hladké funkce výběrového průměru. Takovou funkcí je kromě samotného výběrového průměru i výběrový rozptyl.

Budeme pracovat s odhadem $T_n = g(\bar{X}_n)$ parametru $\theta = g(\mu)$, kde $g: \mathbb{R}^k \to \mathbb{R}$. Nechť dále σ označuje směrodatnou odchylku odhadu $\sqrt{n}\, T_n$.

Zaveďme označení pro k – rozměrný vektor $\boldsymbol{t} = (t^1, \cdots, t^k)$

$$\|\boldsymbol{t}\| = \sqrt{\sum_{j=1}^{k} (t^j)^2}$$

a charakteristickou funkci $\chi(\boldsymbol{t})$ náhodného vektoru $\boldsymbol{X}$

$$\chi(\boldsymbol{t}) = E\left(e^{i \sum_{j=1}^{k} t^j X^j}\right).$$

Věta 3.21

Nechť $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ je náhodný výběr z absolutně spojitého k – rozměrného rozdělení $\boldsymbol{X}$ s distribuční funkcí $\boldsymbol{F}$ a charakteristickou funkcí $\chi(t)$, který splňuje Cramérovu podmínku a $E\left(\|\boldsymbol{X}_1\|^j\right) < \infty$.

Nechť dále funkce $g(\overline{\boldsymbol{X}}_{\boldsymbol{n}}) \in C^{j+2}$ v okolí bodu $E(\boldsymbol{X}_1)$. Pak platí pro $j \geq 1$

$$\mathrm{P}\left(\sqrt{\mathrm{n}}\frac{g(\overline{\boldsymbol{X}}_{\boldsymbol{k}}) - g(\mu)}{\sigma} \leq x\right) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\varphi(x) +$$

$$n^{-1} p_2(x)\varphi(x) + \ldots + n^{-\frac{j}{2}} p_j(x)\varphi(x) + \mathrm{o}\left(\mathrm{n}^{-\frac{\mathrm{j}}{2}}\right), \qquad (3.33)$$

stejnoměrně vzhledem k $\boldsymbol{x}$, kde p_j jsou polynomy stupně nejvýše $3\,j - 1$, liché funkce pro sudé hodnoty j a sudé funkce pro liché hodnoty j, s koeficienty závislými na momentech náhodné veličiny $\boldsymbol{X}$ až do řádu $j + 2$.

Důkaz: Je proveden v Hall (1992), věta 2.2.

Poznámka 3.22

Věta 3.21 se nejvíce užívá pro hodnotu $j = 2$, pak v (3.33) jsou jen dva neznámé polynomy $p_1\, a\, p_2$. Koeficienty těchto polynomů se získají porovnáním s momenty náhodné veličiny X do řádu 4.

Nechť $(X_1, \cdots, X_n)$ je náhodný výběr z náhodné veličiny $\boldsymbol{X}$ s distribuční funkcí F se střední hodnotou μ, rozptylem σ^2 a nechť dále je $E(X_1 - \mu)^4 < \infty$. Nechť dále **X** splňuje Cramérovu podmínku. Označme dále

$$\beta(\boldsymbol{X}) = \frac{E(X_1 - \mu)^3}{\sigma^3} \text{ a } \gamma(\boldsymbol{X}) = \frac{E(X_1 - \mu)^4}{\sigma^4} - 3 \text{ a}$$

$$p_1(x) = \frac{\beta(\boldsymbol{X})\,(2\,x^2 + 1)}{6},$$

$$p_2(x) = -x\left(\frac{\beta^2(\boldsymbol{X})}{18}(x^4 + 2x^2 - 3) - \frac{\gamma(\boldsymbol{X})}{12}(x^2 - 3) + \frac{1}{4}(x^2 + 3)\right).$$

Potom distribuční funkci studentizované statistiky $T_n = \sqrt{n}\,\frac{\bar{X} - \mu}{s}$ můžeme vyjádřit následujícím způsobem $H_n(x) = P(T_n \le x) = \Phi(x) + \frac{p_1(x)\,\varphi(x)}{\sqrt{n}} + \frac{p_2(x)\,\varphi(x)}{n} + o\left(\frac{1}{n}\right)$, stejnoměrně vzhledem k *x*.

Důkaz :

Je proveden v Hall (1987).

Poznámka 3.23

Předchozí věta nám umožňuje odhadnout například rozdělení rozptylu náhodného výběru $(Z_1, \cdots, Z_n)$, definujeme $\boldsymbol{X}_i = (Z_i\,,\ Z_i^2)$, pro $i = 1, \cdots, n$. Nyní je třeba zvolit funkci g takto

$$g(x_1\,, x_2) = x_2 - x_1^2.$$

Odtud totiž vyplývá, že

$g(\mu) = E(Z_1^2) - \big(E(Z_1)\big) = VAR(Z_1)$, $g(\bar{Z}_n) = \overline{Z_n} - (\overline{Z_n})^2 = \frac{1}{n}\left(\sum_{j=1}^{n} Z_j^2\right) - (\overline{Z_n})^2 = s_n^2$.

Věta 3.21 nám bude sloužit pro zpřesnění odhadů mezi původním rozdělením H_n a bootstrapovým rozdělením H_n^*. Budeme dále předpokládat, že jsou splněny předpoklady věty 3.21, potom platí

$$H_n(x) = P(R_n(X_1, \cdots, X_n; F) \leq x) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\varphi(x) + n^{-1} p_2(x)\varphi(x) + o(n^{-1}),$$

bootstrapovou verzi lze zapsat ve tvaru

$$H_n^*(x) = P(R_n^*(X_1^*, \cdots, X_n^*; F_n) \leq x) = \Phi(x) + n^{-\frac{1}{2}} \widehat{p_1}(x)\varphi(x) + n^{-1} \widehat{p_2}(x)\varphi(x) + o_P(n^{-1}).$$

Dále platí stejnoměrně pro všechna *x*

$$H_n(x) - H_n^*(x) = o_P\left(n^{-\frac{1}{2}}\right). \tag{3.34}$$

Všechna tato tvrzení jsou uvedena a dokázána v Hall (1992), strana 83, 84.

Tedy bootstrapová aproximace je přesnější než klasická aproximace normovaným normálním rozdělením. Toto zvýšení přesnosti platí ovšem v případě standardizovaných nebo studentizovaných statistik (Hall (1992), strana 83,84), a to i v případě, kdy rozdělení vykazuje výraznou asymetrii.

Edgeworthův rozvoj se velmi často využívá ke stanovení intervalových odhadů veličin, které jsou asymptoticky rovny normovanému normálnímu rozdělení. Můžeme ho tedy použít také pro stanovení intervalových odhadů indexu EVI γ. Využijeme k tomu výsledek podobný závěrům věty 2.24., který je uveden v článku Peng, Qi (1997).

Věta 3.24

Jestliže budeme předpokládat platnost výrazu (2.26) a $k \to \infty, \frac{k}{n} \to 0$, potom

$$\sqrt{k}\ (\hat{\gamma}_H(n,k) - \gamma) \xrightarrow{d} N(0; \gamma^2) \tag{3.35}$$

právě když $k = o\left(n^{\frac{2(\rho-1)\gamma}{1+2(\rho-1)\gamma}}\right)$.

Potom dále

$$P\left(\frac{\sqrt{k}\ (\hat{\gamma}_H(n,k) - \gamma)}{\gamma} \leq x\right) =$$

$$= \Phi(x) + \varphi(\mathrm{x})\left(\frac{1 - \mathrm{x}^2}{3\sqrt{\mathrm{k}}} + \frac{\frac{\mathrm{c}_2}{\mathrm{c}_1}(\rho - 1)\gamma}{(1 + (\rho - 1)\gamma)\left(\frac{\mathrm{c}_2}{\mathrm{c}_1}\right)^{(1-\rho)\gamma}} \sqrt{\mathrm{k}} \left(\frac{\mathrm{n}}{\mathrm{k}}\right)^{-(\rho-1)\gamma}\right)$$

$$+ \mathrm{o}\left(\frac{1}{\sqrt{\mathrm{k}}} + \sqrt{\mathrm{k}} \left(\frac{\mathrm{n}}{\mathrm{k}}\right)^{-(\rho-1)\gamma}\right)$$

stejnoměrně v R.

Důkaz: Proveden v Peng, Qi (1997) a v Cheng, Pan (1998).

Na základě tohoto tvrzení je možné nalézt oboustranný intervalový odhad parametru γ.

Za předpokladů předchozí věty 3.24 platí pro $k \to \infty, \frac{k}{n} \to 0$

$$P\left(\hat{\gamma}_H(n,k) - \frac{x_\alpha \hat{\gamma}_H(n,k)}{\sqrt{k}} \le \gamma \le \hat{\gamma}_H(n,k) + \frac{x_\alpha \hat{\gamma}_H(n,k)}{\sqrt{k}}\right) = \alpha + \mathrm{o}\left(\frac{1}{\sqrt{\mathrm{k}}} + \sqrt{\mathrm{k}}\left(\frac{\mathrm{n}}{\mathrm{k}}\right)^{-(\rho-1)\gamma}\right).$$

Důkaz: Theorem 1. v Cheng, Peng (2001).

Pomocí velmi složitých výpočtů a odhadů bylo v článku Cheng, Pan (2000) dokázáno velmi podobné tvrzení pro momentový odhad indexu γ. Jde o Theorem 2.1 v uvedeném článku.

Věta 3.25

Nechť je splněna platnost výrazu 2.26 a $k \to \infty, \frac{k}{n} \to 0$, potom

$$P\left(\frac{\sqrt{k}\ (\hat{\gamma}_M(n,k) - \gamma)}{\gamma} \le x\right) = \Phi(x) + \frac{1}{\sqrt{\mathrm{k}}}\left(\mathrm{d}(\gamma)\ (1 - \mathrm{x}^2) - \frac{1}{\sigma(\gamma)}[\rho + \mathrm{e}(\gamma)]\right) + \mathrm{o}\left(\frac{1}{\sqrt{\mathrm{k}}}\right).$$

Hodnoty $d(\gamma), \sigma(\gamma)$ a $e(\gamma)$ jsou specifikovány při důkazu tohoto tvrzení. Jde buď o polynomy nebo racionálně lomené funkce parametru γ.

Důkazy obou předchozích vět jsou technicky velmi náročné. Lze je ovšem využít ke konstrukci intervalových odhadů indexu γ jak pro Hillův odhad, tak i pro momentový odhad.

4. Metoda bootstrap a extremální rozdělení

4.1. Úvod

V předchozí kapitole jsme zavedli metodu bootstrap a vyšetřovali jsme její využití ve statistice. Obecně bohužel nejde samotná metoda bootstrap použít přímo v případě extremálního rozdělení.

V předchozí kapitole jsme uváděli článek Singha (1981), v kterém jsou uvedeny případy, kdy bootstrap nekonverguje. Zobecnění některých výše uvedených případů je uvedeno v článku Athreya (1987). Na myšlenky uvedené v těchto článcích navázal Hall (1990a), v tomto článku se autor zabýval především situací náhodných veličin z Frèchetovské třídy. Ukázal, že v obecném případě bootstrapová posloupnost nekonverguje dokonce ani k náhodné veličině. V dalším článku Hall (1990b) ukázal, že statistika úplného bootstrapu v extremálním rozdělení má nulovou systematickou chybu (bias), jestliže je tato statistika lineární ve svých proměnných. Z toho vyplývá, že úplný bootstrap, který je použitý v případě extremálního rozdělení s nenulovou systematickou chybou, nebude konvergovat. Angus (1993) ukázal, že bootstrapová statistika z extremálního rozdělení nekonverguje k tomuto rozdělení. Jeho limitou bude ve skutečnosti pravděpodobnostní míra. Shao, Tu (1995, ex. 4, strana 123) ukázali, že použijeme – li bootstrap, který není úplný, je již konvergence v pořádku. Dále Geluk, de Haan (2002) ukázali, že jen bootstrap vytvořený z podvýběru může odhadnout správné limitní rozdělení.

Z toho vyplývá, že pro extremální rozdělení je zapotřebí speciálního přístupu k zabezpečení konvergence bootstrapové posloupnosti.

Vlastní metodu uveřejnil poprvé ve svém článku Hall (1990b), v němž se zabýval odhadem MSE a hustoty pro případ rozdělení typu extremálního. Navrhl využít metodu tzv. „sample fraction“. Metoda je popsána ve čtvrté kapitole článku, v níž se autor zabývá odhadem indexu chvostu, a spočívá v tom, že z klasického náhodného výběru o rozsahu n realizujeme postupně další náhodné výběry o takovém rozsahu n_1, pro který platí

$$n_1 \to \infty \text{ a zároveň } \frac{n_1}{n} \to 0 \text{ pro } n \to \infty.$$

Samozřejmě, že hodnotu n_1 se snažíme volit tak, abychom zároveň volili minimální velikost MSE resp. jiný ukazatel konvergence. Odtud název „optimal sample-fraction“. Vlastní metodu budeme používat pro případ využití metody bootstrap pro realizaci odhadů uvedených v předchozí kapitole – Hillův odhad, Pickandsův odhad, momentový odhad. Detaily provedení jsou uvedeny dále.

Jedna ze základních vlastností výše uvedených odhadů indexu chvostu γ je ta, že jejich asymptotická systematická chyba je závislá na velikosti výběru. Asymptotická systematická chyba roste s velikostí počtu prvků z výběru. Na druhou stranu, když je použit příliš malý počet prvků z výběru, je jejich rozptyl příliš veliký. Proto jsme postaveni před otázku optimální volby počtu prvků z výběru (jde hlavně o nejvyšší kvantily).

Výběrem optimální velikosti hodnoty n_1, která je obecně nazývána „optimal sample fraction“, se postupně zabývá celá řada autorů. V dalším uvedeme základní články, které využijeme pro naši práci.

Prvním takovým článkem, který se zabýval použitím metody sample fraction pro případ nalezení optimálního počtu prvků z výběru je Dekkers, de Haan (1993). V tomto článku autoři pracují s momentovým odhadem a zaměřují se na konkrétní případy dat, která pochází z náhodných veličin např. rovnoměrné rozdělení, Cauchyho rozdělení, exponenciálního rozdělení a zobecněného rozdělení extrémních hodnot. Pro všechna tato rozdělení autoři nachází optimální počet prvků podvýběru. Např. pro Cauchyho rozdělení je hodnota $n_1 = \left[2^{-\frac{3}{5}}.3^{\frac{6}{5}}.\left(\frac{n}{\pi}\right)^{\frac{4}{5}}\right]$, kde n je počet prvků ve výběru a $[.]$ označuje celou část daného výrazu. Momentový odhad byl zvolen proto, že ho lze použít na údaje pocházející z náhodných veličin všech tří tříd – Weibulovy, Frèchetovy a Gumbelovy.

Základem myšlenky optimální části výběru je využití vlastností tzv. parametrů podmínek druhého řádu viz např. de Haan, Ferreira (2006), které mimo jiné určují asymptotickou rychlost konvergence základní úlohy extremální statistiky. Rozvedením této myšlenky, speciálně jen pro metodu bootstrap, jsou články Draisma, de Haan, Peng, Pereira (1999) – zde je metoda aplikována na odhad Pickandsův a zároveň i na momentový odhad a články Danielsson, de Haan, Peng, de Vries (2001), Gomes, Oliveira (2001), které řeší stejnou problematiku i pro Hillův odhad.

Všechny uvedené články mají společné rysy. Prvním krokem je definice optimální hodnoty indexu $k(n)$. Při této definici je využit druhý asymptotický moment. Ukazuje se, že zjistit tuto hodnotu přímo je velmi složité, proto u všech odhadů používáme pomocné modifikované odhady $k_i(n)$, o kterých víme, že platí $k(n) \sim k_i(n)$. Tyto modifikované odhady jsou voleny tak, aby určení hodnot $k_i(n)$ bylo relativně jednoduché. Ve většině případů je zapotřebí nalézt také konsistentní odhad parametru druhého řádu ρ. Samozřejmě, že technicky se jednotlivé způsoby nalezení optimálních hodnot i algoritmů k jejich určení liší.

Naznačíme nyní jakým způsobem je obecně metoda aplikována pro dané odhady.

4.2. Optimal sample fraction pro Hillův odhad

Hillův odhad je definován v (2.24) jako

$$\hat{\gamma}_H(n,k) = \frac{1}{k}\sum_{i=0}^{k-1} \log X_{n-i:n} - \log X_{n-k:n},$$

předpokládáme dále samozřejmě, že $k = k(n) \to \infty$, $\frac{k}{n} \to 0$, jestliže $n \to \infty$. Ke studiu odhadu $\hat{\gamma}_H(n,k)$ využijeme podmínky druhého řádu viz (2.21). Pomocí následující věty vytvoříme kritérium pro určení optimální hodnoty *k*.

Věta 4.1

Nechť jsou splněny podmínky (2.21). Potom existují náhodné veličiny P_n^H, které konvergují v distribuci k normálnímu rozdělení a platí

$$\hat{\gamma}_H(n,k) - \gamma = \frac{\gamma}{\sqrt{k}} P_n^H + \frac{1}{1-\rho} A\left(\frac{n}{k}\right) + o_p\left(A\left(\frac{n}{k}\right)\right).$$

Důkaz: de Haan, Peng (1998).

Detaily důkazu jsou provedeny v průběhu důkazu věty 4.7.

V 2. kapitole jsme se mimo jiné zabývali asymptotickou normalitou $\hat{\gamma}_H(n,k)$. Řešení je uvedeno ve větě 2.24. Z tohoto tvrzení vyplývá:

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

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

jestliže navíc $\lim_{n\to\infty} \sqrt{k}\ A\left(\frac{n}{k}\right) = \infty$, potom je

$$\frac{(\hat{\gamma}_H(n,k) - \gamma)}{A\left(\frac{n}{k}\right)} \xrightarrow{P} \frac{1}{1-\rho}.$$

Můžeme tedy zjistit rychlost konvergence $\hat{\gamma}_H(n,k)$ nezávisle na $\lim_{n\to\infty} \sqrt{k}\ A\left(\frac{n}{k}\right)$.

Věta 4.2

Symbolem $AMSE\left(\hat{\gamma}_H(n,k)\right)$ označíme asymptotickou střední kvadratickou chybu. Nechť jsou splněny podmínky (2.21). Potom je hodnota asymptotické střední kvadratické chyby rovna

$$AMSE\left(\hat{\gamma}_H(n,k)\right) = \frac{\gamma^2}{k} + \frac{1}{(1-\rho)^2} A^2\left(\frac{n}{k}\right). \qquad (4.1)$$

Důkaz: de Haan, Peng (1998)

Poznámka 4.3

Optimální hodnota metody „sample fraction„ se určí jako minimum výše uvedené asymptotické střední kvadratické chyby.

Definice 4.4

Optimální hodnotu $k_1(n)$ metody sample fraction pro Hillův odhad označíme jako

$$k_1(n) = \underset{k}{arg\,min} AMSE\left(\hat{\gamma}_H(n,k)\right) \qquad (4.2)$$

Poznámka 4.5

V průběhu práce nalezneme konsistentní odhad pro $k_1(n)$.

Pro využití metody bootstrap se bude více hodit mírně upravená statistika

$$\hat{\gamma}_H^1(n,k) = M_n^{(2)}(k) - 2\left(M_n^{(1)}(k)\right)^2, \tag{4.3}$$

kde hodnota $M_n^{(\alpha)}(k)$ je definovaná v (2.30).

Pro tuto statistiku můžeme nalézt podobný rozvoj jako pro $\hat{\gamma}_H(n,k)$.

$$\hat{\gamma}_H^1(n,k) = \frac{2\gamma^2}{\sqrt{k}} P_n^{H,1} + \frac{2\gamma\rho}{(1-\rho)^2} A\left(\frac{n}{k}\right)\left(1 + o_p(1)\right), \tag{4.4}$$

kde $P_n^{H,1}$ jsou náhodné veličiny konvergující v distribuci k standardnímu normálnímu rozdělení. Odtud je asymptotická střední kvadratická chyba rovna

$$AMSE\left(\hat{\gamma}_H^1(n,k)\right) = \frac{4\gamma^4}{k} + \frac{4\gamma^2\rho^2}{(1-\rho)^4} A^2\left(\frac{n}{k}\right). \tag{4.5}$$

Položme

$$\hat{k}_1(n) = \underset{k}{arg\,min}\, AMSE\left(\hat{\gamma}_H^1(n,k)\right). \tag{4.6}$$

Pro nalezení hodnoty optimální hodnoty metody sample fraction budeme potřebovat několik dalších tvrzení.

Nejdříve zavedeme pomocná označení. Nechť dále jsou $Y_1, \dots, Y_n$ nezávislé náhodné veličiny se společnou distribuční funkcí $G(y) = 1 - \frac{1}{y}$, pro $y \geq 1$. Nechť dále je $Y_{n,1} \leq \cdots \leq Y_{n,n}$ pořádková statistika vytvořená z $Y_1, \dots, Y_n$. Označme dále $\{X_{n,n-i+1}\}_{i=1}^n \overset{d}{=} \{UY_{n,n-i+1}\}_{i=1}^n$, kde U je kvantilová funkce chvostu z definice 2.5.

V následujícím lemmatu, jsou provedeny základní pomocné výpočty, které se využívají při důkazech hlavních vět této části. Zřejmě jsou náhodné veličiny Y_i rovny Paretovu rozdělení s parametry (1,1). Je známo, že transformace $log(Y_i)$ je exponenciální rozdělení s parametrem 1. Tyto znalosti se využijí právě při důkazu následujícího lemmatu.

Lemma 4.6

Nechť $k \in (0,n)$ a $k \to \infty$. Potom platí

1) Pro $n \to \infty$ je $\frac{Y_{n,n-k}}{\frac{n}{k}} \xrightarrow{P} 1$.
2) Pro $n \to \infty$, jsou (P_n, Q_n) asymptoticky normální náhodné veličiny se střední hodnotou rovnou 0, s rozptylem 1 resp. 20 a kovariancí 4, kde

$$P_n = \sqrt{k}\left\{\frac{1}{k}\sum_{i=1}^{k} \log Y_{n,n-i+1} - \log Y_{n,n-k} - 1\right\}$$

a

$$Q_n = \sqrt{k}\left\{\frac{1}{k}\sum_{i=1}^{k}\left(\log Y_{n,n-i+1} - \log Y_{n,n-k}\right)^2 - 2\right\}.$$

Důkaz: Lemma 3.1 Deckers, de Haan (1993).

V následující větě využijeme Lemma 4.6, abychom odvodili vztah (4.1). Využijeme při tom nerovnosti (4.10).

Věta 4.7

Nechť $|A| \in R_\rho\ (\rho < 0)$, $\sigma > 0$ a $d \neq 0$. Definujme

$$k_0(n) = \underset{k}{arg\,min}\left(\frac{\sigma^2}{k} + d^2\, A^2\left(\frac{n}{k}\right)\right). \tag{4.7}$$

Potom je

$$k_0(n) = \frac{n}{s^-\left(\frac{\sigma^2}{n\,d^2}\right)}\left(1 + o(1)\right) \in R_{\frac{-2\rho}{1-2\rho}}, \text{ jestliže } n \to \infty, \tag{4.8}$$

kde s^- je inversní funkce k funkci s, která je definována takto

$$A^2(x) = \int_x^\infty s(t)dt\,\left(1 + o(1)\right), \text{jestliže } x \to \infty.$$

Existence takové monotónní funkce vyplývá z Lemma 2.9 v Dekkers, de Haan (1993). Navíc pro pevné $\delta > 0$ a pro $n \to \infty$ je

$$k_0(n) \sim E(\hat{\gamma}_H(n,k) - \gamma)^2 1_{|\hat{\gamma}_H(n,k)-\gamma|^2 < k^{\delta-\frac{1}{2}}}. \tag{4.9}$$

Důkaz: uveden v de Haan, Peng (1998).

Důkaz této věty provedeme. Využijeme při něm metody použité v článku Dekkers, de Haan (1993). Podotkněme, že i následující věty budeme řešit stejným způsobem.

Nechť platí definice 2.19, potom vztah (2.21) je ekvivalentní tomu, že funkce

$$|\log U(t) - \gamma \log t - c_0|$$

je pravidelně se měnící s indexem ρ pro jistou konstantu c_0 (viz Geluk, de Haan (1987), část II.1). Potom tedy ve vztahu (2.21) je

$$A(t) = \rho\,(\log U(t) - \gamma \log t - c_0)$$

Aplikujeme-li Potterovu nerovnost (Potter (1942)) na funkci A, nalezneme pro každé $\varepsilon \in (0,1)$ kladnou hodnotu $t_0$takovou, že pro $t \geq t_0$ a pro $t\,x \geq t_0$ platí

$$(1-\varepsilon)x^{\rho}e^{-\varepsilon|logx|}-1 \leq \frac{\log U(tx)-\log U(t)-\gamma\log x}{\frac{A(t)}{\rho}} \leq (1+\varepsilon)x^{\rho}e^{\varepsilon|logx|}-1$$

(4.10)

Jestliže nyní zaměníme hodnotu t za $Y_{n,n-k}$ a hodnotu x za $Y_{n,n-i+1}/Y_{n,n-k}$ a použijeme-li předchozí nerovnost pro $i = 1,2,\dots,k$ a dělíme-li hodnotou k získáme následující odhad

$$\hat{\gamma}_H(n,k) \approx \gamma + \frac{\gamma P_n}{\sqrt{k}} + \frac{1}{\rho}A\left(Y_{n,n-k}\right)(1\pm\varepsilon)\left\{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{Y_{n,n-i+1}}{Y_{n,n-k}}\right)^{\rho\pm\varepsilon}-1\right\}.$$

Označme

$$\left\{\frac{Y_{n,n-i+1}}{Y_{n,n-k}}\right\}_{i=1}^{k} \overset{d}{=} \{Y_i\}_{i=1}^{k}$$

dále ze slabého zákona velkých čísel vyplývá, že

$$\hat{\gamma}_H(n,k) \approx \gamma + \frac{\gamma P_n}{\sqrt{k}} + \frac{1}{\rho}(1\pm\varepsilon)\left(\frac{1}{1-\rho\pm\varepsilon}-1\right)A\left(Y_{n,n-k}\right),$$

odtud je

$$\hat{\gamma}_H(n,k) = \gamma + \frac{\gamma P_n}{\sqrt{k}} + \frac{1}{1-\rho}A\left(\frac{n}{k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right).$$

V předchozím vztahu jsme zaměnili $Y_{n,n-k}$ hodnotou $\frac{n}{k}$ a využili jsme toho, že $|A|$ je pravidelně se měnící funkce. Odtud

$$AsyE(\hat{\gamma}_H(n,k)-\gamma)^2 \approx \frac{\gamma^2}{k} + \frac{\left(A\left(\frac{n}{k}\right)\right)^2}{(1-\rho)^2}.$$

Můžeme dále předpokládat (viz Lemma 2.9 v Deckers, de Haan (1993)), že A^2 má monotónní derivaci s, která je proto pravidelně se měnící s indexem $2\rho-1$. Proto $s^{-}\left(\frac{1}{t}\right)$ (s^{-} označíme inversní funkci k s) je pravidelně se měnící s indexem $\frac{1}{1-2\rho}$. První část závěrů věty vyplývá z minimalizace pravé strany předchozí rovnice.

Druhou část závěrů věty dokážeme tak, že výraz o_P je možno zaměnit za výraz o na výběrovém prostoru. Definujeme pro $\delta_0 \in \left(0,\frac{1}{2}\right)$ následující množinu

$$E_n = \left\{\omega; |P_n| < k^{\delta_0-\frac{1}{2}} \wedge |D_n^{\pm}| < k^{\delta_0-\frac{1}{2}} \wedge \left|\frac{k}{n}Y_{n,n-k}-1\right| < k^{\delta_0-\frac{1}{2}}\right\}, \text{ kde}$$

$$D_n^{\pm} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{Y_{n,n-i+1}}{Y_{n,n-k}}\right)^{\rho\pm\varepsilon} - \frac{1}{1-\rho\pm\varepsilon}$$

Dále zvolíme ε a t_0 jako v případě užití Potterovy nerovnosti a za předpokladu $\frac{n}{k}\left(1-k^{\delta_0-\frac{1}{2}}\right) \geq t_0$ je $Y_{n.n-k} \geq t_0$. Protože je A pravidelně se měnící s indexem ρ, platí

$$\left|A(Y_{n.n-k}) - A\left(\frac{n}{k}\right)\right| < 2\,\varepsilon\,A\left(\frac{n}{k}\right)$$

na množině E_n. Nyní využijeme předchozích dvou faktů a Potterovy nerovnosti a máme

$$\left|\hat{\gamma}_H(n,k) - \gamma + \frac{\gamma P_n}{\sqrt{k}} + \frac{1}{1-\rho}\,A\left(\frac{n}{k}\right)\right| < \varepsilon\,A\left(\frac{n}{k}\right)$$

na množině E_n(můžeme tedy zaměnit $o_P(A)$ za $o(A)$). Odtud pro $n \to \infty$ a pro prostřední posloupnost $k(n)$

$$\frac{E(\hat{\gamma}_H(n,k)-\gamma)^2 1_{|\hat{\gamma}_H(n,k)-\gamma|^2<k^{\delta-\frac{1}{2}}} 1_{E_n}}{\frac{\gamma^2}{k} + \frac{\left(A\left(\frac{n}{k}\right)\right)^2}{(1-\rho)^2}} \to 1$$

Dále ukážeme, že výše uvedený druhý moment je asymptoticky roven nule na doplňku množiny E_n.

$$E(\hat{\gamma}_H(n,k)-\gamma)^2 1_{|\hat{\gamma}_H(n,k)-\gamma|^2<k^{\delta-\frac{1}{2}}} 1_{\left\{|P_n|>k^{\delta_0-\frac{1}{2}}\right\}} \leq k^{\delta-\frac{1}{2}}\,P\left(\left\{|P_n| > k^{\delta_0-\frac{1}{2}}\right\}\right)$$

Pomocí Bennettovy nerovnosti(Petrov(1975), kapitola III.5) dále platí

$$P\left(\left\{|P_n| > k^{\delta_0-\frac{1}{2}}\right\}\right) \leq k^{-c}$$

pro jistou hodnotu c. Odtud plyne

$$\lim_{n\to\infty} \frac{E(\hat{\gamma}_H(n,k)-\gamma)^2 1_{|\hat{\gamma}_H(n,k)-\gamma|^2<k^{\delta-\frac{1}{2}}}}{\frac{\gamma^2}{k}+\frac{\left(A\left(\frac{n}{k}\right)\right)^2}{(1-\rho)^2}} = 0$$

Z této rovnosti vyplývá, že

$$E(\hat{\gamma}_H(n,k)-\gamma)^2 1_{|\hat{\gamma}_H(n,k)-\gamma|^2<k^{\delta-\frac{1}{2}}} \approx \frac{\gamma^2}{k} + \frac{\left(A\left(\frac{n}{k}\right)\right)^2}{(1-\rho)^2}$$

Tím je dokázána druhá část věty.

Z této věty a z tvaru $AMSE$ pro odhady $\hat{\gamma}_H(n,k)$ a $\hat{\gamma}_H^1(n,k)$ vyplývá následující věta.

Věta 4.8

Předpokládejme, že platí podmínky druhého řádu (Definice 2.11), potom

$$k_1(n) = \frac{n}{s^{-}\left(\frac{\gamma^2(1-\rho)^2}{n}\right)}\left(1 + o(1)\right), \tag{4.11}$$

$$\hat{k}_1(n) = \frac{n}{s^{-}\left(\frac{\gamma^2(1-\rho)^4}{n\rho^2}\right)}\left(1 + o(1)\right), \tag{4.12}$$

a tedy platí

$$\frac{\hat{k}_1(n)}{k_1(n)} \sim \left(1 - \frac{1}{\rho}\right)^{\frac{1}{1-2\rho}} \tag{4.13}$$

Důkaz: Uveden v Danielsson, de Haan, Peng, de Vries, (2001).

Provedeme rozbor tohoto důkazu. Ze závěrů předchozí věty vyplývá, že

$$\hat{\gamma}_H(n,k) \overset{d}{=} \gamma + \frac{\gamma P_n}{\sqrt{k}} + \frac{1}{1-\rho}\, A\left(Y_{n,n-k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right). \tag{4.14}$$

dále je

$$(\hat{\gamma}_H(n,k))^2 \overset{d}{=} \gamma^2 + \frac{2\gamma^2 P_n}{\sqrt{k}} + \frac{2\gamma}{1-\rho}\, A\left(Y_{n,n-k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right). \tag{4.15}$$

Podobně

$$M_n^{(2)}(k) \overset{d}{=} 2\,\gamma^2 + \frac{2\gamma^2 P_n}{\sqrt{k}} + \frac{2\gamma(1-\rho)}{(1-\rho)^2} A\left(Y_{n,n-k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right). \tag{4.16}$$

Závěr důkazu je technicky totožný s předchozí větou.

Ze vztahu (4.13) plyne, že odhad $k_1(n)$ je ekvivalentní odhadu $\hat{k}_1(n)$, jestliže existuje konsistentní odhad parametru druhého řádu ρ.

V dalším se již zaměříme na tvorbu vlastní bootstrapové procedury. Vytvořme n_1 nezávislých výběrů z empirické d. f. vytvořené z $\boldsymbol{\mathcal{X}_n} = \{X_1, \cdots, X_n\}$. Získáme hodnoty $X_1^*, \cdots, X_{n_1}^*$ a vytvoříme pořádkovou statistiku $X_{1:n_1}^*, X_{2:n_1}^*, \cdots, X_{n_1:n_1}^*$ a pomocí této statistiky sestrojíme

$${}^{*}M^{(i)}(n_1,k_1) = \frac{1}{k_1}\sum_{j=1}^{k_1}\left(log\, X_{n_1-j+1:n_1}^* - log\, X_{n_1-k_1:n_1}^*\right)^i. \tag{4.17}$$

Podle našeho postupu vypočteme hodnotu pomocného odhadu

$$M_*(n_1,k_1) = {}^*M^{(2)}(n_1,k_1) - 2\left({}^*M^{(1)}(n_1,k_1)\right)^2, \tag{4.18}$$

kde $k_1 \to \infty$ a $\frac{k_1}{n_1} \to 0$ a dále $n_1 = O(n^{1-\varepsilon})$, pro hodnotu $0 < \varepsilon < 1$.

Nechť $k_{1,0}^*(n_1)$ je hodnota, kterou se minimalizuje podmíněný druhý moment $E\left(\left(M_*(n_1,k_1)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$ vzhledem k proměnné k_1. Dá se ukázat, že $k_{1,0}^*(n_1)$ má následující asymptotické vlastnosti:

Věta 4.9

Předpokládejme, že platí podmínky druhého řádu (Definice 2.11) a nechť dále je $k_1 \to \infty$ a $\frac{k_1}{n_1} \to 0$ a dále $n_1 = O(n^{1-\varepsilon})$, pro hodnotu $\varepsilon \in (0,1)$. Označme $k_{1,0}^*(n_1)$ takové, že

$$Q(n_1,k_1) = E\left(\left(M_*(n_1,k_1) - 2\left(\gamma_H^*(n_1,k_1)\right)^2\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$$

je minimální. Potom

$$\frac{k_{1,0}^*(n_1)\, s^{-\left(\gamma^2(1-\rho)^4/(n_1\rho^2)\right)}}{n_1} \xrightarrow{P} 1, \text{ jestliže } n \to \infty. \tag{4.19}$$

Důkaz: Uveden v Danielsson, de Haan, Peng, Vries (1997).

Tento důkaz provedeme.

Označme $G_n$empirickou distribuční funkci n nezávislých náhodných veličin typu rovnoměrné rozdělení. Symbolem G_n^- označme inversní funkci k G_n. Nechť n je dostatečně veliké a $n_1 = O(n^{1-\varepsilon})$, potom je (rovnice (10) a (17) v kapitole 10.5 v Shorack, Wellner (1986))

$$\frac{1}{2} \le \sup_{0 < t \le n_1(\log n_1)^2} t\, G_n^-\left(\frac{1}{t}\right) \le 2 \ \ s.v. \tag{4.20}$$

a

$$\sup_{t \ge 2} \left|\sqrt{t}\left(G_n\left(\frac{1}{t}\right) - \frac{1}{t}\right)\right| \le \frac{\log n}{\sqrt{n}} \ \ s.v.$$

Odtud je

$$\sup_{4 < t \le n_1(\log n_1)^2} \left|\sqrt{\frac{1}{G_n^-\left(\frac{1}{t}\right)}}\left[G_n\left(G_n^-\left(\frac{1}{t}\right)\right) - G_n^-\left(\frac{1}{t}\right)\right]\right| \le \frac{\log n}{\sqrt{n}} \ \ s.v.$$

Proto pro všechna $4 < t \le n_1(\log n_1)^2$ je

$$\left| t\, G_n^-\left(\frac{1}{\mathrm{t}}\right) - 1 \right| \le \frac{2\sqrt{\mathrm{t}}\log n}{\sqrt{n}} \quad s.v. \tag{4.21}$$

Označme dále F_n empirickou distribuční funkci $\boldsymbol{\mathcal{X}_n}$, F_n^- inversní funkci k F_n a

$$U_n = \left(\frac{1}{1-F_n}\right)^-.$$

Nyní využijeme předchozích vztahů

$$\begin{cases} |\log y| \le 2|y-1| & \frac{1}{2} \le y \le 2 \\ |y^{-\rho}-1| \le (-\rho)(2^{-\rho-1} \vee 2^{1+\rho}) & \frac{1}{2} \le y \le 2 \end{cases}$$

$$\log U_n(t) = \log F_n^-\left(1-\frac{1}{t}\right) \overset{d}{=} \log F^-\left(G_n^-\left(1-\frac{1}{\mathrm{t}}\right)\right) = \log U\left(\frac{1}{1-G_n^-\left(1-\frac{1}{\mathrm{t}}\right)}\right) =$$

$$\overset{d}{=} \log U\left(\frac{t}{t\, G_n^-\left(\frac{1}{\mathrm{t}}\right)}\right).$$

Odtud vyplývá, že pro každé $\varepsilon \in (0,1)$ existuje $t_0 > 4$ takové, že pro $t_0 < t < n_1(\log n_1)^2$ a $t_0 < t\,x < n_1(\log n_1)^2$ je

$$\frac{\log U_n(tx) - \log U_n(t) - \gamma\log x}{\frac{A(tx)}{\rho}} \overset{d}{=} \frac{\log U\left(\frac{tx}{txG_n^-\left(\frac{1}{\mathrm{tx}}\right)}\right) - \log U(tx) - \gamma\log\left(\frac{1}{txG_n^-\left(\frac{1}{\mathrm{tx}}\right)}\right)}{\frac{A(tx)}{\rho}} \frac{A(tx)}{A(t)} -$$

$$- \frac{\log U\left(\frac{t}{tG_n^-\left(\frac{1}{\mathrm{t}}\right)}\right) - \log U(t) - \gamma\log\left(\frac{1}{tG_n^-\left(\frac{1}{\mathrm{t}}\right)}\right)}{\frac{A(t)}{\rho}} + \frac{\log U(tx) - \log U(t) - \gamma\log x}{\frac{A(t)}{\rho}} +$$

$$+ \frac{\gamma\log\left(\frac{1}{txG_n^-\left(\frac{1}{\mathrm{tx}}\right)}\right)}{\frac{A(t)}{\rho}} - \frac{\gamma\log\left(\frac{1}{tG_n^-\left(\frac{1}{\mathrm{t}}\right)}\right)}{\frac{A(t)}{\rho}}$$

$$\le \left[(1+\varepsilon)\left(txG_n^-\left(\frac{1}{\mathrm{tx}}\right)\right)^{-\rho} e^{\varepsilon\left|\log\left(txG_n^-\left(\frac{1}{\mathrm{tx}}\right)\right)\right|} - 1\right](1+\varepsilon)x^\rho e^{\varepsilon|\log x|} -$$

$$-(1-\varepsilon)\left(tG_n^-\left(\frac{1}{\mathrm{t}}\right)\right)^{-\rho} e^{-\varepsilon\left|tG_n^-\left(\frac{1}{\mathrm{t}}\right)\right|} + 1 + (1+\varepsilon)x^\rho e^{\varepsilon|\log x|} - 1 +$$

$$\left|\frac{\gamma\rho}{A(t)}\right| 2\left(\left|txG_n^-\left(\frac{1}{\mathrm{tx}}\right)-1\right|+\left|tG_n^-\left(\frac{1}{\mathrm{t}}\right)-1\right|\right) \leq$$

$$\leq \left[(-\rho)\,(\max(2^{-\rho+1}, 2^{3+\rho}))+2\left|\frac{\gamma\rho}{A(t)}\right|\right]\frac{2\sqrt{t}\log n}{\sqrt{n}}\left(\sqrt{x}+1\right)$$
$$+\ (1+9\varepsilon)(1+\varepsilon)x^{\rho}e^{\varepsilon|\log x|}-$$
$$-1+7\varepsilon\ s.v. \qquad (4.22)$$

Podobně platí

$$\frac{\log U_n(tx)-\log U_n(t)-\gamma\log x}{\frac{A(tx)}{\rho}} \geq$$

$$\geq \left[(-\rho)\,(\max(2^{-\rho+1}, 2^{3+\rho}))+2\left|\frac{\gamma\rho}{A(t)}\right|\right]\frac{2\sqrt{t}\log n}{\sqrt{n}}\left(\sqrt{x}+1\right)$$
$$+(1-9\varepsilon)(1-\varepsilon)x^{\rho}e^{-\varepsilon|\log x|}-$$
$$-1-7\varepsilon\ \ s.v. \qquad (4.23)$$

Provedeme opět substituci hodnoty t za Y_{n_1,n_1-k_1} a hodnoty tx za Y_{n_1,n_1-i+1} $(i=1,\dots,k_1)$. Odtud vyplývá, že

$$4 \leq Y_{n_1,n_1-i+1} \leq Y_{n_1,n_1}\ \ (i=1,\dots,k_1)$$ v pravděpodobnosti

a

$$\frac{Y_{n_1,n_1}}{(n_1(\log n_1)^2)} \to 0$$ v pravděpodobnosti

pro $n_1 \to \infty$ a $\frac{k_1}{n_1} \to 0$.

Nyní budeme minimalizovat

$$E\left(\left(M_*(n_1,k_1)-2\left(\gamma_H^*(n_1,k_1)\right)^2\right)^2\middle|\boldsymbol{X_n}\right).$$

Poznamenejme, že podmíněně vzhledem k $\boldsymbol{X_n}$, je P_{n_1}je normalizováno pomocí jistého počtu nezávislých stejně rozdělených exponenciálních náhodných veličin. Proto, když n_1 roste, náhodná veličina P_{n_1}se blíží k normálnímu rozdělení. Podobné platí i o náhodné veličině Q_{n_1}.

Při důkazu dále postupujeme stejně jako v předchozí větě, užijeme dále

$$\gamma_H^*(n_1,k_1)\overset{d}{=}\gamma+\frac{\gamma P_{n_1}}{\sqrt{k_1}}+\frac{1}{1-\rho}\,A\left(Y_{n_1,n_1-k_1}\right)+o_P\left(A\left(\frac{n_1}{k_1}\right)\right)+O\left(\frac{\log n\sqrt{\frac{n_1}{k_1}}}{\sqrt{n}}\right),$$

$$\left(\gamma_H^*(n_1,k_1)\right)^2 \overset{d}{=} \gamma^2 + \frac{2\gamma^2 P_{n_1}}{\sqrt{k_1}} + \frac{2\gamma}{1-\rho}\, A\left(Y_{n_1,n_1-k_1}\right) + o_P\left(A\left(\frac{n_1}{k_1}\right)\right) + O\left(\frac{\log n\sqrt{\frac{n_1}{k_1}}}{\sqrt{n}}\right)$$

a

$$M_*(n_1,k_1) \overset{d}{=} 2\gamma^2 + \frac{\gamma^2 Q_{n_1}}{\sqrt{k_1}} + \frac{2\gamma(2-\rho)}{(1-\rho)^2}\, A\left(Y_{n_1,n_1-k_1}\right) + o_P\left(A\left(\frac{n_1}{k_1}\right)\right) + O\left(\frac{\log n\sqrt{\frac{n_1}{k_1}}}{\sqrt{n}}\right)$$

Poznamenejme, že výraz $\frac{\log n\sqrt{\frac{n_1}{k_1}}}{\sqrt{n}} = o\left(\frac{1}{\sqrt{k_1}}\right)$, při minimalizaci ho tedy můžeme zanedbat.

Výše uvedené dvě věty hrají podstatnou roli v konstrukci konzistentního bootstrapového odhadu pro $\hat{k}_1(n)$. Předpokládejme, že dále platí podmínky druhého řádu a že navíc je $A(x) = c\,x^{\rho}$. Pak je možné zvolit funkci s takto: $s(x) = -2\,\rho\,c^2 x^{2\rho-1}$. Odtud vyplývá, že $s^{-}(x) = (2\,\rho\,c^2)^{\frac{1}{1-2\rho}}\, x^{\frac{-1}{1-2\rho}}$. Přesněji:

Věta 4.10

Předpokládejme, že platí podmínky druhého řádu (Definice 2.11) pro $A(t) = c\,t^{\rho}, t \to \infty$ a nechť dále je $k_1 \to \infty$ a $\frac{k_1}{n_1} \to 0$ a dále $n_1 = O(n^{1-\varepsilon})$, pro hodnotu $\varepsilon \in (0,1)$. Potom

$$\left(\frac{n_1}{n}\right)^{\frac{2\rho}{1-2\rho}} \frac{k_1^*(n_1)}{\hat{k}_1(n)} \xrightarrow{P} 1, \text{ kdykoli } n \to \infty. \qquad (4.24)$$

Důkaz: Danielsson, de Haan, Peng, de Vries, (2001).

Důkaz vyplývá z věty 4.7 a věty 4.9 a ze skutečnosti, že

$$t^{\frac{1}{2\rho-1}} s^{-}\left(\frac{1}{t}\right) \to (-2\rho c^2)^{\frac{1}{1-2\rho}}.$$

Věta 4.11

Nechť $n_1 = O(n^{1-\varepsilon})$ pro $\varepsilon \in (0,\frac{1}{2})$ a $n_2 = \frac{(n_1)^2}{n}$. Předpokládejme dále, že platí podmínky druhého řádu (Definice 2.11) pro $A(t) = c\,t^{\rho}, t \to \infty$ a nechť dále je $k_i \to \infty$ a $\frac{k_i}{n_i} \to 0$ ($i = 1, 2$). Určeme $k_{i,0}^*$ takové, že

$$E\left(\left(M_*(n_i,k_i) - 2\left(\gamma_H^*(n_i,k_i)\right)^2\right)^2 \middle| \boldsymbol{X_n}\right)$$

je minimální (i = 1, 2). Potom

$$\frac{\frac{\left(k_{1,0}^{*}(n_1)\right)^2}{k_{2,0}^{*}(n_2)}\left(\frac{\left(\log k_{1,0}^{*}(n_1)\right)^2}{\left(2\log n_1-\log k_{1,0}^{*}(n_1)\right)^2}\right)^{\frac{\log n_1-\log k_{1,0}^{*}(n_1)}{\log n_1}}}{k_0(n)} \xrightarrow{P} 1,$$

jestliže $n \to \infty$.

Důkaz:

Podle Proposition 1.7.1 Geluk, de Haan (1987) je ($k_{1,0}^{*}$ je pravidelně se měnící funkce v nekonečnu s indexem $\frac{-2\rho}{1-2\rho}$)

$$\frac{\log k_{1,0}^{*}}{\log n_1} \xrightarrow{P} \frac{-2\rho}{1-2\rho}$$

a odtud jednoduchou úpravou je

$$\frac{\log k_{1,0}^{*}}{-2\log n_1+2\log k_{1,0}^{*}} \xrightarrow{P} \rho. \tag{4.25}$$

K dalšímu postupu použijeme závěry věty 4.10 pro $k_{1,0}^{*}\ a\ k_{2,0}^{*}$

$$\left(\frac{n_1}{n}\right)^{\frac{2\rho}{1-2\rho}}\frac{k_{1,0}^{*}}{\hat{k}_0} \xrightarrow{P} 1,$$

$$\left(\frac{n_2}{n}\right)^{\frac{2\rho}{1-2\rho}}\frac{k_{2,0}^{*}}{\hat{k}_0} \xrightarrow{P} 1.$$

Úpravou těchto dvou limit a využitím vztahu $n_2 = \frac{(n_1)^2}{n}$ vyplývá, že

$$\frac{\hat{k}_0 k_{2,0}^{*}}{\left(k_{1,0}^{*}\right)^2} \xrightarrow{P} 1. \tag{4.26}$$

Podle závěrů věty 4.8 je dále

$$\frac{\left(k_{1,0}^{*}(n_1)\right)^2}{k_{2,0}^{*}(n_2)k_0(n)} \to \left(1-\frac{1}{\rho}\right)^{\frac{2}{1-2\rho}}.$$

Důkaz věty dokončíme pomocí vztahu (4.25).

Věta 4.11

Předpokládejme, že platí předpoklady věty 4.10. Definujme

$$\hat{k}_0(n) = \frac{\left(k^*_{1,0}(n_1)\right)^2}{k^*_{2,0}(n_2)}\left(\frac{\left(\log k^*_{1,0}(n_1)\right)^2}{\left(2\log n_1 - \log k^*_{1,0}(n_1)\right)^2}\right)^{\frac{\log n_1 - \log k^*_{1,0}(n_1)}{\log n_1}}.$$

Potom $\gamma_H(n, \hat{k}_0(n))$ je stejně asymptoticky vydatný jako $\gamma_H(n, k_0(n))$.

Důkaz:

Platí

$$\lim_{n\to\infty} \frac{\hat{k}_0(n)}{k_0(n)} \xrightarrow{P} 1.$$

Podle věty 4.1 v Hall, Welsh (1985) $\gamma_H(n, \hat{k}_0(n))$ konverguje stejně rychle jako $\gamma_H(n, k_0(n))$.

Bootstrapová procedura:

I. Krok: Položme $n_1 = [n^{1-\varepsilon}]$ pro hodnotu $\varepsilon \in (0; \frac{1}{2})$, kde $[x]$ označuje celou část čísla *x*. Vyberme z výběru o *n* hodnotách nový bootstrapový výběr o délce n_1. Budeme dále počítat $E\left(\left(M_*(n_1, k)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$. Nalezneme hodnotu $k_1^*(n_1)$, pro který je předchozí moment minimální.

II. krok: Položíme $n_2 = \left[{n_1}^2/_n\right]$ a zopakujeme krok I. s tím, že nalezneme hodnotu $k_1^*(n_2)$, pro kterou je $E\left(\left(M_*(n_2, k)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$ minimální.

III. krok: Odhadneme parametr ρ tímto konzistentním odhadem: $\hat{\rho}_n = \frac{-\log k_1^*(n_1)}{-2\log n_1 + 2\log k_1^*(n_1)}$

IV. krok: nyní definujeme odhad $k_1(n)$,

$$\hat{k}_1(n) = \frac{\left(k_1^*(n_1)\right)^2}{k_1^*(n_2)}\left(1 - \frac{1}{\hat{\rho}_n}\right)^{\frac{1}{2\hat{\rho}_n - 1}}.$$

Tato procedura byla popsána v Danielsson (2001). Efektivnější procedura byla uveřejněna v Gomes, Oliveira (2001).

4.3. Optimal sample fractions pro momentový odhad

Nechť $F \in D(G_\gamma)$ pro $\gamma \in R$. Momentový odhad $\hat{\gamma}^{(M)}$ pro index γ jsme zavedli v definici 2.31 takto

$$\hat{\gamma}^{(M)}(n,k) = M^{(1)}(n,k) + 1 - \frac{1}{2}\left(1 - \frac{\left(M^{(1)}(n,k)\right)^2}{M^{(2)}(n,k)}\right)^{-1}, \text{ kde}$$

$$M^{(i)}(n,k) = \frac{1}{k}\sum_{j=1}^{k}\left(log\ X_{n-j+1:n} - log\ X_{n-k:n}\right)^i.$$

Dále ukážeme, že optimální volba velikosti výběru optimal sample fraction v odhadu parametru γ je závislá na minimalizaci střední kvadratické chyby $\hat{\gamma}^{(M)} - \gamma$.

Nechť dále v dalším je $\gamma_- = min(\gamma, 0)$ a $\gamma_+ = max(\gamma, 0)$.

Budeme hledat minimum $E\left(\hat{\gamma}^{(M)}(n,k) - \gamma\right)^2$ v asymptotickém smyslu (druhý moment asymptotického rozdělení). Rozdíl mezi $\hat{\gamma}^{(M)}(n,k) - \gamma$ můžeme rozdělit na dvě části - nedegenerovanou část s asymptoticky normálním rozdělením a degenerovanou část s nulovou systematickou chybou.

Řekneme, že index $k_1(n)$ je optimální hodnotou metody optimal sample fraction pro momentový odhad, jestliže

$$k_1(n) = \underset{k}{arg\,min}\ AMSE\left(\hat{\gamma}^{(M)}(n,k) - \gamma\right). \tag{4.27}$$

Naším cílem je nalézt konzistentní odhad $k_1(n)$. V dalším budeme vycházet z článku (Draisma et al. 1999). Využijeme v něm uvedenou větu 2.1. Základní myšlenky důkazů následujících tvrzení lze ovšem nalézt již v článku Dekkers, Einmahl, de Haan (1989).

Věta 4.12

Předpokládejme, že $F \in D(G_\gamma)$ a že platí podmínky druhého řádu a tvrzení věty 2.20 o funkci $H(x)$ pro $\rho < 0$, dále položme

$k_0(n) = \underset{k}{arg\,min}\ E\left(\hat{\gamma}^{(M)}(n,k) - \gamma\right)^2$. Potom, jestliže $n \to \infty$,

$$k_0(n) \Big/ \left(\frac{V^2(\gamma)}{b^2(\gamma,\rho)}\right)^{\frac{1}{1-2\rho^*}} . \frac{n}{s^-\left(\frac{1}{n}\right)} \to 1, \tag{4.28}$$

kde s^- je inversní funkce klesající funkce s, splňující následující vztah

$$A_0^2(x) = \left(1 + o(1)\right)\int_x^\infty s(t)dt,$$

a

$$\rho^* = \begin{cases} \rho & \gamma > 0 \\ \gamma & \rho < \gamma < 0 \\ \rho & \gamma < \rho \end{cases}. \tag{4.29}$$

a dále

$$V^2(\gamma) = \begin{cases} \gamma^2 + 1 & if\ \gamma \geq 0, \\ \frac{(1-\gamma)^2(1-2\gamma)(6\gamma^2-\gamma+1)}{(1-3\gamma)(1-4\gamma)} & if\ \gamma < 0, \end{cases} \tag{4.30}$$

$$b(\gamma, \rho) = \begin{cases} \frac{\gamma}{\rho(1-\rho)} + \frac{1}{(1-\rho)^2} & if\ \gamma > 0, \\ \frac{1}{1-\gamma} & if\ \rho < \gamma < 0, \\ \frac{(1-\gamma)(1-2\gamma)}{(1-\rho-\gamma)(1-\rho-2\gamma)} & if\ \gamma < \rho \end{cases} \tag{4.31}$$

a

$$A_0(t) = \begin{cases} A(t) & if\ \gamma > 0, \\ \frac{a(t)}{U(t)} & if\ \rho < \gamma < 0, \\ A(t) & if\ \gamma < \rho. \end{cases} \tag{4.32}$$

Navíc

$$AMSE\left(\hat{\gamma}^{(M)}(n,k)\right) = \frac{V^2(\gamma)}{k} + b^2(\gamma,\rho)A_0^2\left(\frac{n}{k}\right), \tag{4.33}$$

Důkaz: Theorem 3.1 (Draisma et al. 1999).

Vlastní důkaz je založen na třech pomocných tvrzeních. V prvním jsou definována pomocná rozdělení podobně jako v Lemma 4.6.

Lemma 4.13

Nechť $Y_1, \dots, Y_n$ jsou nezávislé náhodné veličiny se společnou distribuční funkcí F(x) = 1-1/x (x > 1). Nechť dále je $Y_{n,1} \leq \cdots \leq Y_{n,n}$ pořádková statistika vytvořená z $Y_1, \dots, Y_n$. Předpokládejme dále, že $k \to \infty$ a $\frac{k}{n} \to 0$. Potom

i) $\frac{Y_{n,n-k}}{\frac{n}{k}} \xrightarrow{P} 1$

ii) Definujme

$$P_n = \frac{1}{k}\sum_{i=1}^{k}\frac{\left(Y_{n,n-i+1}/Y_{n,n-k}\right)^{\gamma_-}-1}{\gamma_-} - \frac{1}{1-\gamma_-}$$

$$Q_n = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\left(Y_{n,n-i+1}/Y_{n,n-k}\right)^{\gamma_-}-1}{\gamma_-}\right)^2 - \frac{2}{(1-\gamma_-)(1-2\gamma_-)}$$

$$R_n = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\left(Y_{n,n-i+1}/Y_{n,n-k}\right)^{\gamma_-}-1}{\gamma_-}\right)^3 - \frac{6}{(1-\gamma_-)(1-2\gamma_-)(1-3\gamma_-)}$$

Rozdělení $\sqrt{k}(P_n, Q_n, R_n)$ konverguje v distribuci k rozdělení (P, Q, R), které je normálně rozdělené se střední hodnotou nula a kovarianční maticí :

$$EP^2 = \frac{1}{(1-\gamma_-)^2(1-2\gamma_-)},$$

$$EQ^2 = \frac{4(5-11\gamma_-)}{(1-\gamma_-)^2(1-2\gamma_-)^2(1-3\gamma_-)(1-4\gamma_-)},$$

$$ER^2 = \frac{36(19-105\gamma_- + 146{\gamma_-}^2)}{(1-\gamma_-)^2(1-2\gamma_-)^2(1-3\gamma_-)^2(1-4\gamma_-)(1-5\gamma_-)(1-6\gamma_-)},$$

$$E(PQ) = \frac{4}{(1-\gamma_-)^2(1-2\gamma_-)(1-3\gamma_-)},$$

$$E(PR) = \frac{18}{(1-\gamma_-)^2(1-2\gamma_-)(1-3\gamma_-)(1-4\gamma_-)},$$

$$E(QR) = \frac{12(9-21\gamma_-)}{(1-\gamma_-)^2(1-2\gamma_-)^2(1-3\gamma_-)(1-4\gamma_-)(1-5\gamma_-)}.$$

Navíc

$$k\,E\,P_n^2 \to E\,P^2$$
$$k\,E\,Q_n^2 \to E\,Q^2$$
$$k\,E\,R_n^2 \to E\,R^2$$

iii) Definujme pro *j=1,2,3*

$$d_n^{(j)} = \frac{1}{k}\sum_{i=1}^{k} j\,H\left(Y_{n,n-i+1}/Y_{n,n-k}\right)\left(\frac{\left(Y_{n,n-i+1}/Y_{n,n-k}\right)^{\gamma_-}-1}{\gamma_-}\right)^{j-1}.$$

Potom podle slabého zákona velkých čísel je

$$d_n^{(j)} \to d_j = \int_1^{\infty} j\,H(u)\left(\frac{u^{\gamma_-}-1}{\gamma_-}\right)^{j-1}\frac{du}{u^2}, \qquad j = 1,2,3,$$

tedy

$$d_1 = \frac{1}{(1-\gamma_-)(1-\rho-\gamma_-)},$$

$$d_2 = \frac{2(3-2\rho-4\gamma_-)}{(1-\gamma_-)(1-2\gamma_-)(1-\rho-\gamma_-)(1-\rho-2\gamma_-)},$$

$$d_3 = \frac{18\,{\gamma_-}^2 - 22\gamma_- + 15\rho\gamma_- + 3\rho^2 - 8\rho + 6}{(1-\gamma_-)(1-2\gamma_-)(1-3\gamma_-)(1-\rho-\gamma_-)(1-\rho-3\gamma_-)}.$$

Důkaz:

Je podobný důkazu věty 3.4 s využitím lemma 3.1 v Dekkers (1989). Výpočet vychází z transformace původních náhodných veličin $Y_{n,i}$ a z určení jejich typu.

Druhým pomocným tvrzením je

Lemma 4.14

Nechť je f měřitelná funkce. Předpokládejme, že existuje reálný parametr α a funkce $a_1(t) > 0$ a funkce $A_1(t) \to 0$ takové, že pro všechna x > 0 je

$$\lim_{t\to\infty} \frac{\frac{f(tx)-f(t)}{a_1(t)} - \frac{x^\alpha - 1}{\alpha}}{A_1(t)} = H_1(x)$$

kde

$$H_1(x) = \frac{1}{\beta}\left(\frac{x^{\alpha+\beta}-1}{\alpha+\beta} - \frac{x^\alpha - 1}{\alpha}\right) \ (\beta \le 0).$$

Potom pro každé $\varepsilon > 0$ existuje kladné číslo t_0 takové, že pro všechna $t \ge t_0, tx \ge t_0$, je

$$\left|\frac{\frac{f(tx)-f(t)}{a_1(t)} - \frac{x^\alpha - 1}{\alpha}}{A_1(t)} - H_1(x)\right| \le \varepsilon\left(1 + x^\alpha + 2x^{\alpha+\beta}e^{\varepsilon|\log x|}\right).$$

Důkaz:

Vyplývá z tvrzení uvedených v článcích de Haan, Stadtmuller (1996), de Haan, Peng (1996), z článku Omey, Willekens (1988) a Proposition 1.19.4 ve studii de Haan, Geluk (1987).

V tomto lemma je uveden odhad vzdálenosti funkce $H_1(x)$ od změněné funkce z podmínky druhého řádu. Je potřebný pro odhad části výrazů vzdálenosti podmínky druhého řádu a funkce $H(x)$.

Třetí pomocné tvrzení je obdobné postupu při důkazu věty 4.8.

Lemma 4.15

Nechť platí podmínky druhého řádu (Definice 2.21) a tvrzení věty 2.20 o funkci $H(x)$ pro $\rho < 0$ a $n_1 = O(n^{1-\varepsilon_0})$ pro $\varepsilon_0 \in (0,1)$. Potom pro každé $\varepsilon \in (0,1)$ existuje $t_0 > 0$ takové, že pro všechny hodnoty $t_0 \le t \le n_1(\log n_1)^2$ a $t_0 \le tx \le n_1(\log n_1)^2$ platí

$$\left|\frac{\frac{\log U_n(tx) - \log U_n(x)}{a(t)/U(t)} - \frac{x^\gamma - 1}{\gamma_-}}{A(t)} - H(x)\right| \le$$

$$\le \left(\frac{\sqrt{tx}\log n}{n} + \varepsilon\right) d(\gamma_-,\rho)x^{\gamma+\rho}e^{\varepsilon|\log x|} + \left(\frac{\sqrt{t}\log n}{n} + \varepsilon\right) d(\gamma_-,\rho) +$$

$$+ \varepsilon\left(1 + x^\gamma + 2x^{\gamma_-+\rho}e^{\varepsilon|\log x|}\right) + \frac{d(\gamma_-,\rho)}{|A(t)|}\frac{\sqrt{t}\log n}{n}\left(\sqrt{x}+1\right),$$

kde $d(\gamma_-, \rho) > 0$ je konstanta závislá jen na γ_- a na ρ.

Důkaz:

Je obdobný důkazu pomocného tvrzení ve větě 4.8.

Důkaz vlastní věty 4.12 je založen na třech předchozích lemmatech a na stejném postupu uvedeném v důkazu věty 4.8. Hodnotu *t* v lemma 4.15 nahradíme $Y_{n,n-k}$ a hodnotu *x* nahradíme $\frac{Y_{n,n-i}}{Y_{n,n-k}}$. Tyto jsme dosadily do předchozí nerovnosti uvedené v lemma 4.15. Po úpravě získáváme

$$\frac{M^{(1)}(n,k)}{a(Y_{n,n-k})/U(Y_{n,n-k})} = \frac{1}{1-\gamma_-} + P_n + \frac{A(n/k)}{(1-\gamma_-)(\rho^*-\gamma_-)} + o\left(A\left(\frac{n}{k}\right)\right), \tag{4.34}$$

při důkazu tohoto tvrzení užíváme stejné postupy jako ve větě 4.8. Umocněním (4.34) je

$$\left(\frac{M^{(1)}(n,k)}{a(Y_{n,n-k})/U(Y_{n,n-k})}\right)^2 = \frac{1}{(1-\gamma_-)^2} + \frac{2P_n}{(1-\gamma_-)} + \frac{2A(n/k)}{(1-\gamma_-)^2(\rho^*-\gamma_-)} + o\left(A\left(\frac{n}{k}\right)\right). \tag{4.35}$$

Podobně využijeme náhodné veličiny Q_n pro úpravu výrazu $M^{(2)}(n,k)$ a získáme

$$\frac{M^{(2)}(n,k)}{\left(a(Y_{n,n-k})/U(Y_{n,n-k})\right)^2} =$$

$$= \frac{2}{(1-\gamma_-)(1-\gamma_-)} + Q_n + \frac{2A\left(\frac{n}{k}\right)(2\rho^* + 1 - 4\gamma_-)}{(1-\gamma_-)(1-2\gamma_-)(\rho^*-\gamma_-)(\rho^*-2\gamma_-)} + o\left(A\left(\frac{n}{k}\right)\right).$$

Oba předchozí výrazy dosadíme do vztahu pro momentový odhad $\hat{\gamma}^{(M)}(n,k)$ a využijeme toho, že $a(Y_{n,n-k})/U(Y_{n,n-k}) \xrightarrow{P} \gamma_+$. Tedy

$$\hat{\gamma}^{(M)}(n,k) = M^{(1)}(n,k) + \frac{M^{(2)}(n,k) - 2\left(M^{(1)}(n,k)\right)^2}{2M^{(2)}(n,k) - 2\left(M^{(1)}(n,k)\right)^2} =$$

$$= (\gamma_+ + o_P(1))\left(\frac{1}{(1-\gamma_-)^2} + \frac{2P_n}{(1-\gamma_-)} + \frac{2A(n/k)}{(1-\gamma_-)^2(\rho^*-\gamma_-)}\right) +$$

$$+\gamma_- + (1-\gamma_-)^2(1-2\gamma_-)\left(\left(\frac{1}{2} - \gamma_-\right)Q_n - 2P_n\right) +$$

$$+A\left(\frac{n}{k}\right)\frac{(1-\gamma_-)(1-2\gamma_-)}{(\rho^*-\gamma_-)(\rho^*-2\gamma_-)} + \left(A\left(\frac{n}{k}\right)\right).$$

Dále je

$$\hat{\gamma}^{(M)}(n,k) - \gamma =$$

$$= M^{(1)}(n,k) - \gamma_+ + \frac{M^{(2)}(n,k) - 2\left(M^{(1)}(n,k)\right)^2}{2M^{(2)}(n,k) - 2\left(M^{(1)}(n,k)\right)^2} - \gamma_- =$$

$$= \left(\gamma_+ + \frac{a(Y_{n,n-k})}{U(Y_{n,n-k})} - \gamma_+\right)\left(\frac{1}{1-\gamma_-} + P_n + d_1 A\left(\frac{n}{k}\right)\right) - \gamma_+ +$$

$$+ \frac{(1-\gamma_-)^2(1-2\gamma_-)}{2}\left((1-2\gamma_-)Q_n - 4P_n + (d_2 - 2\gamma_- d_2 - 4d_1)A\left(\frac{n}{k}\right)\right) + o_P\left(A\left(\frac{n}{k}\right)\right) =$$

$$= \left(\frac{a(Y_{n,n-k})}{U(Y_{n,n-k})} - \gamma_+\right)\frac{1}{1-\gamma_-} + \left(\frac{a(Y_{n,n-k})}{U(Y_{n,n-k})} - \gamma_+\right)\left(P_n + d_1 A\left(\frac{n}{k}\right)\right) +$$

$$+ \frac{(1-\gamma_-)^2(1-2\gamma_-)}{2}\left(\left(\frac{2\gamma_+}{(1-\gamma_-)^2(1-2\gamma_-)} - 4\right)P_n + (1-2\gamma_-)Q_n\right.$$

$$\left. + \left(d_2 - 2\gamma_- d_2 - 4d_1 + \frac{2\gamma_+ d_1}{(1-\gamma_-)^2(1-2\gamma_-)}\right)A\left(\frac{n}{k}\right)\right) + o_P\left(A\left(\frac{n}{k}\right)\right).$$

Užijeme – li nyní výsledky Lemma 4.13 a věty 2 část iii) z de Haan, Stadtmuller (1996), potom platí následující hodnoty limit

$$\frac{\frac{a(t)}{U(t)} - \gamma_+}{A(t)} = \begin{cases} \dfrac{\frac{a(t)}{U(t)} - \gamma}{A(t)} \to \dfrac{\gamma}{\rho} & \gamma > 0 \\ \dfrac{\frac{a(t)}{U(t)}}{|A(t)|} \to \infty & \rho < \gamma \leq 0' \\ \dfrac{\frac{a(t)}{U(t)}}{A(t)} \to 0 & \gamma < \rho \end{cases}$$

Využijeme tohoto vztahu a výsledků z lemma 4.13 je

$$\text{AMS}E\left(\hat{\gamma}^{(M)}(n,k) - \gamma\right)^2 = \left(V^2(\gamma)/k + b^2(\gamma,\rho)A^2\left(\frac{n}{k}\right)\right)\left(1 + o(1)\right) =$$

$$= \left(V^2(\gamma)r/n + b^2(\gamma,\rho)A^2(r)\right)\left(1 + o(1)\right),$$

kde $r = \frac{k}{n}$. Nyní se použije postup uvedený v Dekkers, de Hann (1993) věty 3.2, 3.3 a 3.4 tj. budeme hledat kritické body uvedené funkce. Stále předpokládáme, že *k(n)* je prostřední posloupnost. Musíme dokázat, že $k_0(n)$ je skutečně minimum. Ovšem pro každou hodnotu $k(n)$, pro kterou je $\frac{k(n)}{k_0(n)} \to 0$ nebo $\frac{k(n)}{k_0(n)} \to \infty$ je druhý asymptotický moment $\hat{\gamma}^{(M)}(n,k) - \gamma$ příliš velký, jakmile $k(n) \to \infty$ nebo $\frac{k(n)}{n} \to 0$. Abychom zůstali ve správných mezích, můžeme přidat omezení $\log n \leq k(n) \leq \frac{n}{\log n}$. Odtud již vyplývají závěry věty 4.12.

Při hledání optimální hodnoty k u Hillova odhadu jsme použili alternativní odhad a pomocí něho jsme nalezli základní vlastnosti takovéto optimální hodnoty k. Podobný postup využijeme nyní i pro momentový odhad. Definujme alternativní odhad pro parametr γ

$$\hat{\gamma}^{(M1)}(n,k) = \sqrt{\frac{M^{(2)}(n,k)}{2}} + 1 - \frac{2}{3}\left(1 - \frac{M^{(1)}(n,k)M^{(2)}(n,k)}{M^{(3)}(n,k)}\right)^{-1} \tag{4.36}$$

Ukážeme, že statistika $\hat{\gamma}^{(M)}(n,k) - \hat{\gamma}^{(M1)}(n,k)$ má podobné asymptotické chování jako $\hat{\gamma}^{(M)}(n,k) - \gamma$.

Věta 4.16

Nechť jsou splněny předpoklady věty 4.12. Položme

$$k_2(n) = \underset{k}{arg\,min}\; AMSE\left(\hat{\gamma}^{(M)}(n,k) - \hat{\gamma}^{(M1)}(n,k)\right)^2. \tag{4.37}$$

Potom

$$k_2(n) \sim \left(\frac{\bar{V}^2(\gamma)}{\bar{b}^2(\gamma,\rho)}\right)^{\frac{1}{1-2\rho^*}} . \frac{n}{s^{-}\left(\frac{1}{n}\right)}\,, \tag{4.38}$$

kde

$$\bar{V}^2(\gamma) = \begin{cases} \frac{1}{4}\,(\gamma^2+1) & \gamma > 0, \\ \frac{1}{4}\frac{(1-\gamma)^2\left(1-8\gamma+48\gamma^2-154\gamma^3+263\gamma^4-222\gamma^5+72\gamma^6\right)}{(1-2\gamma)(1-3\gamma)(1-4\gamma)(1-5\gamma)(1-6\gamma)} & \gamma < 0, \end{cases} \tag{4.39}$$

$$\bar{b}\,(\gamma,\rho) = \begin{cases} -\frac{\gamma(1-\rho)+\rho}{2(1-\rho)^3} & \gamma > 0, \\ \frac{(1-2\rho)-\sqrt{(1-\rho)(1-2\rho)}}{(1-\rho)(1-2\rho)} & \rho < \gamma < 0, \\ \frac{1}{2}\frac{-\rho(1-\gamma)^2}{(1-\rho-\gamma)(1-\rho-2\gamma)(1-\rho-3\gamma)} & \gamma < \rho. \end{cases} \tag{4.40}$$

Důkaz: Draisma (1999), Theorem 3.2.

Důkaz tvrzení je položen na stejných postupech jako důkaz věty 4.15. V postupu je nutné zaměnit neznámou hodnotu γ hodnotou odhadu $\hat{\gamma}^{(M1)}(n,k)$. Navíc je možno ukázat (postupem stejným jako v důkazu předchozí věty), že

$$\frac{M^{(3)}(n,k)}{\left(a(Y_{n,n-k})/U(Y_{n,n-k})\right)^3} = \frac{6}{(1-\gamma_-)(1-2\gamma_-)(1-3\gamma_-)} + R_n + d_3 A\left(\frac{n}{k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right)$$

a

$$\frac{M^{(1)}(n,k)M^{(2)}(n,k)}{\left(a(Y_{n,n-k})/U(Y_{n,n-k})\right)^3} = \frac{2}{(1-\gamma_-)^2(1-2\gamma_-)} + \frac{2P_n}{(1-\gamma_-)(1-2\gamma_-)} + \frac{Q_n}{(1-\gamma_-)} +$$

$$+\left(\frac{2d_1}{(1-\gamma_-)(1-2\gamma_-)} + \frac{d_2}{(1-\gamma_-)}\right) A\left(\frac{n}{k}\right) + o_P\left(A\left(\frac{n}{k}\right)\right).$$

Dále se v důkazu odhadne nejprve rozdíl $\hat{\gamma}^{(M1)}(n,k) - \gamma$ a využije se výsledek předchozí věty, odhad $\hat{\gamma}^{(M)}(n,k) - \gamma$. Pomocí těchto dvou odhadů již můžeme odhadnout rozdíl $\hat{\gamma}^{(M)}(n,k) - \hat{\gamma}^{(M1)}(n,k)$. Detaily výpočtu jsou uvedeny Draisma (1999).

Dále položme pro hodnotu $\delta > 0$

$${}^*\hat{\gamma}\,(n_1,k_1) = \left(\hat{\gamma}^{(M^*)}(n_1,k_1) - \hat{\gamma}^*(n_1,k_1)\right) I\left(\left|\hat{\gamma}^{(M^*)}(n_1,k_1) - \hat{\gamma}^*(n_1,k_1)\right| < {k_1}^{\delta-\frac{1}{2}}.\right)$$

Pro tento odhad můžeme vyslovit následující větu.

Věta 4.17

Nechť jsou splněny podmínky věty 4.12. Předpokládejme dále, že $\rho < 0$. Určeme $k_{0,1}^* = k_{0,1}^*$

$$\frac{k_{0,1}^*(n)}{k_0(n)} \to 1.$$

Proto

$$\frac{k_{0,1}^*(n)}{\left\{\left(\frac{\bar{V}^2(\gamma)}{\bar{b}^{\,2}(\gamma,\rho)}\right)^{\frac{1}{1-2\rho^*}} \frac{n}{s^-\left(\frac{1}{n}\right)}\right\}} \xrightarrow{P} 1.$$

Důkaz: Uveden v Draisma (1999) jako věta 3.3.

Upozorněme, že daná věta platí pro libovolné $\delta > 0$.

V dalším budeme podobně jako v případě Hillova odhadu definovat základní kroky bootstrapové procedury.

Vytvořme n_1 nezávislých výběrů z empirické d.f. vytvořené z $\boldsymbol{\mathcal{X}_n} = \{X_1, \cdots, X_n\}$. Získáme hodnoty $X_1^*, \cdots, X_{n_1}^*$ a vytvoříme pořádkovou statistiku $X_{n_1,1}^*, X_{n_1,2}^*, \cdots, X_{n_1,n_1}^*$ a pomocí této statistiky

$${}^*M^{(i)}(n_1,k_1) = \frac{1}{k_1}\sum_{j=1}^{k_1}\left(log\, X_{n_1,n_1-j+1}^* - log\, X_{n_1,n_1-k_1}^*\right)^i. \qquad (4.41)$$

Pro hodnoty $k_1 < n_1$ definujeme

$$\hat{\gamma}^{(M^*)}(n_1,k_1) = {}^*M^{(1)}(n_1,k_1) + 1 - \frac{1}{2}\left(1 - \frac{\left({}^*M^{(1)}(n_1,k_1)\right)^2}{{}^*M^{(2)}(n_1,k_1)}\right)^{-1} \tag{4.42}$$

a

$$\hat{\gamma}^*(n_1,k_1) = \sqrt{\frac{{}^*M^{(2)}(n_1,k_1)}{2}} - 1 + \frac{2}{3}\left(1 - \frac{{}^*M^{(1)}(n_1,k_1)\,{}^*M^{(2)}(n_1,k_1)}{{}^*M^{(3)}(n_1,k_1)}\right)^{-1}, \tag{4.43}$$

a předpokládáme dále, že $k_1 \to \infty$, $\frac{k_1}{n_1} \to 0$ a $n_1 = O(n^{1-\varepsilon})$ pro hodnotu $0 < \varepsilon < 1$.

Věta 4.18

Předpokládejme, že jsou splněny podmínky věty 4.12 a $n_1 = O(n^{1-\varepsilon})$ pro $0 < \varepsilon < \frac{1}{2}$. Definujme $k_{0,1}^* = \underset{k_1}{arg\,min}\; E\left(\left({}^*\hat{\gamma}\,(n_1,k_1)\right)^2 \middle| \boldsymbol{\mathcal{X}}_{\mathbf{n}}\right)$. Potom pro $n \to \infty$ je

$$\frac{k_{0,1}^*(n_1)}{\left\{\left(\frac{\bar{V}^2(\gamma)}{\bar{b}^{\,2}(\gamma,\rho)}\right)^{\frac{1}{1-2\rho^*}} \frac{n_1}{s^{-}\left(\frac{1}{n_1}\right)}\right\}} \xrightarrow{P} 1. \tag{4.44}$$

Důkaz: Draisma ((1999), věta 3.4.

Důkaz je založen na rozkladu

$${}^*M^{(1)}(n_1,k_1) \overset{d}{=} \frac{1}{k_1}\sum_{i=1}^{k_1} \log U_n\left(Y_{n_1,n_1-i+1}\right) - \log U_n\left(Y_{n_1,n_1-k_1}\right),$$

kde $\left\{Y_{n_1,i}\right\}_{i=1}^{n_1}$ jsou pořádkové statistiky se společnou distribuční funkcí $1 - \frac{1}{x}$ $(x > 1)$ a nezávislé na $\boldsymbol{\mathcal{X}}_{\mathbf{n}}$. pomocí stejné metody jako při důkazu věty 4.12 můžeme dokázat, že

$$\frac{{}^*M^{(1)}(n_1,k_1)}{a(Y_{n_1,n_1-k_1})/U(Y_{n_1,n_1-k_1})} = \frac{1}{1-\gamma_-} + P_{n_1} + \frac{A(n_1/k_1)}{(1-\gamma_-)(\rho^*-\gamma_-)} + o_P\left(A\left(\frac{n_1}{k_1}\right)\right) + O_P\left(\frac{\sqrt{\frac{n_1}{k_1}}\log n}{\sqrt{n}}\right)$$

Podobně můžeme vyjádřit i ${}^*M^{(2)}(n_1,k_1)$ a ${}^*M^{(3)}(n_1,k_1)$. Pokud nyní použijeme výsledky vět 4.12, 4.16 a 4.17 získáme i platnost věty 4.18.

Věta 4.19

Předpokládejme, že jsou splněny podmínky věty 3.37 a nechť dále je $A_0(t) = ct^{\rho^*}$, kde $c \neq 0$ a $\rho^* < 0$, potom je

$$\frac{k_0(n)}{k_{0,1}^*}\left(\frac{n_1}{n}\right)^{\frac{-2\rho^*}{1-2\rho^*}} \xrightarrow{P} 1. \tag{4.45}$$

Důkaz: Draisma ((1999), Corollary 3.1.

Důkaz je založen na předchozích větách a na skutečnosti, že $\lim_{t\to\infty} t^{-\gamma}a(t)/U)t)$ je kladná konstanta v případě, že $\rho < \gamma < 0$ (de Haan, Stadtmuller(1993)). Tedy $\lim_{t\to\infty} \frac{A_0(t)}{c_0 t^{\rho^*}} = 1$, proto je

$$\lim_{t\to\infty} \frac{s^-(1/t)}{(-2c_0\rho^*)^{\frac{1}{1-2\rho^*}} t^{\frac{1}{1-2\rho^*}}} = 1$$

V dalším potřebujeme odstranit faktor $\left(\frac{n_1}{n}\right)^{\frac{-2\rho^*}{1-2\rho^*}}$. K tomu použijeme podobně jako u ostatních odhadů druhou bootstrapovou posloupnost o rozsahu n_2.

Věta 4.20

Nechť jsou splněny podmínky věty 4.19 a platí $n_2 = \frac{(n_1)^2}{n}$. Nechť dále je $k_{0,1}^*(n_2) = \underset{k_1}{arg\ min}\ E\left(\left({}^*\hat{\gamma}\ (n_2, k_1)\right)^2 \middle| \boldsymbol{\mathcal{X}}_{\mathbf{n}}\right)$.

Potom je

$$\frac{k_0(n)}{\left(\frac{\left(k_{0,1}^*(n_1)\right)^2}{k_{0,1}^*(n_2)}\right)} \xrightarrow{P} 1.$$

Důkaz:

Vyplývá z předchozí věty.

Důsledek 4.21

Předpokládejme, že jsou splněny podmínky předchozí věty 4.20. Potom

$$k_0(n)\Big/\left(\frac{\left(k_{0,1}^*(n_1)\right)^2}{k_{0,1}^*(n_2)}\left(\frac{V^2(\gamma)\bar{b}^2(\gamma,\rho)}{\bar{V}^2(\gamma)b^2(\gamma,\rho)}\right)^{\frac{1}{1-2\rho^*}}\right) \xrightarrow{P} 1.$$

Důkaz:

Vyplývá z předchozích vět 4.12, 4.18 a 4.20.

Důsledek 4.22

Předpokládejme, že jsou splněny podmínky věty 4.20. Definujme

$$\hat{k}_0(n) = \frac{\left(k^*_{0,1}(n_1)\right)^2}{k^*_{0,1}(n_2)} \left(\frac{V^2(\hat{\gamma}_n)\bar{b}^2(\hat{\gamma}_n, \hat{\rho}_n)}{\bar{V}^2(\hat{\gamma}_n)\, b^2(\hat{\gamma}_n, \hat{\rho}_n)} \right)^{\frac{1}{1-2\hat{\rho}_n}},$$

kde hodnoty $k^*_{0,1}(n_1)$ a $k^*_{0,1}(n_2)$ jsou definovány v předchozích větách 4.18 a 4.20. Nechť dále je $\hat{\gamma}_n$ libovolný konsistentní odhad γ (např. $\hat{\gamma}^{(M)}(n,k)$ s posloupností $k = k(n)$, kde pro $k \to \infty$ je $\frac{k}{n} \to 0$) a $\hat{\rho}_n$ je konsistentní odhad ρ^* například

$$\hat{\rho}_n = \frac{\log k^*_{0,1}(n_1)}{-2\log n_1 + 2\log k^*_{0,1}(n_1)}.$$

Potom

$$\frac{\hat{k}_0(n)}{k_0(n)} \xrightarrow{P} 1,$$

navíc druhý asymptotický moment $\hat{\gamma}^{(M)}\left(n, \hat{k}_0(n)\right) - \gamma$ je asymptoticky roven druhému asymptotickému momentu $\hat{\gamma}^{(M)}(n, k_0(n)) - \gamma$.

Důkaz:

Vyplývá z Důsledku 4.21. Odhad $\hat{\rho}_n$ vyplývá ze vztahu

$$\log k^*_{0,1}(n_1)/\log n_1 \xrightarrow{P} \frac{-2\rho^*}{1-2\rho^*}$$

Při tvorbě bootstrapové posloupnosti budeme dále používat stejnou proceduru jako u Hillova odhadu a využijeme předchozích výsledků. Volíme hodnoty $n_1 = \mathrm{O}(\mathrm{n}^{1-\varepsilon})$ pro $0 < \varepsilon < 1/2$ a n_2 , kde $n_2 = \left[\frac{n_1^2}{n}\right]$.

Bootstrapová procedura:

I. Krok. Položme $n_1 = [n^{1-\varepsilon}]$ pro hodnotu $\varepsilon \in (0; \frac{1}{2})$, kde $[x]$ označuje celou část čísla x. Vyberme z výběru o n hodnotách nový bootstrapový výběr o délce n_1. Budeme dále počítat $E\left(\left(M_*(n_1,k)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$. Nalezneme hodnotu $k_1^*(n_1)$, pro který je předchozí moment minimální.

II. krok: Položíme $n_2 = \left[{n_1}^2/_n\right]$ a zopakujeme krok I. s tím, že nalezneme hodnotu $k_1^*(n_2)$, pro kterou je $E\left(\left(M_*(n_2,k)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$ minimální.

III. krok: Odhadneme parametr ρ tímto konzistentním odhadem:
$\hat{\rho}_n = \frac{-\log k_1^*(n_1)}{-2\log n_1 + 2\log k_1^*(n_1)}$.

IV. krok: nyní definujeme odhad $k_1(n)$,

$$\hat{k}_1(n) = \frac{\left(k_1^*(n_1)\right)^2}{k_1^*(n_2)} \left(1 - \frac{1}{\hat{\rho}_n}\right)^{\frac{1}{2\hat{\rho}_n - 1}}.$$

4.4. Optimal sample fraction pro Pickandsův odhad

Pickandsův odhad je definován v (2.36) jako

$$\hat{\gamma}_P(n,k) = \frac{1}{log2}\, log\left(\frac{X_{n-k:n} - X_{n-2k:n}}{X_{n-2k:n} - X_{n-4k:n}}\right)$$

V této části budeme na tento odhad aplikovat metodu optimal sample fraction a navíc nalezneme optimální hodnotu parametru *k*. Požadujeme dále splnění podmínek druhého řádu z definice 2.19 a navíc zavedeme nový požadavek diferencovatelnosti distribučních funkcí z dané sféry přitažlivosti.

Definice 4.23 (Pickands (1986))

Řekneme, že distribuční funkce F je v diferencovatelné sféře přitažlivosti G_γ (označujeme $F \in D_{dif}(G_\gamma)$), právě když je F diferencovatelné v levém okolí $x_\infty = sup\{x; F(x) < 1\}$ a existují konstanty $a_n > 0$ a $b_n \in R$ takové, že

$$\lim_{n\to\infty} \frac{\partial}{\partial x}[F^n(a_n x + b_n)] = G_\gamma'(x) \qquad (4.46)$$

lokálně stejnoměrně.

Následující věta charakterizuje diferencovatelnou sféru přitažlivosti G_γ.

Věta 4.24

$F \in D_{dif}(G_\gamma)$ pro $\gamma \in R$, právě když $U(x)$ (kvantilová funkce chvostu – definice 2.5) je diferencovatelná pro všechna dostatečně velká *x* a $U'(x) \in R_{\gamma-1}$ (pravidelně se měnící funkce s indexem $\gamma - 1$).

Důkaz: Pickands (1986).

V následujících větách jsou využity vlastnosti funkce logaritmus a podmínky druhého řádu. Pomocí nich jsou nalezeny asymptotické vlastnosti předkládaných výrazů.

Poznámka 4.25

Označme dále F_n empirickou distribuční funkci $\boldsymbol{X_n}$, F_n^- inversní funkci k F_n a

$$U_n = \left(\frac{1}{1 - F_n}\right)^{-1}.$$

Lemma 4.26

Nechť platí (4.46) a $n_1 = O(n^{1-\epsilon_0})$ pro nějaké $\varepsilon_0 \in (0,1)$. Potom pro každé $\varepsilon \in (0,1)$ existuje $t_0 > 0$ takové, že kdykoli $t_0 \le t \le n_1 (\log n_1)^2$ a zároveň

$t_0 \le tx \le n_1 (\log n_1)^2$, pak platí

$$\left|\frac{\frac{U_n(tx)-U_n(t)}{a(t)}-\frac{x^\gamma-1}{\gamma}}{A(t)}-\frac{1}{\rho}\left(\frac{x^{\rho+\gamma}-1}{\rho+\gamma}-\frac{x^\gamma-1}{\gamma}\right)\right| \leq$$

$$\leq \left(\frac{\sqrt{tx}\log n}{n}+\varepsilon\right)D(\gamma,\rho)x^{\gamma+\rho}e^{\varepsilon|\log x|}+\left(\frac{\sqrt{t}\log n}{n}+\varepsilon\right)D(\gamma,\rho)+$$

$$+\varepsilon\left(1+x^\gamma+2x^{\gamma+\rho}e^{\varepsilon\,|\log x|}\right)+\frac{D(\gamma,\rho)}{|A(t)|}\frac{\sqrt{t}\log n}{n}\left(\sqrt{x}+1\right),$$

kde $D(\gamma,\rho)>0$ je konstanta závislá jen na hodnotách γ a ρ.

Důkaz: Jeho provedení je obdobné důkazu lemmatu 4.15.

Lemma 4.27

Jestliže $F \in D_{dif}(G_\gamma)$, potom podmínka (4.46) platí pro $a_n = nU'(n)$ a $b_n = U(n)$ a pro každé $k \to \infty, \frac{k}{n} \to 0$ a $\theta \in (0,1>$ stochastický proces

$$W_{n,k}(\theta) = \sqrt{k}\,\frac{X_{n-[k\theta]:n}-U\left(\frac{n}{k\theta}\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)}$$

konverguje k Gaussovskému procesu $w(\theta)$, který má střední hodnotu 0 a kovarianční strukturu

$$Cov\left(w(\theta_1),w(\theta_2)\right) = \theta_1{}^{-\gamma}\theta_2{}^{-\gamma-1}, 0 < \theta_1 < \theta_2 \leq 1.$$

Důkaz: Uveden v Cooil (1985), věta 2.3.

V následujícím tvrzení je určena teoreticky optimální hodnota $k_0(n)$.

Věta 4.28

Nechť $F \in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x) = c\,x^\rho$, kde $c \neq 0$ a $\rho < 0$. Označme index $k_0(n)$ takový, že hodnota $E(\hat{\gamma}_P(n,k) - \gamma)^2$ je minimální. Potom

$$\lim_{n\to\infty} k_0(n) \Bigg/ \left\{\left(\frac{1+2^{-2\gamma-1}}{4\rho c^2\left(\frac{1-2\rho}{\rho}\right)^2\left(\frac{2^{-\gamma-\rho-1}}{\gamma+\rho}\right)^2 2^{-2\rho}}\right)^{\frac{1}{1-2\rho}} n^{\frac{-2\rho}{1-2\rho}}\right\} = 1. \qquad (4.47)$$

Podobně jako u ostatních námi šetřených odhadů nalezneme postupně asymptotické odhady neznámé hodnoty γ v předchozím výrazu a nahradíme ji.

Důkaz:Proveden v Draisma, de Haan, Peng, Pereira (1999).

V následujícím textu tento důkaz provedeme:

$$\sqrt{k}(\hat{\gamma}_P(n,k)-\gamma)=\sqrt{k}\left(\frac{1}{-log2}\ log\left(\frac{X_{n-2k:n}-X_{n-4k:n}}{X_{n-k:n}-X_{n-2k:n}}\right)-\gamma\right)=$$

$$=\frac{\sqrt{k}}{-log2}\log\left(1+2^{\gamma}\frac{X_{n-2k:n}-X_{n-4k:n}}{X_{n-k:n}-X_{n-2k:n}}-1\right)\overset{d}{=}$$

$$\overset{d}{=}\frac{\sqrt{k}}{-log2}\ \frac{-X_{n-2k:n}+X_{n-4k:n}-2^{-\gamma}(-X_{n-k:n}+X_{n-2k:n})}{2^{-\gamma}(X_{n-k:n}-X_{n-2k:n})}\left(1+o_P(1)\right)=$$

$$=\left(\frac{\sqrt{k}}{-log2}\frac{X_{n-4k:n}-U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})\left(X_{n-2k:n}-U\left(\frac{n}{2k}\right)\right)+2^{-\gamma}\left(X_{n-k:n}-U\left(\frac{n}{k}\right)\right)}{2^{-\gamma}(X_{n-k:n}-X_{n-2k:n})}\right.$$

$$\left.+\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})U\left(\frac{n}{2k}\right)+2^{-\gamma}U\left(\frac{n}{k}\right)}{2^{-\gamma}(X_{n-k:n}-X_{n-2k:n})}\right)\left(1+o_P(1)\right)=$$

$$\left(\frac{X_{n-2k:n}-X_{n-k:n}}{\frac{n}{k}U'\left(\frac{n}{k}\right)}\xrightarrow{\boldsymbol{P}}\frac{2^{-\gamma}-1}{\gamma}\right)$$

$$\overset{d}{=}\left(\frac{\sqrt{k}}{-log2}\frac{X_{n-4k:n}-U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})\left(X_{n-2k:n}-U\left(\frac{n}{2k}\right)\right)+2^{-\gamma}\left(X_{n-k:n}-U\left(\frac{n}{k}\right)\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}\right.$$

$$\left.+\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})U\left(\frac{n}{2k}\right)+2^{-\gamma}U\left(\frac{n}{k}\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}\right)\left(1+o_P(1)\right)$$

$$\overset{d}{=}\frac{1}{-log2}\ \frac{1}{2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}(w(4)-(1+2^{-\gamma})w(2)+2^{-\gamma}w(1))+o_P(1)$$

$$+\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})U\left(\frac{n}{2k}\right)+2^{-\gamma}U\left(\frac{n}{k}\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}\left(1+o_P(1)\right)$$

Funkce *w* je převzata z Lemmatu 4.27.

Tedy asymptotický rozptyl $\sqrt{k}(\hat{\gamma}_P(n,k)-\gamma)$ je roven

$$\frac{\gamma^2\frac{1}{2}(1+2^{-2\gamma-1})}{(\log 2)^2(2^{-\gamma}-1)^2}$$

a asymptotický $Bias\left(\sqrt{k}(\hat{\gamma}_P(n,k)-\gamma)\right)$ je roven

$$\sqrt{k}A\left(\frac{n}{k}\right)\frac{2^{-\rho}}{-\log 2}\,\frac{\gamma}{2^{-\gamma}-1}\,\frac{1-2^{\rho}}{\rho}\,\frac{2^{-\gamma-\rho}-1}{\gamma+\rho}.$$

Nyní využijeme $A(t)=ct^{-\rho}$ a užijeme stejné postupy jako v závěru důkazu věty 4.12.

$$\lim_{n\to\infty}{}^{k_0(n)}\Bigg/\left\{\left(\frac{1+2^{-2\gamma-1}}{4\rho c^2\left(\frac{1-2^{\rho}}{\rho}\right)^2\left(\frac{2^{-\gamma-\rho}-1}{\gamma+\rho}\right)^2 2^{-2\rho}}\right)^{\frac{1}{1-2\rho}} n^{\frac{-2\rho}{1-2\rho}}\right\}\to 1$$

Věta 4.29

Nechť $F\in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x)=c\,x^{-\rho}$, kde $c\neq 0$ a $\rho<0$. Označme index $\hat{k}_0(n)$ takový, že hodnota $E\left(\hat{\gamma}_P(n,k)-\hat{\gamma}_P(n,4k)\right)^2$ je minimální. Potom

$$\lim_{n\to\infty}{}^{\hat{k}_0(n)}\Bigg/\left\{\left(\frac{1+2^{-2\gamma-1}}{4\rho c^2\left(\frac{1-2^{\rho}}{\rho}\right)^2\left(\frac{2-\gamma-\rho-1}{\gamma+\rho}\right)^2 2^{-2\rho}}\cdot\frac{\frac{5}{4}}{(1-2^{-2\rho})^2}\right)^{\frac{1}{1-2\rho}} n^{\frac{-2\rho}{1-2\rho}}\right\}=1. \qquad (4.48)$$

Pomocí obou vět již můžeme vztah mezi oběma výše uvedenými indexy.

Důkaz: Draisma, de Haan, Peng, Pereira (1999).

Při důkazu budeme postupovat stejně jako u předchozí věty. Tedy

$$\sqrt{k}\left(\hat{\gamma}_P(n,k)-\hat{\gamma}_P(n,4k)\right)=\sqrt{k}(\hat{\gamma}_P(n,k)-\gamma)-\sqrt{k}(\hat{\gamma}_P(n,4k)-\gamma)=$$

$$\begin{aligned}\overset{d}{=}\Bigg(&\frac{\sqrt{k}}{-log2}\frac{X_{n-4k:n}-U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})\left(X_{n-2k:n}-U\left(\frac{n}{2k}\right)\right)+2^{-\gamma}\left(X_{n-k:n}-U\left(\frac{n}{k}\right)\right)}{2^{-\gamma}(X_{n-k:n}-X_{n-2k:n})}\\&+\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})U\left(\frac{n}{2k}\right)+2^{-\gamma}U\left(\frac{n}{k}\right)}{2^{-\gamma}(X_{n-k:n}-X_{n-2k:n})}\\&-\frac{\sqrt{k}}{-log2}\frac{X_{n-16k:n}-U\left(\frac{n}{16k}\right)-(1+2^{-\gamma})\left(X_{n-8k:n}-U\left(\frac{n}{8k}\right)\right)+2^{-\gamma}\left(X_{n-4k:n}-U\left(\frac{n}{4k}\right)\right)}{2^{-\gamma}(X_{n-8k:n}-X_{n-4k:n})}\\&-\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{16k}\right)-(1+2^{-\gamma})U\left(\frac{n}{8k}\right)+2^{-\gamma}U\left(\frac{n}{4k}\right)}{2^{-\gamma}(X_{n-8k:n}-X_{n-4k:n})}\Bigg)\left(1+o_P(1)\right)=\end{aligned}$$

Podobně jako v předchozím důkazu využijeme

$$\left(\frac{X_{n-8k:n}-X_{n-4k:n}}{\frac{n}{k}U'\left(\frac{n}{k}\right)} \xrightarrow{P} 2^{-2\gamma}\frac{2^{-\gamma}-1}{\gamma}\right)$$

$$\begin{aligned}\overset{d}{=}&\frac{1}{-log2}\frac{1}{2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}(w(4)-(1+2^{-\gamma})w(2)+2^{-\gamma}w(1))+o_P(1)\\&+\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{4k}\right)-(1+2^{-\gamma})U\left(\frac{n}{2k}\right)+2^{-\gamma}U\left(\frac{n}{k}\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)2^{-\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}\left(1+o_P(1)\right)\\&-\frac{1}{-log2}\frac{1}{2^{-3\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}(w(16)-(1+2^{-\gamma})w(8)+2^{-\gamma}w(4))+o_P(1)\\&-\frac{\sqrt{k}}{-log2}\frac{U\left(\frac{n}{16k}\right)-(1+2^{-\gamma})U\left(\frac{n}{8k}\right)+2^{-\gamma}U\left(\frac{n}{4k}\right)}{\frac{n}{k}U'\left(\frac{n}{k}\right)2^{-3\gamma}\left(\frac{2^{-\gamma}-1}{\gamma}\right)}\left(1+o_P(1)\right)\end{aligned}$$

Z těchto rovností již můžeme určit asymptotický rozptyl $\sqrt{k}\left(\hat{\gamma}_P(n,k)-\hat{\gamma}_P(n,4k)\right)$:

$$\frac{\gamma^2\frac{1}{2}\frac{3}{4}(1+2^{-2\gamma-1})}{(\log 2)^2(2^{-\gamma}-1)^2}$$

a asymptotický $Bias\ \sqrt{k}\left(\hat{\gamma}_P(n,k)-\hat{\gamma}_P(n,4k)\right)$:

$$\sqrt{k}A\left(\frac{n}{k}\right)\frac{2^{-\rho}}{-\log 2}\frac{\gamma}{2^{-\gamma}-1}\frac{1-2^{\rho}}{\rho}\frac{2^{-\gamma-\rho}-1}{\gamma+\rho}(1-2^{-2\rho})$$

Závěr důkazu bude probíhat stejně jako u předchozí věty. Za $A(t)=ct^{-\rho}$ a použijeme stejný postup jako v závěru důkazu věty 4.12.

Důsledek 4.30

Nechť $F\in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x)=c\,x^{-\rho}$, kde $c\neq 0$ a $\rho<0$. Označme index $k_0(n)$ takový, že hodnota $E(\hat{\gamma}_P(n,k)-\gamma)^2$ je minimální, index $\hat{k}_0(n)$ takový, že hodnota $E\left(\hat{\gamma}_P(n,k)-\hat{\gamma}_P(n,4k)\right)^2$ je minimální. Potom

$$\lim_{n\to\infty}\frac{\hat{k}_0(n)}{k_0(n}=\left(\frac{\frac{5}{4}}{(1-2^{-2\rho})^2}\right)^{\frac{1}{1-2\rho}}. \qquad (4.47)$$

Důkaz: Draisma, de Haan, Peng, Pereira (1999).

V dalším se již zaměříme na tvorbu vlastní bootstrapové procedury. Vytvořme $n_1 < n$ nezávislých výběrů z empirické d. f. vytvořené z $\boldsymbol{X_n} = \{X_1, \cdots, X_n\}$. Získáme hodnoty $X_1^*, \cdots, X_{n_1}^*$ a vytvoříme pořádkovou statistiku $X_{n_1,1}^*, X_{n_1,2}^*, \cdots, X_{n_1,n_1}^*$. Definujeme

$$\hat{\gamma}_P^*(n_1, k_1) = \frac{1}{log2}\, log\left(\frac{X_{n_1,n_1-k_1}^* - X_{n_1,n_1-2k_1}^*}{X_{n_1,n_{1-2k_1}}^* - X_{n_1,n_1-4k_1}^*}\right).$$

Podobně si budeme počínat s indexem $n_2 < n$,

Nyní uvedeme dvě věty, které umožňují využít metodu bootstrap pro Pickandsův odhad.

Věta 4.31

Nechť $F \in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x) = c\, x^{-\rho}$, kde $c \neq 0$ a $\rho < 0$. Nechť dále je $n_1 = O(n^{1-\varepsilon})$ pro $\varepsilon \in (0; 1)$. Označme index $k_{1,0}^*(n_1)$ takový, že $E\left(\left(\hat{\gamma}_P^*(n_1, k_1) - \hat{\gamma}_P^*(n_1, 4k_1)\right)^2 \middle| \boldsymbol{X_n}\right)$ je minimální. Potom pro $n \to \infty$ je

$$k_{1,0}^*(n_1) \Bigg/ \left\{\left(\frac{1 + 2^{-2\gamma-1}}{4\rho c^2 \left(\frac{1-2^\rho}{\rho}\right)^2 \left(\frac{2-\gamma-\rho-1}{\gamma+\rho}\right)^2 2^{-2\rho}} \cdot \frac{\frac{5}{4}}{(1-2^{-2\rho})^2}\right)^{\frac{1}{1-2\rho}} n_1^{\frac{-2\rho}{1-2\rho}}\right\} \xrightarrow{P} 1.$$

Důkaz: Draisma, de Haan, Peng, Pereira (1999).

Podobně můžeme počítat i hodnotu indexu $k_{2,0}^*(n_2)$. Posledním krokem bude nahrazení užití těchto hodnot v odhadu $k_0(n)$.

Věta 4.32

Nechť $F \in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x) = c\, x^{-\rho}$, kde $c \neq 0$ a $\rho < 0$. Nechť dále je $n_1 = O(n^{1-\varepsilon})$ pro $\varepsilon \in (0; 1/2)$ a dále $n_2 = \left[\frac{n_1^2}{n}\right]$, kde $[x]$ označuje celou část z hodnoty *x*. Označme index $k_{i,0}^*(n_i)$ takový, že $E\left(\left(\hat{\gamma}_P^*(n_i, k_i) - \hat{\gamma}_P^*(n_i, 4k_i)\right)^2 \middle| \boldsymbol{X_n}\right)$ je minimální (i=1,2). Označme dále funkci $f(\rho) = \left(\frac{\frac{5}{4}}{(1-2^{-2\rho})^2}\right)^{\frac{1}{1-2\rho}}$. Potom pro $n \to \infty$ je

$$\frac{\left(k_{1,0}^{*}(n_1)\right)^2}{\left(k_{2,0}^{*}(n_2)\right)^2 f\left(\frac{\log k_{1,0}^{*}(n_1)}{2\left(\log k_{1,0}^{*}(n_1) - log(n_1)\right)}\right)} \Big/ k_0(n) \xrightarrow{P} 1 \qquad (4.48).$$

Důkaz: Draisma, de Haan, Peng, Pereira (1999).

Důkazy všech předchozích vět a tvrzení jsou založeny na využití podmínek druhého řádu a asymptotického rozvoje funkce logaritmus a jsou obdobné větám (4.12), (4.16), (4.20) a (4.21).

Bootstrapová procedura:

I. krok. Položme $n_1 = [n^{1-\varepsilon}]$ pro hodnotu $\varepsilon \in (0; \frac{1}{2})$, kde $[x]$ označuje celou část čísla *x*. Vyberme z výběru o *n* hodnotách nový bootstrapový výběr o délce n_1. Budeme dále počítat $E\left(\left(\hat{\gamma}_P^*(n_1, k_1) - \hat{\gamma}_P^*(n_1, 4k_1)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$. Nalezneme hodnotu $k_{1,0}^{*}(n_1)$, pro který je předchozí výraz minimální.

II. krok: Položíme $n_2 = \left[{n_1}^2/_n\right]$ a zopakujeme krok I. s tím, že nalezneme hodnotu $k_{2,0}^{*}(n_2)$, pro kterou je $E\left(\left(\hat{\gamma}_P^*(n_2, k_2) - \hat{\gamma}_P^*(n_2, 4k_2)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$ minimální.

III. krok: Nyní vypočítáme odhad $k_0(n)$,

$$k_0(n) = \frac{\left(k_{1,0}^{*}(n_1)\right)^2}{\left(k_{2,0}^{*}(n_2)\right)^2 f\left(\frac{\log k_{1,0}^{*}(n_1)}{2\left(\log k_{1,0}^{*}(n_1) - log(n_1)\right)}\right)}$$

Všechna tato tvrzení a definice jsme uvedli proto, abychom mohli popsat asymptotické chování rozdílů jednotlivých odhadů od skutečné hodnoty γ. Detaily vyjádření ponecháme v příslušných článcích uvedených výše. Nalézt optimální velikost k_1 volby bootstrapového výběru je stejné jako najít minimální hodnotu výrazu

$k_1(n) = \underset{k}{arg\,min} AMSE\left(\hat{\gamma}^{(o)}(n,k) - \gamma\right)$, kde AMSE je asymptotický druhý moment a

$\hat{\gamma}^{(o)}(n,k)$ je kterýkoli z námi uvedených odhadů v části 2. Tuto hodnotu nemůžeme určit přímo z výše uvedeného výrazu, protože se v něm vyskytuje neznámá hodnota γ, ale využijeme pro daný odhad stejného indexu γ asymptotického chování pro upravené odhady.

Nalezneme pro daný typ odhadu sadu jednoduše změněných původních odhadů ($\gamma_H, \gamma_M, \gamma_P$) pro něž umíme najít hodnotu

$$k_2(n) = \underset{k}{arg\,min} AMSE\left(\hat{\gamma}^{(o)}(n,k) - \tilde{\gamma}\right), \text{ kde } \tilde{\gamma}$$

je výše uvedený změněný odhad a prokáže se, že $\frac{k_1(n)}{k_2(n)} \to 1$ v pravděpodobnosti. Dále se vytvoří další upravený odhad $\bar{\gamma}$, pro který vypočteme také hodnotu

$$k_3(n) = \underset{k}{arg\,min} AMSE(\bar{\gamma} - \tilde{\gamma})$$

Jestliže navíc zvolíme $n_2 = \frac{(n_1)^2}{n}$, potom platí

$$\frac{k_1(n)}{\left(\frac{(k_3(n_1))^2}{k_3(n_2)}\right)} \to 1$$

a $AMSE\left(\hat{\gamma}^{(o)}(n,k) - \tilde{\gamma}\right)$ je asymptoticky rovný $AMSE\left(\hat{\gamma}^{(o)}(n,k) - \gamma\right)$. Detaily výpočtů je možné nalézt pro momentový a Pickandsův odhad v (Draisma , de Haan , Peng , Pereira, 1999) a pro odhad Hillův v práci (Gomes, Oliveira, 2001).

Z takových závěrů je možné vytvořit algoritmus pro zpracování jednotlivých typů odhadů pomocí bootstrapu.

Dále tento algoritmus uvedeme:

1. Nechť velikost výběrového souboru je *n*, hodnotu první výběrové části volíme $n_1 = [n^{1-\epsilon}]$ pro $\epsilon \in \left(0, \frac{1}{2}\right)$ kde $[x]$ označuje celou část *x* a zvolme dále velikost druhé části $n_2 = \frac{n_1^2}{n}$. Dále zvolíme hodnotu počátečního indexu $k_{aux} = \left[2\sqrt{n}\right]$ (Drees, Kaufmann, 1998)
2. Pro výběr $(x_1, x_2, \cdots, x_n)$, spočítáme $\hat{\gamma}^{(o)}(n,k)$, *k*=1,2,…,*n-1*.
3. Pro *i* od 1 do *B* vygenerujeme nezávisle *B* bootstrapových výběrů $\left(x_1^*, x_2^*, \cdots, x_{n_2}^*\right)$ a $\left(x_1^*, x_2^*, \cdots, x_{n_2}^*, x_{n_{2+1}}^*, \cdots, x_{n_1}^*\right)$ o velikostech n_2 a n_1, z empirické d.f.

$$F_n^*(x) = \frac{\sum_{j=1}^{n} I\left[X_j \le x\right]}{n}$$

získáme $\left(t_{n_1,i}^*(k), t_{n_2,i}^*(k)\right)$, $1 \le i \le B$ vyšetřovaných hodnot statistik

$$T_{n_j}^*(k) = (\hat{\gamma}^{(o)}\left(n_j, k\right) - \hat{\gamma}^{(o)}(n, k_{aux}) | \boldsymbol{X_n}).$$

4. Pro obě vyšetřované hodnoty *j*=1, 2vypočteme

$$\widehat{MSE}^*(n_j,k) = \frac{\sum_{i=1}^{B}\left(t^*_{n_j,i}(k)\right)^2}{B} \quad k=1,2,\ldots,n\text{-}1\ ,\ j=1,2$$

5. Zvolíme $k_0^*(n_j) = \underset{1 \le k \le n-1}{arg\ min}\ \widehat{MSE}^*(n_j,k)$, j =1,2
6. Vypočteme $\hat{k}_0(n,k_{aux},n_1) = \left(\frac{(k_0^*(n_1))^2}{k_0^*(n_2)}\right)$. Jestliže $\hat{k}_0 \notin < 1,n)$, ukončíme výpočet, jinak pokračujeme dalším krokem.
7. Získáme $\hat{\gamma}^{(o)}(n_1,k_{aux}) = \hat{\gamma}^{(o)}\left(n_1,\hat{k}_0(n,k_{aux},n_1)\right)$.

V dalším textu uvedeme několik poznámek k uvedenému algoritmu.

- Obecná doporučená volba pro algoritmy tohoto typu je $\varepsilon = 0{,}15$. Ukazuje se, že menší volba vede k velikostem bootstrapových výběrů srovnatelných s *n*. V bootstrapovém souboru o velikosti $n_1 + n_2$ pak máme obecně mnoho duplicit a ty mohou výsledek nepříjemně ovlivnit. Jestliže bychom chtěli vybírat tak, aby výběry o délkách $n_1 + n_2$ byly menší než *n*, pak pro hodnotu $\varepsilon = 0{,}1$ musí *n* být aspoň 112 a pro hodnotu $\varepsilon = 0{,}05$ musí být dokonce nejméně 15 115. Pro hodnotu $\varepsilon = 0{,}15$ je rozsah výběru nejméně 15. Proto bývá tato hodnota doporučována pro výpočet. Naše simulační studie ale vede k jiným výsledkům.
- Jestliže v 5. kroku algoritmu je $\hat{k}_0 \notin < 1,n)$, potom většinou vygenerujeme nový výběrový soubor a provedeme výpočet znova.
- Jestliže máme ve výběrovém souboru záporné hodnoty, přepíšeme je jejich absolutními hodnotami (důvodem je výpočet hodnoty logaritmu v příslušném odhadu).
- V krocích 3. a 4. použijeme klasickou metodu Monte Carlo
- Je zřejmé, že nejdůležitější počáteční hodnotou je správná volba n_1. Tuto volbu spolu s volbou *B* budeme studovat pro jednotlivé odhady v části práce věnované simulační studii.

4.5. Optimal sample fraction pro PORT odhady

V této části budeme studovat využití metody sample fraction na skupinu semiparametrických odhadů, které se nazývají „PORT“ odhady. Vlastní název i metodologie těchto odhadů je vytvořena v základním článku Araújo (2006).

V mnoha aplikacích je zapotřebí mít k dispozici odhady, které jsou invariantní vzhledem k měřítku, ale především vzhledem k posunutí. To základní odhady typu Hillův odhad nebo momentový odhad nesplňovaly. Právě ve výše uvedeném článku se autoři zaměřili na tvorbu odhadů, které jsou obdobné např. Hillovu nebo momentovému odhadu, ale zároveň jsou invariantní vzhledem k měřítku a k posunutí.

Při použití techniky PORT pracujeme s upraveným výběrem, kdy původní hodnoty zmenšíme o náhodně stanovenou mez $X_{n_q:n}$ (následující zavedení je uvedeno v (2.49)

$$X^q = (X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \dots, X_{n_q+1:n} - X_{n_q:n})$$

kde $n_q = [nq] + 1$ a

1. $0 < q < 1$, pro d. f. F s konečným nebo nekonečným levým krajním bodem $x_* = inf\{x; F(x) > 0\}$, a proto je stanovená mez rovna empirickému kvantilu,
2. $q = 0$ s konečným levým krajním bodem, potom je náhodná mez volena jako výběrové minimum.

V předchozích částech jsme prokázali, že v případě, že $k = k(n)$ má následující dvě vlastnosti: 1. $\lim_{n\to\infty} k(n) \to \infty$ a 2. $\lim_{n\to\infty} \frac{k(n)}{n} = 0$ (pro další text budeme zjednodušeně označovat $k(n) = k$) a jsou zároveň splněny podmínky druhého řádu z definice 2.19 platí pro Hillův odhad (podle Věty 4.1)

$$\hat{\gamma}_H(n,k) - \gamma \overset{d}{=} \frac{\gamma}{\sqrt{k}} P_n^H + \frac{1}{1-\rho}\left(\frac{n}{k}\right) + o_p\left(A\left(\frac{n}{k}\right)\right)$$

a pro momentový odhad (důkaz Věty 4.12)

$$\hat{\gamma}^{(M)}(n,k) - \gamma \overset{d}{=} \frac{\sqrt{\gamma^2+1}}{\sqrt{k}} P_n^M + \frac{(\gamma(1-\rho)+\rho)A\left(\frac{n}{k}\right)}{\gamma(1-\rho)^2}\left(1 + o_P(1)\right),$$

kde P_n^H a P_n^M jsou náhodné veličiny asymptoticky rovné normovanému normálnímu rozdělení.

V následující větě rozšíříme tyto asymptotické vlastnosti i na PORT odhady. Nejdříve ale zavedeme ještě potřebná pomocná označení. Pro účely této části bude

$$\hat{\gamma}^{H(q)} = \hat{\gamma}_H(n,k)(X^q),\ \hat{\gamma}^{M(q)} = \hat{\gamma}^{(M)}(n,k)(X^q), 0 \le q < 1 \qquad (4.49)$$

Dále symbolem x_q^* označíme q – tý kvantil náhodné veličiny s d. f. *F*, tedy

$$F(x_q^*) = q,$$

speciálně je $x_0^* = x_F$. Potom je

$$X_{n_q:n} \xrightarrow{P} x_q^*, n \to \infty, 0 \le q < 1.$$

Pro zavedené odhady v (4.49) platí asymptotická representace uvedená v následující větě. Tato věta je uvedena ve výše uvedeném článku Araújo (2006) jako věta 2.1.

Věta 4.33

Nechť *k* je optimální hodnota pro metodu sample fraction a nechť jsou splněny podmínky druhého řádu 2.19., nechť dále je $0 \le q < 1$ a nechť dále T označuje obecně *H* nebo *M*. Potom platí následující asymptotická representace

$$\hat{\gamma}^{T(q)} - \gamma \overset{d}{=} \frac{\sigma_T}{\sqrt{k}} P_n^T + \left(c_T A\left(\frac{n}{k}\right) + d_T \frac{x_q^*}{U\left(\frac{n}{k}\right)} \right)\left(1 + o_P(1)\right) \qquad (4.50)$$

kde P_n^T jsou náhodné veličiny asymptoticky rovné normálnímu normovanému rozdělení,

$$\sigma_H^2 = \gamma^2, c_H = \frac{1}{1-\rho}, d_H = \frac{\gamma}{\gamma+1}, \qquad (4.51)$$

$$\sigma_M^2 = \gamma^2 + 1, c_M = \frac{\gamma(1-\rho)+\rho}{\gamma(1-\rho)^2}, d_M = \left(\frac{\gamma}{\gamma+1}\right)^2. \qquad (4.52)$$

Provedeme dále důkaz této věty. Obsahuje totiž rysy podobných vět, které jsme uvedli v části sample fraction pro momentový odhad a pro Hillův odhad.

Důkaz věty 2.1 bude proveden pomocí dvou lemmat.

Lemma 4.34

Nechť *F* je d.f. náhodné veličiny X a předpokládejme, že *U* je kvantilová funkce, která vystupuje v podmínkách druhého řádu v definici 2.19. Vyšetřujme náhodnou veličinu $X_q = X - x_q^*$ s d. f. $F_q(x) = F(x) + x_q^*$ a k ní kvantilovou funkci $U_q(t) = F_q^{\leftarrow}\left(1 - \frac{1}{t}\right) = U(t) - x_q^*$.

Potom U_q splňuje také podmínky druhého řádu

$$\lim_{t\to\infty} \frac{\frac{U_q(tx)}{U_q(t)} - x^\gamma}{A_q(t)} = x^\gamma \left(\frac{x^{\rho_q} - 1}{\rho_q}\right) \qquad (4.53)$$

pro $x > 0$, $\rho_q \le 0$, kde

$$\left(A_q(t), \rho_q\right) = \begin{cases} (A(t), \rho) & \rho > -\gamma, \\ \left(A(t) + \frac{\gamma x_q^*}{U(t)}, -\gamma\right) & \rho = \gamma, \\ \left(\frac{\gamma x_q^*}{U(t)}, -\gamma\right) & \rho < -\gamma. \end{cases} \tag{4.54}$$

Důkaz:

Nejdříve zjistíme hodnotu podílu funkcí U_q s argumenty *t x* a *x*. Tedy

$$\frac{U_q(tx)}{U_q(t)} = \frac{U(tx) - x_q^*}{U(t) - x_q^*} = \frac{U(t\,x)}{U(t)}\left(\frac{1 - {x_q^*}/{U(tx)}}{1 - {x_q^*}/{U(t)}}\right) =$$

$$= \frac{U(t\,x)}{U(t)}\left(1 + x_q^* \frac{{1}/{U(t)} - {1}/{U(tx)}}{1 - {x_q^*}/{U(t)}}\right) =$$

$$= \frac{U(t\,x)}{U(t)}\left(1 + \frac{x_q^*}{U(t)}\left(1 - \frac{U(t)}{U(tx)}\right)\left(\frac{1}{1 - {x_q^*}/{U(t)}}\right)\right) =$$

$$= \frac{U(t\,x)}{U(t)}\left(1 + \frac{x_q^*}{U(t)}\left(1 - \frac{U(t)}{U(tx)}\right)(1 + o(1))\right) =$$

$$= x^\gamma\left(1 + \frac{x^\rho - 1}{\rho}\,A(t)(1 + o(1))\right)\left(1 + \frac{\gamma x_q^*}{U(t)}\frac{x^{-\gamma} - 1}{-\gamma}(1 + o(1))\right) =$$

$$= x^\gamma\left(1 + \frac{x^\rho - 1}{\rho}\,A(t) + \frac{\gamma x_q^*}{U(t)}\frac{x^{-\gamma} - 1}{-\gamma} + o(A(t)) + o\left(\frac{1}{U(t)}\right)\right)$$

Jestliže nyní dosadíme do vztahu (4.53), za podmínek (4.54) a provedeme limitu, ověříme platnost Lemma 4.34.

Další pomocné tvrzení je obdoba tvrzení věty 4.7, která jsme používali při důkazu asymptotického rozvoje Hillova odhadu a momentového odhadu. Využijeme při něm vlastnosti Paretova rozdělení a zároveň slabého zákona velkých čísel.

Lemma 4.35

Nechť $M_n^{(r,q)} = \frac{1}{k}\sum_{j=1}^{k}\left(log\,\frac{X_{n-j+1:n} - X_{n_q:n}}{X_{n-k:n} - X_{n_q:n}}\right)^r, r = 1, 2$ a $0 \le q < 1$. Nechť dále k je optimální hodnota pro metodu sample fraction.

Potom pro tuto hodnotu *k* a při platnosti podmínek druhého řádu uvedených v definici 2.19 a pro libovolné $0 \leq q < 1$ platí

$$M_n^{(r,q)} - \frac{1}{k}\sum_{j=1}^{k}\left(log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*}\right)^r = o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right), r = 1,2.$$

Důkaz:

Nejdříve zvolme $r = 1$. Využijeme nyní následující aproximace: $\log(1+x) \sim x$, pro $x \to 0$ a dále je $X_{n_q:n} = x_q^*(1 + o_P(1))$. Na základě těchto aproximací budeme dále upravovat:

$$M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*} =$$

$$= \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - X_{n_q:n}}{X_{n-k:n} - X_{n_q:n}} - \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*} =$$

$$= \frac{1}{k}\sum_{j=1}^{k} log\frac{1 - \frac{X_{n_{q:n}}}{X_{n-j+1:n}}}{1 - \frac{X_{n_{q:n}}}{X_{n-k:n}}} - log\frac{1 - \frac{x_q^*}{X_{n-j+1:n}}}{1 - \frac{x_q^*}{X_{n-k:n}}} =$$

$$= \frac{1}{k}\sum_{j=1}^{k}\left(\frac{X_{n_q:n}}{X_{n-k:n}} - \frac{X_{n_q:n}}{X_{n-j+1:n}} + \frac{x_q^*}{X_{n-j+1:n}} - \frac{x_q^*}{X_{n-k:n}}\right)\left(1 + o_P(1)\right) =$$

$$= \frac{X_{n_q:n} - x_q^*}{X_{n-k:n}}\frac{1}{k}\sum_{j=1}^{k}\left(1 - \frac{X_{n-k:n}}{X_{n-j+1:n}}\right)\left(1 + o_P(1)\right) =$$

$$= \frac{o_P(1)}{X_{n-k:n}}\frac{1}{k}\sum_{j=1}^{k}\left(1 - \frac{X_{n-k:n}}{X_{n-j+1:n}}\right)\left(1 + o_P(1)\right)$$

Dále použijeme metodu shodnou s postupem důkazu. Označíme $\{Y_j\}_{j=1}^{k}$ n.n.v. *Y* typu Paretovo rozdělení s d. f. $F_Y(t) = 1 - \frac{1}{t}$, pro $t > 1$ a $\{Y_{j:k}\}_{j=1}^{k}$ pořadovou statistiku.

Protože je $X_{n-k:n} \overset{d}{=} U(Y_{n-k:n})$, kde $Y_{n-k:n}$ je *(n-k)* pořadová statistika asociovaná s Paretovým rozdělením o velikosti *n* a dále je $\left(\frac{k}{n}\right) Y_{n-k:n} \xrightarrow{P} 1$ pro optimální hodnotu *k*. Potom je tedy $\frac{X_{n-k:n}}{U\left(\frac{n}{k}\right)} \xrightarrow{P} 1$. Proto je

$$\left\{\frac{Y_{n-j+1:n}}{Y_{n-k:n}}\right\}_{j=1}^{k} \overset{d}{=} \{Y_{n-j+1:k}\}_{j=1}^{k}$$

Vrátíme se dále k úpravám:

$$M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*} =$$

$$= \frac{o_P(1)}{U(Y_{n-k:n})}\frac{1}{k}\sum_{j=1}^{k}\left(1 - \frac{U(Y_{n-k:n})}{U\left(\frac{Y_{n-j+1:n}}{Y_{n-k:n}}\, Y_{n-k:n}\right)}\right)\left(1 + o_P(1)\right) =$$

$$= o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right)\frac{1}{k}\sum_{j=1}^{k}\left(1 - Y_{k-j+1:k}^{-\gamma}\right)\left(1 + o_P(1)\right) =$$

$$= \frac{1}{k}\sum_{j=1}^{k}\left(1 - Y_j^{-\gamma}\right)o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right)\left(1 + o_P(1)\right)$$

Pro další úpravu použijeme slabý zákon velkých čísel Štěpán (1987) IV.2.6. Upozorněme, že $E(Y^{-\gamma}) = \frac{1}{\gamma+1}$

$$M_n^{(1,q)} - \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*} =$$

$$= \frac{\gamma}{\gamma+1}\left(1 + o_P\left(1/\sqrt{k}\right)\right)\, o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right) = o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right)$$

Pro hodnotu $r = 2$ provedeme důkaz obdobně.

Poznámka 4.36

Jestliže $0 < q < 1$. Potom platí

$$X_{n_q:n} - x_q^* = O_P\left(\frac{1}{\sqrt{n}}\right)$$

a pro $r = 1, 2$ je

$$\sqrt{k}\left[M_n^{(r,q)} - \frac{1}{k}\sum_{j=1}^{k}\left(log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*}\right)^r\right] = O_P\left(\frac{\sqrt{\frac{k}{n}}}{U\left(\frac{n}{k}\right)}\right) = o_P(1)$$

Nyní dokončíme důkaz věty 4.33. Nejdříve dokážeme tvrzení pro PORT – Hillův odhad.

Podle Lemma 4.35 a závěrů Věty 4.1 pro data $X_q = X - x_q^*$ je

$$\hat{\gamma}^{H(q)} = \hat{\gamma}_H(n,k)(X^q) = \frac{1}{k}\sum_{j=1}^{k} log\frac{X_{n-j+1:n} - x_q^*}{X_{n-k:n} - x_q^*} + o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right),$$

dále podle Lemma 4.34 již můžeme psát

$$\hat{\gamma}^{H(q)} \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}}\ P_n^H + \frac{A_q\left(\frac{n}{k}\right)}{1-\rho_q}\left(1 + o_P(1)\right) + o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right),$$

kde hodnoty s indexem jsou uvedeny v (4.54).

Podobně pro PORT – momentový odhad podle Lemma 4.35 a důkazu Věty 4.12 pro data $X_q = X - x_q^*$ je

$$\hat{\gamma}^{M(q)} \overset{d}{=} \gamma + \frac{\sqrt{\gamma^2 + 1}}{\sqrt{k}}\ P_k^M + \frac{\left(\gamma\left(1-\rho_q\right)\right)A_q\left(\frac{n}{k}\right)}{\gamma\left(1-\rho_q\right)^2}\left(1 + o_P(1)\right) + o_P\left(\frac{1}{U\left(\frac{n}{k}\right)}\right)$$

Poznámka 4.37

Předpokládejme, že $\boldsymbol{X}$ jsou data, která jsou popsána pomocí d. f. F. Nechť c je konstanta, označme $\boldsymbol{X}_c = \boldsymbol{X} + c$. Potom data $\boldsymbol{X}_c$ jsou popsána pomocí d. f. $F_c(x) = F(x-c)$ a jejich kvantilová funkce $U_c(t) = c + U(t)$. Potom (4.50) je splněno, když zaměníme $\hat{\gamma}^{H(q)}$ za $\hat{\gamma}_H(n,k)|c$ (Hillův odhad uplatněný na data posunutá o hodnotu c) a zaměníme dále x_q^* za –c. Podobně si můžeme počínat i v případě momentového odhadu.

Jestliže platí podmínky druhého řádu s d. f. F a s funkcí $A(t) = A_c(t)$ z podmínek druhého řádu, potom

$$\frac{U_c(t\,x)}{U_c(t)} = \frac{U(tx)}{U(t)}\left(1 - \frac{c\,\gamma}{U(t)}\left(\frac{x^{-\gamma}-1}{-\gamma}\right) + o\left(\frac{1}{U(t)}\right)\right).$$

Tedy odtud je

$$\frac{U_c(t\,x)}{U_c(t)} - x^{\gamma} = x^{\gamma}\left(A(t)\left(\frac{x^{\rho}-1}{\rho}\right) - \frac{c\,\gamma}{U(t)}\left(\frac{x^{-\gamma}-1}{-\gamma}\right) + o(A(t)) + o\left(\frac{1}{U(t)}\right)\right).$$

Z toho můžeme nyní nalézt následující representaci např. Hillova odhadu s posunutím c (budeme označovat $\hat{\gamma}_{n,k}^{H/c}$)

$$\hat{\gamma}_{n,k}^{H/c} = \gamma + \frac{\sigma_H}{\sqrt{k}} P_n^H + \left(c_H A\left(\frac{n}{k}\right) - d_H \frac{c}{U\left(\frac{n}{k}\right)} \right)\left(1 + o_P(1)\right),$$

podobným způsobem můžeme nalézt representaci momentového odhadu $\hat{\gamma}_{n,k}^{M/c}$ s posunutím *c*.

Uvedeme nejdříve ještě jeden důsledek předchozí věty 4.33.

Věta 4.38

Nechť jsou μ_1 a μ_2 reálné hodnoty a nechť dále *T* označuje postupně *H* (Hillův odhad) a *M* (momentový odhad).

a) Pro $\gamma > -\rho$,

$$\hat{\gamma}^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}} P_n^T + c_T A_q\left(\frac{n}{k}\right)\left(1 + o_P(1)\right),$$

Jestliže $\sqrt{k}\, A\left(\frac{n}{k}\right) \to \mu_1$, potom

$$\sqrt{k}\left(\hat{\gamma}^{T(q)} - \gamma\right) \underset{n \to \infty}{\overset{d}{\to}} N(\mu_1 c_T, \sigma_T^2).$$

b) Pro $\gamma < -\rho$,

$$\hat{\gamma}^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}} P_n^T + d_T \frac{x_q^*}{U\left(\frac{n}{k}\right)}\left(1 + o_P(1)\right),$$

Jestliže $\sqrt{k}\, U\left(\frac{n}{k}\right) \to \mu_2$, potom

$$\sqrt{k}\left(\hat{\gamma}^{T(q)} - \gamma\right) \underset{n \to \infty}{\overset{d}{\to}} N\left(\mu_2 d_T x_q^*, \sigma_T^2\right).$$

c) Pro $\gamma = -\rho$,

$$\hat{\gamma}^{T(q)} \overset{d}{=} \gamma + \frac{\sigma_T}{\sqrt{k}} P_n^T + \left(c_T A_q\left(\frac{n}{k}\right) + d_T \frac{x_q^*}{U\left(\frac{n}{k}\right)} \right)\left(1 + o_P(1)\right).$$

Jestliže $\sqrt{k}\, A\left(\frac{n}{k}\right) \to \mu_1$ a $\sqrt{k}\, U\left(\frac{n}{k}\right) \to \mu_2$, potom

$$\sqrt{k}\left(\hat{\gamma}^{T(q)} - \gamma\right) \underset{n \to \infty}{\overset{d}{\to}} N\left(\mu_1 c_T + \mu_2 d_T x_q^*, \sigma_T^2\right).$$

V dalším se zaměříme na upravený typ odhadu PORT, který je vytvořen pro náhodné veličiny tzv. Hall-Welshova typu a zároveň je tvaru, který minimalizuje rozptyl odchylek od skutečného řešení (třída takovýchto odhadů se nazývá *MVRB*). Takovéto typy odhadu byly zavedeny v článcích Gomes, de Haan, Henriques Rodrigues (2004), Caeiro, Gomes, Pestana (2005). Jsou založeny na odhadech parametru ρ ($\rho < 0$), který se vyskytuje v podmínkách druhého řádu (2.21) a parametru β ($\beta \neq 0$), který se vyskytuje ve velmi široké třídě uvedené v následující definici 4.39.

Definice 4.39

Třídou Hall-Welshova typu nazveme všechny náhodné veličiny, pro které je kvantilová funkce *U* rovna

$$U(t) = C\ t^{\gamma}\left(1 + \frac{A(t)}{\rho} + o(t^{\rho})\right),\ A(t) = \gamma\ \beta\ t^{\rho}, \qquad (4.55)$$

jakmile $t \to \infty$, kde $C > 0, \gamma > 0, \rho < 0, \beta \neq 0$.

Z definice 4.39 vyplývá, že existuje pomalu měnící funkce v nekonečnu $L(t)$taková, že $U(t) = t^{\gamma} L(t)$. Předpoklad platnosti (4.55) je ekvivalentní volbě $A(t) = \gamma\ \beta\ t^{\rho}, \rho < 0$ v mnohem obecnějších podmínkách druhého řádu v (2.21).

Definice 4.40

Pro klasické odhady Hillův a momentový zavedeme odhad typu, který minimalizuje rozptyl a redukuje bias (obecné označení třídy těchto odhadů je MVRB) odhadu. Pro Hillův odhad ho označíme

$$\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho}) = \hat{\gamma}_{H}(n, k)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) \qquad (4.56)$$

a dále pro momentový odhad

$$\hat{\gamma}_{\bar{M}}(n, k, \hat{\beta}, \hat{\rho}) = \hat{\gamma}_{M}(n, k)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) - \left(\frac{\hat{\beta}\hat{\rho}}{(1-\hat{\rho})^{2}}\right)\left(\frac{n}{k}\right)^{\hat{\rho}}, \qquad (4.57)$$

kde $\hat{\beta}$ a $\hat{\rho}$ jsou konzistentní odhady parametrů druhého řádu, které jsou uvedené v (4.55).

V základním článku Caieiro, Gomes, Pestana (2005) jsou dokázána tvrzení o základních vlastnostech odhadů uvedených v (4.56) a (4.57). V následující větě je uvedeno tvrzení o asymptotickém chování těchto odhadů.

Věta 4.41

Nechť jsou splněny předpoklady definice 4.39. Nechť $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19). Potom platí

$$\hat{\gamma}_{\bar{H}}(n,k,\beta,\rho) \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}} Z_k^{(1)} + R_k, \tag{4.58}$$

kde $R_k = o_p\left(A\left(\frac{n}{k}\right)\right)$ a $Z_k^{(1)}$ je asymptoticky normované normální rozdělení. Navíc platí dále

$$\sqrt{k}(\hat{\gamma}_{\bar{H}}(n,k,\beta,\rho) - \gamma) \xrightarrow{d} N(0,\gamma^2), \tag{4.59}$$

kdykoli $\sqrt{k}\, A\left(\frac{n}{k}\right) \to \lambda$, kde λ je libovolné reálné číslo. Konvergence je chápána pro $n \to \infty$.

Důkaz:

Vycházíme z předpokladu, že všechny potřebné parametry jsou známy. Potom z klasické věty (4.1) pro Hillův odhad platí

$$\hat{\gamma}_H(n,k) \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}} Z_k^{(1)} + \frac{1}{1-\rho} A\left(\frac{n}{k}\right)(1 + o_P((1))$$

Tuto rovnost budeme nyní aplikovat na $\hat{\gamma}_{\bar{H}}(n,k,\beta,\rho)$.

$$\hat{\gamma}_{\bar{H}}(n,k,\beta,\rho) \overset{d}{=} \left(\gamma + \frac{1}{1-\rho} A\left(\frac{n}{k}\right)(1 + o_P((1))\right)\left(1 - \frac{A\left(\frac{n}{k}\right)}{\gamma(1-\rho)}\right) \overset{d}{=}$$

$$\overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}} Z_k^{(1)} + o_P\left(A\left(\frac{n}{k}\right)\right)$$

Tím je tvrzení (4.58) dokázáno. Druhá část tvrzení věty je dokázána v článku Caieiro, Gomes, Pestana (2005).

Je zřejmé, že pro práci s odhadem $\hat{\gamma}_{\bar{H}}(n,k,\hat{\beta},\hat{\rho})$ nebo $\hat{\gamma}_{\bar{M}}(n,k,\hat{\beta},\hat{\rho})$ bude nutné nalézt konzistentní odhady parametrů β a ρ .

Odhad parametru ρ.

V článku Alves (2003) byla zavedena třída odhadů parametru ρ z podmínek druhého řádu, které mají pro $\rho < 0$ vlastnost asymptotické normality. Tato třída je parametrizována pomocí parametru $\tau \geq 0$ a je definována následně

$$\hat{\rho}_\tau(k) = \hat{\rho}_n^{(\tau)} = -\left|\frac{3\left(T_n^{(\tau)}(k)-1\right)}{T_n^{(\tau)}(k)-3}\right|, \tag{4.60}$$

kde

$$T_n^{(\tau)}(k) = \begin{cases} \dfrac{\left(M_n^{(1)}(k)\right)^{\tau} - \left(M_n^{(2)}(k)/2\right)^{\tau/2}}{\left(M_n^{(2)}(k)/2\right)^{\tau/2} - \left(M_n^{(3)}(k)/6\right)^{\tau/3}} & \tau > 0 \\ \dfrac{ln\left(M_n^{(1)}(k)\right) - \frac{1}{2} ln\left(M_n^{(2)}(k)/2\right)}{\frac{1}{2} ln\left(M_n^{(2)}(k)/2\right) - \frac{1}{3} ln\left(M_n^{(3)}(k)/6\right)} & \tau = 0, \end{cases}$$

hodnotu výrazu $M_n^{(j)}(k)$ definujeme takto,

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} \left(ln \frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^j , j \geq 1.$$

V následující větě 4.42, která je uvedená v článku Alves (2003) jako součást vět 2.1 a 3.1, popíšeme základní vlastnosti tohoto odhadu parametru ρ .

Věta 4.42

Nechť jsou splněny podmínky druhého řádu a zároveň $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19), kde $\rho < 0$ a nechť dále platí $\sqrt{k}\, A(n/k) \rightarrow \infty$ pro $n \rightarrow \infty$. Potom

$$\hat{\rho}_n^{(\tau)} \xrightarrow{P} \rho,$$

pro $n \rightarrow \infty$ a pro libovolné τ.

Důkaz je proveden ve výše uvedeném článku.

Poznámka 4.43

V článku Alves (2003) je studována volba hodnoty τ. Ukazuje se, že v praxi je vhodné volit hodnotu $\tau = 0$, jestliže parametr $\rho \in <-1, 0)$ a $\tau = 1$, jestliže parametr $\rho \in (-\infty, -1)$. Zároveň je vhodná počáteční volba $k_0 = \min\left(n-1, \frac{2n}{\ln(\ln(n))}\right)$. Doporučuje se zvolit několik hodnot τ a na základě nich i možných odhadů parametru ρ, abychom zajistili vyšší přesnost pro velké hodnoty *k*. Pro volbu hodnoty k_1, která je menší než k_0. Doporučuje se volit hodnotu k takto

$$k = min\left(n-1, \left(\frac{2n^{0,995}}{\ln(\ln(n))}\right)\right) \qquad (4.61)$$

Potom pro $\hat{\rho} - \rho = O_P((n^{0,005}\ln(\ln(n)))^{\rho})$ (pokud $\rho > -49{,}75$) a tudíž pro prostřední posloupnost $k = k(n)$ splňující předpoklady (2.18) a (2.19), je $(\hat{\rho} - \rho)\, ln\left(\frac{n}{k}\right) = o_P(1)$ a jestliže navíc platí $\sqrt{k}\, A\left(\frac{n}{k}\right) \rightarrow \lambda$, kde λ je reálné číslo, potom je $\sqrt{k}\, A\left(\frac{n}{k}\right)\, (\hat{\rho} - \rho)\, ln\left(\frac{n}{k}\right) = o_P(1)$.

Odhad parametru β

Označme $\hat{\rho}$ odhad parametru ρ typu (4.60) počítaného pomocí hodnoty k z (4.61). Potom je odhad parametru β (definován v článku Gomes, Martins (2002)) stanoven takto

$$\hat{\beta}_{\hat{\rho}}(k) = \left(\frac{k}{n}\right)^{\hat{\rho}} \frac{\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}\right)\left(\frac{1}{k}\sum_{i=1}^{k}U_i\right) - \left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}U_i\right)}{\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}\right)\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}U_i\right) - \left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-2\hat{\rho}}U_i\right)}, \tag{4.62}$$

kde $U_i = i\left(ln(X_{n-i+1:n}) - ln(X_{n-i:n})\right)$, pro $1 \le i \le k$.

V následující větě, která je dokázána v článku Gomes, Martins (2002), je šetřena konzistence odhadu (4.62) a zároveň jsou uvedeny jeho asymptotické vlastnosti.

Věta 4.44

Nechť jsou splněny podmínky druhého řádu (2.21) s $A(t) = \gamma\,\beta\,t^{\rho}$ a zároveň $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19), kde $\rho < 0$ a nechť dále platí $\sqrt{k}\,A(n/k) \to \infty$ pro $n \to \infty$. Potom $\hat{\beta}_{\hat{\rho}}(k)$stanovený v (4.62), je konzistentním odhadem β, kdykoli $(\hat{\rho} - \rho)\,o_P(1/\ln(n))$. Navíc jestliže je hodnota parametru ρ známa, platí

$$\hat{\beta}_{\hat{\rho}}(k) \overset{d}{=} \beta + \frac{\gamma\,\beta\,(1-\rho)\sqrt{1-2\rho}}{\rho\sqrt{k}\,A(n/k)}\,W_k^B + R_k^B, \tag{4.63}$$

kde $R_k^B = o_P(1)$ a W_k^B je asymptotické s normovaným normálním rozdělením. Přesněji platí

$$W_k^B = \frac{(1-\rho)\sqrt{1-2\rho}}{|\rho|}\left(\frac{Z_k^{(1)}}{1-\rho} - \frac{Z_k^{(1-\rho)}}{\sqrt{1-2\rho}}\right), \tag{4.64}$$

kde $Z_k^{(\alpha)} = \sqrt{(2\alpha-1)\,k}\,\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{\alpha-1}E_i - \frac{1}{\alpha}\right)$, E_i jsou klasické nezávislé náhodné veličiny typu exponenciálního.

Důkaz tohoto tvrzení je založen na využití Réniyho representace exponenciálních pořádkových statistik podobně jako v důkazu Lemma 4.6. Asymptotická representace (4.63) platí, jakmile je $\hat{\beta}_{\hat{\rho}}(k)$ počítáno pro hodnoty k uvedené v (4.61). Jestliže navíc $\sqrt{k}\,A(n/k)R_k^B \to \lambda_k^B$ je reálné číslo, je $\hat{\beta}_{\hat{\rho}}(k)$ asymptoticky normální a jestliže vyjdeme ze vztahu (4.63), potom

$$\hat{\beta}_{\hat{\rho}}(k) - \beta \overset{d}{\sim} -\beta\,ln(n/k)\,(\hat{\rho}_{\tau}(k) - \rho). \tag{4.65}$$

Ve výše uvedeném článku je také dokázáno, že pokud budeme brát hodnotu k podle (4.61) jak pro odhad β, tak i pro odhad ρ (odhady (4.60) a (4.62)) bude řád konvergence $\left(\hat{\beta} - \beta\right)$ roven $O(ln(n)\,(n^{0,005}\ln(\ln(n)))^{\rho})$.

V další části se budeme zabývat aplikací dvou předchozích odhadů v odhadech (4.56) a (4.57). Následující dvě věty popisují asymptotické chování (4.56).

Věta 4.45

Nechť jsou splněny předpoklady definice 4.39. Nechť $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19). Uvažujme odhad $\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho})$, kde $\hat{\beta}$ a $\hat{\rho}$ jsou odhady (4.60) a (4.62), oba počítané pomocí hodnoty k z výrazu ((4.61) a takové, že platí $\hat{\rho} - \rho = o_P(1/\ln(n)))$. Potom $\sqrt{k}\left(\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho}) - \gamma\right)$ je asymptoticky normální se střední hodnotou 0 a rozptylem γ^2 a dále je $\sqrt{k}\, A(n/k) \to \lambda$, kde λ je libovolné reálné číslo.

Důkaz: (proveden v Caieiro, Gomes, Pestana (2007))

Využijeme dále Taylorův rozvoj a převedeme $\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho})$ na vyšetřování $\hat{\gamma}_{\bar{H}}(n, k, \beta, \rho)$

$$\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho}) \overset{d}{=} \hat{\gamma}_H(n, k)\left(1 - \frac{\beta}{1-\rho}\left(\frac{n}{k}\right)^{\rho} - (\hat{\beta} - \beta)\frac{1}{1-\rho}\left(\frac{n}{k}\right)^{\rho}(1 + o_P(1))\right.$$

$$\left. - \frac{\beta}{1-\rho}(\hat{\rho} - \rho)\left(\frac{n}{k}\right)^{\rho}\left(\frac{1}{1-\rho} + ln(n/k)\right)(1 + o_P(1))\right) \overset{d}{=}$$

$$\overset{d}{=} \hat{\gamma}_{\bar{H}}(n, k, \beta, \rho) - \frac{A(\frac{n}{k})}{1-\rho}\left(\frac{\hat{\beta} - \beta}{\beta} + (\hat{\rho} - \rho)\, ln(n/k)\right)(1 + o_P(1))$$

Dále $\hat{\beta}$ a $\hat{\rho}$ jsou konzistentní odhady parametrů β a ρ, $(\hat{\rho} - \rho)\ln\left(\frac{n}{k}\right) = o_P(1)$, sčítance $\left(\hat{\beta} - \beta\right)$ a $(\hat{\rho} - \rho)$ v předposlední závorce výrazu na pravé straně jsou oba $o_P(A\left(\frac{n}{k}\right))$ a podle předpokladů je $\sqrt{k}\, A(n/k) \to \lambda$. Z těchto skutečností vyplývají závěry věty.

Jestliže budeme hledat odhady parametrů β a ρ pomocí stejné hodnoty k, můžeme očekávat, že se zvětší variabilita výsledného odhadu $\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}, \hat{\rho})$. Následující věta tuto informaci upřesňuje.

Věta 4.46

Nechť jsou splněny předpoklady definice 4.39. Nechť $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19) a nechť $\sqrt{k}\, A(n/k) \to \lambda$, kde λ je libovolné reálné číslo, potom

$$\sqrt{k}\left(\hat{\gamma}_{\bar{H}}(n, k, \hat{\beta}_{\hat{\rho}}(k), \hat{\rho}) - \gamma\right) \underset{n \to \infty}{\overset{d}{\to}} N\left(0, \sigma^2 = \gamma^2\left(\frac{1-\rho}{\rho}\right)^2\right) \qquad (4.66)$$

tedy asymptotický rozptyl se zmenšil o násobek $\left(\frac{1-\rho}{\rho}\right)^2$, který je menší než 1, protože $\rho \leq 0$.

Důkaz: (proveden v Caieiro, Gomes, Pestana (2007))

Podle (4.56) je

$$\hat{\gamma}_{\bar{H}}\left(n,k,\hat{\beta}_{\hat{\rho}}(k),\hat{\rho}\right) = \hat{\gamma}_H(n,k)\left(1-\frac{\hat{\beta}_{\hat{\rho}}(k)}{1-\hat{\rho}}\right)\left(\frac{n}{k}\right)^{\hat{\rho}},$$

provedeme úpravu podobnou jako v předchozí větě 4.45

$$\hat{\gamma}_{\bar{H}}\left(n,k,\hat{\beta}_{\hat{\rho}}(k),\hat{\rho}\right) = \hat{\gamma}_{\bar{H}}(n,k,\beta,\rho) - \frac{A(\frac{n}{k})}{1-\rho}\left(\frac{\hat{\beta}_{\hat{\rho}}(k)-\beta}{\beta} + (\hat{\rho}-\rho)\,ln(n/k)\right)(1+o_P(1))$$

Protože výraz $\frac{\hat{\beta}_{\hat{\rho}}(k)-\beta}{\beta}$ je řádu konvergence $1/\left(\sqrt{k}\,A(\frac{n}{k})\right)$, je po roznásobení první sčítanec řádu $1/\sqrt{k}$. Pro další úpravu použijeme vztahy (4.58), (4.59), (4.63) a (4.64),

$$\frac{\gamma}{\sqrt{k}}\left(Z_k^{(1)} + \frac{(1-\rho)(1-2\rho)}{\rho^2}\left(\frac{Z_k^{(1)}}{1-\rho} - \frac{Z_k^{(1-\rho)}}{\sqrt{1-2\rho}}\right)\right),$$

po úpravě je

$$\frac{\gamma}{\sqrt{k}}\left(\left(\frac{1-\rho}{\rho}\right)^2 Z_k^{(1)} + \frac{(1-\rho)\sqrt{1-2\rho}}{\rho^2} Z_k^{(1-\rho)}\right),$$

kde $Z_k^{(\alpha)}$ jsou náhodné veličiny, které jsou asymptotické s normovaným normálním rozdělením. Dále z článku Caieiro, Gomes, Pestana (2007) vyplývá, že asymptotická kovariance mezi $Z_k^{(1)}$ a $Z_k^{(1-\rho)}$ je rovna $\sqrt{1-2\rho}\,/(1-\rho)$. Využijeme dále i platnost $\sqrt{k}\,A\left(\frac{n}{k}\right)(\hat{\rho}-\rho)\ln\left(\frac{n}{k}\right) \to 0$. Odtud již vyplývají závěry věty.

K tomu, abychom mohli studovat AMSE dané metody, musíme umět odhadnout asymptotické chování daných odhadů. Ukáže se, že toto chování je velmi podobné jak klasickým odhadům Hillovým a momentovým, ale i jejich PORT verzím.

Věta 4.47

Nechť jsou splněny předpoklady definice 4.39. Nechť $k = k_n$ je prostřední posloupnost splňující předpoklady (2.18) a (2.19). Potom existuje posloupnost náhodných veličin Z_k^C, které jsou asymptotické normovanému normálnímu rozdělení (symbolem C označujeme oba možné odhady (4.56) a (4.57) z definice 4.40) a reálná čísla $\sigma_C > 0$ a čísla $b_{C,1}$ ($b_{\bar{H},1} = \frac{1}{1-\rho}$ resp. $b_{\bar{M},1} = \frac{1}{1-\rho} + \frac{\rho}{\gamma\,(1-\rho)}$) a $b_{C,2}$ taková, že platí následující asymptotická rovnost

$$C(k) \overset{d}{=} \gamma + \frac{\sigma_C Z_k^C}{\sqrt{k}} + b_{C,1} A\left(\frac{n}{k}\right) + b_{C,2} A^2\left(\frac{n}{k}\right) (1 + o_P(1)) \tag{4.67}$$

Jestliže budeme mít k dispozici konzistentní odhady parametrů β a ρ například $\hat{\beta}$ a $\hat{\rho}$ takové, že platí $\hat{\rho} - \rho = o_P(1/\ln(n)))$, můžeme garantovat existenci dvojice reálných čísel $\left(b_{\bar{C},1}, b_{\bar{C},2}\right)$ takových, že pro odpovídající hodnoty k takové, že jestliže $\sqrt{k}\, A^4\left(\frac{n}{k}\right) \to \lambda_A$ je reálné číslo, potom

$$\bar{C}(k) \overset{d}{=} \gamma + \frac{\sigma_C Z_k^C}{\sqrt{k}} + b_{\bar{C},1} A\left(\frac{n}{k}\right) + b_{\bar{C},2} A^2\left(\frac{n}{k}\right) (1 + o_P(1)) \tag{4.68}$$

Důkaz:

Provedeme stejnou metodou jako ve větě 4.45. Detaily důkazu jsou uvedeny například v Gomes, Figueiro, Neves (2012).

K tomu, abychom mohli spojit obě studované metody PORT a MVRB, musíme v první řadě uvést algoritmy pro způsob výpočtu konzistentních odhadů parametrů β a ρ . Bez znalosti takových odhadů bychom nemohli metodu MVRB vůbec provádět. V praxi se užívá algoritmus navržený v článku Alves (2003).

Algoritmus pro odhady parametrů druhého řádu β a ρ

I. krok:

Nechť $(X_1, X_2, \dots, X_n)$ je náhodný výběr s realizací $(x_1, x_2, \dots, x_n)$. Provedeme odhad parametru ρ takto:

$$\hat{\rho}_\tau(k) = \hat{\rho}_n^{(\tau)} = -\left|\frac{3\left(T_n^{(\tau)}(k) - 1\right)}{T_n^{(\tau)}(k) - 3}\right|,$$

kde

$$T_n^{(\tau)}(k) = \begin{cases} \dfrac{\left(M_n^{(1)}(k)\right)^\tau - \left(M_n^{(2)}(k)/2\right)^{\tau/2}}{\left(M_n^{(2)}(k)/2\right)^{\tau/2} - \left(M_n^{(3)}(k)/6\right)^{\tau/3}} & \tau > 0 \\ \dfrac{ln\left(M_n^{(1)}(k)\right) - \frac{1}{2} ln\left(M_n^{(2)}(k)/2\right)}{\frac{1}{2} ln\left(M_n^{(2)}(k)/2\right) - \frac{1}{3} ln\left(M_n^{(3)}(k)/6\right)} & \tau = 0, \end{cases}$$

hodnotu výrazu $M_n^{(j)}(k)$ definujeme takto,

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=1}^{k} \left(ln \frac{X_{n-i+1:n}}{X_{n-k:n}}\right)^j, j \geq 1$$

II. krok:

Uvažujme $\hat{\rho}_\tau(k)$ pro hodnoty $k \in \mathcal{K} = ([n^{0,995}], [n^{0,999}])$ a nalezněme medián tohoto odhadu, který označíme χ_τ, pro $\tau = 0, 1$. Zvolíme dále

$$\tau^* = \begin{cases} 0 \quad jestliže \quad \sum_{k \in \mathcal{K}} (\hat{\rho}_0(k) - \chi_0)^2 \leq \sum_{k \in \mathcal{K}} (\hat{\rho}_1(k) - \chi_1)^2 \\ 1 \qquad\qquad\qquad\qquad\qquad\qquad jinak \end{cases}$$

III. krok:

Pro hodnotu $k_1 = [n^{0,995}]$, vypočítáme nyní odhady $\hat{\rho}^* = \hat{\rho}_{\tau^*}(k_1)$ a odhad $\hat{\beta}^* = \widehat{\beta}_{\hat{\rho}^*}(k_1)$, kde

$$\hat{\beta}_{\hat{\rho}}(k) = \left(\frac{k}{n}\right)^{\hat{\rho}} \frac{\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}\right)\left(\frac{1}{k}\sum_{i=1}^{k} U_i\right) - \left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}} U_i\right)}{\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}}\right)\left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-\hat{\rho}} U_i\right) - \left(\frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{-2\hat{\rho}} U_i\right)},$$ kde hodnoty U_i jsou brány z výrazu (4.62).

Podle praktických zkušeností velmi často je hodnota $\tau^* = 0$ pro $|\rho| \leq 1$ a volba $\tau^* = 1$ je pro ostatní možnosi $|\rho| > 1$.

Nyní propojíme obě studované metody PORT a MVRB a aplikujeme je na Hillův a momentový odhad. Základem budou výsledky uvedené v Gomes, Figueeiredo, Neves (2012), Gomes, Henrigueés-Rodrigues (2011) a Caeiro, Gomes (2014).

Pomocí metodologie tvorby odhadu PORT a tvorby odhadu MVRB upravíme jednoduše odhad MVRB následujícím způsobem. Předpokládejme, že je již provedená volba náhodné hranice $X_{n_q:n}$, tak že je

$$X^q = (X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \dots, X_{n_q+1:n} - X_{n_q:n})$$

kde $n_q = [nq] + 1$ a

1. $0 < q < 1$, pro d.f. F s konečným nebo nekonečným levým krajním bodem $x_* = inf\{x; F(x) > 0\}$, a proto je stanovená mez rovna empirickému kvantilu,
2. $q = 0$ s konečným levým krajním bodem, potom je náhodná mez volena jako výběrové minimum.

Tyto hodnoty jednoduše implementujeme do odhadů (4.56) a (4.57), tak že za hodnoty x_i dosadíme hodnoty $x_i - x_q$, kde x_q je náhodná mez, závisející na hodnotách kvantilů daného výběru.

Tedy bude

$$\hat{\gamma}_{\bar{H},q}(n,k,\hat{\beta},\hat{\rho}) = \hat{\gamma}_H(n,k)(X^q)\left(1-\frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) \tag{4.69}$$

a dále pro momentový odhad

$$\hat{\gamma}_{\bar{M},q}(n,k,\hat{\beta},\hat{\rho}) = \hat{\gamma}_M(n,k)(X^q)\left(1-\frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) - \left(\frac{\hat{\beta}\hat{\rho}}{(1-\hat{\rho})^2}\right)\left(\frac{n}{k}\right)^{\hat{\rho}}, \tag{4.70}$$

kde $\hat{\beta}$ a $\hat{\rho}$ jsou konzistentní odhady parametrů druhého řádu, které jsou uvedené v (4.55).

Nyní již máme dostatek výsledků, abychom dokončili konstrukci metody sample fraction pro takovéto odhady.

Označíme symbolem $\hat{\gamma}$, kterýkoli z odhadů (4.69) a (4.70). V dalším zavedeme $AMSE(\hat{\gamma})$ - asymptotickou střední kvadratickou chybu a ukážeme dva možné přístupy k nalezení odhadu $\hat{\gamma}$. Jednou z možností je právě metoda sample fraction, kterou budeme využívat především v okamžiku zdůraznění MVRB a druhá bude specifická metoda používaná v případě zdůraznění přístupu pomocí PORT. V obou případech jsme již udělali první tři kroky v algoritmu: Nalezli jsme odhady parametrů druhého řádu β a ρ, které jsou uvedené v předchozích krocích I. – III..

A. Algoritmus pro nalezení optimální hodnoty k pomocí metody sample fraction

Podobně jako v případech ostatních odhadů, na které tuto metodu aplikujeme, budeme hledat hodnotu $k_0^{\hat{\gamma}}$, která minimalizuje MSE. Provedeme to stejnou metodou jako v ostatních případech, kdy využíváme tuto metodu. Tedy

$$k_0^{\hat{\gamma}}(n) = \underset{n}{argmin}\, MSE(\hat{\gamma}(n)) \tag{4.71}$$

Podle závěrů věty 4.47 můžeme určit hodnotu AMSE těchto odhadů takto

$$k_{0|\hat{\gamma}}(n) = \underset{n}{argmin}\, MSE\big(\hat{\gamma}(n)\big) =$$

$$= \underset{n}{argmin}\left(\frac{{\sigma_C}^2}{k} + {b_{\bar{C},2}}^2 A^4\left(\frac{n}{k}\right)\right) = k_0^{\hat{\gamma}}(n)\big(1+o(1)\big) \tag{4.72}$$

V článcích Draisma (1999), Danielson (2001) je uvedeno, že $k_{0|\hat{\gamma}}(n)$ je konzistentní odhad $k_0^{\hat{\gamma}}(n)$. V dalším postupu použijeme pomocnou statistiku

$$T_{k,n|\hat{\gamma}} = T\,(k|\hat{\gamma}) = \hat{\gamma}([k/2]) - \hat{\gamma}(k),\ \ k = 2,\cdots,n-1 \tag{4.73}$$

Tato statistika konverguje v pravděpodobnosti k nule, pro prostřední posloupnost. Použijeme – li vztah (4.68), můžeme (4.73) zjednodušit

$$T\,(k|\hat{\gamma}) = \frac{\sigma_{\hat{\gamma}} P_k^{\hat{\gamma}}}{\sqrt{k}} + \, b_{\hat{\gamma}}\,(2^{2\rho} - \,1)\,A^2(n/k)\big(1 + o_P(1)\big),$$

kde $P_k^{\hat{\gamma}}$ je asymptoticky normované normální rozdělení. Jestliže dále označíme $k_{0|T}(n) = \underset{k}{argmin}\, AMSE(T_{k,n})$, platí

$$k_{0|\hat{\gamma}}(n) = \, k_{0|T}(n)\,(1 - 2^{2\rho})^{\frac{2}{1-4\rho}}(1 + o(1)) \tag{4.74}$$

Popíšeme nyní klasický postup při užití této metody sample fraction, který je podobný již dříve uvedeným postupům u Hillova, momentového nebo Pickandsova odhadu. Mějme daný výběr $\underline{X}_n = (X_1, \cdots, X_n)$ s neznámou d. f. F a nechť $T_{k,n}$je statistika uvedená v (4.73). Uvažujme pro $n_1 = O(n^{1-\epsilon})$, kde $0 < \epsilon < 1$ bootstrapový výběr

$$\underline{X}^*_{n_1} = \big(X_1^*, \cdots, X_{n_1}^*\big),$$

který je řízený pomocí empirické d.f.

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \le x]},$$

která je asociovaná s daným náhodným výběrem $\underline{X}_n$.

Pro uvedenou hodnotu n_1 použijeme pomocnou statistiku $T^*_{k_1,n_1}$, pro hodnoty $1 < k_1 < n_1$. Potom jestliže označíme $k^*_{0|T}(n_1) = \underset{k_1}{argmin}\, AMSE(T^*_{k_1,n_1})$, je

$$\frac{k^*_{0|T}(n_1)}{k_{0|T}(n)} = \left(\frac{n_1}{n}\right)^{-\frac{4\rho}{1-4\rho}}(1 + o(1)).$$

Tedy pro jiný výběr o délce $n_2 = [n_1^2/n]$ platí,

$$\frac{\left(k^*_{0|T}(n_1)\right)^2}{k_{0|T}(n_2)} = \left(\frac{n_1^2}{n^2}\frac{n}{n_2}\right)^{-\frac{4\rho}{1-4\rho}} k_{0|T}(n)\big(1 + o(1)\big). \tag{4.75}$$

Na základě vztahu (4.75) zkonstruujeme konzistentní odhad $k_{0|T}$ a na základě vztahu (4.74) nalezneme konzistentní odhad parametru $k_{0|\hat{\gamma}}$. Tento odhad je na základě vztahu (4.7) konzistentním odhadem parametru $k_0^{\hat{\gamma}}(n)$. Označme dále $\hat{k}^*_{0|T}$ výběrovou hodnotu $k^*_{0|T}$ a $\hat{\rho}$ odpovídající odhad parametru ρ, potom je

$$\hat{k}^*_{0|\gamma} = \, k_0^{\hat{\gamma}}(n, n_1) = min\left(n - 1, \left(\left(1 - 2^{2\hat{\rho}}\right)^{\frac{2}{1-4\hat{\rho}}} \frac{\left(\hat{k}^*_{0|T}(n_1)\right)^2}{\left(\hat{k}_{0|T}\left(\left[n_1^2/n+1\right]\right)\right)}\right) + 1\right), \tag{4.76}$$

Odtud je odhad parametru γ daný vztahem

$$\hat{\gamma}^* = \hat{\gamma}^*_{n,n_1|T} = \hat{\gamma}\left(k_0^{\hat{\gamma}}(n,n_1)\right) \tag{4.77}$$

Tento výše uvedený postup dále užijeme. Předpokládáme, že jsme v algoritmu již učinily první tři kroky I., II. a III.

IV. krok:

Vypočteme hodnotu $\hat{\gamma}(k)$, pro hodnoty $k = 1,2,\cdots,n-1$

V. krok:

Nyní podobně jako v ostatních algoritmech stanovíme hodnotu délky $n_1 = [n^{1-\epsilon}] = [n^{0,955}]$ a vypočteme druhou délku $n_2 = [n_1^2/n] + 1$

VI. krok:

Máme již stanovený počet bootstrapových výběrů *B*. Podobně jako u ostatních typů odhadů budeme nyní počítat každý jednotlivý bootstrapový výběr a vytvářet jej takto

$\left(x_1^*,\cdots,x_{n_2}^*\right)$ a $\left(x_1^*,\cdots,x_{n_2}^*,x_{n_2+1}^*,\cdots,x_{n_1}^*\right)$

VII. krok:

Označíme $T_{k,n}^*$ bootstrapovou realizaci $T_{k,n}$ ze (4.73) a pro každé $1 \le l \le B$ stanovíme hodnotu $t_{k,n_1,l}^*$ pro $1 < k < n_1$ a hodnotu $t_{k,n_2,l}^*$ pro $1 < k < n_2$. Na základě těchto hodnot určíme

$$MSE^*(n_i,k) = \frac{1}{B}\sum_{i=1}^{B}\left(t_{k,n_i,l}^*\right)^2, k = 2,\cdots,n_i - 1$$

Navíc pro hodnoty $k = 2,\cdots,n_i - 1$ a *i* = 1,2 vypočítáme

$$BIAS^*(n_i,k) = \frac{1}{B}\sum_{i=1}^{B} t_{k,n_i,l}^*$$

VIII. krok:

V této části algoritmu určíme hodnoty $k_{0|T}^*(n_i)$, pomocí minimalizace hodnot MSE^*. Tedy $k_{0|T}^*(n_i) = \underset{1<k<n_i}{argmin}\, MSE^*(n_i,k)$, pro hodnoty $i = 1,2$.

IX. krok:

Ze vztahu (4.76) určíme $\hat{k}_{0|\gamma}^* = k_0^{\hat{\gamma}}(n,n_1)$.

X. krok:

Vypočteme $\hat{\gamma}^* = \hat{\gamma}^*_{n,n_1|T} = \hat{\gamma}\left(k_0^{\hat{\gamma}}(n,n_1)\right)$.

V následující poznámce uvedeme některé omezující podmínky.

Poznámka 4.48

1. Všechny záporné hodnoty z výběru musíme odstranit.

2. V II. kroku algoritmu užíváme především hodnotu $\tau = 0$. Hodnota $\tau = 1$ se užívá jen v případě, že $\rho = 0$.

3. V prvních třech krocích můžeme použít i jiné druhy odhadů parametru ρ. Viz například Goegenbeur (2008), Ciuperca, Mercadier (2010). Podobně pro parametr β můžeme použít odhady uvedené v článku Caiero, Gomes (2006).

4. Podobně jako v algoritmech pro výpočet odhadu γ pomocí bootstrapové metody pro Hillův odhad, momentový odhad, můžeme diskutovat hodnoty B – zde se doporučuje jako nejmenší hodnota 250, dále je možné volit i jiné počáteční hodnoty nastavení n_1 - vede k jiné volbě hodnoty ε v kroku V.

Uvedeme dále algoritmus pro volbu odhadu, který je zároveň typu PORT i MVRB

B. Algoritmus pro heuristickou volbu hodnoty k a hodnoty q

I. krok – III. krok.:

Tyto kroky použijeme z předchozího algoritmu.

IV. krok

Mějme daný náhodný výběr $\underline{X}_n = (X_1, \cdots, X_n)$ s neznámou d. f. F, nechť $\underline{x}_n = (x_1, \cdots, x_n)$ je realizace tohoto náhodného výběru. Uvažujme pro hodnoty $q = -\frac{1}{n}; 0; 0{,}05; 0{,}5$ náhodný výběr $\underline{X}_n^{(q)}$s jeho realizací $\underline{x}_n^{(q)}$ a vypočtěme $\hat{\rho} = \hat{\rho}_{0,q} = \hat{\rho}_0\left(k_1, \underline{x}_n^{(q)}\right)$ a $\hat{\beta} = \hat{\beta}_{\hat{\rho}}\left(k_1, \underline{x}_n^{(q)}\right)$, podle výše uvedených kroků I. – III. Hodnotu k_1 volíme podle doporučení $k_1 = [n^{1-\epsilon}]$, kde hodnota $\epsilon = 0{,}001, 0{,}005$.

V. krok:

Pro hodnoty $k = 1, 2, \cdots, n - [n\,q] - 1$ vypočteme dále $E_1^{(q)}(n) = \hat{\gamma}_H(n,k)(X^q)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right)$ a$E_2^{(q)}(n) = \hat{\gamma}_M(n,k)(X^q)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) - \left(\frac{\hat{\beta}\hat{\rho}}{(1-\hat{\rho})^2}\right)\left(\frac{n}{k}\right)^{\hat{\rho}}$. Pro hodnotu $q = -1/n$ jsou dané předchozí hodnoty rovny $\hat{\gamma}_H(n,k)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right)$ a $\hat{\gamma}_M(n,k)\left(1 - \frac{\hat{\beta}}{1-\hat{\rho}}\left(\frac{n}{k}\right)^{\hat{\rho}}\right) - \left(\frac{\hat{\beta}\hat{\rho}}{(1-\hat{\rho})^2}\right)\left(\frac{n}{k}\right)^{\hat{\rho}}$.

VI. krok:

V této části eliminujeme všechny hodnoty q z kroku I., pro které je hodnota $E_i^{(q)}([n^{0,95}]) - E_i^{(q)}([n^{0,05}])$, záporná nebo nula. Tato podmínka vychází z praxe, odpovídá většině praktických situací.

VII. krok:

Pro každé q a pro $k > \hat{k}_0 = \frac{1}{2}\left(\frac{(1-\hat{\rho})^2 n^{-2\hat{\rho}}}{-2\hat{\rho}\hat{\beta}^2}\right)^{1/(1-2\hat{\rho})}$ upravíme tvar odhadu následovně: $\tilde{\gamma}_{\bar{C},q}(n,k,\hat{\beta},\hat{\rho}) = \begin{cases} \hat{\gamma}_{\bar{C},q}(n,k,\hat{\beta},\hat{\rho}) & , k \le \hat{k}_0 \\ max\left(\hat{\gamma}_{\bar{H},q}(n,k-1,\hat{\beta},\hat{\rho}), \hat{\gamma}_{\bar{H},q}(n,k,\hat{\beta},\hat{\rho})\right) & , k > \hat{k}_0 \end{cases}$. Za $\bar{C}$ postupně dosadíme H(Hillův odhad) nebo M(momentový odhad).

VIII. krok:

Pro každé q nyní nalezneme „největší běh“, který označíme jako $\tilde{\gamma}_{\bar{C},q}(n,k_{q,1},\hat{\beta},\hat{\rho})$ s délkou běhu $m_q = k_{q,2} - k_{q,1} + 1$, kde platí $\tilde{\gamma}_{\bar{C},q}(n,k_{q,1},\hat{\beta},\hat{\rho}) = \tilde{\gamma}_{\bar{C},q}(n,k_{q,1}+1,\hat{\beta},\hat{\rho}) = \cdots = \tilde{\gamma}_{\bar{C},q}(n,k_{q,2},\hat{\beta},\hat{\rho})$.

IX. krok:

Nalezneme $q^{**} = \underset{q}{argmax}\, m_q$

X. krok:

Zvolíme $k^{**} = k_{q^{**},2}$. Jako odhad bude poté $\hat{\gamma} = \tilde{\gamma}_{\bar{C},q^{**}}(n,k^{**},\hat{\beta},\hat{\rho})$.

4.6. Simulační studie

V této části jsou uvedeny výsledky simulačních studií, které byly provedeny na základě naznačených bootstrapových procedur pro Hillův, momentový i Pickandsův odhad parametru EVI. Numerická realizace těchto odhadů byla zpracována na dvoujádrovém počítači s operační pamětí 4GB. Výpočty byly realizovány v programu Mathematica verze 8.1, resp. 9.0 a zároveň ověřovány v programu R verze 3.0.2. V části výpočtů byly použity speciální algoritmy paralelních výpočtů resp. paralelního programování.

Při realizaci studie jsme nejdříve vymezili náhodné veličiny, s kterými budeme pracovat. Výběr jsme zúžili na základní představitele jednotlivých sfér přitažlivosti. Ve výběru je samozřejmě základní rozdělení Paretovo s různými parametry, základní forma je rozdělení s parametry 1 a 1. Dalšími zkoumanými typy rozdělení z této oblasti přitažlivosti jsou například F-rozdělení, Burrovo rozdělení, Cauchyho rozdělení a Studentovo rozdělení. Ve sféře přitažlivosti Gumbelově jsme používali rozdělení normální, exponenciální a gamma. Konečně ve sféře přitažlivosti Weibullově jsme užívali rovnoměrné rozdělení a beta rozdělení.

Dále jsme připravili algoritmy pro výběr náhodných hodnot z výše uvedených náhodných veličin. K tomu jsme především použili generátor náhodných hodnot, který je integrální součástí programu Mathematica. Provedli jsme jen jednoduché úpravy, abychom získávali lehčeji dané hodnoty – omezení na počet hodnot resp. zjednodušení výběru náhodné veličiny, z níž budeme generovat náhodné hodnoty.

Dalším samostatným pomocným programem je procedura založená na výběru posloupností dat o délkách n_1 a n_2 tak, jak je uvedeno v požadavcích metody sample fraction. Pro jednotlivé typy odhadů byly sestrojeny speciální programy – procedury, které vychází ze závěrů části 4.1. Z dat jsou tedy nalezeny konzistentní odhady parametru ρ. Tento odhad se posléze využívá v závěrečné fázi tvorby všech typů odhadů. Dále jsou v procedurách odhadů – s výjimkou PORT odhadu – uvedeny podobné kroky – proveden výběr obou částí dat o délce n_1 a n_2 a je přistoupeno ke klasickým krokům při bootstrapových procedurách, provedli jsme daná opakování a zapisovali při nich údaje, které byly pro nás užitečné – počet prvků v základním výběru – označováno n, velikost první části - n_1, velikost druhé části-n_2, počet opakování v proceduře bootstrap – B, parametr – ε - rozhoduje o velikosti jednotlivých částí, volitelně jsou uvedeny parametry, které umožňují redukovat počet prvků v základním výběru na hodnoty v jistém intervalu resp. umožňují výsledky uložit do předem zadaného souboru pro další zpracování. Výstupy z takovýchto programů jsou seznamy, v nichž jsou uvedené vstupní parametry, ale i hodnota daného odhadu, stanovení optimálního poměru vzhledem k n, výběrově hodnota MSE, doba zpracování jedné konkrétní hodnoty.

Pro metodu PORT byly vytvořeny speciální programy, protože se v tomto případě nejedná o přímo metodu typu „optimal sample fraction“. Bylo nutné konzistentně odhadnout parametry ρ a β - k tomu jsou uvedeny v části 4.1 procedury a dále postupovat pomocí správné volby q a hodnoty k. V podkapitole 4.1 je popsán detailně algoritmus pro správný postup pomocí metody PORT.

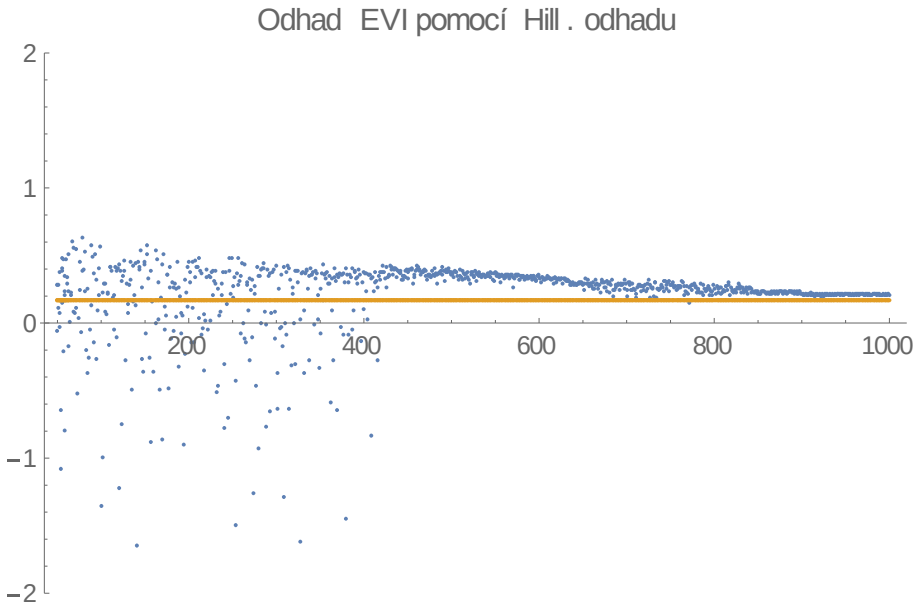
Všechny výsledky jsou postupně uváděny pro jednotlivé odhady a v závěru této části jsou uvedeny porovnávací studie celé bootstrapové procedury.

Jak je zřejmé z předchozí části, bootstrapová procedura závisí na několika parametrech. Jde především o parametr ε, pomocí něhož stanovujeme velikost počátečního výběru. Dále je podstatný parametr B – počet opakování výběru. Volba těchto parametrů je studována v následujícím textu.

4.6.1. Metoda bootstrap a Hillův odhad.

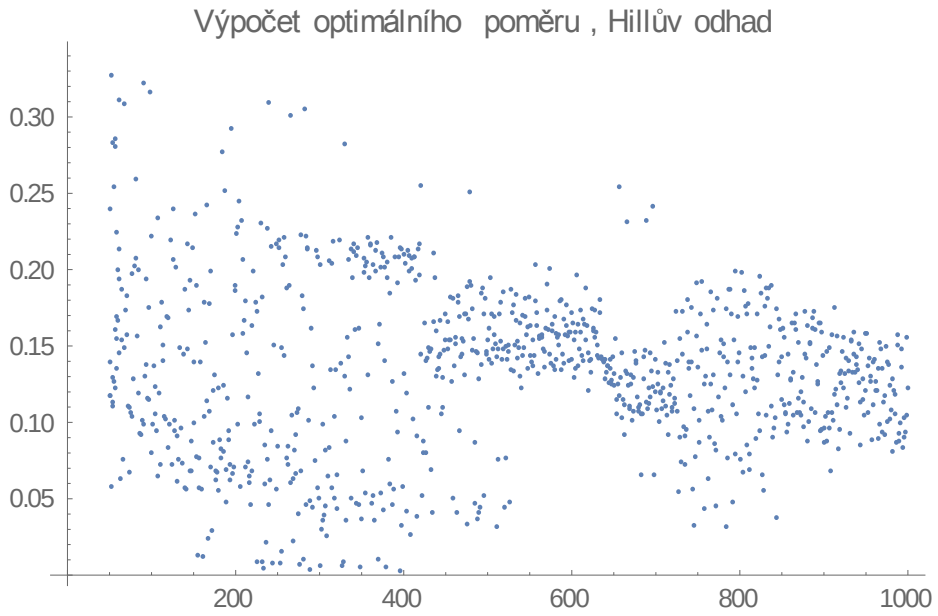
Uvedeme dále realizaci metody optimal sample fraction na Hillův odhad a data pocházející z rozdělení, které patří do Frèchetovy sféry přitažlivosti.

Burrovo rozdělení s parametry (1;2;3). Postupně jsou zkoumány výběry s rozsahy od 50 do 1000 (po jedné hodnotě), zobrazeny jsou odhad EVI a zároveň i poměr k_0/n.



Graf 5 - Odhad EVI pomocí Hillova odhadu.

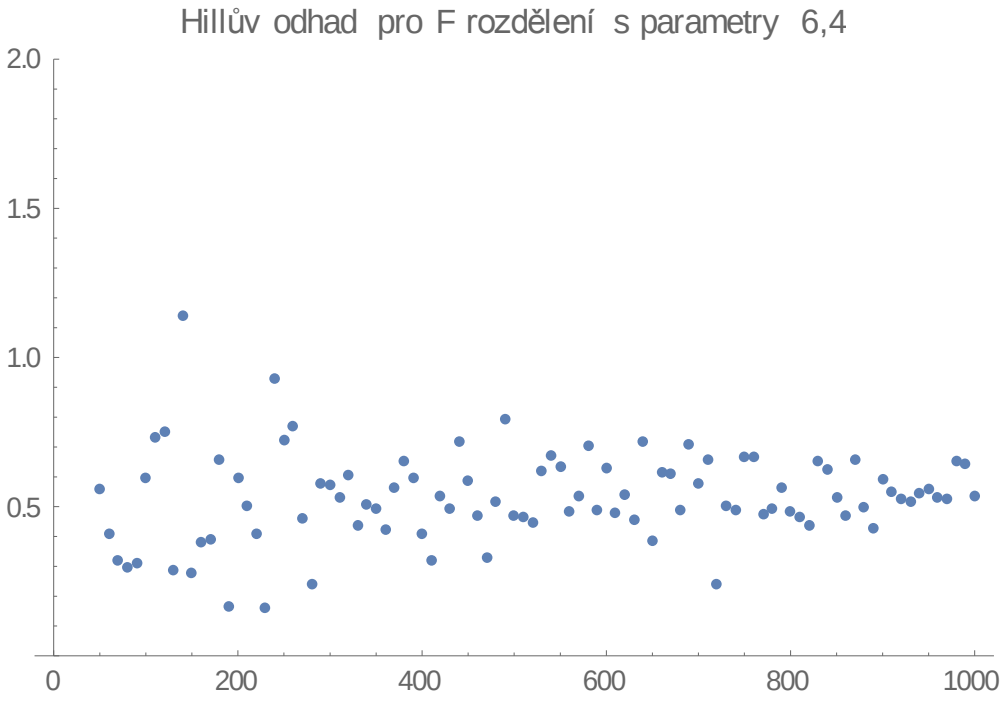
Z grafu je vidět, že mohou být vypočítány i odhady záporné. V případě, že základní výběr malý, pak je odhad i velmi rozptýlený. Naopak pro větší hodnoty počítáme velmi přesnou hodnotu odhadu indexu. Nalézt tyto odhady není zcela jednoduché především z časových důvodů.



Graf 6 - Výpočet optimálního poměru Hillova Odhadu.

Z grafu je vidět, že optimální poměr k_0/n se nachází v intervalu od 0,02 až do 0,3. S rostoucí hodnotou se jeho rozsah zužuje na interval od 0,08 do 0,18.

F rozdělení s parametry 6,4. Velikost výběru byla volena od 50 do 1000 s krokem 10.



Graf 7 - Hillův odhad pro F rozdělení s parametry 6;4.

Hodnota skutečného indexu EVI je v tomto případě rovna 0,5. Odhad je proto realizován velmi přesně. S rostoucí velikostí výběrového souboru je stále přesnější.

Graf 8 - Volba optimální části - Hill. odhad

Optimální část je na úrovni 0,05, což odpovídá správnosti výpočtů. Pro tuto náhodnou veličinu je optimální hodnota indexu *k* velmi dobře vypočítána.

Graf 9 - Doba zpracování Hillova odhadu.

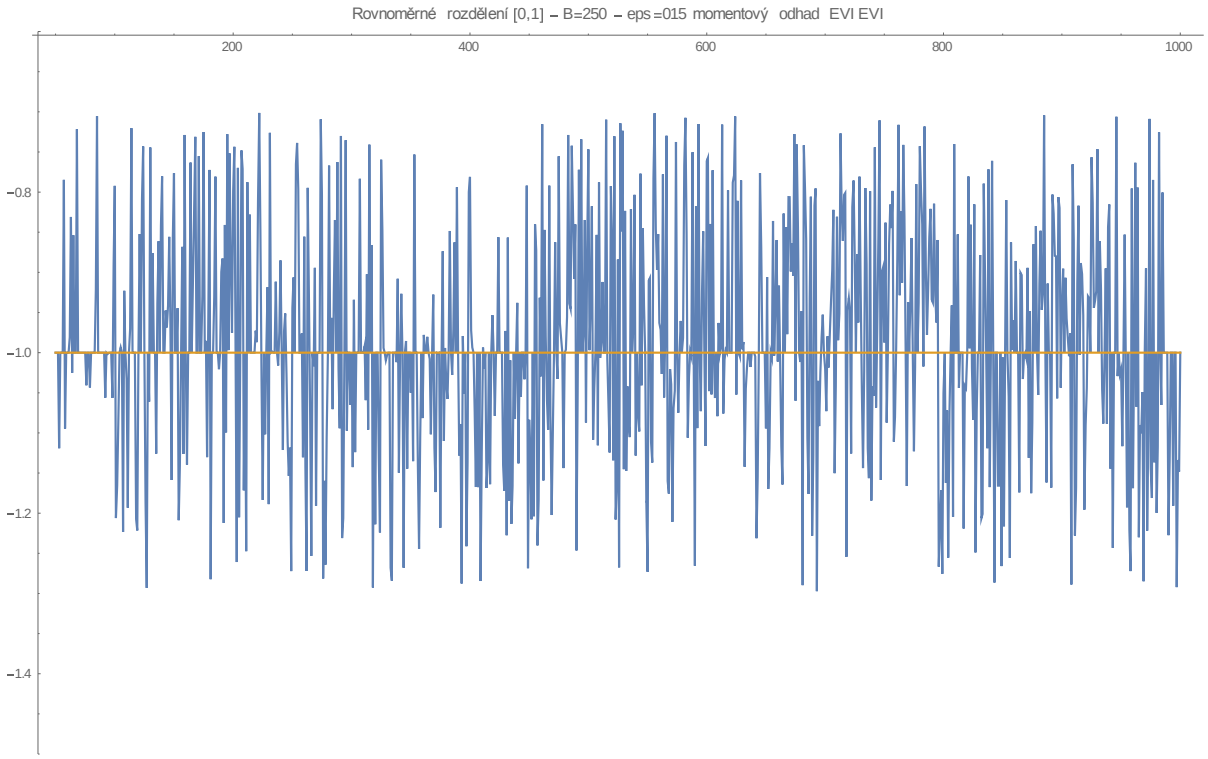
Tento graf znázorňuje čas potřebný k výpočtu optimal sample fraction. Ani pro hodnotu *n=1000* není nijak dramatický. Jeho hodnota nedosahuje ani 200s.

4.6.2. Metoda bootstrap a momentový odhad

V této části budeme vyšetřovat užití metody optimal sample fraction na momentový odhad. Jako vstupní údaje nám poslouží postupně rozdělení ze všech třech různých sfér přitažlivosti.

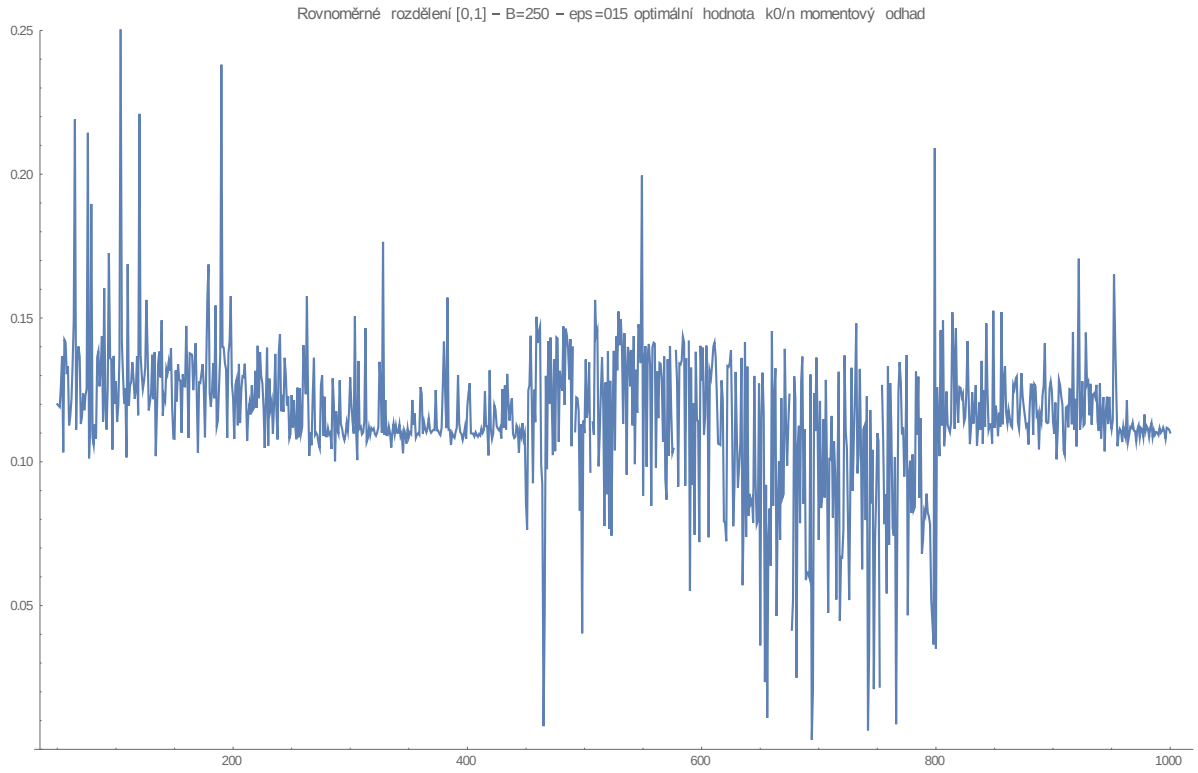
Jako první zpracujeme údaje ze sféry přitažlivosti Weibull (EVI < 0).

Rovnoměrné rozdělení na intervalu (0;1).



Graf 10 - Momentový odhad EVI pro rovnoměrné rozdělení γ = -1.

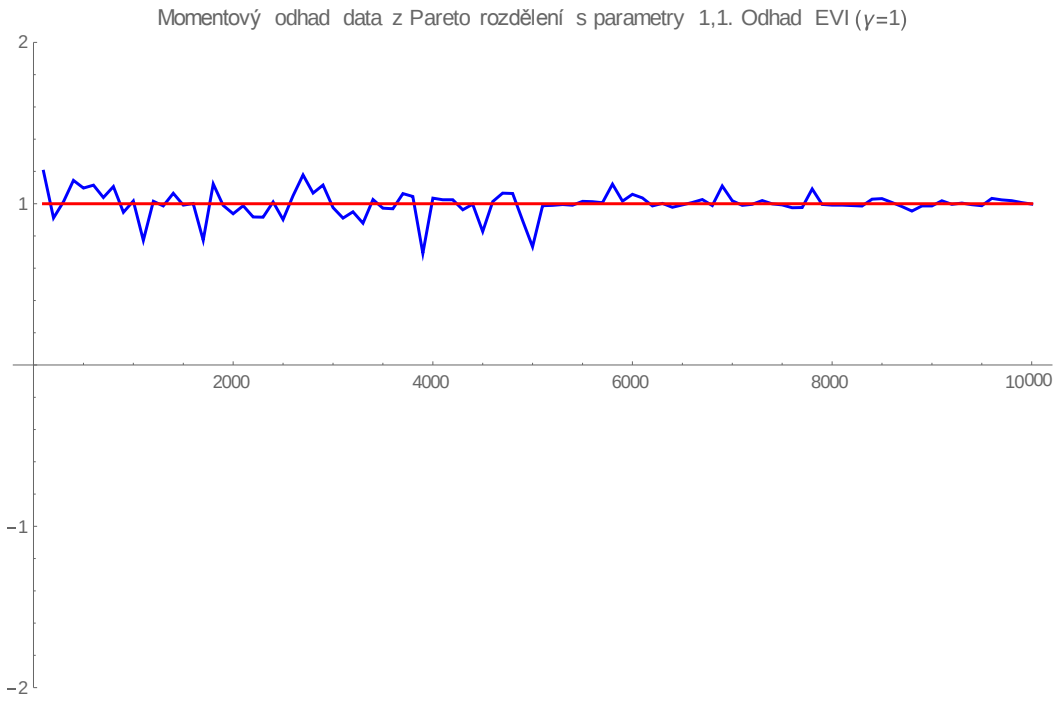
Je zřejmé, že hodnoty odhadů kolísají kolem skutečné hodnoty EVI = -1. Graf je proveden pro velké zvětšení a proto se zdají hodnoty velmi kolísavé. Byla zvolena počáteční hodnota $B=250$ a $\varepsilon = 0{,}15$. Odhady EVI jsou pro tyto malé hodnoty z intervalu (1,3;-0,7). Jsou tedy dostatečně přesné.



Graf 11 - Optimální hodnota k0/n pro momentový odhad.

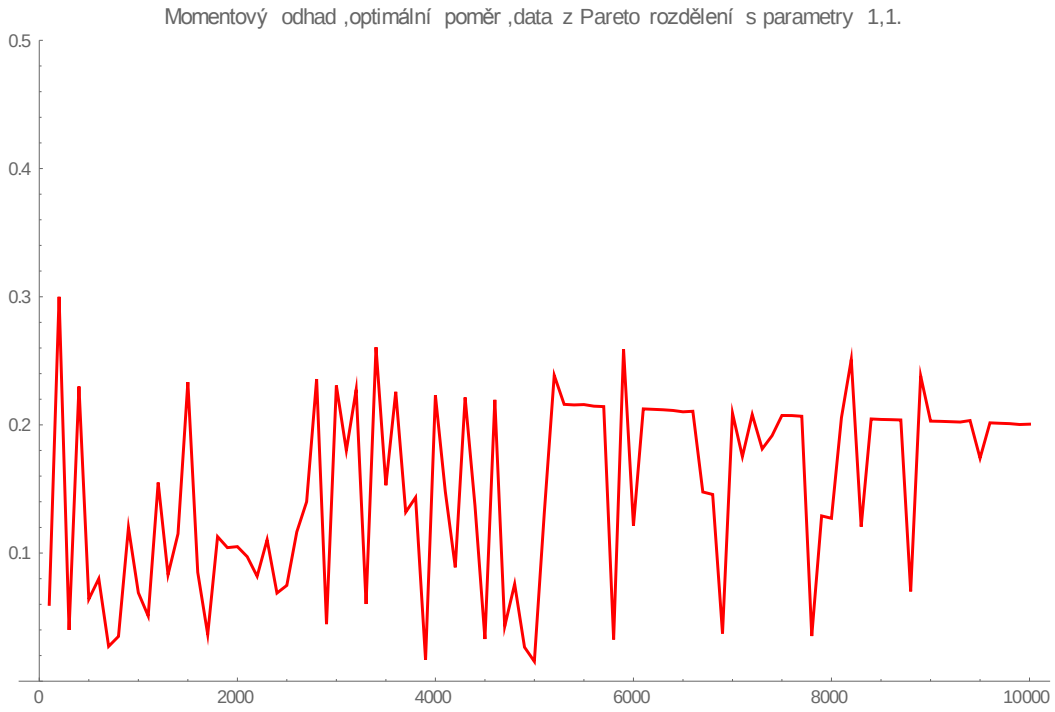
Optimální poměr se v tomto případě ustálil zhruba na hodnotě 0,12. Proti předchozím rozdělením je optimální poměr celkově větší.

Dalším typem rozdělení je **Pareto** s parametry (1;1). Patří do sféry přitažlivosti Frèchet , jeho EVI $\gamma = 1$.



Graf 12 - Odhad EVI pro rozdělení Pareto pomocí momentového odhadu.

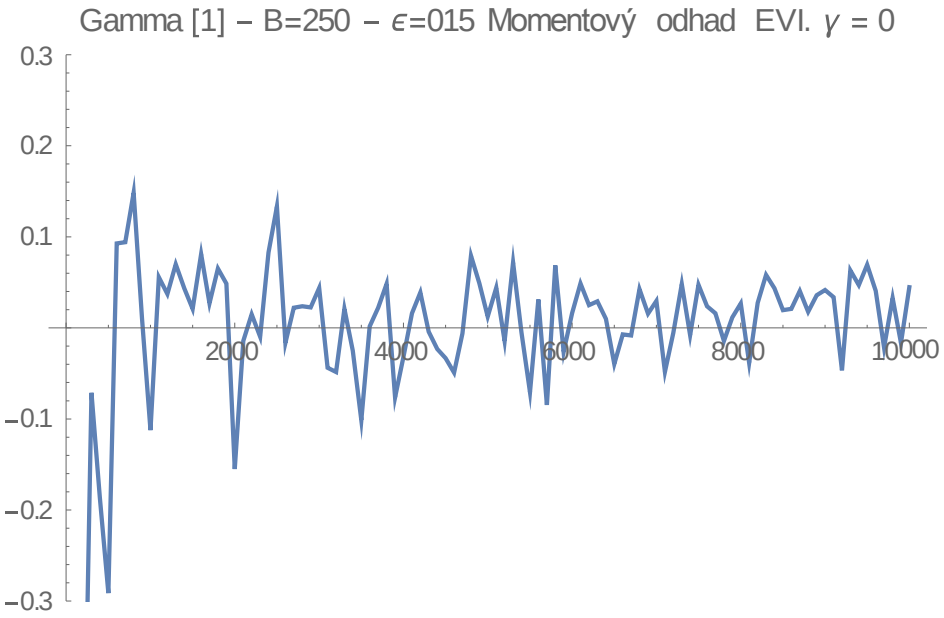
Z grafu je patrné, že s rostoucím rozsahem výběru *n* výběru je odhad přesnější, ale i pro hodnoty *n* menší než 5000 není chyba odhadu příliš veliká.



Graf 13 - Optimální poměr pro rozdělení Pareto a momentový odhad

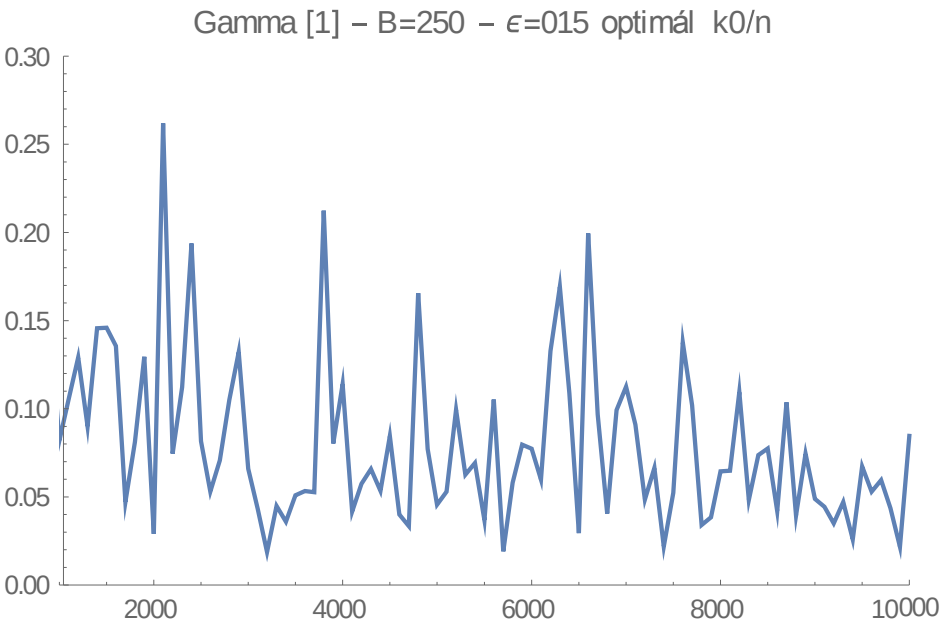
Na tomto grafu je patrné velké rozkolísání hodnot optimálního poměru od 0,02 až po hodnoty větší než 0,2. Zároveň je daný poměr kolem hodnoty 0,15, tak jak se většinou obecně uvažuje.

Posledním typem je příklad náhodné veličiny s rozdělením **Gamma** s parametry 1,1. Je zařazena do sféry přitažlivosti Gumbelovy třídy, její hodnota EVI $\gamma = 0$.



Graf 14 - Odhad EVI pro Gamma rozdělení a momentový odhad

Vlastní odhad parametru γ tohoto rozdělení je relativně kvalitní, neboť má pro většinu hodnot *n* chybu maximálně 0,1.



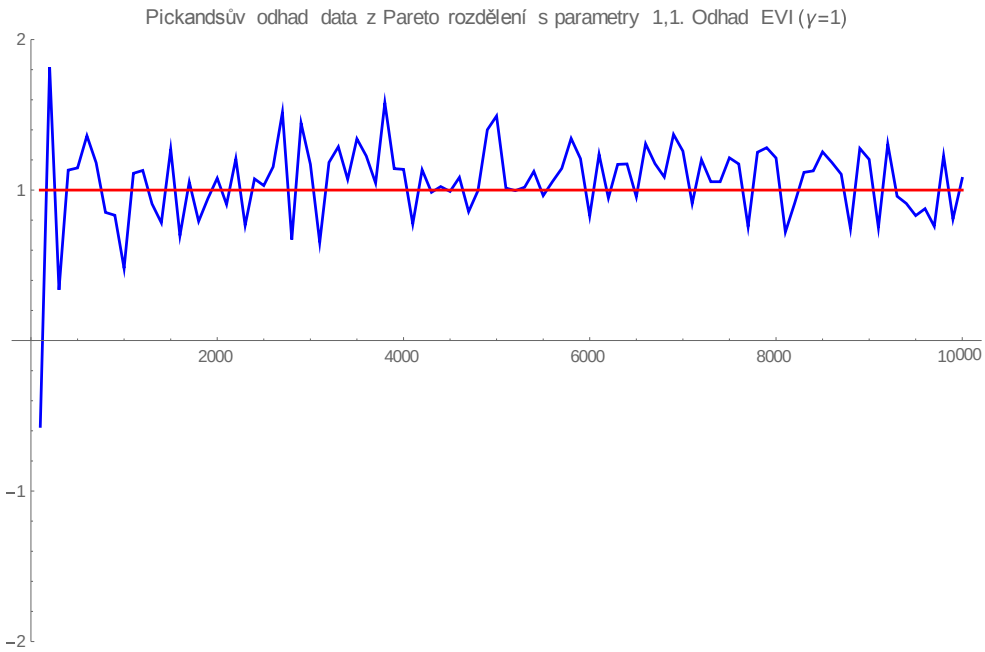
Graf 15 - Optimální poměr pro rozdělení Gamma a momentový odhad

Je zjevný trend zmenšování optimálního poměru. Navíc leží v teoreticky odhadnutém intervalu<0,05;0,2>.

4.6.3. Pickandsův odhad

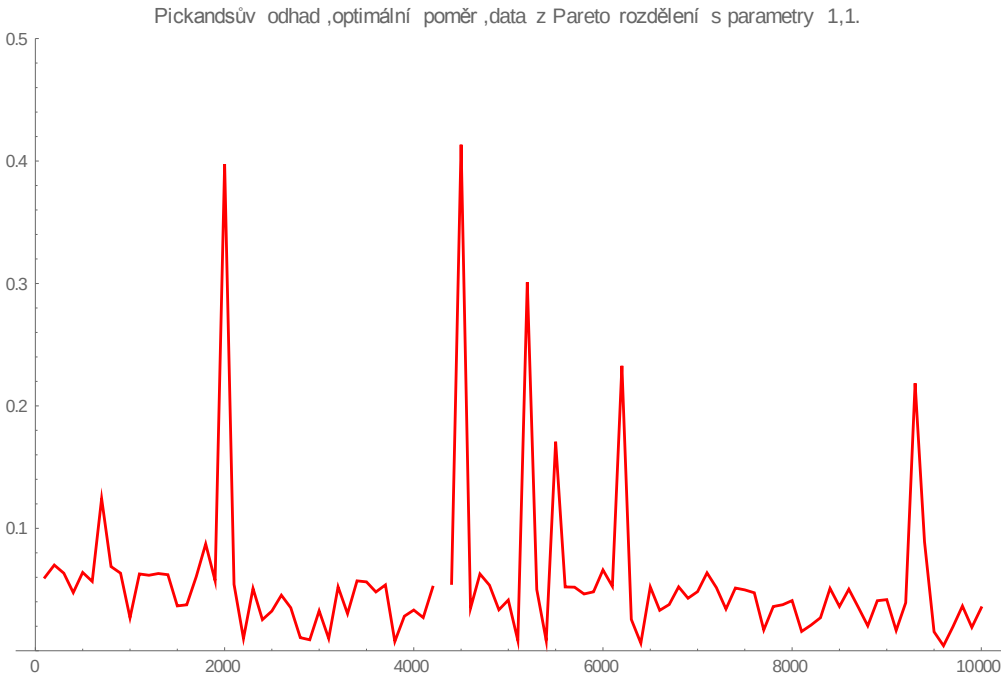
Postupně uvedeme podobně jako v případě momentového odhadu data pocházející z rozdělení ze všech různých typů sfér přitažlivosti. Pro tento odhad jsme zvolili poměrně rozsáhlá data. Jde o údaje o velikosti výběru od 100 do 10 000, krok jsme nastavili na 100.

Prvním rozdělením je Pareto rozdělení s parametry (1;1). Je známo, že hodnota $\gamma = 1$.



Graf 16 - Odhad EVI pro rozdělení Pareto(1;1) a Pickandsův odhad

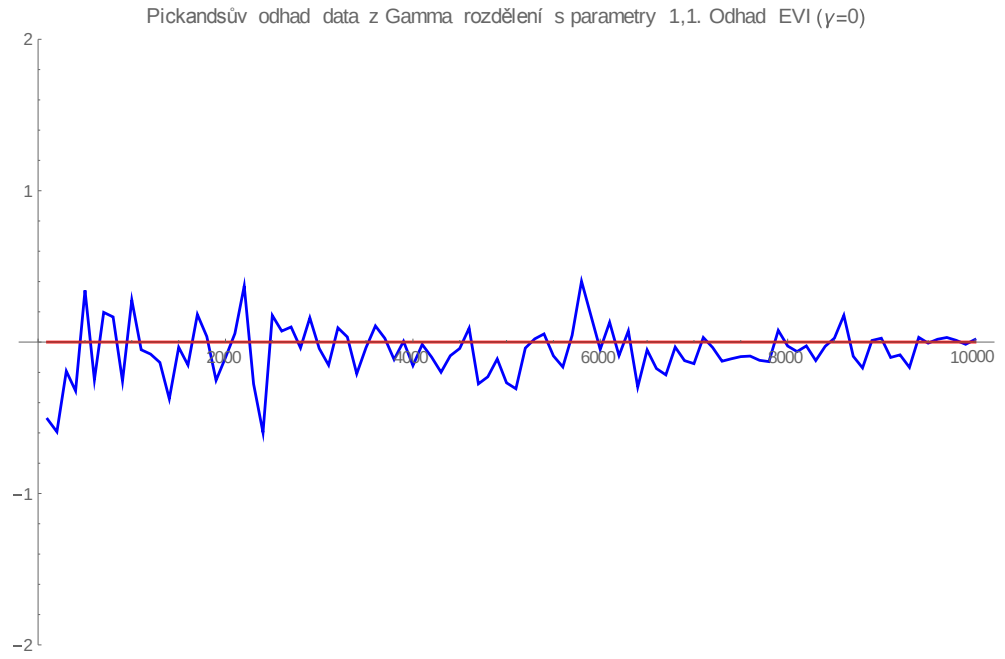
Odhad indexu γ se pro toto rozdělení pohybuje od 0,08 až po 1,5. Důležité ovšem je, že vlastní správnou hodnotu $\gamma=1$ v podstatě nenabývá. Navíc rozkolísanost je stejná pro malé i velké hodnoty *n*.



Graf 17 - Optimální poměr pro rozdělení Pareto a Pickandsův odhad

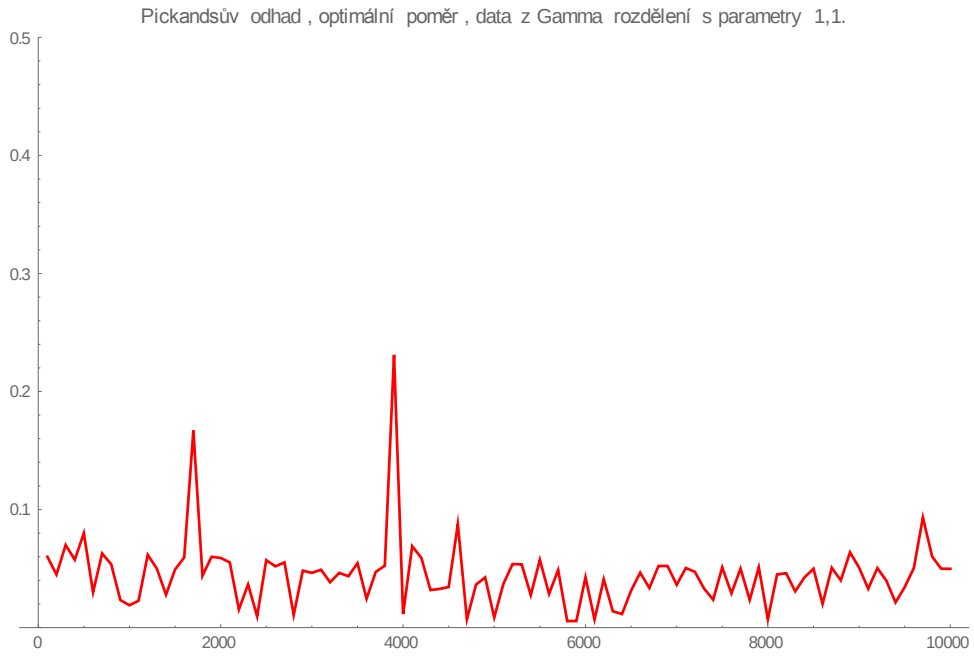
Optimální poměr v tomto případě ani zdaleka není v teoretickém rozmezí. Naopak je možno nahlédnout, že v několika hodnotách je příliš veliký (větší než 0,3).

Druhou náhodnou veličinou je rozdělení Gamma s parametry (1;1). Je známo, že hodnota $\gamma = 0$.



Graf 18 - Odhad EVI pro rozdělení Gamma a Pickandsův odhad

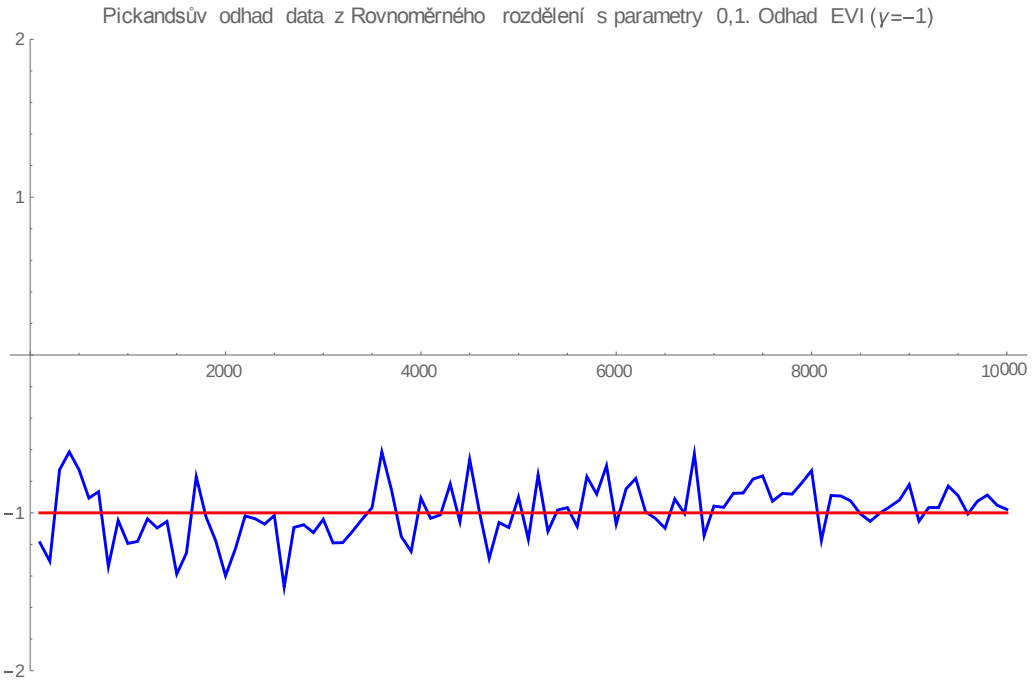
V tomto případě již byla velikost *n* až do 10 000. Hodnota chyby odhadu je pro většinu hodnot *n* menší než 0,25. S rostoucí hodnotou *n* se zlepšuje odhad.



Graf 19 - Optimální poměr pro rozdělení Gamma a Pickandsův odhad

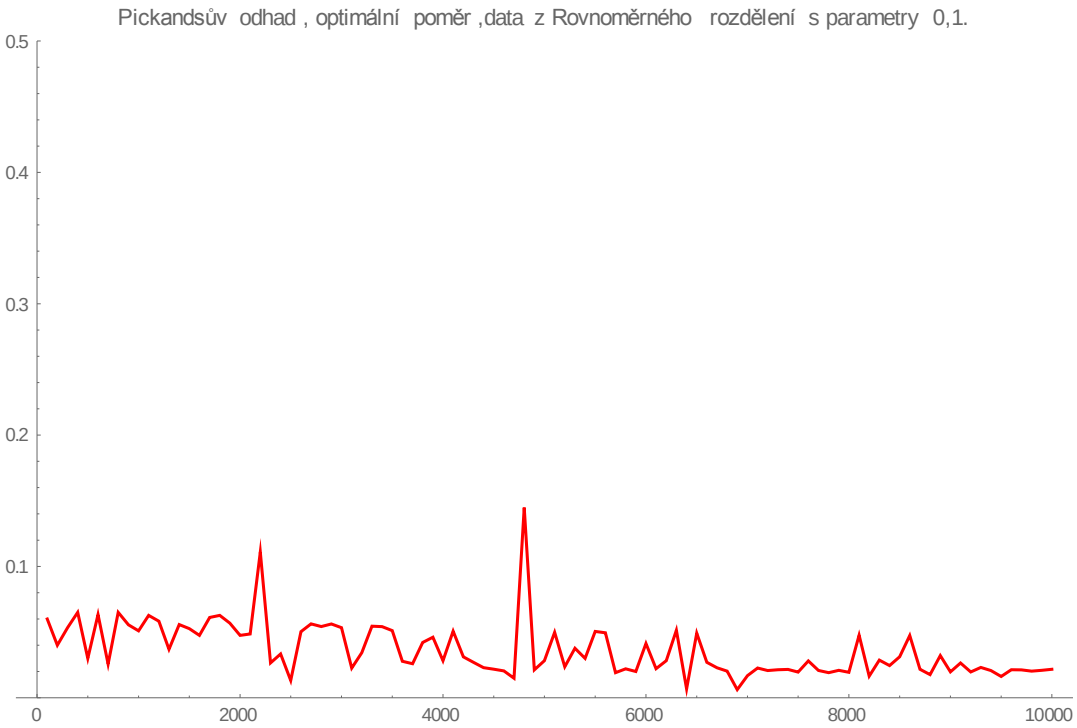
V tomto případě je optimální poměr pro většinu hodnot *n* odhadnut na maximálně 0,1.

Třetím šetřeným rozdělením je rovnoměrné rozdělení na intervalu <0;1>. Je známo, že hodnota $\gamma = -1$.



Graf 20 - Index EVI pro rovnoměrné rozdělení a Pickandsův odhad

Odhad indexu γ je zdařilý, především pro větší hodnoty *n*. Chyba je nejvýše 0,25.



Graf 21 - Optimální poměr pro rovnoměrné rozdělení a Pickandsův odhad

Podobně jako v předchozím případě je optimální poměr odhadnut ve většině případů na maximálně 0,8. Navíc s rostoucí hodnotou *n* se stabilizuje na zhruba 0,04.

4.6.4. Analýza jednotlivých odhadů a jejich porovnání

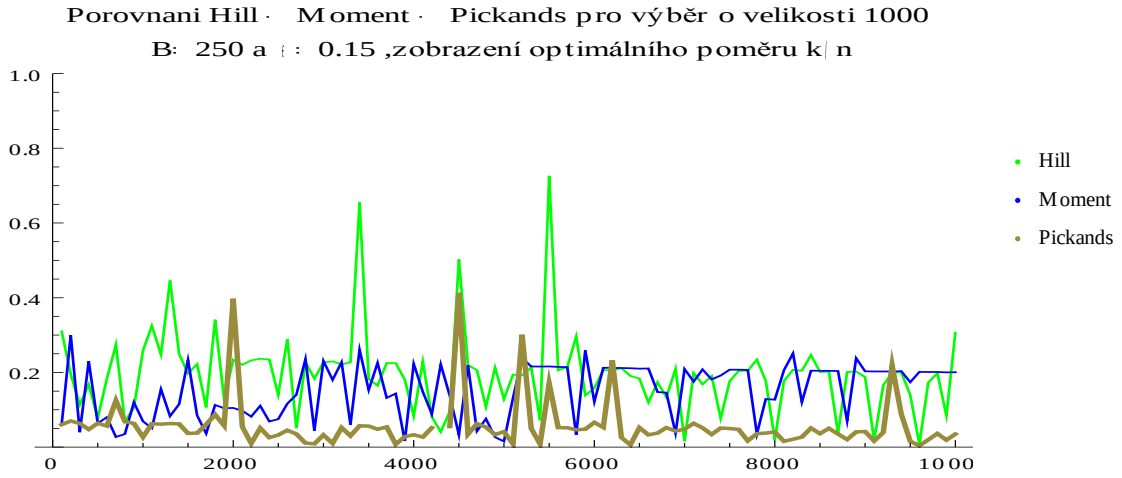
V této části jsou uvedeny hlavní výsledky. Jde o nalezení správného poměru posledních *k* pořadových statistik. Z pohledu bootstrapové techniky jde spíše o nalezení postupu pro výpočet odhadu parametru γ. Počáteční hodnota $k_{aux} = \lfloor 2\sqrt{n} \rfloor$ je zvolena velmi malá, abychom ji mohli jednoduše zvětšit. Udává se, že správná hodnota *k/n* uváděná ve všech třech vyšetřovaných odhadech by měla ležet mezi 10% až 20%. Z toho se snadno odvodí, že hodnota k_{aux} ji může být jen pro výběry o velikostech $100 \leq n \leq 400$. Jakmile je výběr větší nebo menší, hodnota optimálního *k* musí být nutně jiná.

Z uvedených šetření je zřejmé, že Pickandsův odhad je nejméně náročný z hlediska výpočetních operací, druhý v náročnosti je Hillův odhad a nejnáročnější je momentový odhad. V dalším uvedeme některá porovnávání, která budeme komentovat. Úkolem simulační studie je pro dané odhady nalézt optimální volbu poměru a tím i optimální hodnotu *k*.

K tomu, abychom mohli porovnávat všechny tři typy odhadů je nutno vhodně zvolit data. Je to proto, že Hillův odhad pracuje jen pro $\gamma > 0$. Vybrali jsme proto klasické rozdělení Paretovo s parametry (1;1). Pro toto rozdělení je známé, že hodnota $\gamma = 1$.

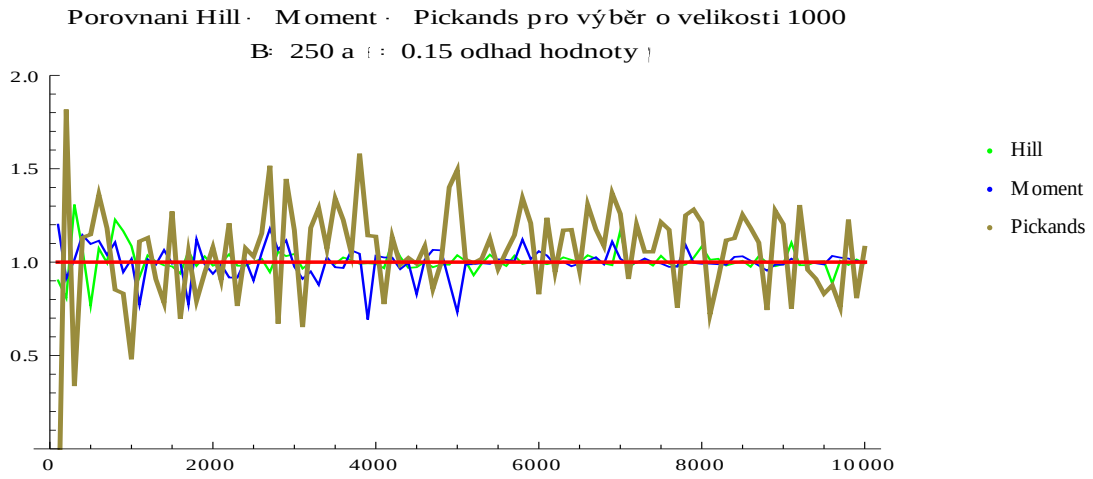
V této části se snažíme nalézt nejlepší kombinaci hodnot *B* a ε z algoritmu uvedeného v předchozí části.

Pro jednodušší práci jsme simulovali výběrové soubory o velikosti 500 a 1000 prvků. Nejdříve jsme použili doporučované hodnoty tedy $B = 250$ a $\varepsilon = 0{,}15$. Získali jsme následující výsledky:



Graf 22 - Porovnání optim. poměru pro Hill - Moment - Pickands

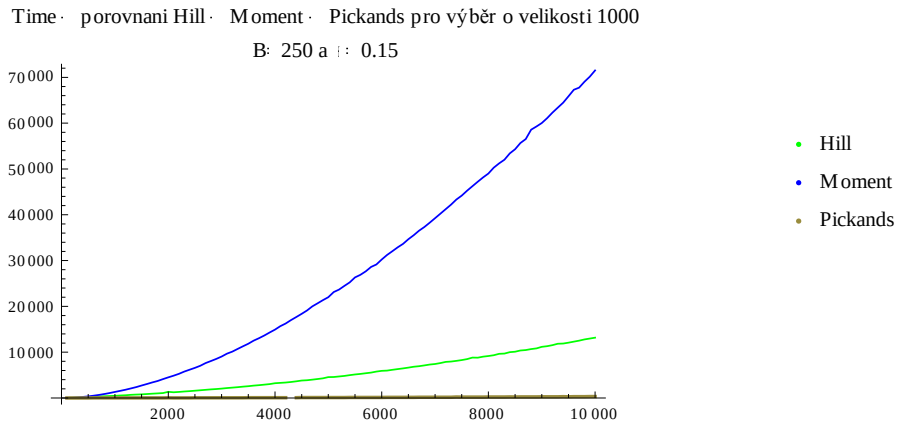
Postupně byly volené hodnoty rozsahu výběru od 100 do 10000 a zjištěné optimální hodnoty podle algoritmu. Je zřejmé, že všechny hodnoty mají klesající tendenci. U Pickandsova odhadu jsou na úrovni 5%, u druhých dvou odhadů klesají jen k 20%.



Graf 23 - EVI pro Hill - Moment - Pickands

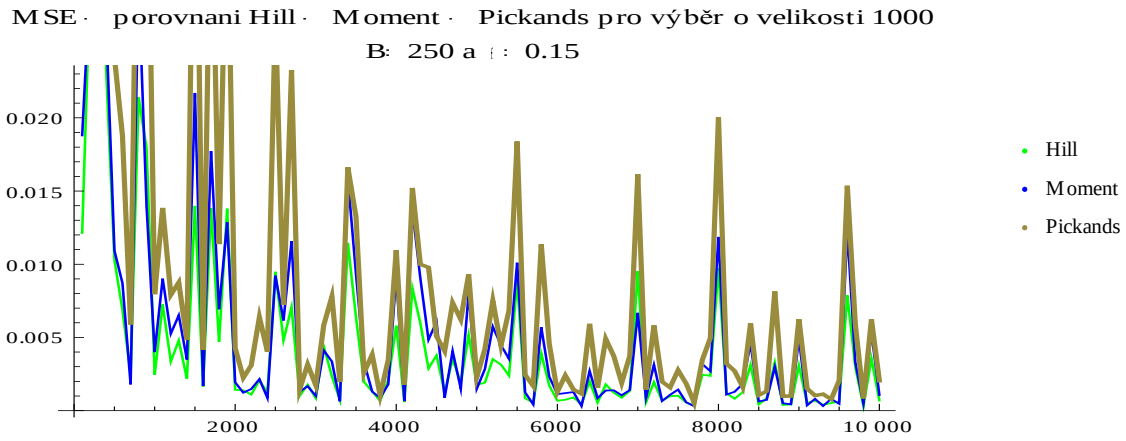
Z grafu vyplývá, že odhady Hillův a momentový jsou zhruba stejně efektivní. Výrazně zaostává Pickandsův odhad. Dané šetření plně podporuje kvalitu jednotlivých odhadů EVI.

Pro vlastní chod algoritmu je také určující doba práce. Tu si zobrazíme na dalším grafu. Jasně je zřejmá velká časová náročnost především u momentového odhadu. Například pro $n = 10000$ je na úrovni 20 hodin .



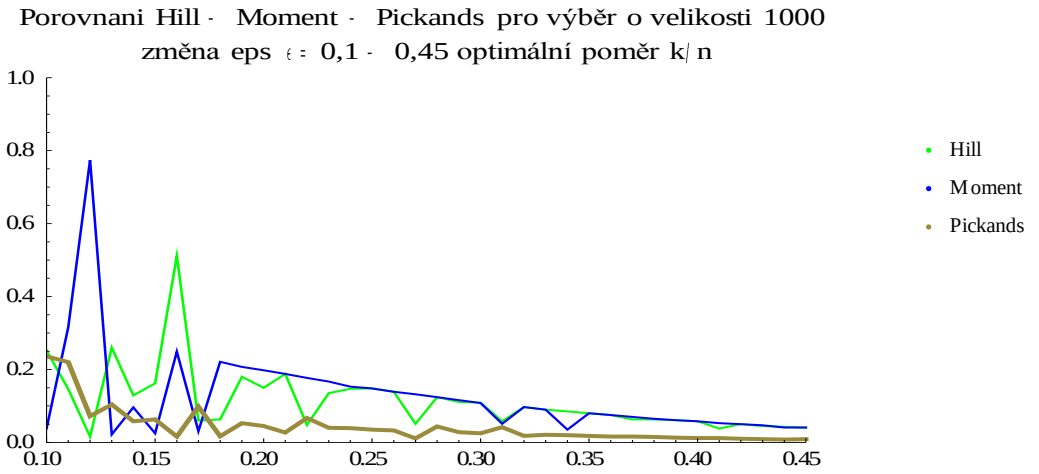
Graf 24 - Doba zpracování pro Hill - Moment - Pickands

Poslední obrázek je porovnání MSE jednotlivých odhadů. Z tohoto pohledu se opět vymyká odhad Pickandsův, je méně kvalitní než druhé dva. Ty mají MSE zhruba stejné, přesto je zajímavé, že pro jisté hodnoty jsou MSE „velké“ pro všechny tři odhady.

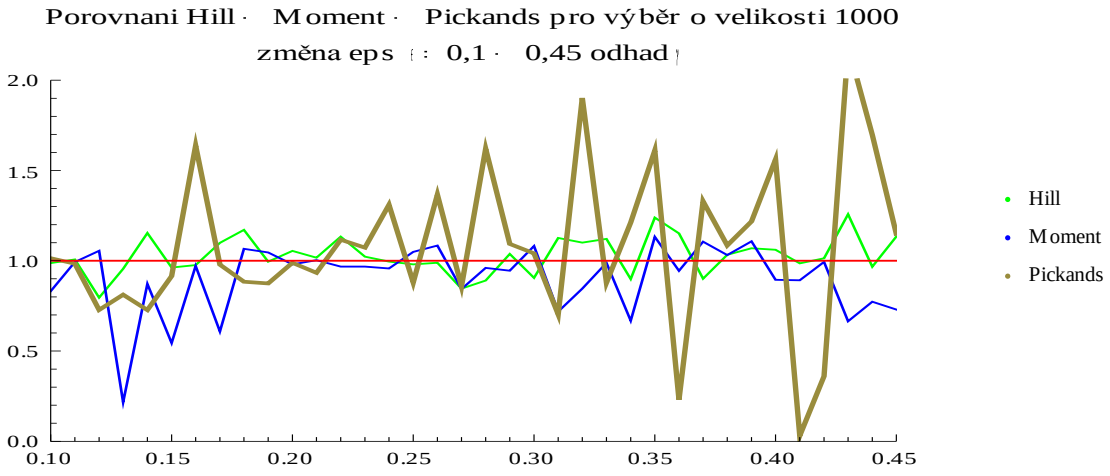


Graf 25 - MSE pro Hill - Moment - Pickands pro n do 1000, B=250 a ε= 0.15

V dalším jsme provedli změnu ε od 0,1 do 0,45. Hodnoty *B* jsme ponechali na 250 resp. 500. Vzhledem k omezenému prostoru ukážeme jen část výsledků (podobných výsledků máme k dispozici několik desítek).



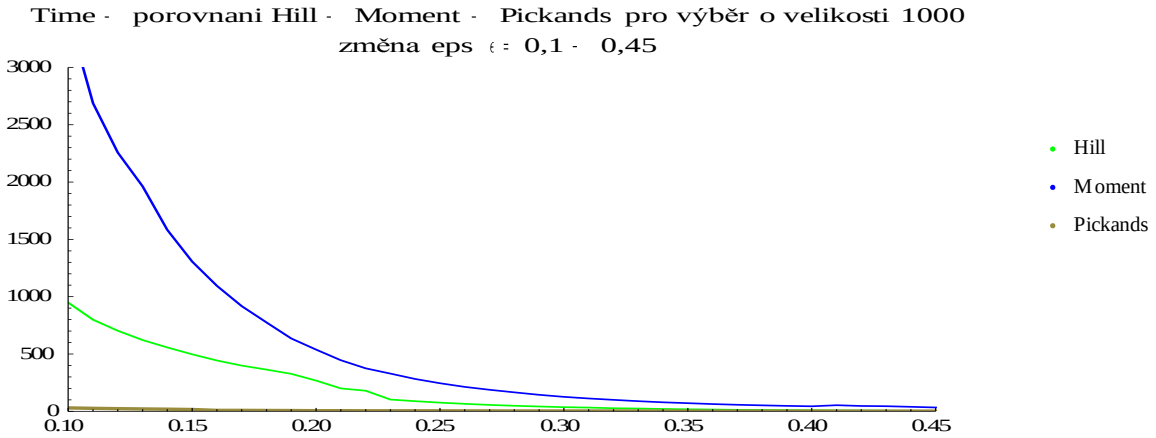
Graf 26 - Optimální poměr při porovnání odhadů pro od 0,1 do 0,45



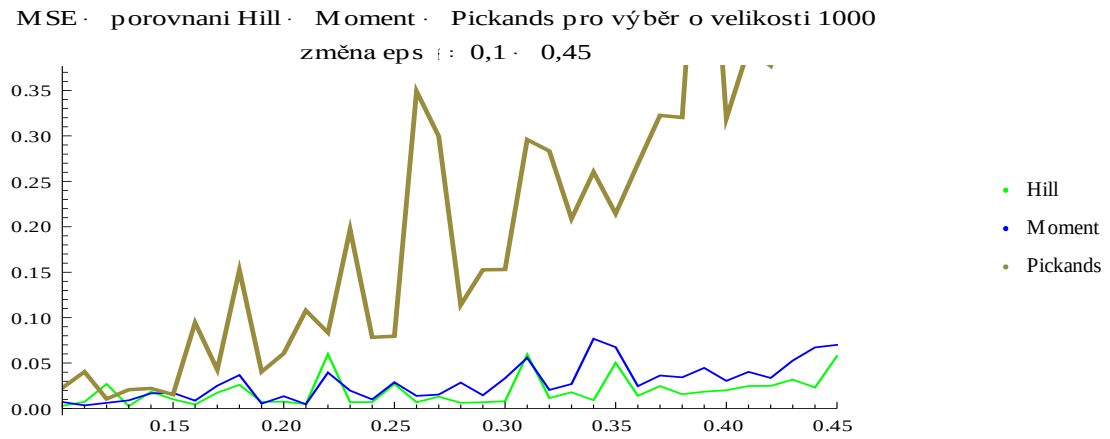
Graf 27 - Porovnání odhadů pro EVI při od 0,1 do 0,45

Z prvního obrázku je zřejmé, že s rostoucím ε klesá hodnota poměru *k/n*, tak jak jsme očekávali. Druhý obrázek prokazuje opět odlišnost Pickandsova odhadu v přesnosti.

Mezi Hillovým odhadem a momentovým není výrazný rozdíl. Nejpřesnější jsou odhady v oblasti ε od 0,18 do 0,25.



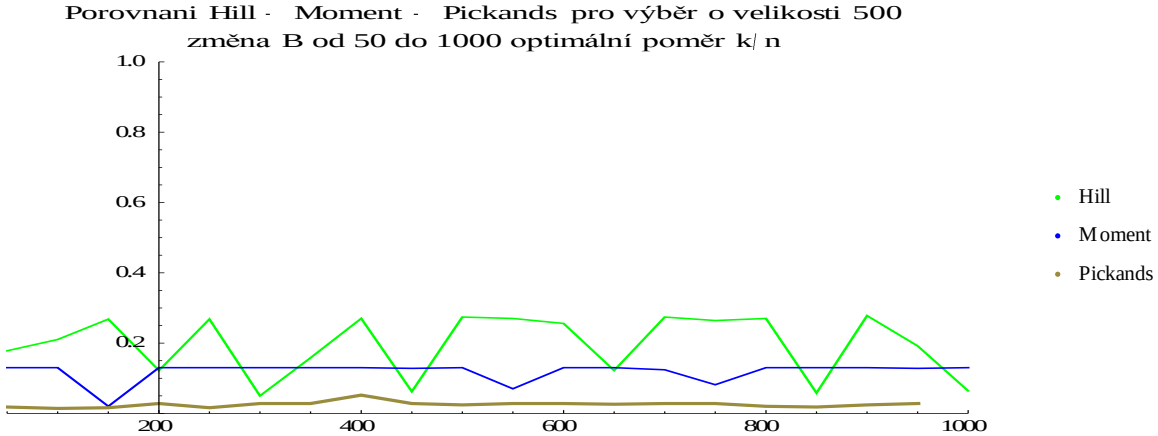
Graf 28 - Porovnání doby zpracování pro odhady celkem



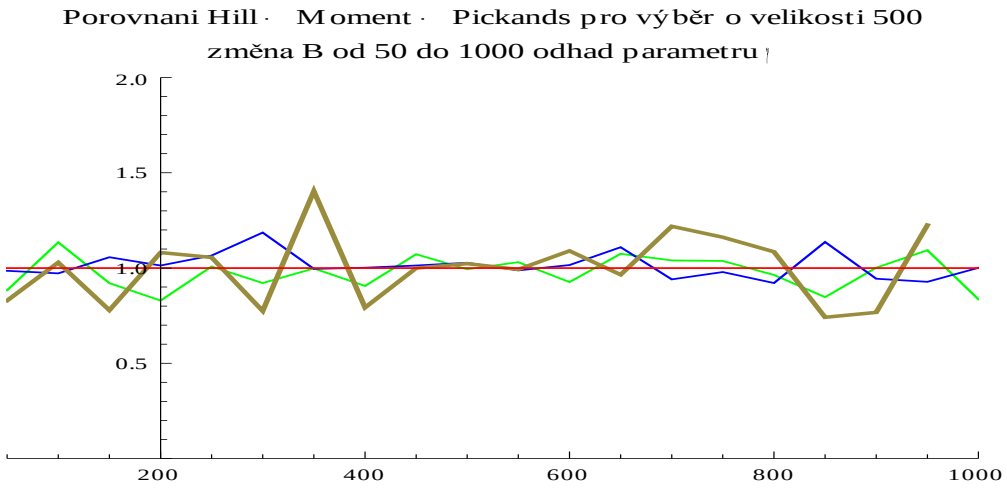
Graf 29 - Porovnání MSE pro odhady celkem

Z obrázků je patrný prudce se snižující výpočetní čas od cca 3000 vteřin u momentového odhadu u ε = 0,1 až k řádově desítkám sekund u ε = 0,45. Zároveň je patrná tendence růstu MSE u Pickandsova odhadu. Nejmenší hodnoty MSE nabývají Hillův a momentový odhad pro ε = 0,19 až ε = 0,21.

Podobně jsme porovnávali chod algoritmu pro měněné hodnoty *B* a konstantní ε .

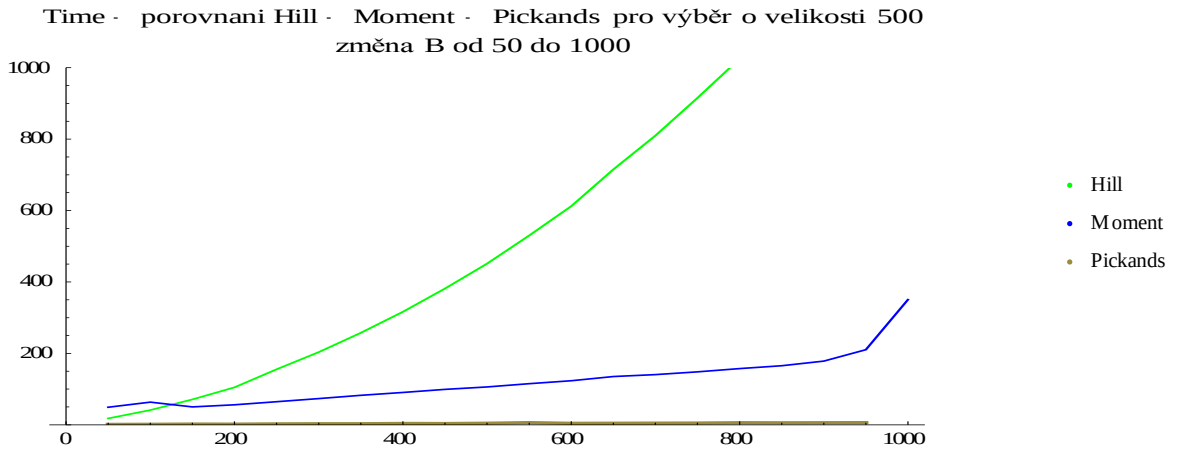


Graf 30 - Optimální poměr pro odhady s různou hodnotou B

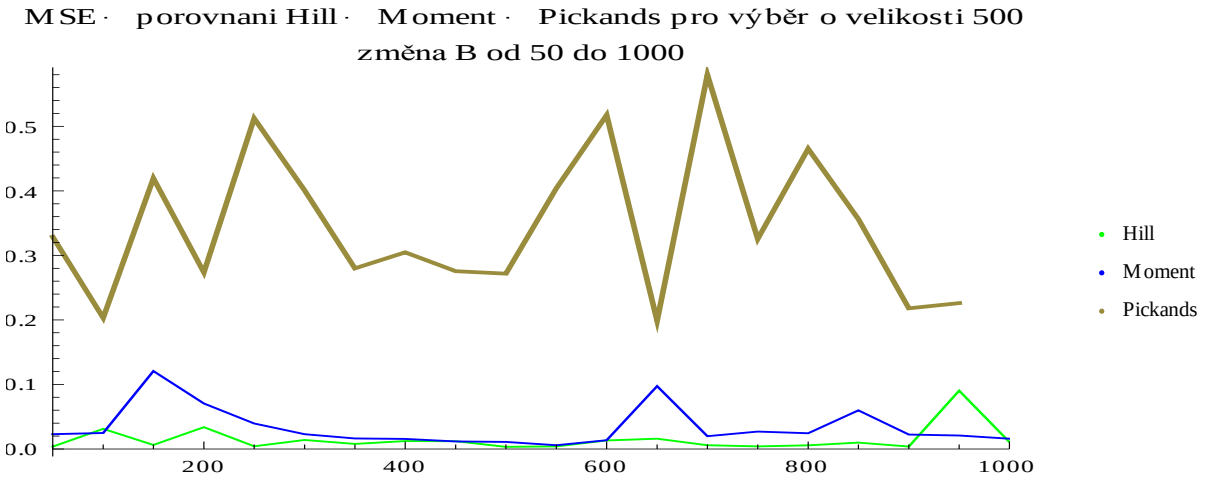


Graf 31 - Grafy pro konstantní ε=0.25 a proměnné B od 50 do 1000

Z grafů je patrno, že se optimální poměry málo mění, jsou zhruba na 15%. Podstatné je, že nejpřesnější odhady jsou ustáleny od *B*= 400 až do 500.



Graf 32 - Výpočetní náročnost ε=0.25 a proměnné B od 50 do 1000

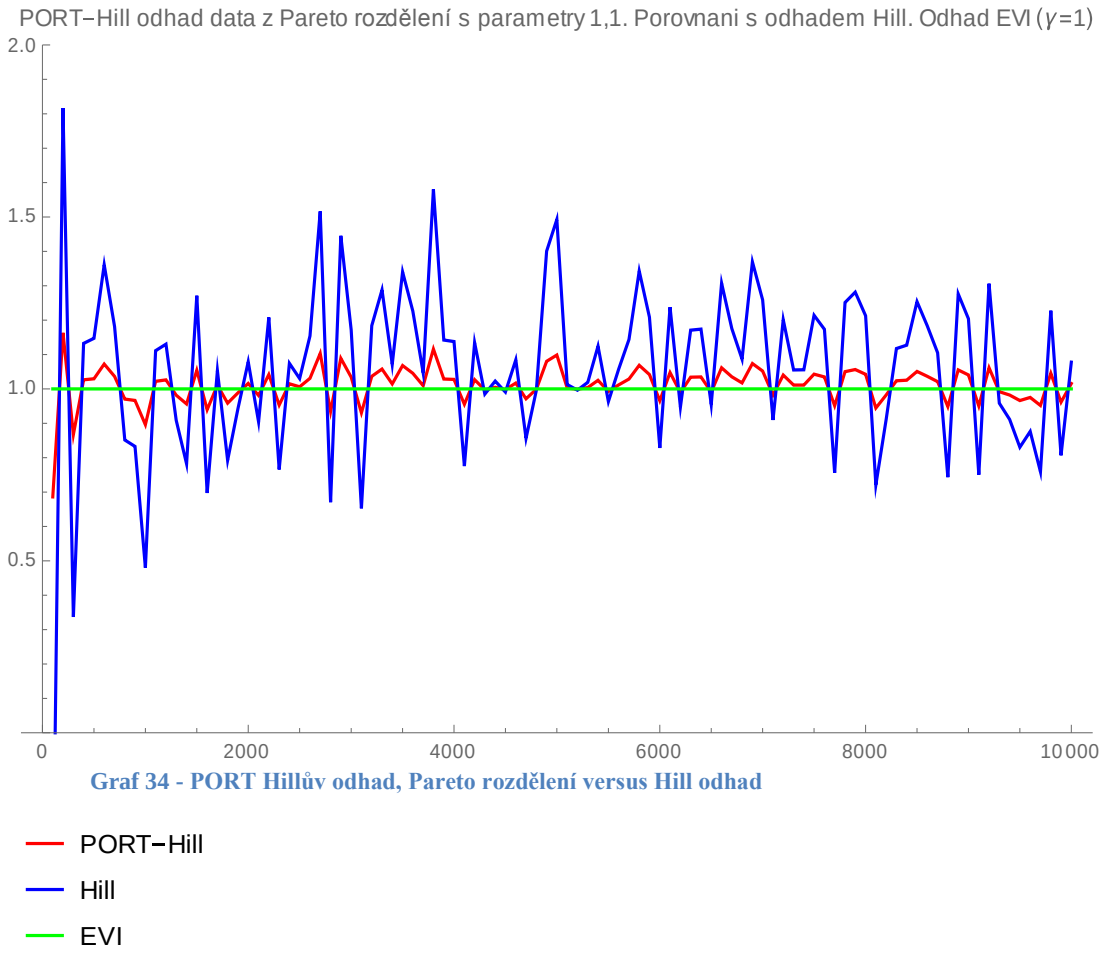


Graf 33 - MSE Pro konstantní ε=0.25 a proměnné B od 50 do 1000

Z grafů je na první pohled zřejmá část minimální MSE v oblasti *B* od 350 do 500. Na první pohled je zřejmý také velký rozdíl časové náročnosti proti předchozím grafům.

V následujících grafech jsou uvedeny odhady, které nejsou klasické. Jde o odhady PORT a MVRB. Prvním typ odhadu PORT byl poprvé zveřejněn v roce 2006. Jeho hlavní výhoda spočívá v tom, že je podobně jako Pickandsův odhad invariantní vzhledem k posunutí a násobení konstantou. Jeho nevýhodou je to, že jeho realizace je mnohem náročnější než klasické odhady.

Druhý typ odhadu – odhad MVRB je odhadem, který zlepšuje výkonnost odhadu, snižuje odchylku (bias) a zmenšuje také rozptyl. Je tedy přesnější než stejné typy odhadů např. Hill, moment. Poprvé byl konstruován kolem roku 2008. Tento odhad můžeme řešit metodou „optimal sample fraction“, tuto metodu je možné i zjednodušit. Je výpočetně dosti náročná.

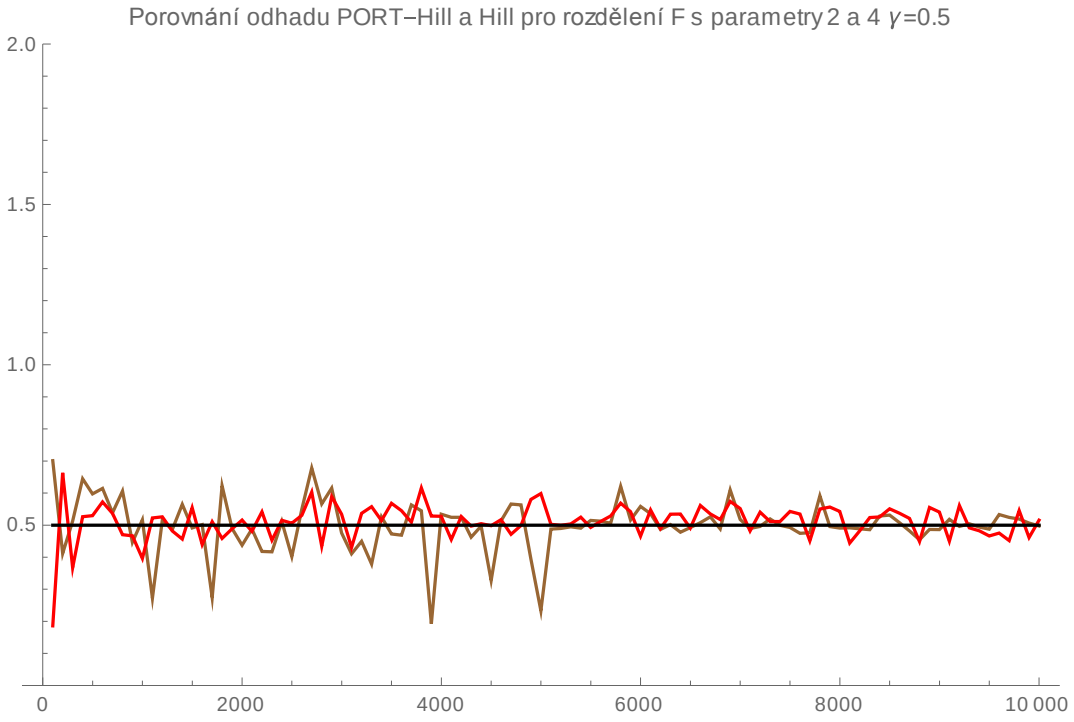


Graf 34 - PORT Hillův odhad, Pareto rozdělení versus Hill odhad

,

Na výše uvedeném grafu jsou zobrazeny výsledky odhadů Hillova a PORT-Hilova. Soubor dat, s nímž jsme pracovali byl generován z náhodné veličiny typu Paretovo rozdělení s parametry 1,1.

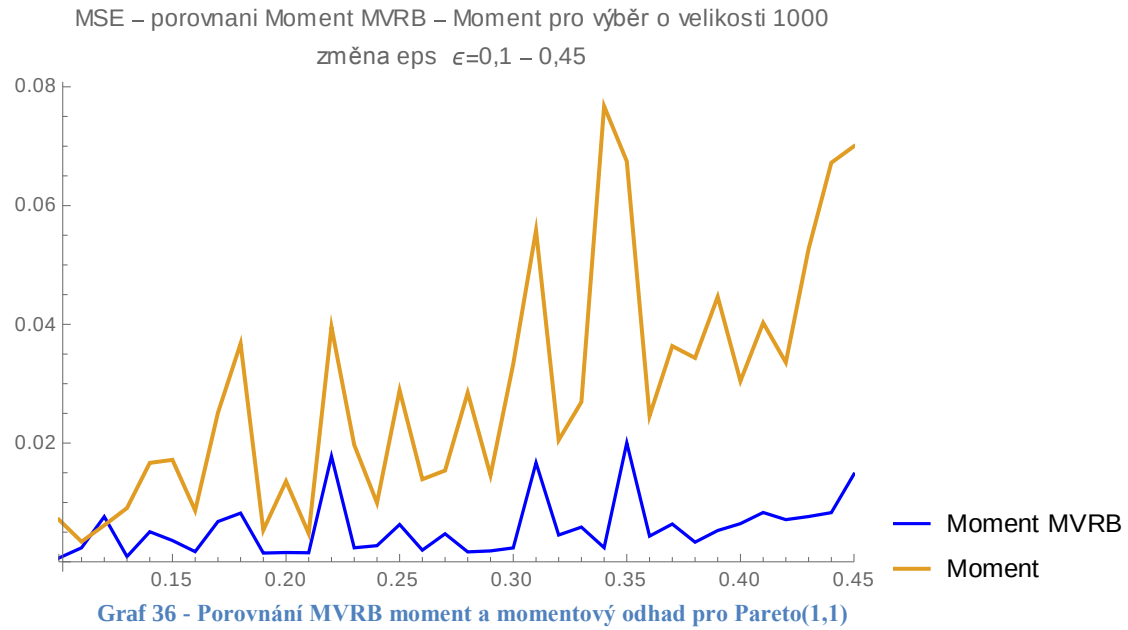
Vlastní výpočty probíhaly podobně jako u ostatních porovnání na stejných datech pro oba typy odhadu. Rozsah počtu členů ve výběru byl od 50 do 10 000. Na první pohled je zřejmé, že odhad PORT – Hill je mnohem přesnější. Absolutní odchylka v celém intervalu nepřesáhla hodnotu 0,3. Velká přesnost výpočtu je ovšem vynahrazena také velkou výpočetní náročností. Zatímco u Hillova odhadu vystačíme s metodou „optimal sample fraction“, u odhadu typu PORT musíme ještě navíc volit správnou hodnotu parametru q.



Graf 35 - Porovnání PORT Hill a Hill pro F rozdělení

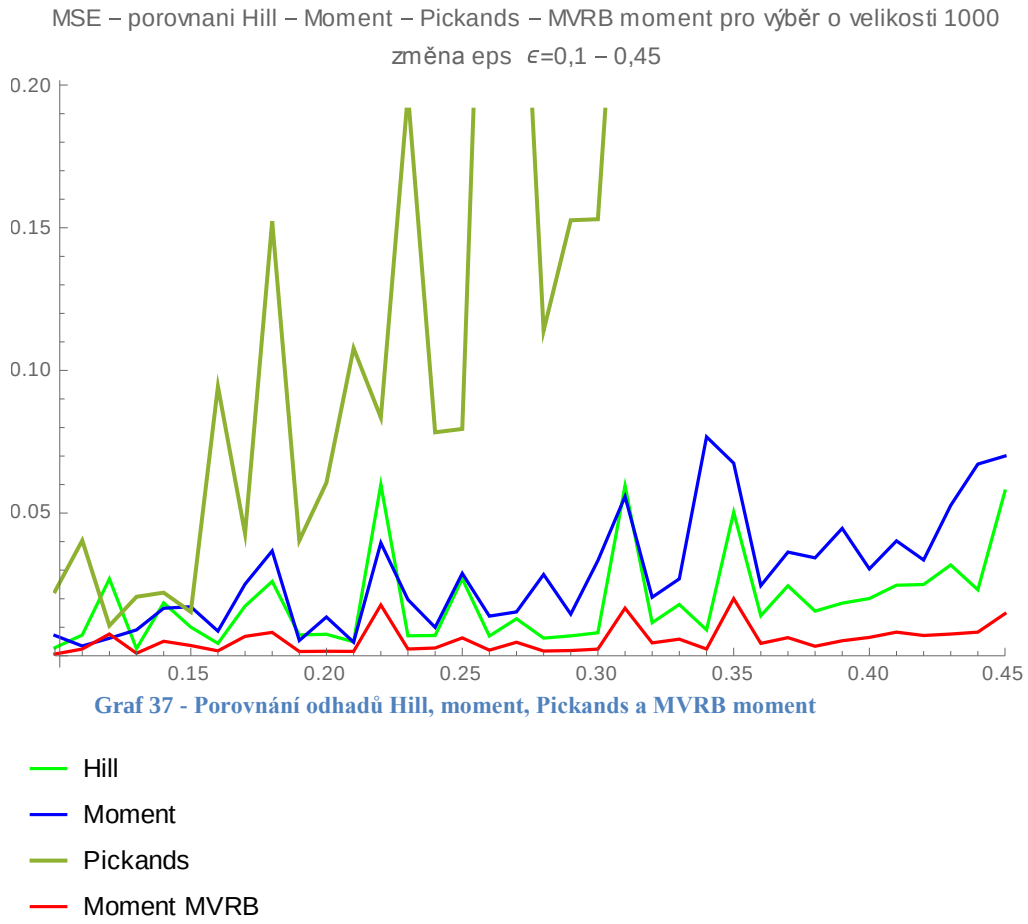
,

Podobně jako v předchozím případě porovnáváme postupně výběry od 50 prvkového výběru až do výběru o velikosti 10 000. Základem je rozdělení, z něhož jsou data generována. Jde o náhodnou veličinu typu F s parametry 2 a 4. Skutečná hodnota EVI je 0,5. Opět jsme počítali odhad Hillův pomocí metody „sample fraction“, odhad PORT je počítán pomocí algoritmu uvedeného v předchozí části. Při porovnání výsledků zjistíme, že oba výpočty jsou zhruba stejně kvalitní. Především pro počet prvků ve výběru od zhruba 5 000 prvků je odhad pomocí obou metod stejně kvalitní.



Graf 36 - Porovnání MVRB moment a momentový odhad pro Pareto(1,1)

Na předchozím grafu jsou uvedeny dva různé odhady – momentový odhad a momentový MVRB odhad. Protože oba typy odhadů se počítají pomocí metody „optimal sample fraction“ máme možnost srovnávat oba typy odhadu i jinak než jen podle kvality odhadu. Byl zvolen základní rozsah 5 000 prvků ve výběru, typ a pomocí volby ε jsou zvoleny části – „fraction“, které jsou základem celé bootstrapové metody. Na grafu je patrné, že existuje pro oba typy odhadů část, kdy je hodnota MSE nejmenší. Pro momentový odhad je to část kolem ε = 0,15, pro momentový odhad MVRB jde o hodnoty od 0,22 do 0,28. Dále můžeme na první pohled pozorovat, že se skutečně podařilo hodnotu MSE u odhadu typu MVRB znatelně snížit.



Graf 37 - Porovnání odhadů Hill, moment, Pickands a MVRB moment

,

Na závěr jsme zvolili možnost porovnat všechny odhady, které jsme vyšetřovali. Opět jsme zvolili podobně jako v předchozím případě velikost výběru 5 000, náhodné hodnoty jsou generovány z rozdělení Pareto s parametry 1,1. Je to proto, abychom mohli všechny odhady vůbec počítat. Na první pohled je opět zřejmé, že se vymyká odhad Pickandsův, ten je znatelně nejméně kvalitní (z pohledu MSE) než ostatní. U ostatních odhadů je zřetelná oblast od zhruba 0,22 do 0,28, kdy je hodnota MSE minimální. To by byla zřejmě optimální oblast pro volbu hodnoty ε, pomocí níž konstruujeme obě části (fraction). Zároveň je z grafu také patrné, že odhad typu MVRB skutečně minimalizuje hodnotu MSE tak, jak je v návrhu tohoto odhadu uvedeno. V této části jsme se zaměřili jen na jeden aspekt tvorby odhadu, pokud bychom vyšetřovali například časovou náročnost výpočtů odhadů, získali bychom v podstatě

„převrácené“ hodnoty – nejméně časově náročný je odhad Pickandsův a nejvíce náročný je odhad MVRB.

4.6.5. Shrnutí výsledků simulační studie

V této části jsme v první řadě na základě mnoha simulací ověřili známou vlastnosti klasických odhadů EVI. Jednoznačně z nich vyplynula doporučení používat ve většině reálných situací odhad Hillův. Odhad Pickandsův je nejen mnohem méně přesný (vyplývá to již z jeho konstrukce), ale především se jeho přesnost nijak nezlepšuje s rostoucím počtem n. Jedinou jeho výhodou je menší časová náročnost než ostatní odhady. Momentový odhad má výhodu universálního použití pro všechny sféry přitažlivosti, jeho nevýhodou je ovšem mnohem větší časová náročnost algoritmu. Momentový odhad je stejně přesný jako Hillův , v některých situacích je dokonce přesnější. Právě proto se jeví jako možná alternativa pro Hillův odhad tam, kde Hillův odhad nelze použít.

Hillův odhad je nejužívanější volba, ovšem má svá omezení plynoucí z jeho použití pro jednu sféru přitažlivosti. Naštěstí většina aplikací je zaměřena na rozdělení z Frèchetovy sféry přitažlivosti.

V práci jsme ovšem studovali i vlastnosti nových typů odhadů – typu PORT nebo MVRB. V obou případech jde o postupy, které jsou velmi náročné na rychlost výpočetní techniky a rychlost přístupu k paměti. Podle diskuse algoritmů uvedené v podkapitole 4.5 je zřejmé, že musíme nejdříve zjisti hodnoty konzistentního odhadu parametru $\hat{\rho}$ – to je však společné pro všechny odhady. Navíc ovšem je zapotřebí u metody PORT nalézt vhodnou hodnotu parametru q a u metody MVRB dalšího parametru $\hat{\beta}$. Tyto kroky jsou výpočetně velmi náročné. Je vhodné mít k tomu speciálně přizpůsobenou výpočetní techniku nebo výpočty řešit např. paralelním programováním.

Ve všech situacích jsme chtěli ověřit doporučené hodnoty pro základní bootstrapové procedury $B = 250$ a $\varepsilon = 0{,}15$.

Z uvedených simulací je ovšem zřejmé, že je pro jednotlivé odhady optimální volit hodnotu B spíše větší než 250. Nejvhodnější volbou se jeví B=500 a ε naopak na úrovni 0,25. Tím podstatně zrychlíme výpočet algoritmu, ale neztratíme přesnost výsledku. Hodnotu $B = 500$ jsme ověřili i pro nové typy odhadů PORT a MVSB. U nich ovšem doporučujeme v souladu s literaturou hodnotu $\varepsilon = 0{,}005$.

V průběhu simulace jsme dále ověřili, že rychlost výpočtu záleží na mnoha faktorech. Kromě zmíněných parametrů B, ε záleží velmi podstatně na typu odhadu. Jak jsme již uvedli nejrychleji je vypočítán Pickandsův odhad. Odhad Hillův je počítán mnohem pomaleji, v některých případech řádově pomaleji. Ještě pomalejší výpočet je prováděn u momentového odhadu. U Hillova a momentového odhadu je to dáno množstvím výpočtů, které musíme provést. O další řád pomalejší jsou odhady PORT resp. MVRB. Tyto odhady mají sice mimořádně dobré vlastnosti, ale výpočet je skutečně náročný vzhledem k množství pomocných parametrů, které je zapotřebí v každém kroku nalézt.

5. Kvantilová regrese a extremální statistiky

V Koenker (2015) je uveden příklad prvního využití regrese, která nesla známky tzv. mediánové regrese. Zajímavé na tom je to, že tento případ se udál kolem roku 1757. Tedy dříve než první využití klasické Gaussovy regrese. Jezuita Josip Boskovič v Chorvatsku se pokoušel odhadnout tvar zeměkoule pomocí regresní funkce minimalizující součet absolutních odchylek. Dnes tento typ regrese nazýváme mediánová regrese. Popis experimentu i detaily datového souboru nalezneme ve výše uvedeném článku.

Významnou statistickou úlohou je zkoumání a hledání závislosti proměnných. Jeden z nejpoužívanějších statistických nástrojů je regresní analýza. Cílem regresní analýzy je nalézt a popsat funkční vztahy mezi proměnnými. Nalézt model, který popíše závislost vysvětlované náhodné veličiny $\boldsymbol{Y}$ na vysvětlujících veličinách $X_1, \cdots, X_k$.

Pomocí naměřených hodnot $\boldsymbol{X} = (X_1, \cdots, X_k)'$ nalezneme predikci o závisle proměnné Y – chceme nalézt vhodnou funkci $h(\boldsymbol{X})$, která závisle proměnnou v předem definovaném smyslu definuje.

Kvalita takovéto predikce se potom posuzuje pomocí kriteriální funkce *L(u)*, která se nazývá ztrátová funkce. Za vhodnou optimální volbu *h(**X**)* se pak uvažuje funkce, která minimalizuje střední hodnotu takové ztrátové funkce.

5.1. Klasický regresní lineární model

Jestliže bude zvolena kriteriální funkce $L(u) = u^2$, potom mluvíme o klasické regresi.

Věta 5.1

Nechť $Y, X_1, \cdots, X_k$ jsou náhodné veličiny, $\boldsymbol{X} = (X_1, \cdots, X_k)^T$ a nechť $EY^2 < \infty$. Potom pro každou měřitelnou funkci h: $\mathbb{R}^k \rightarrow \mathbb{R}$ platí

$$E\big(Y - E(Y|\boldsymbol{X})\big)^2 \leq E(Y - h(\boldsymbol{X}))^2. \tag{5.1}$$

Rovnost v (5.1) nastává právě když

$$P\big((h(X)) = E(Y|\boldsymbol{X})\big) = 1.$$

V klasické regresi nejčastěji volíme funkci h jako lineární funkci,

$$h(\boldsymbol{X}) = \beta_1 X_1 + \cdots + \beta_k X_k.$$

Parametry $\{\beta_i\}_{i=1}^k$ se nazývají regresní koeficienty. Za pomocí prostředků statistické analýzy tyto koeficienty budeme z existujících dat odhadovat. Bude tedy platit

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.2}$$

vektor $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$ je vektorem náhodných veličin splňující podmínky

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}. \tag{5.3}$$

pokud jsou navíc ještě náhodné veličiny $\{\epsilon_i\}_{i=1}^n$ normálně rozdělené nezávislé náhodné veličiny, označujeme tento model jako klasický lineární regresní model. Potom je tedy $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ (jako v předchozích kapitolách označujeme symbolem $N(\mu, \sigma^2)$ jednorozměrné normální rozdělení s parametry μ, σ a symbolem $N(\boldsymbol{a}, \boldsymbol{B})$ nazveme n rozměrné normální rozdělení s parametry $(\boldsymbol{a}, \boldsymbol{B})$), potom je dále $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$

Neznámé parametry můžeme odhadnout pomocí známé metody nejmenších čtverců. Požadujeme, aby byl výraz $(\boldsymbol{Y} - \boldsymbol{X\beta})^T (\boldsymbol{Y} - \boldsymbol{X\beta})$ minimální.

Věta 5.2

Odhady metodou nejmenších čtverců jsou dány vztahem

$$\boldsymbol{b} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y}. \tag{5.4}$$

Soustavě rovnic $\boldsymbol{X^T X\, b} = \boldsymbol{X^T Y}$ říkáme normální rovnice. Dále uvedeme některé vlastnosti odhadu $\boldsymbol{b}$.

Věta 5.3

Odhad $\boldsymbol{b}$ je nestranný, tj.

$$\boldsymbol{E\, b} = \boldsymbol{\beta} \tag{5.5}$$

a má variační matici

$$VAR(\boldsymbol{b}) = \sigma^2 (\boldsymbol{X^T X})^{-1} \tag{5.6}$$

a dále platí

$$\boldsymbol{b} \sim N(\beta, \sigma^2 (\boldsymbol{X^T X})^{-1})$$

5.2. Úvod do kvantilové regrese

V předchozí části jsme zopakovali základní pojmy z prostředí lineární regrese. Kvantilová regrese se od klasické regrese (využívající k optimalizaci metodu nejmenších čtverců)liší tím, že sledujeme jinou podmíněnou charakteristiku náhodné veličiny $\boldsymbol{Y}$. V klasické regresi, jak bylo popsáno v předchozí části, vyšetřujeme minimum střední čtvercové odchylky $arg_{t \in \mathbb{R}} \min E((\boldsymbol{Y} - t)^2 | \boldsymbol{X} = \boldsymbol{x})$. Pomocí tohoto minima můžeme tedy odhadnout podmíněnou střední hodnotu vysvětlované proměnné $\boldsymbol{Y}$ neboli

$$E(\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}) = t. \tag{5.7}$$

U kvantilové regrese odhadujeme podmíněnou kvantilovou funkci. Kvantilová regrese vychází z poznatku, τ - tý kvantil náhodné veličiny Y lze na základě pozorování $y_1, \cdots, y_n$

odhadnout pomocí empirického τ – tého kvantilu neboli je možné ho najít jako argument minima (vzhledem k proměnné u_α) dále uvedené empirické rizikové funkce

$$\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\,(y_i - u_\alpha)$$

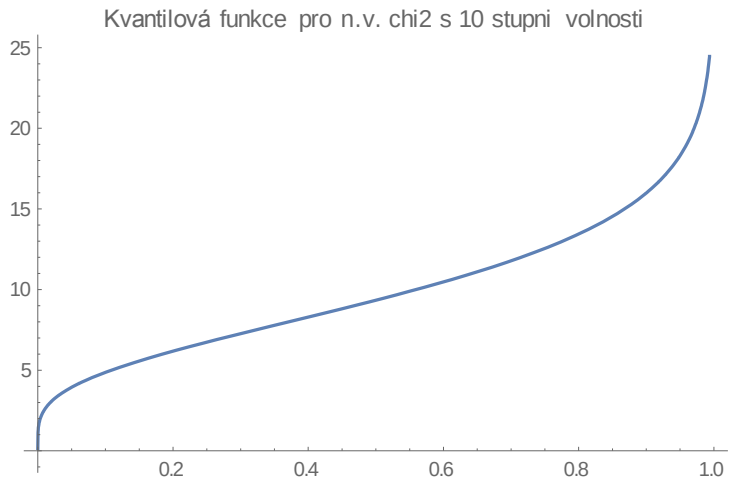
Definice 5.4

Nechť F je distribuční funkce spojité náhodné veličiny Y a $\tau \in (0,1)$. Potom funkce

$$Q(\tau) = F^{-1}(\tau) = q_\tau = inf(x \in \mathbb{R};\ \tau > F(x)), \qquad (5.8)$$

se nazývá kvantilová funkce a číslo $q_\tau = Q(\tau)$ se nazývá τ - kvantil náhodného rozdělení Y.

Příklad kvantilové funkce náhodné veličiny $Y \sim \chi^2(10)$ je uveden na následujícím obrázku.



Graf 38 - Kvantilová funkce pro n.v. chi kvadrat s 10 stupni volnosti

Nyní budeme postupovat podobně jako v části 5.1. nejdříve uvedeme další možnou reprezentaci kvantilové funkce.

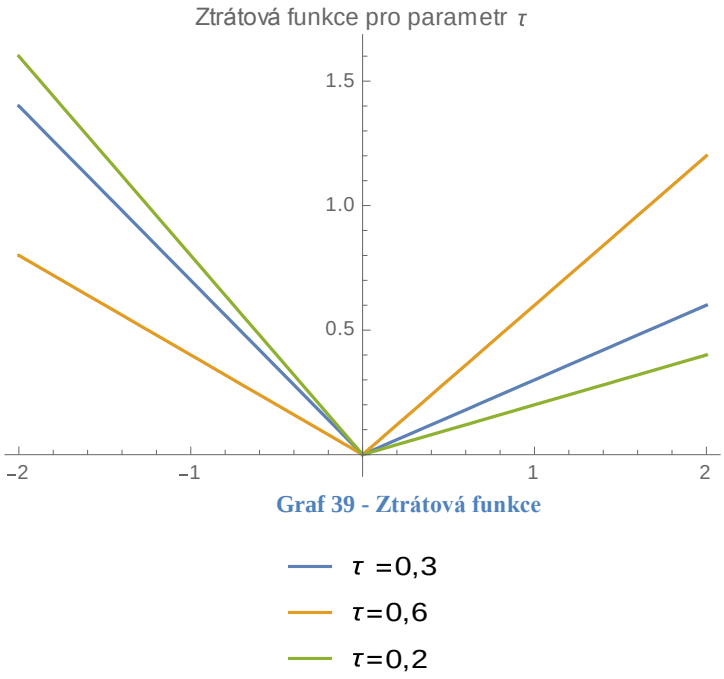
Poznámka 5.5

Kvantilovou funkci $Q(\tau)$ můžeme také zavést jako řešení optimalizační úlohy

$$Q(\tau) = q_\tau = arg_{t\in\mathbb{R}}\min\ E\big(\rho_\tau(\boldsymbol{X} - t)\big), \qquad (5.9)$$

kde $\rho_\tau(\boldsymbol{X} - t)$ je tzv. **ztrátová** funkce, která je definována následujícím předpisem

$$\rho_\tau: x \mapsto \begin{cases} \tau & ,\quad x > 0, \\ 1-\tau & ,\quad x \le 0. \end{cases}$$



Graf 39 - Ztrátová funkce

Nejdříve dále upravíme hodnotu $E\left(\rho_\tau(\boldsymbol{X}-t)\right)$

$$H(t) = E\left(\rho_\tau(\boldsymbol{X}-t)\right) = (1-\tau)\int_{-\infty}^{t}(t-x)\,dF(x) + \tau\int_{t}^{\infty}(x-t)dF(x).$$

Dále budeme postupovat stejně jako u věty 5.1. Při hledání řešení minimalizačního problému (5.9), nalezneme kritický bod výše uvedené funkce $H(t)$

$$H'(t) = \frac{\partial}{\partial t}\left((1-\tau)\int_{-\infty}^{t}(t-x)\,dF(x) + \tau\int_{t}^{\infty}(x-t)dF(x)\right) = 0$$

Protože jsou splněny podmínky pro záměnu pořadí derivace a integrálu, můžeme ji provést. Tedy získáme po úpravě

$$(1-\tau)\int_{-\infty}^{t}\frac{\partial}{\partial t}(t-x)\,dF(x) + \tau\int_{t}^{\infty}\frac{\partial}{\partial t}(x-t)dF(x) = 0,$$

$$(1-\tau)F(t) - \tau\,(1-F(t)) = 0,$$

$$F(t) = \tau,$$

$$t = F^{-1}(\tau).$$

Řešení dané minimalizační úlohy je tedy τ - tý kvantil náhodné veličiny $\boldsymbol{X}$. proto je podmíněná kvantilová funkce náhodné veličiny $\boldsymbol{Y}$ řešením optimalizační úlohy

$$Q_{\boldsymbol{Y}}(\tau|\boldsymbol{X}=\boldsymbol{x}) = arg_{t\in\mathbb{R}}\min E(\rho_\tau(\boldsymbol{Y}-t)|\boldsymbol{X}=\boldsymbol{x}). \qquad (5.10)$$

Pokud budeme studovat plně lineární model $Q_{\boldsymbol{Y_i}}(\tau|\boldsymbol{X}=\boldsymbol{x}) = \boldsymbol{x_i^T}\,\boldsymbol{\beta}(\tau)$, můžeme získat i vlastnosti ztrátové funkce kvantilové regrese pro tento případ. Výše uvedeným postupem získáme odhad $\hat{\beta}(\tau)$ minimalizací ztrátové funkce

$$\hat{\beta}(\tau) = \; arg_{\beta\in\mathbb{R}^k} min \sum_{j=1}^{n} \rho_\tau\left(\boldsymbol{y_j} - x_j^T\beta\right). \tag{5.11}$$

Funkce $S(\beta) = \sum_{j=1}^{n} \rho_\tau\left(\boldsymbol{y_j} - x_j^T\beta\right)$ je spojitá, po částech lineární a diferencovatelná až na j bodů, v nichž se anulují jednotlivé argumenty funkce ρ_τ. Ovšem i v těchto bodech nalezneme Gateauxův diferenciál v každém směru. Můžeme proto počítat tyto diferenciály v každém směru. Funkce $S(\beta)$ bude v daném bodě nabývat minima, jestliže pro libovolný směr bude hodnota Gateauxova diferenciálu v tomto bodě kladná. Analýzu daného postupu lze včetně algoritmu k nalezení tohoto minima nalézt v Koenker (2005).

Na rozdíl od klasické lineární regrese neexistuje jednoduché analytické vyjádření regresních parametrů kvantilové regrese. Naštěstí v Koenker (2005) lze nalézt postup, který umožňuje tyto odhady provést pro malé až středně velké soubory dat. Základní technikou řešení je lineární programování. Detaily o způsobech práce, algoritmech a způsobech řešení problémů je možno nalézt například v Vanderbei (2013).

Pro úlohu (5.10) s *k* vysvětlujícími proměnnými, který je uvedený ve tvaru

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \; \epsilon,$$

je možné přepsat tuto minimalizační úlohu na tvar viz Koenker (2005)

$$min\left\{\begin{matrix}\tau\; 1_n^T\,\epsilon_+ + \; (1-\tau)\; 1_n^T\,(-\epsilon)_+ | \boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \epsilon_+ + \; (-\epsilon)_+, \tau \in (0,1), \\ \boldsymbol{\beta} \in \mathbb{R}^k, [\epsilon_+, (-\epsilon)_+] \in \mathbb{R}_+^{2\,n}\end{matrix}\right\} \tag{5.12}$$

kde $\epsilon_+ = \begin{cases}\epsilon & \epsilon > 0 \\ 0 & jinak\end{cases}$.

Pokračujeme – li v postupu navrženém v Kroenker (2005), přepíšeme úlohu (5.12) do kanonického tvaru (je nutno vyjádřit hodnoty y_i pomocí nezáporných výrazů). Pro ukázku vyjádříme

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_{j,\alpha} + \; \epsilon_{i,\alpha} = \sum_{j=1}^{k} x_{ij}\left(\beta_{j,\alpha}^{+} - \; \beta_{j,\alpha}^{-}\right) + \; \left(\epsilon_{i,\alpha}^{+} - \; \epsilon_{i,\alpha}^{-}\right),$$

kde $a^+ = \max(0,a)$, $a^- = \max(0,-a)$, $\rho_\alpha(t) = \; \alpha\; t^+ + \; (1-\alpha)t^-$

Odtud obecně

$$\min_{x}\{\boldsymbol{c}^T\,\boldsymbol{x}|\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \; \in \boldsymbol{S}\}, \tag{5.13}$$

kde

$$\boldsymbol{c} = \left(\mathbf{0}_k^T, \mathbf{0}_k^T, \tau\, \mathbf{1}_{n,}^T, (1-\tau)\mathbf{1}_n^T\right)^{\boldsymbol{T}}$$

$$\boldsymbol{x} = \ (\boldsymbol{b}^T, \boldsymbol{\epsilon}_+, (-\epsilon)_+)^{\boldsymbol{T}}$$

$$\boldsymbol{A} = \ (\boldsymbol{x} - \boldsymbol{x}\, \boldsymbol{I_n} - \ \boldsymbol{I_n})$$

$$\boldsymbol{b} = \ \boldsymbol{y}^{\boldsymbol{T}}$$

$$S = \ \mathbb{R}^k \ \times \ \mathbb{R}_+^{2\,n}.$$

Z takového kanonického tvaru lze odvodit duální úlohu

$$\max_{y}\{\boldsymbol{b}^{\boldsymbol{T}}\, \boldsymbol{y} | \boldsymbol{c} - \ \boldsymbol{A}^T\, \boldsymbol{y}, \boldsymbol{y} \in (0,1)^{\boldsymbol{n}}, \cdots\} \tag{5.14}$$

Postup odvození obou úloh je nastíněn v Kroenker (2005). Zde autor zároveň doporučuje užívat pro malé a středně velké výběry simplexovou metodu a pro velké výběry například Frisch – Newtonovu metodu vnitřního bodu. Obě výše uvedené úlohy jsou řešeny pomocí věty 6.1 a 6.2 v uvedené monografii Kroenker (2005).

5.3. Vlastnosti kvantilové regrese

Dále uvedené vlastnosti uvádíme bez důkazů. Všechny důkazy lze nalézt v monografii Koenker (2005). První z vlastností, které uvedeme, budou vlastnosti odhadů parametrů $\boldsymbol{\beta}$.

5.3.1. Invariance vůči transformacím modelu

Dané vlastnosti jsou uvedené ve větě 2.3 v Koenker (2005).

Věta 5.6 (Koenker, Basset, 1978)

Nechť A je regulární matice typu k xk, $\gamma \in \mathbb{R}^k$a $\alpha > 0$. Potom pro libovolné $\tau \in< 0{,}1 >$ platí

(i) $$\hat{\beta}(\tau;\ \alpha\, \boldsymbol{Y}\, , \boldsymbol{X}) = \ \alpha\, \widehat{\beta}\, (\tau; \boldsymbol{Y}, \boldsymbol{X}) \tag{5.15}$$

(ii) $$\hat{\beta}(\tau; -\alpha\, \boldsymbol{Y}\, , \boldsymbol{X}) = \ -\alpha\, \widehat{\beta}\, (1-\tau; \boldsymbol{Y}, \boldsymbol{X}) \tag{5.16}$$

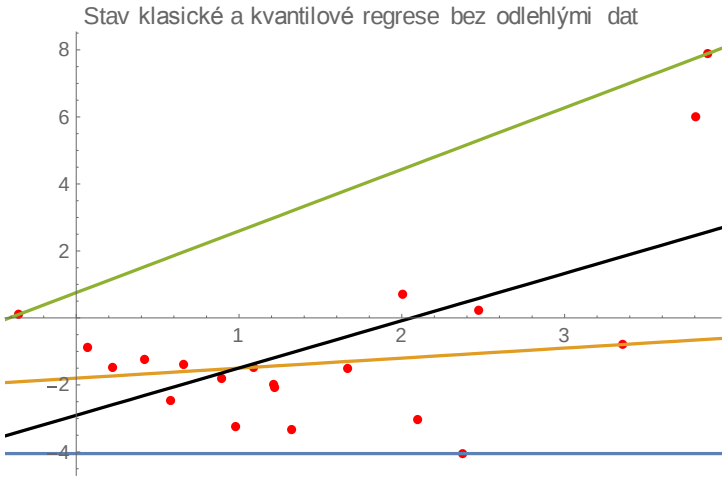
(iii) $$\hat{\beta}(\tau;\ \boldsymbol{Y}\ + \boldsymbol{X}\, \gamma\, , \boldsymbol{X}) = \ \widehat{\beta}\, (\tau; y, \boldsymbol{X}) + \ \gamma \tag{5.17}$$

(iv) $$\hat{\beta}(\tau;\ \ y\, , \boldsymbol{X}\, \boldsymbol{A}) = \ \boldsymbol{A}^{-1}\widehat{\beta}\, (\tau; y, \boldsymbol{X}). \tag{5.18}$$

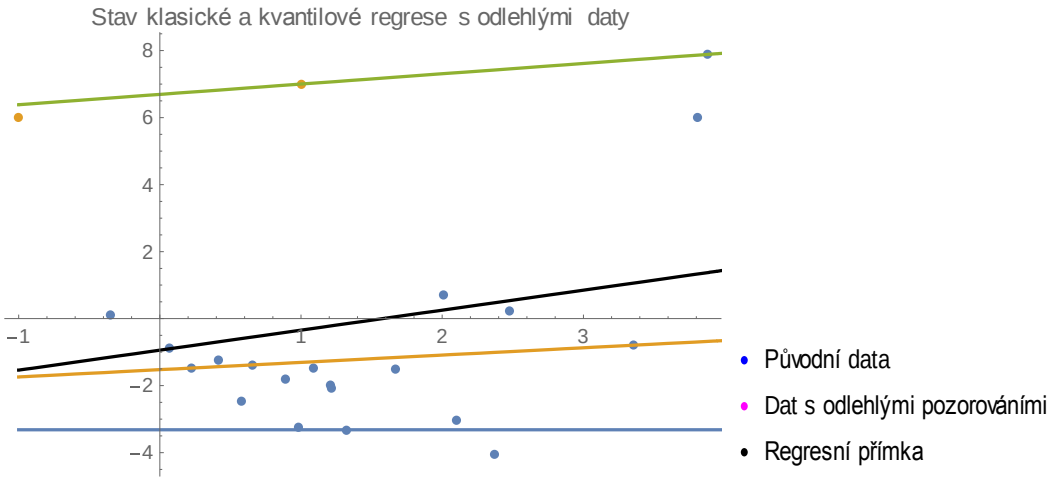
Vlastnosti (i), (ii) jsou nazývány equivariancí vzhledem k homogenitě y. Vlastnost (iii) je označována jako equivariantní vzhledem k posunutí v proměnné y a konečně poslední vlastnost (iv) je equivariantní vzhledem ke změně tvaru $\boldsymbol{X}$.

5.3.2. Robustnost

Kvantilová regrese je robustní vzhledem k extrémním pozorováním. Pojem robustnosti vychází ze skutečnosti, že množinou pozorování prokládáme nadroviny a zjišťujeme, že určité procento τ pozorování leží pod hadrovinou a $1-\tau$ procent pozorování leží pod touto hadrovinou. Podstatné je, že sledujeme počet takových pozorování a není podstatná velikost odchylek. Je proto přirozené, že extrémní pozorování nijak odhad nezmění. Například pro metodu nejmenších čtverců tato vlastnost neplatí. Výsledný odhad provedený pomocí kvantilové regrese je ve srovnání s metodou nejmenších čtverců výrazně méně citlivý na extrémní (odlehlá) pozorování. Na následujících obrázcích je možno pozorovat chování klasické regrese a kvantilové regrese vzhledem k odlehlým pozorováním.



Graf 40 - Robustnost - bez odlehlých dat



Graf 41 - Robustnost - včetně odlehlých údajů

5.3.3. Nestrannost

Na rozdíl od metody nejmenších čtverců jsou odhady pomocí kvantilové regrese slabě vychýlené viz Tasche (2001). Autor dokazuje, že odhady získané kvantilovou regresí jsou nestranné za předpokladu, náhodné složky pochází ze stejného rozdělení se spojitou a symetrickou hustotou. Jestliže jsou ale náhodné složky asymetrické nebo nepochází ze stejného unimodálního rozdělení, nejsou odhady pomocí kvantilové regrese nestranné. Symetrie a unimodálnost jsou rozhodující pro nestrannost odhadů v kvantilové regresi. Metoda nejmenších čtverců i metoda nejmenších absolutních odchylek jsou nestranné. Tedy tyto metody jsou vhledem k nestrannosti mnohem robustnější.

5.3.4. Vydatnost

Vydatnost odhadu spočívá ve skutečnosti, že při porovnání s ostatními nestrannými odhady vykazuje nejmenší možný rozptyl. Formálně odhad $\hat{\beta}$ je vydatný, jestliže

$$VAR\ \hat{\beta}\ \leq VAR\ \beta^{*} \tag{5.19}$$

pro všechny nestranné odhady β^{*}. Pro případ vícerozměrných parametrů je definice vydatnosti založena na kovariančních maticích odhadu viz Anděl (1978). Odhad $\hat{\beta}$ je vydatnější než odhad β^{*} jestliže $VAR\ \hat{\beta}\ - VAR\ \beta^{*}$je pozitivně definitní matice.

V Newey, Powell (1990) je ukázáno, že odhad kvantilové regrese není vydatným odhadem pro β_{τ}. Jsou zde také uvedeny konstrukce vydatných odhadů β.

Asymptotickou kovarianční matici lze odhadnout pomocí následujícího vztahu

$$VAR\ \hat{\beta}_{\alpha} = (\boldsymbol{X}^{T}\boldsymbol{X})^{-1}\ \boldsymbol{X}^{T}\ \boldsymbol{D}\ \boldsymbol{X}\ (\boldsymbol{X}^{T}\boldsymbol{X})^{-1}, \tag{5.20}$$

kde $\boldsymbol{D}$ je diagonální matice s prvky definovanými

$$d_{i} = \begin{cases} \left(\frac{\tau}{f(0)}\right)^{2}, pokud\ (y_{i} - x_{i}\beta_{\alpha}) > 0, \\ \left(\frac{1-\tau}{f(0)}\right)^{2}, pokud\ (y_{i} - x_{i}\beta_{\alpha}) \leq 0, \end{cases} \tag{5.21}$$

kde $f(0)$ je hodnota hustoty odchylek modelu v bodě 0.

V případě, že jsou pozorování nezávislá a stejně rozdělená, můžeme diagonální prvky matice $\boldsymbol{D}$ nahradit konstantou $c = \frac{\tau(1-\tau)}{\left(f\left(F^{-1}(\tau)\right)\right)^{2}}$. Pro mediánovou regresi (případ $\tau = 0{,}5$) můžeme tuto konstantu vyjádřit $c = \left(\frac{0{,}5}{f(0)}\right)^{2}$. Potom můžeme upravit vztah pro asymptotickou kovarianční matici na jednodušší tvar

$$VAR\ \hat{\beta}_{\tau} = c(\boldsymbol{X}^{T}\boldsymbol{X})^{-1} \tag{5.22}$$

Tento postup je uveden v Koenker (2005).

Problém se pak redukuje na vyjádření hustoty reziduí – chyb modelu. Protože tato hustota skoro vždy není známa a musíme proto nalézt její odhad. Postup při tvorbě tohoto odhadu je uveden v Koenker, Bassett (1978, 1982). V práci Koenker, Bassett (1982) je odvozen asymptotický vzorec. Ten ovšem poskytuje ve většině případů velmi podhodnocené odhady chyb. Proto se v tomto případě používá k získání odhadu hustoty metoda bootstrap. Podrobnosti lze nalézt v Greene (2008) , Hušek, Pelikán (2003).

5.3.5. Konzistence a asymptotická normalita

V článcích Koenker, Bassett (1978, 1982), Buchinsky (1998), Powell (1984) řešili autoři problém asymptotického normalitu odhadu kvantilové regrese a také jeho konzistenci. Označíme – li $\lim_{k\to\infty} \frac{1}{k} (\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{V}$ a jestliže můžeme náhodné složky popsat pomocí distribuční funkce F s ryze kladnou hustotou v τ ($f(F^{-1}(\tau)) > 0$), pak je odhad $\hat{\beta}_\tau$ asymptoticky normální

$$\sqrt{k}\left(\hat{\beta}_{\tau k} - \beta_\tau\right) \xrightarrow{P} N(\mathbf{0}, \boldsymbol{V}_\tau), \tag{5.23}$$

kde

$$\boldsymbol{V}_\tau = \frac{\tau(1-\tau)}{f^2\left(F^{-1}(\tau)\right)} \boldsymbol{V^{-1}}. \tag{5.24}$$

Znamená to, že odhad $\hat{\beta}_{\alpha k}$ je asymptoticky normální se střední hodnotou β_τ a s kovarianční maticí $\frac{1}{k}\boldsymbol{V}_\tau$. Tedy platí

$$\hat{\beta}_{\tau k} \sim N\left(\beta_\tau, \frac{1}{k}\boldsymbol{V}_\tau\right) \tag{5.25}$$

5.3.6. Zachování monotónní transformace vysvětlované proměnné

V případě monotónní transformace $h(.)$ vysvětlované proměnné $\boldsymbol{Y}$ platí, že

$$F(y) = P(\boldsymbol{Y} \le y) = P(h(\boldsymbol{Y}) \le y) = F(h(y))$$

z tohoto vztahu můžeme vyvodit, že platí

$$Q_\tau(h(y_i)|\boldsymbol{X} = x_i) = h\left(Q_\tau(y_i|\boldsymbol{X} = x_i)\right) \tag{5.26}$$

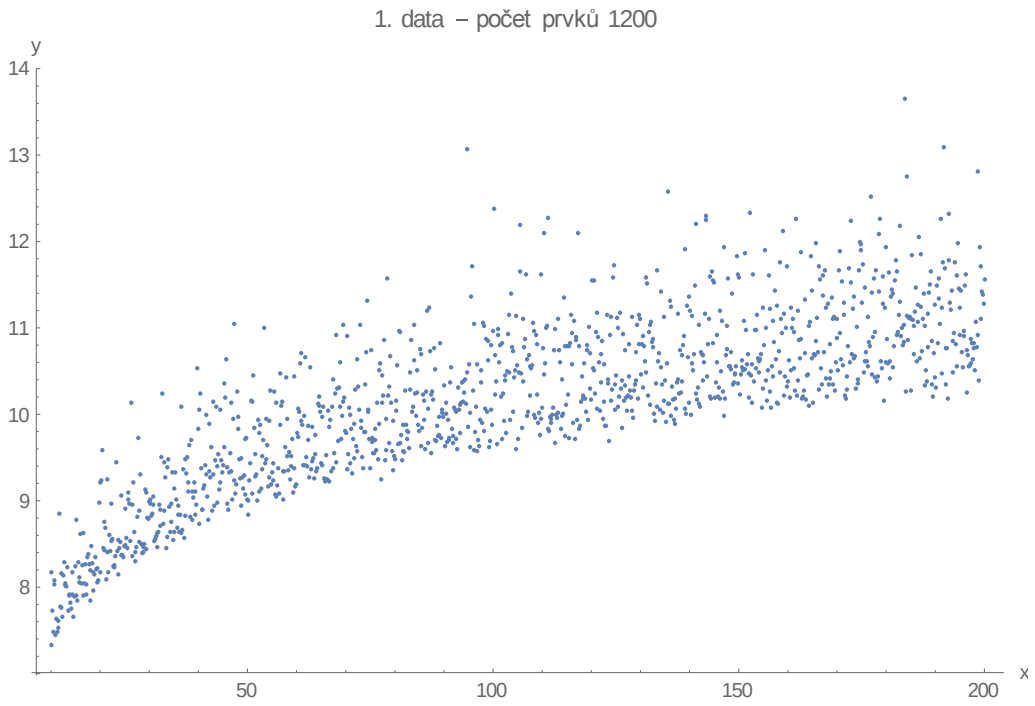
Jestliže tedy chceme zjistit například podmíněný kvantil logaritmu hmotnosti, stačí nalézt příslušný kvantil náhodné veličiny popisující hmotnost.

5.4. Příklad na kvantilovou regresi

V další části této kapitoly zobrazíme s rozborem jeden příklad na kvantilovou regresi. Především jsme se soustředili na grafickou interpretaci. Programově jsme využili aparátu programu Wolfram Mathematica, který budeme také vzužívat v simulacích v této kapitole. Podobně jako v jazyku R jsme použili některé známé procedury pro řešení problémů lineárního programování.

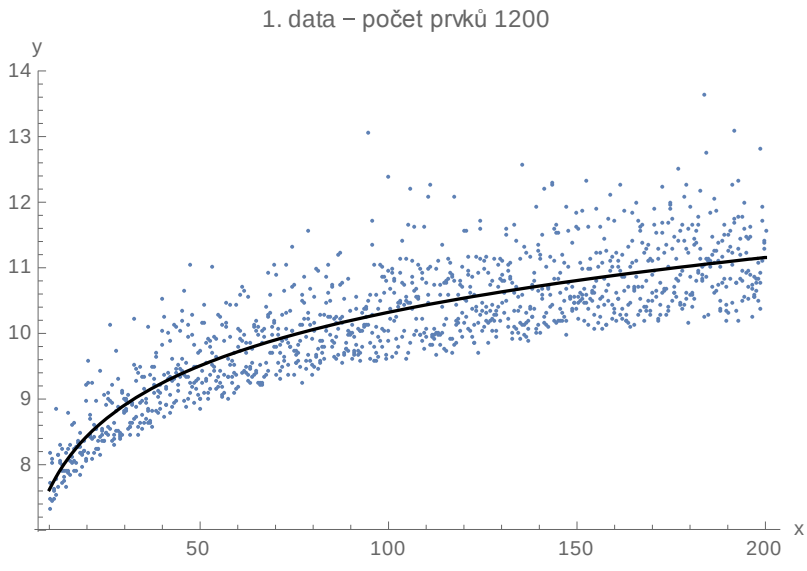
Příklad

Na následujícím obrázku jsou zobrazena data použitá pro náš příklad. Mají zhruba logaritmickou tendenci.



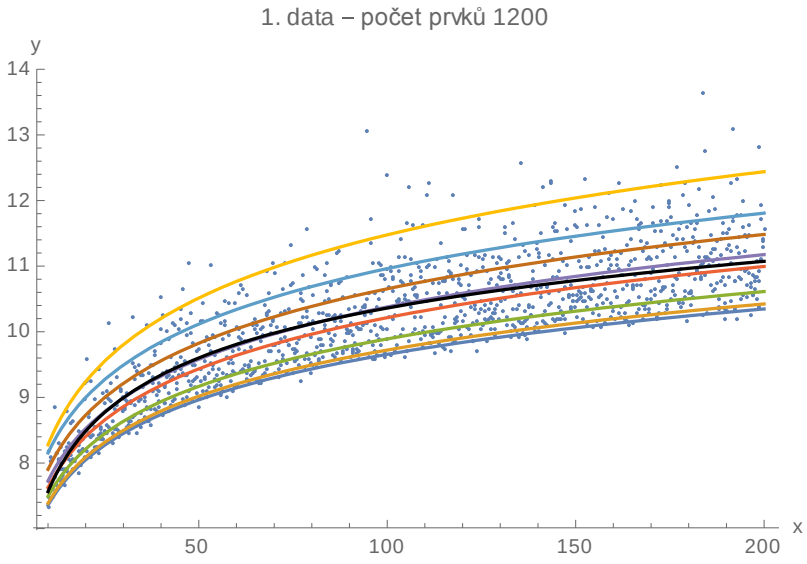
Graf 42 - Rozložení dat

Nejdříve proložíme klasickou lineární regresi.



Lineární regrese

Graf 43 - Lineární regrese a data



Kvantilová a lineární regrese

Graf 44 - Data a kvantilová a lineární regrese

Základní množina funkcí, pomocí nichž budeme vyjadřovat jednotlivé regresní odhady: $1, x, \sqrt{x}, \text{Log}[x]$

Jeden z kvantilových odhadů pro 5%:

$$5.00461345265681 + 1.253495511986674 \times 10^{-8}\sqrt{x} + 0.0001786470235543353x + 1.008294618092901\text{Log}[x]$$

Klasická lineární regrese:

$$4.894539552826588 - 0.03140589557964107\sqrt{x} + 0.0011971037942532525x + 1.2206786125643387\text{Log}[x]$$

Porovnáním snadno zjistíme, že se obě vyjádření velmi podobají, liší se v části lineární a dále v části s $\sqrt{x}$. Podobné výsledky bychom zjistili i pro ostatní kvantily.

Uvedeme dále tabulku, z níž bude zřejmé, jak přesná je kvantilová regrese:

kvantily	**počet prvků pod křivkou v procentech**
0.05	95.0833 %
0.1	90. %
0.25	74.8333 %
0.5	50. %
0.6	40. %
0.75	25. %
0.85	15. %
0.95	4.91667 %

5.5. Zprůměrované regresní kvantily

V této části budou uvedeny nové výsledky. Jurečková, Picek (2014) zavedli zprůměrovaný regresní kvantil

$$\bar{\boldsymbol{B}}_{\boldsymbol{n}}(\boldsymbol{\tau}) = \bar{\boldsymbol{x}}_{\boldsymbol{n}}^{\boldsymbol{T}}\,\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}),\ \ \bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_{nj} \qquad \textbf{(5.27)}$$

a studovali jeho vlastnosti a vztahy k dalším statistikám. Některé vlastnosti $\bar{\boldsymbol{B}}_{\boldsymbol{n}}(\boldsymbol{\tau})$ jsou velmi zajímavé, z pohledu této práce je to především asymptotická ekvivalence s obvyklými kvantily v modelu polohy, kterou uvádíme v následující větě.

Věta 5.7

Předpokládejme, že distribuční funkce F je spojitá a dvakrát diferencovatelná v okolí $F^{-1}(\tau)$ a že $\partial F(F^{-1}(\tau)) = f\left(F^{-1}(\tau)\right) > 0,\ 0 < \tau < 1.$

Potom za podmínek

(A.1) $\lim_{\boldsymbol{n}\to\infty} \boldsymbol{Q}_{\boldsymbol{n}} = \boldsymbol{Q}$, kde $\boldsymbol{Q}_{\boldsymbol{n}} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{X}_{\boldsymbol{n}}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{n}}$ a $\boldsymbol{Q}$ je pozitivně definitní matice. (5.28)

(A.2) $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^4 = O(1),\ n \to \infty, \text{pro } j = 1, \cdots, p.$ (5.29)

(A.3) $x_{i1} = 1,\ \ i = 1, \cdots, n.$ (5.30)

platí

$$n^{1/2}\left[\bar{x}_n^T\left(\hat{\beta}(\tau) - \beta\right) - \varepsilon_{[n\tau]:n}\right] = O_P\left(n^{-\frac{1}{4}}\right), n \to \infty, \qquad (5.31)$$

kde $\varepsilon_{1:n} \le \cdots \le \varepsilon_{n:n}$ jsou pořádkové statistiky odpovídající $\varepsilon_1, \cdots, \varepsilon_n$.

Toto tvrzení nás vede k tomu, že techniky použité v předcházející kapitole by bylo možné aplikovat na regresní kvantily v modelech s rušivou regresí.

Využijeme k tomu postup, který navrhl Drees (1998) a později de Haan a Ferreira (2006). Jeho cílem je nalézt společný teoretický základ pro běžné používané odhady v oblasti teorie extrémních hodnot. Tímto základem je práce s třídou hladkých funkcionálů invariantních vzhledem k poloze a měřítku kvantilového procesu chvostu. Pro tuto třídu Drees (1998) odvozuje asymptotické vlastnosti a zároveň ukazuje, že většina používaných odhadů může být reprezentována těmito funkcionály. Dienstbier (2011) ukázal podobnost kvantilového procesu chvostu v modelu polohy a zprůměrovaného regresně kvantilového procesu chvostu v modelu lineární regrese. Můžeme tedy poměrně snadno výsledky popsané v předcházejících kapitolách rozšířit na analogické odhady, které jsou založené na zprůměrovaných regresních kvantilech, do lineárního regresního modelu.

Zkusme tuto myšlenku popsat. Nejprve definujme váhovou funkci, která je svázána s empirickou kvantilovou funkcí chvostu v 0:

$$\tilde{h}(t) := (t/\widetilde{\log\log}(\tfrac{3}{t}))^{1/2}\ ,\ \text{pro}\ \ t\ \in [0,1] \tag{5.32}$$

a odpovídající vhodný prostor pomocných váhových funkcí:

$$\boldsymbol{\mathcal{H}} := \left\{h\colon [0,1] \to [0,\infty) |\ h\ \in C[0,1], \lim_{t\downarrow 0} \tfrac{h(t)}{\tilde{h}(t)} = 0\ \right\}, \tag{5.33}$$

kde $C[0,1]$ je prostor spojitých funkcí na uzavřeném intervalu $[0,1]$. Pro každé $\gamma \in R$ **a** $\boldsymbol{h}\ \in \boldsymbol{\mathcal{H}}$ definujeme seminormu $\|\cdot\|_{\boldsymbol{\gamma},\boldsymbol{h}}$ na prostoru reálných funkcí na intervalu $[0,1]$ následovně

$$\|z\|_{\gamma,h} := \sup\{t\ \in [0,1];\ t^{\gamma} h(t)|z(t)|\}. \tag{5.34}$$

Nakonec uvažujme prostor obsahující všechny možné empirické kvantilové funkce chvostu a jejich teoretické protějšky, což je prostor reálných funkcí na jednotkovém intervalu se seminormou $\|\cdot\|_{\boldsymbol{\gamma},\boldsymbol{h}}$:

$$D_{\gamma,h} := \left\{z\colon [0,1] \to \mathbb{R};\ \lim_{t\downarrow 0} t^{\gamma} h(t) z(t) = 0, \left(t^{\gamma} h(t) z(t)\right)_{t\ \in\ [0,1]} \in D[0,1]\right\} \tag{5.35}$$

Každý smysluplný odhad $\boldsymbol{\gamma}$ je funkcí horních pořádkových statistik a může být i popsán pomocí členů empirické kvantilové funkce chvostu $Q_{n,k}(t) := F_n^{-1}\left(1 - \frac{k_n}{n} t\right) = X_{n-[k_n t]:n}$. Tedy každý odhad může být uvažován jako nějaký funkcionál aplikovaný na empirickou kvantilovou funkci chvostu $T\left(Q_{n,k}(t)\right)$.

Aby bylo možné odvodit konzistenci odhadů a další vlastnosti uvažovaných odhadů budeme uvažovat třídu funkcionálů splňující následující podmínky

(T.1) $T_{|D_\gamma}$ je borelovsky měřitelný, (5.36)

(T.2) $T(az+b) = T(z)$ pro vš. $z\ \in\ D_{\gamma,h}, a\ >0, b\ \in R$, (5.37)

(T.3) $T(z_\gamma) = \gamma$ (5.38)

Jako příklad funkcionálu uvedených vlastností můžeme uvést funkcionál generující Pickandův odhad

$$T_{Pick}(z) = \frac{1}{log2}\ log\left(\frac{z\left(\frac{1}{4}\right) - z\left(\frac{1}{2}\right)}{z\left(\frac{1}{2}\right) - z(1)}\right) I\left[\frac{z\left(\frac{1}{4}\right) - z\left(\frac{1}{2}\right)}{z\left(\frac{1}{2}\right) - z(1)} > 0\right]. \tag{5.39}$$

Abychom mohli využít výsledky odvozené Dienstbierem (2011) v modelu lineární regrese, budeme předpokládat následující podmínky regresní matice $\boldsymbol{X}$ a distribuční funkce chyb *F:*

(F.1) F je absolutně spojitá s hustotou f kladnou na $(\boldsymbol{x}_*, \boldsymbol{x}^*)$. Derivace f' je ohraničená a druhá derivace f'' existuje na nějakém levém okolí $\boldsymbol{x}^*$. (5.40)

(F.2) $sup\left\{x_* < x < x^*; F(x)(1-F(x))\left|\frac{f^!(x)}{f^2(x)}\right| \leq K_\gamma\right\}$, pro nějaké K_γ kladné. (5.41)

(F.3) $$\lim_{x\to x^*} \frac{(1-F(x))f^!(x)}{f^2(x)} = -1-\gamma^* \quad (5.42)$$

a

$$\lim_{x\to x_*} \frac{(1-F(x))f^!(x)}{f^2(x)} = -1-\gamma_*, \quad (5.43)$$

pro nějaké $\gamma_*, \gamma^* \in R$.

(F.4) $\gamma := \min\{\gamma_*, \gamma^*\} > -1/2.$ (5.44)

(X.1) $x_{i1} = 1,\ i = 1, \cdots, n.$

(X.2) $\lim_{\boldsymbol{n}\to\infty} \boldsymbol{D_n} = \boldsymbol{D}$, kde $\boldsymbol{D_n} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{X_n^T X_n}$ a $\boldsymbol{D}$ je pozitivně definitní matice. (5.45)

(X.3) $\frac{1}{n}\sum_{i=1}^n \left\|x_i^4\right\| = O(1),\ n \to \infty.$

(X.4) $\max_{1\leq i\leq n} \|x_i\| = O(C_n^\Delta),\ n \to \infty$ pro nějaké $\Delta > 0,\ C_n = C(\log\log n)^{\frac{1}{2}}$, (5.46)

$$0 < C < \infty .$$

Dienstbier (2011) ukázal podobnost kvantilového procesu chvostu v modelu polohy a zprůměrovaného regresně kvantilového procesu chvostu v modelu lineární regrese.

Věta 5.8

V modelu lineární regrese (5.2) předpokládejme platnost podmínek (F.1)-(F.4.), (X.1) – (X.4). Předpokládejme také, že distribuční funkce *F* splňuje podmínku druhého řádu

$$\lim_{t\to\infty} \frac{\frac{U(xt)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = \begin{cases} \frac{1}{\gamma}\left(x^\gamma log x - \frac{x^\gamma - 1}{\gamma}\right) & , \rho = 0 \neq \gamma \\ \frac{1}{\rho}\left(\frac{x^\rho - 1}{\rho} - log x\right) & , \rho \neq 0 = \gamma, \\ \frac{1}{2}(log\ x)^2 & \rho = 0 = \gamma. \end{cases}$$

s kvantilovou funkcí chvostu $U(t) = inf\left\{y; F(y) \geq 1 - \frac{1}{t}\right\}, t \geq 1$. Potom můžeme definovat Wienerovy procesy $\{W_n(s)\}, s > 0$ tak, že pro vhodně vybranné funkce A, a, z_γ a

prostor $\boldsymbol{D_{\gamma,h}}$ se seminormou $\|z\|_{\gamma,h} \coloneqq \sup\{t \in [0,1];\ t^{\gamma}h(t)|z(t)|\}$ definovanou pro libovolné $\varepsilon > 0$ na prostoru váhových funkcí

$$\boldsymbol{\mathcal{H}} \coloneqq \left\{ h\colon [0,1] \to [0,\infty) |\ h \in C[0,1], \lim_{t\downarrow 0} \frac{h(t)}{\tilde{h}(t)} = 0 \right\}$$

funkcí $h(t) = t^{\frac{1}{2}+\varepsilon}, t \in [0,1]$

platí

$$\left\| k^{1/2} \left(\frac{\bar{x}_n^T(\hat{\beta}(\tau)-\beta)}{F^{-1}\left(1-\frac{k}{n}\right)} - t^{-\gamma} \right) - \gamma\, t^{-\gamma-1} W_n(t) - \sqrt{k}\, A\left(k/n\right) t^{-\gamma} \frac{t^{\rho}-1}{\rho} \right\|_{\gamma,h} \leq$$

$$\left\| \gamma\, t^{-\gamma} \left(\frac{n}{k}\right)^{\frac{1}{2}} \bar{x}_n^T\, \boldsymbol{D_n^{-1}} \boldsymbol{Z_n} \left(1 - \frac{tk}{n}\right) \right\|_{\boldsymbol{\gamma,h}} + o_P(1), \tag{5.47}$$

$n \to \infty, k = k(n) \to \infty, \frac{k}{n} \to 0,\ \sqrt{k}\, A\left(k/n\right) = O(1)$ a $k \geq \log \Delta(1 \vee \gamma)$ s $\Delta > 1/6$,

kde $Z_n(\tau) = n^{-1/2}\big(\tau(1-\tau)\big)^{-1/2} \sum_{i=1}^{n} x_i \varphi_\tau\big(\epsilon_i - F^{-1}(\tau)\big)$ a $\varphi_\tau(z) = \tau - I[z < 0]$.

Důkaz: proveden v Dienstbier (2011).

Ačkoliv Věta 5.8 neposkytuje stejný výsledek (pouze horní hranici v pravděpodobnosti) jako analogická věta Dreese (1998), lze ji analogicky využít pro odvození některých vlastností odhadů v regresi (jako např. konzistence) pomocí hladkých funkcionálů zprůměrovaných regresních kvantilů aplikovaných regresně kvantilový proces chvostu. Detaily lze vyhledat v Dienstbier (2011). Řadu vlastností odhadů v modelu polohy tak lze úspěšně přenést do lineárního regresního modelu. V dalším textu se zaměříme pouze na Pickandsův odhad jako vhodnou ilustraci problematiky odhadu v lineární regresi.

5.5.1. Regresní Pickandsův odhad a optimal sample fraction

V lineárním regresním modelu Pickandsův odhad můžeme definován podobně jako v (2.36):

$$\hat{\gamma}_{P,RQ}(n,k) = \frac{1}{log 2}\, log\left(\frac{\bar{x}_n^T \widehat{\beta_n}(\tau_{m-\left[\frac{k}{4}\right]}) - \bar{x}_n^T \widehat{\beta_n}(\tau_{m-\left[\frac{k}{2}\right]})}{\bar{x}_n^T \widehat{\beta_n}(\tau_{m-\left[\frac{k}{2}\right]}) - \bar{x}_n^T \widehat{\beta_n}(\tau_{m-k})} \right), \tag{5.48}$$

kde $0 < \tau_{1<\cdots<}\tau_m < 1$ odpovídá $m_n(\boldsymbol{Y},\boldsymbol{X}) = m_n$ řešením $\hat{\beta}(\tau_i), i = 1,\cdots,m_n$ minimalizačního problému (5.10). Tedy regresně kvantilová funkce chvostu

$$\hat{Q}_{\bar{x},n,k}(t) \coloneqq \left\{ \bar{x}_n^T \left(\hat{\beta}\left(1 - \frac{kt}{n}\right)\right)\right\}_{t\in[0,1]} \tag{5.49}$$

je skokovitá funkce.

Věta 5.9

Za předpokladů věty 5.8 je $\hat{\gamma}_{P,RQ}(n,k)$ konzistentním odhadem γ.

Důkaz:

Nechť $Z_n(\tau) = n^{-1/2}\big(\tau(1-\tau)\big)^{-1/2}\sum_{i=1}^{n} x_i\varphi_\tau\big(\epsilon_i - F^{-1}(\tau)\big)$ a $\varphi_\tau(z) = \tau - I[z < 0]$, viz Věta 5.8,

Podle Lemmatu 2.3.5 v Diensbier (2011) platí, že $\left(\frac{n}{k}\right)^{1/2} D_n^{-1}Z_n \to D_n^{-1}W$ slabě na $D[0,1]^p$, kde W je Wienerovův proces. Tedy $\left(\frac{n}{k}\right)^{1/2} \bar{x}_n^T D_n^{-1}Z_n \to \bar{x}_n^T D_n^{-1}W$ slabě na $D[0,1]$, z čehož plyne $t^{-\gamma}\left(\frac{n}{k}\right)^{1/2} \bar{x}_n^T D_n^{-1}Z_n \to t^{-\gamma}\bar{x}_n^T D_n^{-1}W$ na $D_{\gamma,h}$ a tedy

$$\gamma t^{-\gamma}\left(\frac{n}{k}\right)^{1/2} \bar{x}_n^T D_n^{-1}Z_n\left(1-\frac{kt}{n}\right) \to t^{-\gamma}\bar{x}_n^T D_n^{-1}W(t)$$

Z věty 5.8 pak plyne, že

$$\frac{\hat{Q}_{\bar{x},n,k} - \bar{x}_n^T\beta}{F^{-1}\left(1-\frac{k}{n}\right)} \xrightarrow{P} z_\gamma + 1/\gamma, \;\; n \to \infty,$$

slabě na $\boldsymbol{D_{\gamma,h}}$. Funkcionál

$$T_{Pick}(z) = \frac{1}{log2}\, log\left(\frac{z\left(\frac{1}{4}\right) - z\left(\frac{1}{2}\right)}{z\left(\frac{1}{2}\right) - z(1)}\right) I\left[\frac{z\left(\frac{1}{4}\right) - z\left(\frac{1}{2}\right)}{z\left(\frac{1}{2}\right) - z(1)} > 0\right].$$

je spojitý v $\boldsymbol{z_\gamma}$, invariantní vzhledem k poloze a měřítku, z čehož už trvzení plyne.

Užitím Hadamardovi diferencovatelnosti pak můžeme věty 5.8 využít k odvození asymptotické normality $\hat{\gamma}_{P,RQ}(n,k)$.

Odhad $\hat{\gamma}_{P,RQ}(n,k)$ je identický pro nějaké $\tilde{k}$ s odhadem

$$\hat{\gamma}_{P,RQ}(n,\tilde{k}) = \frac{1}{log2}\, log\left(\frac{\bar{x}_n^T\widehat{\beta_n}\left(1-\frac{\tilde{k}}{4n}\right) - \bar{x}_n^T\widehat{\beta_n}\left(1-\frac{\tilde{k}}{2n}\right)}{\bar{x}_n^T\widehat{\beta_n}\left(1-\frac{\tilde{k}}{2n}\right) - \bar{x}_n^T\widehat{\beta_n}\left(1-\frac{\tilde{k}}{n}\right)}\right),$$

Odhady se liší je ve své parametrizaci „prostřední“ posloupnosti.

Věta 5.9 nám dává i oporu pro formulaci obdobných vět týkající Pickandsova odhadu a hledání optimální hodnoty parametru k z kapitoly 4 nyní pro Pickandsův odhad založený na zprůměrovaných regresních kvantilech:.

Například uvedeme větu týkající se optimální hodnoty $k_0(m_n)$:

Věta 5.10

Nechť $F \in D_{dif}(G_\gamma)$ a jsou splněny podmínky druhého řádu pro hodnotu $A(x) = c\,x^\rho$, kde $c \neq 0$ a $\rho < 0$. Označme index $k_0(m_n)$ takový, že hodnota $E\left(\hat{\gamma}_{P,RQ}(m_n,k) - \gamma\right)^2$ je minimální. Potom

$$\lim_{n\to\infty} k_0(m_n) \Bigg/ \left\{ \left(\frac{1+2^{-2\gamma-1}}{4\rho c^2 \left(\frac{1-2^\rho}{\rho}\right)^2 \left(\frac{2^{-\gamma-\rho}-1}{\gamma+\rho}\right)^2 2^{-2\rho}} \right)^{\frac{1}{1-2\rho}} n^{\frac{-2\rho}{1-2\rho}} \right\} = 1.$$

Podobným způsobem lze v tomto případě budovat **bootstrapovou proceduru:**

1. krok. Položme $n_1 = [m_n^{1-\varepsilon}]$ pro hodnotu $\varepsilon \in (0;\frac{1}{2})$, kde $[x]$ označuje celou část čísla *x*. Vyberme z výběru o m_n hodnotách nový bootstrapový výběr o délce n_1. Budeme dále počítat $E\left(\left(\hat{\gamma}^*_{P,RQ}(n_1,k_1) - \hat{\gamma}^*_{P,RQ}(n_1,4k_1)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$. Nalezneme hodnotu $k^*_{1,0}(n_1)$, pro který je předchozí výraz minimální.
2. krok: Položíme $n_2 = \left[{n_1}^2/_{m_n}\right]$ a zopakujeme krok I. s tím, že nalezneme hodnotu $k^*_{2,0}(n_2)$, pro kterou je $E\left(\left(\hat{\gamma}^*_{P,RQ}(n_2,k_2) - \hat{\gamma}^*_{P,RQ}(n_2,4k_2)\right)^2 \middle| \boldsymbol{\mathcal{X}_n}\right)$ minimální.
3. krok: Nyní vypočítáme odhad $k_0(m_n)$,
4. $k_0(m_n) = \dfrac{\left(k^*_{1,0}(n_1)\right)^2}{\left(k^*_{2,0}(n_2)\right)^2 f\left(\frac{\log k^*_{1,0}(n_1)}{2\left(\log k^*_{1,0}(n_1) - log(n_1)\right)}\right)}$

5.6. Simulační studie – kvantilová regrese

5.6.1. Postup simulace

V následujících stránkách je popsána situace, kdy máme k dispozici dvě náhodné proměnné x_1 a x_2 a lineární regresní model $y = a\,x_1 + b\,x_2 + c + e$. Hodnoty a, b, c jsme zvolili pro jednoduchost postupně $a = 3, b = 4, c = 1$. Hodnoty x_1 jsou brány z rovnoměrného rozdělení na intervalu (0;2) a hodnoty x_2 jsou brány z rovnoměrného rozdělení na intervalu (-8;8). Chyby e jsou generovány specifickým rozdělením z Frèchetovy sféry přitažlivosti. Jednotlivé výsledky jsou šetřeny v etapách takto:

1. etapa simulace hodnot x_1, x_2 a y pro jednotlivé případy kontaminace :
   a. Burr(1,1,1)
   b. Cauchy(0,1)
   c. F rozdělení(6,4)
   d. Pareto(1,1)

   Pro každý z těchto typů je vždy vygenerováno postupně 100, 200, 500 a 1000 hodnot, které jsou kromě běžného zápisu také zobrazeny graficky. Navíc je vždy graficky zobrazena hodnota absolutního členu s kontaminovanými hodnotami.

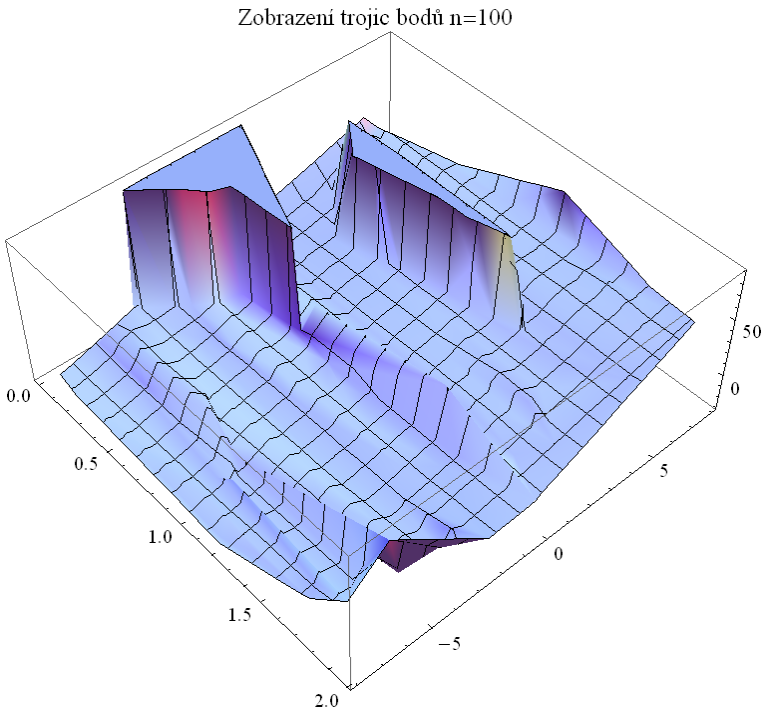
2. etapa zpracování kvantilové regrese pomocí Koenkerových algoritmů v programu R a v programu Mathematica. Výsledkem je pro každý výše uvedený případ výstup 100, 200, 500 nebo 1000 kvantilů, které jsme získali z kvantilové regrese jako popis absolutního členu. Tyto výsledky jsou zobrazeny graficky. Je z nich zřejmé, že poslední kvantily jsou v mnoha případech proti ostatním hodnotám mnohonásobně větší.
3. etapa budeme předpokládat, že chceme najít odhad parametru γ rozdělení kontaminace absolutního členu. K tomu použijeme odhad, který je location/scale invariant. Použijeme odhad Pickandsův.

5.6.2. Simulace hodnot.

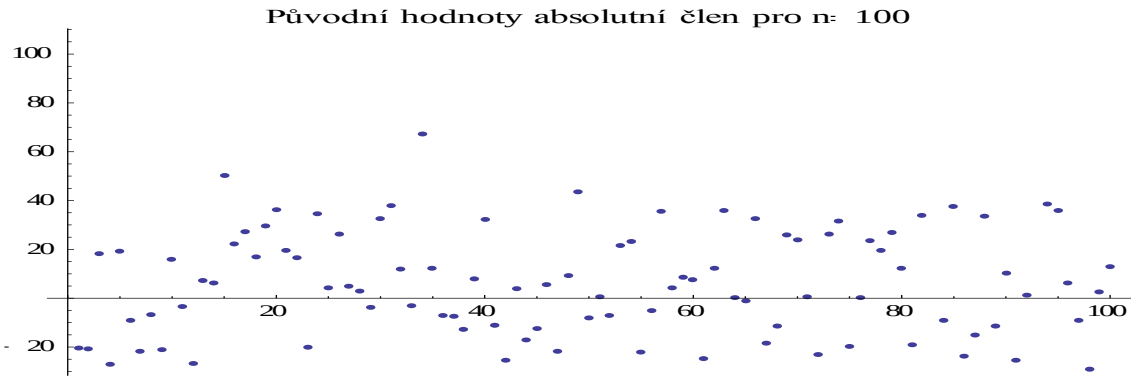
Vzhledem k velkému množství získaných údajů uvedeme v této části jen simulované hodnoty z rozdělění Burr(1,1,1).

Burr(1,1,1) – Podobně jako u ostatních jsme postupně volili simulaci 100, 250, 500 a 1000 hodnot. Grafické znázornění hodnot y a absolutního členu je provedeno zde:

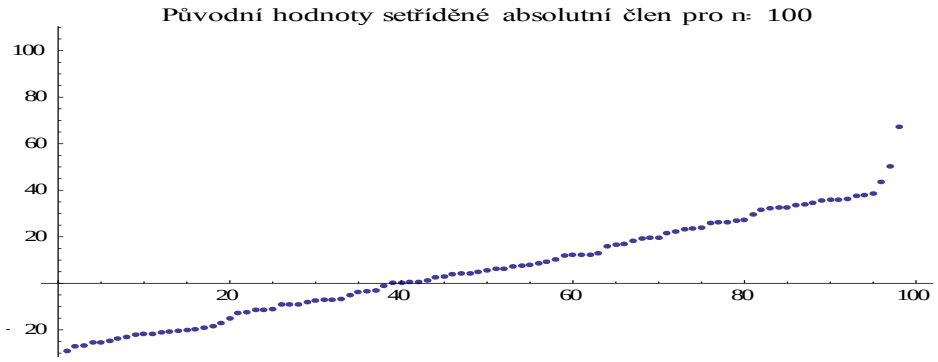
100 hodnot:



Graf 45 - Trojice bodů Burr - 100 bodů



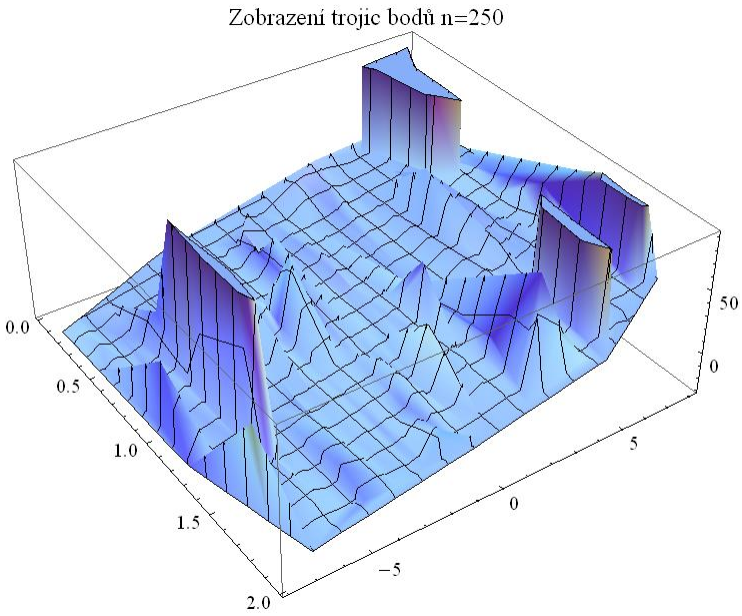
Graf 46 - Původní hodnoty c - Burr, 100 bodů



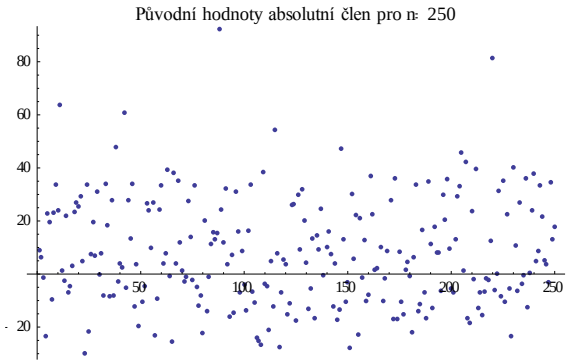
Graf 47 - Setříděné hodnoty c - Burr, 100 hodnot

Z prvního obrázku Graf 45 je patrno, že rozdělení Burr(1,1,1), značně mění hodnoty výsledku y. Na obrázku Graf 47 je zřejmé, že medián hodnot absolutního členu je skutečně roven jedné.

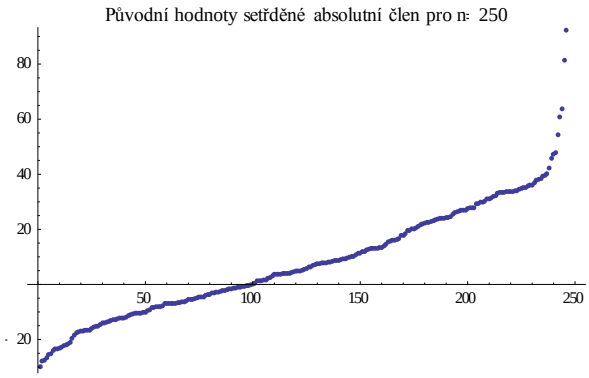
250 hodnot:



Graf 48 - Trojice hodnot-Burr,250 hodnot



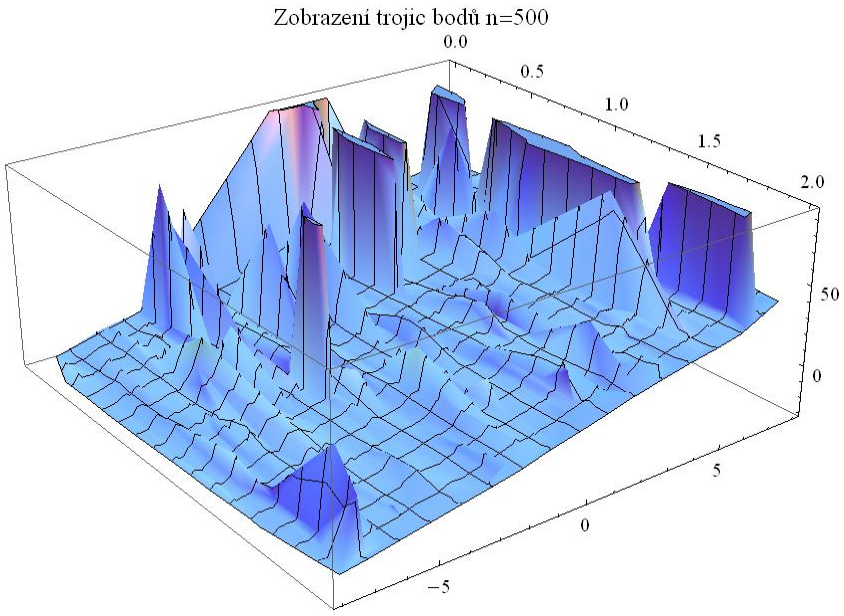
Graf 49 - Absolutní člen, Burr, n=250



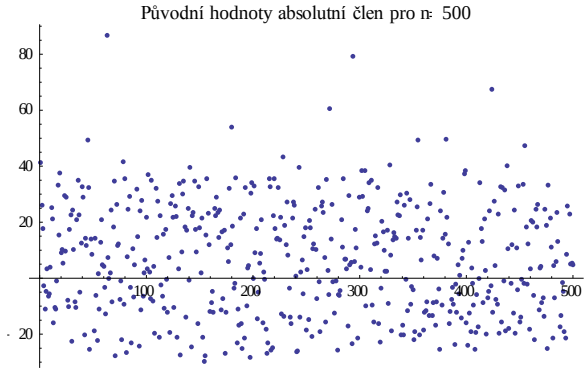
Graf 50 - Setříděný abs. člen, Burr, n=250

Na prvním obrázku jsou zřejmé velké změny, které vytváří rozdělení Burr. Změny jsou jak pozitivní (maxima), tak i negativní (minima). Podobně jako v předchozím je medián absolutních hodnot (intercept) roven přibližně 1.

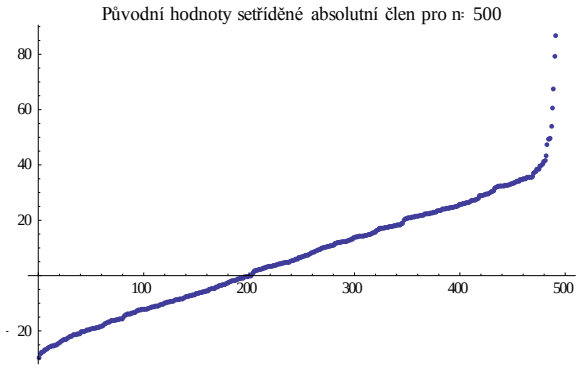
500 hodnot:



Graf 51 - Zobrazení trojic-Burr,500 hodnot



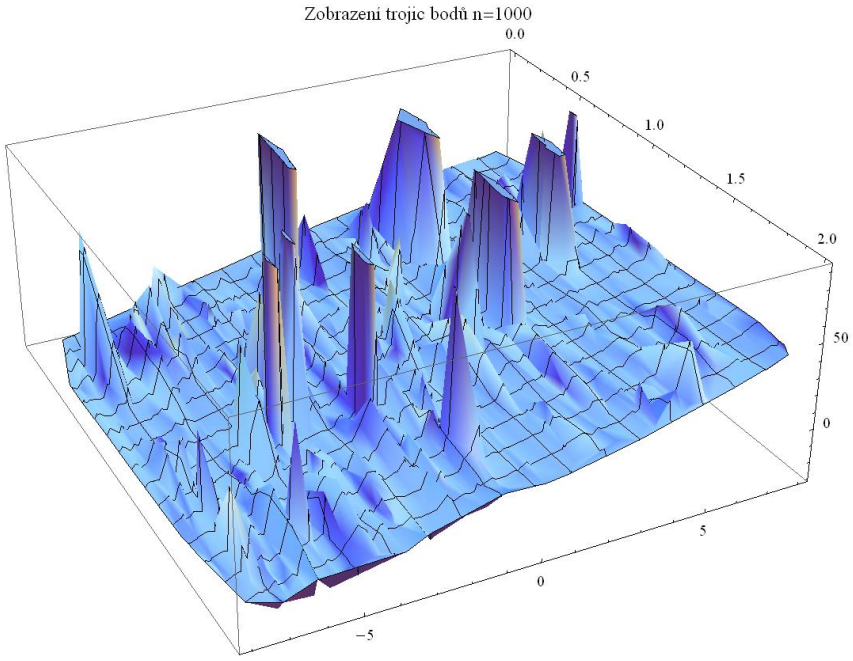
Graf 52 - Abs. člen nesetříděný, Burr, n=500



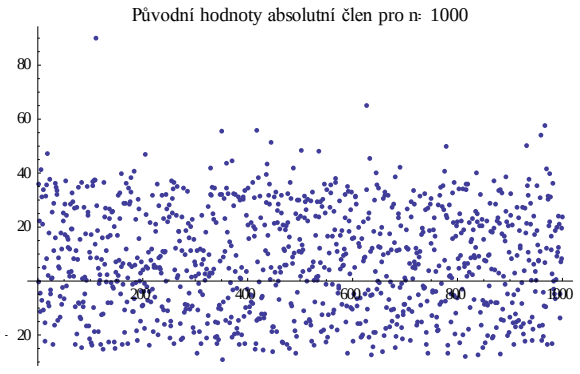
Graf 53 - Setříděné hodnoty abs. členu, Burr, n=500

Nyní je patrno, že hodnoty y jsou mnohem více ovlivněny příměsí – rozdělením Burr. V obrázku Graf 51 je vidět mnoho maxim a minim, která leží mimo předpokládanou oblast roviny. Medián hodnot absolutního členu je opět správně přibližně roven 1.

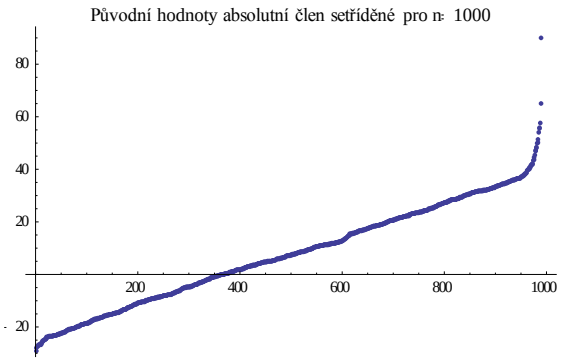
1000 hodnot:



Graf 54 - Trojice hodnot,Burr,1 000 hodnot



Graf 55 - Nesetříděné hodnoty abs. člen, Burr, n=1000



Graf 56 - Setříděné hodnoty abs. člen, Burr, n=1000

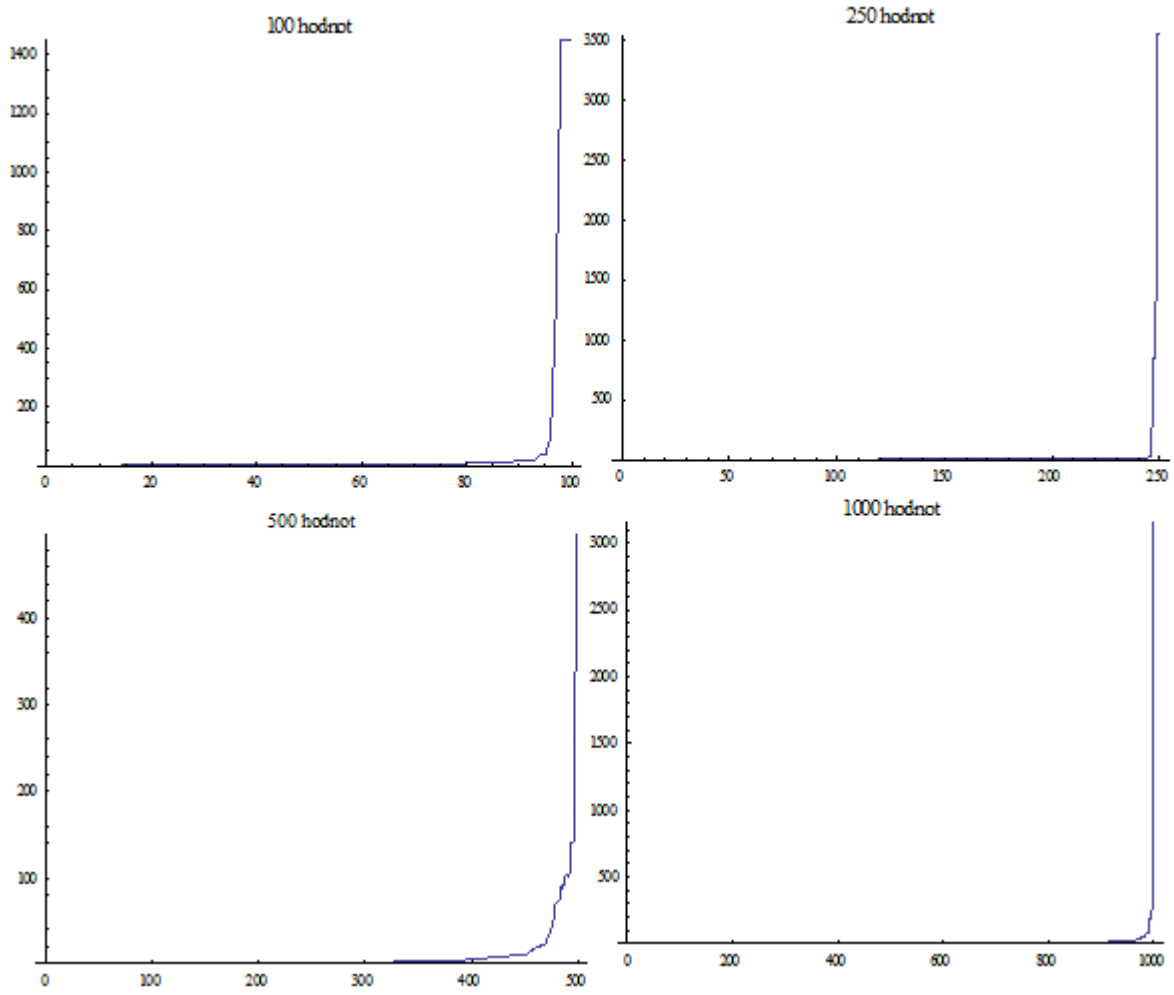
Na prvním obrázku Graf 58 jsou patrná maxima i minima vzniklá z důvodu příměsi Burr. Podobně jako v předchozích případech je medián hodnoty absolutního členu roven přibližně 1.

Pro ostatní náhodné veličiny uvedené v 1. etapě jsou grafické výsledky podobné.

5.6.3. Provedení Koenkerových algoritmů

V této části jsou výše uvedená data zpracována pomocí algoritmů Koenkera v programu R a programu Mathematica. Výsledkem jsou kvantily pro členy a, b, c. Nás bude zajímat nejvíce absolutní člen c. Výstup programu je upraven tak, abychom získali právě stejný počet kvantilů jako je počet vstupních hodnot (kvantily jsou počítány pravidelně jako k*(1/počet)). V následujícím textu jsou zobrazeny výsledky těchto kvantilů pro jednotlivá rozdělení a různé velikosti nasimulovaných hodnot.

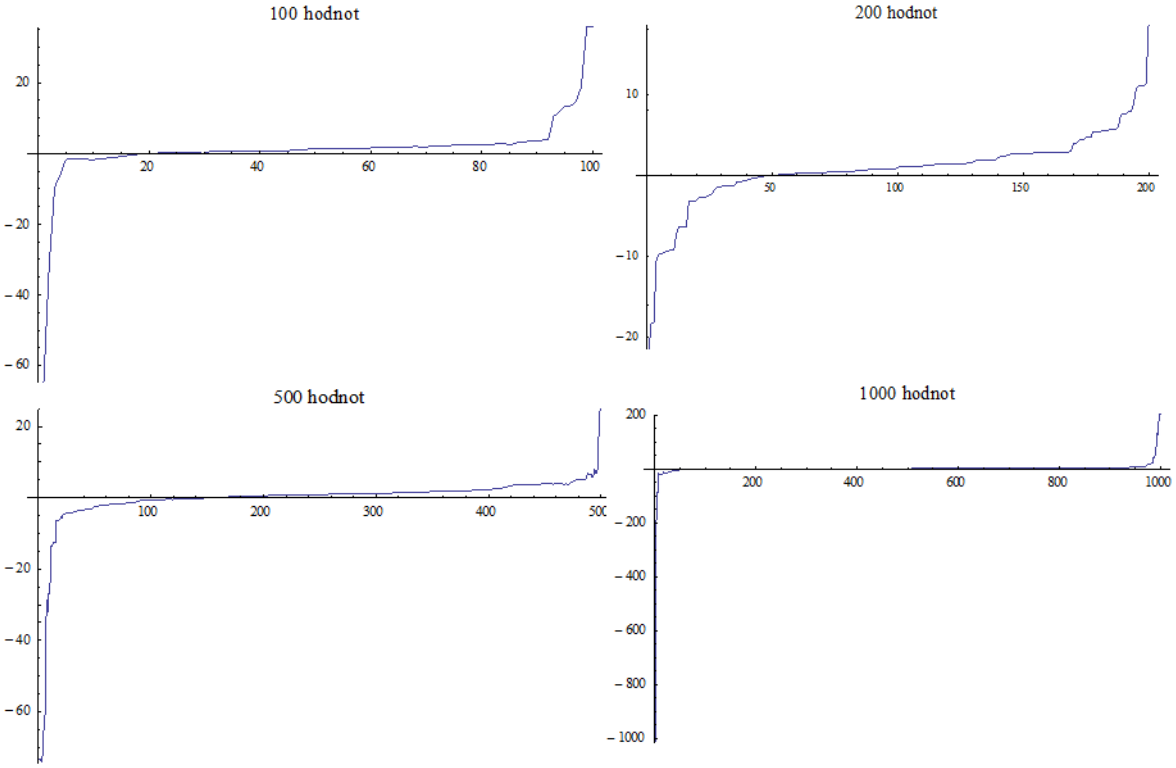
a. **Burr(1,1,1).** Jednotlivé kvantily jsou vytvářeny podle počtu hodnot v základním souboru tedy 100, 200(250), 500 nebo 1000 položek (tedy i kvantilů).



Graf 57 - Hodnoty nejvyšších kvantilů pro rozdělení Burr(1,1,1)

Z grafu pro $n = 100$, je zřejmé, že se kvantily velmi výrazně mění pro kvantily větší než 0,8. Graf pro $n = 250$ ukazuje výraznější změnu hodnot kvantilu pro hodnoty od 0,7. Pro n=500 je výraznější změna patrná od hodnot kvantilů 0,8. Konečně pro $n = 1000$ se kvantily výrazněji mění od hodnot 0,9. Pro $n = 100$ jsou nejvyšší kvantily rovny cca 1400, pro hodnotu $n = 250$ jsou velikosti cca 3500 podobně jako pro $n = 1000$. Hodnota $n = 500$ má nejvyšší kvantily výrazně menší cca 500.

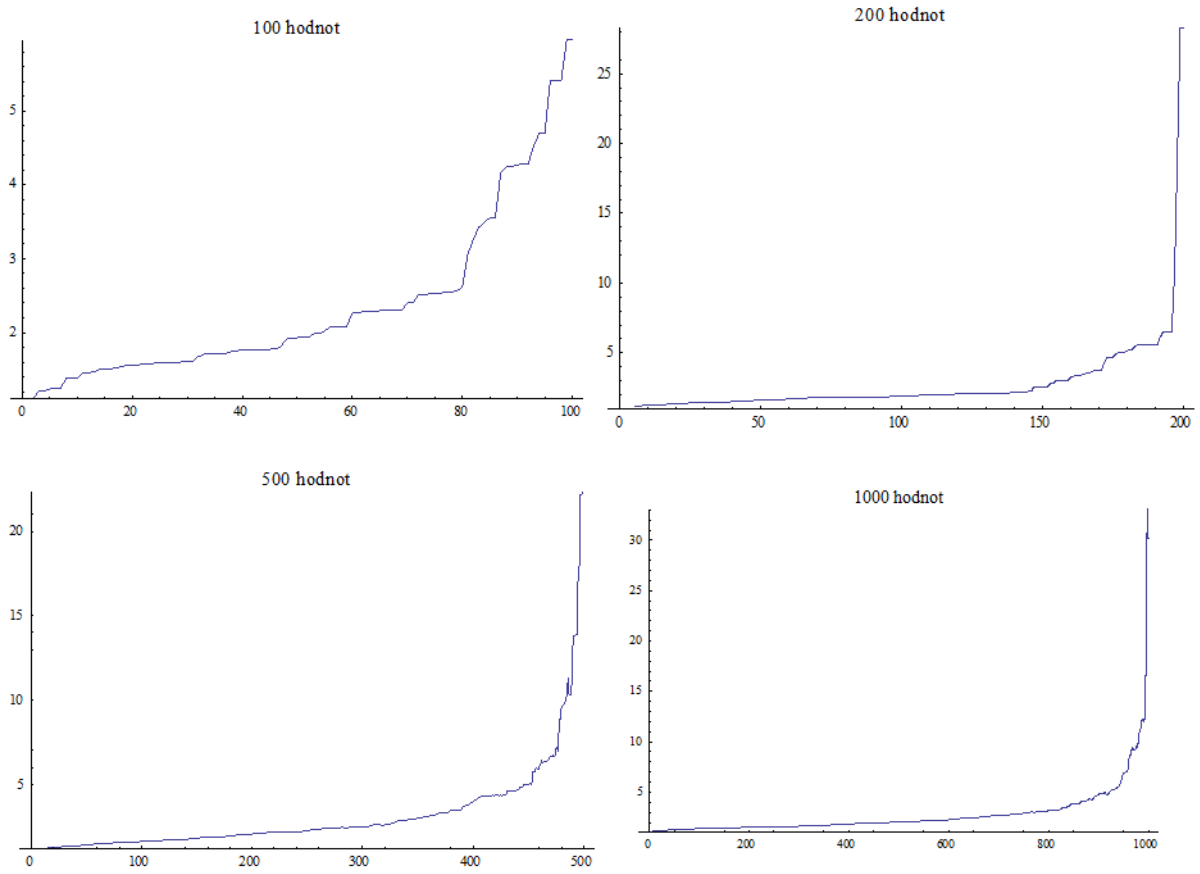
b. **Cauchy(0,1)**. Jednotlivé kvantily jsou vytvářeny podle počtu hodnot v základním souboru tedy 100, 200(250), 500 nebo 1000 položek (tedy i kvantilů).



Graf 58 - Nejvyšší kvantily pro rozdělení Cauchy(0,1)

Z výše uvedených grafů je zřejmé, že hodnoty kvantilů se nejvíce mění od 0 do 0,1 a poté pro hodnoty větší než 0,8 – 0,9. Hodnota nejvyšších kvantilů je pro $n = 100, n = 200$ a $n = 500$ rovna cca 20, naopak nejnižší kvantily jsou hodnotově na úrovni -60. Pro $n = 1000$ jsou nejvyšší kvantily o hodnotě kolem 200 a nejnižší kvantily mají hodnotu od -1000.

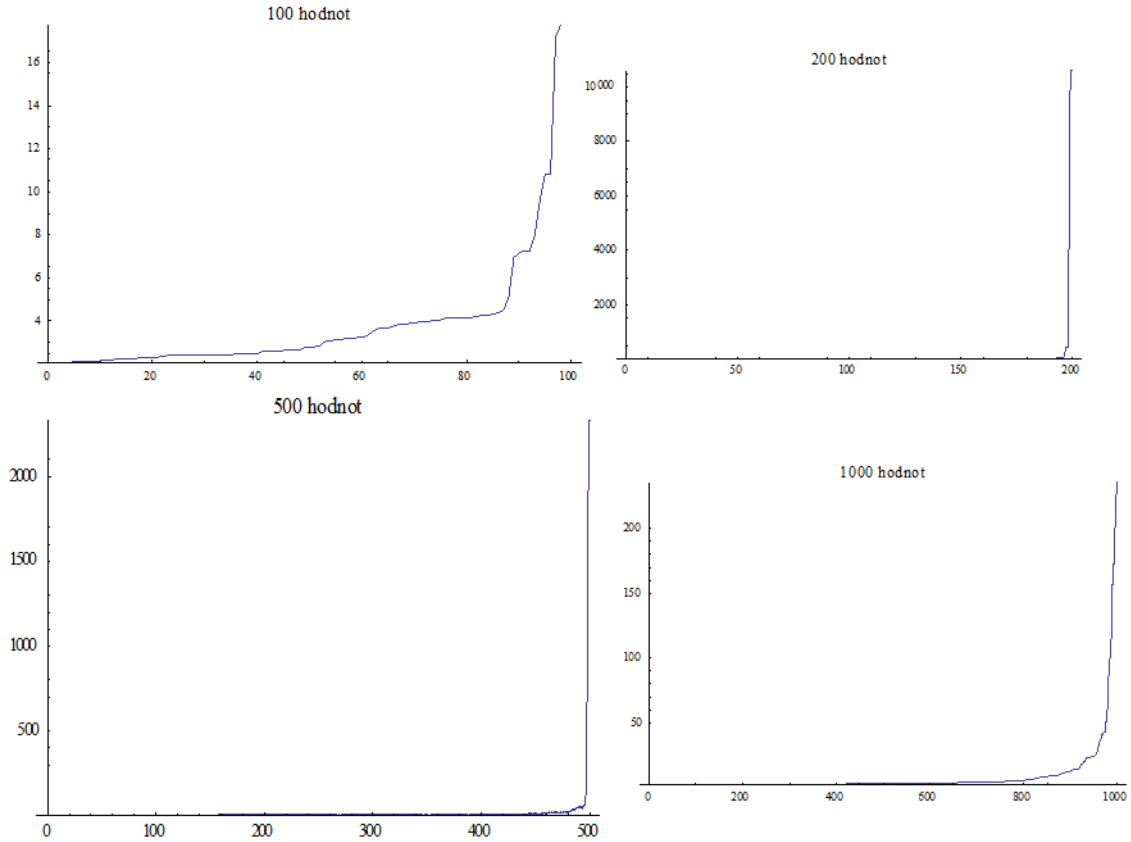
c. **F rozdělení** s parametry 6 , 4. Jednotlivé kvantily jsou vytvářeny podle počtu hodnot v základním souboru tedy 100, 200(250), 500 nebo 1000 položek (tedy i kvantilů).



Graf 59 - Nejvyšší kvantily pro rozdělení F s parametry 6,4

U tohoto rozdělení jsou vidět největší změny hodnot kvantilů pro kvantily větší než 85%. Hodnotově jsou nejvyšší kvantily na úrovni 20 – 30 (s výjimkou n=100, kdy jsou nejvyšší kvantily hodnotově na úrovni 5).

d. **Pareto(1,1).** Jednotlivé kvantily jsou vytvářeny podle počtu hodnot v základním souboru tedy 100, 200(250), 500 nebo 1000 položek (tedy i kvantilů).



Graf 60 - Nejvyšší kvantily rozdělení Pareto(1,1)

U tohoto rozdělení dochází k největším změnám kvantilů od 85%. Hodnotově je mezi danými simulacemi výrazný rozdíl. Velikosti nejvyšších kvantilů pro *n*=100 jsou na úrovni 20. Pro *n*=1000 na úrovni 250. Pro hodnotu *n*=500 dosahují nejvyšší kvantily hodnot cca 2500, konečně pro hodnotu n=200 dokonce hodnotu cca 10000.

5.6.4. Výpočet odhadu parametru γ pomocí Pickandsova odhadu

a. Burr(1;1;1). Skutečná hodnota γ = 1.

Výsledky jsou pro jednotlivé datové soubory rovny:

Tabulka 1 - souhrnné výsledky pro rozdělení Burr(1,1,1)

Počet hodnot v simulaci	Odhad parametru γ	Bootstrap – počet opakování	Volba parametru metody sample fraction	Parametr bootstrapu – sample fraction
100	4,43318	250	0,15	1/50
250	0,854229	250	0,1	39/250
500	1,15672	250	0,15	17/250
1000	1,04059	250	0,15	9/100

Při použití metody optimal sample fraction jsme užili doporučené hodnoty - opakování 250x a hodnotu počátečního indexu na úrovni 0,1 – 0,15. Výpočet jsme s danými hodnotami vždy několikrát provedli a výsledky jsou zobrazeny v tabulce. V tomto případě je procento užitých posledních kvantilů na úrovni cca posledních 10%. Vypočtená hodnota je skutečně přibližně rovna skutečnému indexu γ.

b. Cauchy(0;1). Skutečná hodnota $\gamma = 1$

Výsledky jsou pro jednotlivé datové soubory rovny:

Tabulka 2 - souhrnné výsledky pro rozdělení Cauchy(0,1)

Počet hodnot v simulaci	Odhad parametru γ	Bootstrap – počet opakování	Volba parametru metody sample fraction	Parametr bootstrapu – sample fraction
100	0,03675	250	0,15	7/100
200	0,46903	250	0,13	4/25
500	0,610495	250	0,08	61/500
1000	1,135	250	0,15	17/1000

Při použití metody optimal sample fiction jsme užili doporučené hodnoty - opakování 250x a hodnotu počátečního indexu na úrovni 0,08 – 0,15. Výpočet jsme s danými hodnotami vždy několikrát provedli a výsledky jsou zobrazeny v tabulce. V tomto případě je procento užitých posledních kvantilů na úrovni posledních cca 2% - 10%. Vypočtená hodnota je skutečně přibližně rovna skutečnému indexu γ pro vysokou hodnotu n=1000.

c. F rozdělení s parametry (6;4). Skutečná hodnota $\gamma = 1/8$.

Výsledky jsou pro jednotlivé datové soubory rovny:

Tabulka 3 - souhrnné výsledky pro F-rozdělení F(6,4)

Počet hodnot v simulaci	Odhad parametru γ	Bootstrap – počet opakování	Volba parametru metody sample fraction	Parametr bootstrapu – sample fraction
100	1,56431	250	0,15	1/25
200	0,46902	250	0,14	2/25
500	0,152998	250	0,11	17/250
1000	0,31082	250	0,13	17/250

Při použití metody sample fraction jsme užili doporučené hodnoty - opakování 250x a hodnotu počátečního indexu na úrovni 0,11 – 0,15. Výpočet jsme s danými hodnotami vždy několikrát provedli a výsledky jsou zobrazeny v tabulce. V tomto případě je procento užitých posledních kvantilů na úrovni cca posledních 2% - 4%. Vypočtená hodnota je přibližně rovna skutečnému indexu γ především pro n=500.

d. Pareto rozdělení s parametrem 1. Skutečná hodnota γ = 1.

Výsledky jsou pro jednotlivé datové soubory rovny:

Tabulka 4 - souhrnné výsledky pro rozdělení Pareto P(1,1)

Počet hodnot v simulaci	Odhad parametru γ	Bootstrap – počet opakování	Volba parametru metody sample fraction	Parametr bootstrapu – sample fraction
100	1,77447	250	0,15	4/25
200	1,65081	250	0,1	1/25
500	1,63674	250	0,08	9/125
1000	0,95715	250	0,15	39/1000

Při použití metody sample fraction jsme užili doporučené hodnoty - opakování 250x a hodnotu počátečního indexu na úrovni 0,08 – 0,15. Výpočet jsme s danými hodnotami vždy několikrát provedli a výsledky jsou zobrazeny v tabulce. V tomto případě je procento užitých posledních kvantilů na úrovni cca posledních 4% - 8%. Vypočtená hodnota je přibližně rovna skutečnému indexu γ především pro poslední n=1000.

5.6.5. Shrnutí výsledků simulace kvantilové regrese

V předložené simulační studii jsme využili platnosti jak vět 5.8 a 5.9, tak i větu 5.10. Všechna tato tvrzení nám umožnila použít klasické postupy vhodné pro využití Pickandsova odhadu absolutního členu v lineárním regresním modelu, v kterém je jeden člen ovlivněn hodnotami rozdělení z Frèchetovy sféry přitažlivosti. Ačkoli máme k dispozici všechny údaje pro jednotlivá vybraná rozdělení, zobrazili jsme v první etapě jen jedno – Burrovo. Je to především díky značné velikosti daných údajů. V druhé etapě jsme již uvedli všechna čtyři rozdělení. Graficky jsme u nich zkoumali polohu nejvyšších kvantilů.

V poslední etapě jsme shrnuly veškeré výsledky do tabulek. V nich jsme pro jednotlivá rozdělení zjistili hodnotu Pickandsova odhadu parametru γ, dále jsme jsme zde uvedli používanou hodnotu ε, která slouží ke stanovení počátečních hodnot n_1, n_2. Závěrem celého postupu při aplikaci metody sample fraction při Pickandsově odhadu je stanovení hodnoty užitých nejvyšších kvantilů – v souladu se simulační studií uvedenou v kapitole 4. je tato hodnota na úrovni posledních 4% - 10%.

Celkově jsme si tedy ověřili, že v případě ovlivnění jednoho členu v lineárním regresním modelu rozděleními z Frèchetovy sféry přitažlivosti, je možné použít jak postupy vyplývající z Koenkerových algoritmů pro stanovení jednotlivých hodnot regresních křivek, tak i postupy sloužící k odhadu parametru EVI neznámého rozdělení. Pickandsův odhad byl použit především proto, že je „*location/scale invariant*“. Tuto vlastnost využíváme právě u lineární regrese a pomocí ní určujeme hodnotu neznámého ovlivnění. Na výsledné hodnoty dále užíváme jak Koenkerovy algoritmy, tak i právě Pickandsův odhad.

6. Závěr

Hlavním cílem této práce bylo prostudovat používané neparametrické postupy založené na pořádkových statistikách a zároveň použít metody založené na bootstrapu.

Celá práce je rozdělena do celkem pěti částí. V první úvodní části práce jsou uvedeny odkazy na mnoho zajímavých prací z různých aplikačních oblastí, které se řeší pomocí teorie extremálních statistik. Zároveň je zde uvedena krátká historie úloh, jejichž řešení vedlo na založení celé nové vědní oblasti v matematické statistice – oblasti zabývající se extremálními úlohami. Protože v práci řešíme především neparametrické postupy – semiparametrické postupy, jsou v první kapitole uvedeny alespoň odkazy především na parametrické metody řešení extremálních úloh.

Druhá část obsahuje základní výsledky teorie extremálního rozdělení. Je v ní uvedena základní věta teorie extremálního rozdělení. Jejím důsledkem je rozdělení náhodných veličin do tří disjunktních tříd - sfér přitažlivosti :

a) Gumbelovu
b) Frèchetovu
c) Weibulovu.

Dělení jednotlivých náhodných veličin do těchto sfér přitažlivosti se zjednodušuje zavedením základního parametru extremálního rozdělení – EVI (extremal value index). V této části jsou uvedeny základní tvrzení o vztahu EVI k jednotlivým sférám. Dále jsou zde uvedeny i von Miesesovy postačující podmínky, aby daná náhodná veličina patřila do dané sféry přitažlivosti. S tím souvisí zavedení základního nástroje – pravidelně se měnící funkce. Pomocí vlastností takovýchto funkcí jsou dále odvozeny tzv. podmínky prvního a druhého řádu, které hrají velmi významnou roli v odvození základních semiparametrických odhadů EVI. V souvislosti těmito odhady je zároveň zaveden pojem prostřední posloupnosti (intermediate), pomocí níž je většina takových odhadů realizována. Dále jednotlivé základní typy odhadů odvozujeme. Jde o Hillův odhad, momentový odhad, Pickandsův odhad. Jsou zkoumány jejich základní vlastnosti jako je např. silná a slabá konzistence. Dále je vyšetřována asymptotická normalita odhadů. V závěru kapitoly jsou uvedeny specifické odhady, které mají některé podstatné vlastnosti např. invariance vůči lineárním transformacím. Jedním z důležiých tříd takovýchto odhadů jsou odhady třídy PORT, které jsou posléze využity v poslední kapitole. Celkově bylo odvozeno mnoho desítek různých typůodhadů parametru EVI. Například jsou zaváděny podmínky třetího řádu a z nich vyvozovány jiné typy odhadů, jsou vytvářeny odhady typu smíšeného. Do budoucna je vhodné se zaměřit spíše na tvorbu odhadů vysokých kvantilů, abychom mohli lépe predikovat extremální stavy.

V třetí části jsou uvedeny základy metody bootstrap. Nejdříve jsou uvedeny historické kontexty založení myšlenek metody bootstrap od článku Huback (1923) až k článku Efron (1979). Dále se samozřejmě zabýváme obecným zavedením metody. Přitom jsou využity především výsledky z monografie Hall (1989). Ukážeme postupně konvergenci s.j. rozdílu

bootstrapové posloupnosti a klasické standardizační posloupnosti k nule. Dále se zabýváme rychlostí konvergence bootstrapové posloupnosti a ověřujeme, že pro jisté typy náhodných veličin je tato aproximace lepší než aproximace normálním rozdělením.

Ve zbývajícím prostoru třetí kapitoly se věnujeme teoretickým a praktickým aspektům tzv. Edgeworthova rozvoje. Tento rozvoj je základním aparátem pro zkoumání bootstrapové konvergence resp. Pro zjištění rychlosti konvergence této posloupnosti. V závěru kapitoly je uveden příklad aplikace Edgeworthova rozvoje na Hillův odhad. Ukazuje se, že tento způsob vyjadřování jednotlivých druhů odhadů je výpočetně velmi náročný.

V kapitole čtvrté jsou uvedeny důvody proč obecně bootstrap v extremálním rozdělení nekonverguje. Proto byla vytvořena speciální metoda „optimal sample fraction“, která byla dále v této kapitole studována.

Postupně jsou dokazována tvrzení o konvergenci metody „optimal sample fraction“ pro jednotlivé odhady EVI. Jsou uvedeny detaily důkazů, které jsou založeny na asymptotické normalitě jednotlivých odhadů. V rámci těchto důkazů jsou zavedeny modifikované odhady, pro které je jednodušší nalézt správnou volbu optimální hodnoty k(n). Ukazuje se, že tyto optimální hodnoty závisí na jistých konzistentních odhadech prametru ρ druhého řádu. Zároveň jsou vytvořeny algoritmy pro tvorbu odhadů EVI pomocí metod bootstrapu.

Z postupů uvedených v této části čtvrté kapitoly je zřejmé, že je možné aplikovat metodu „optimal sample fraction“ na většinu vytvořených odhadů EVI. Výsledkem takových postupů je prověření konvergence metody bootstrap aplikované na daný odhad a vytvoření algoritmu k nalezení takového odhadu.

V druhé části této kapitoly je uvedena rozsáhlá simulační studie, v níž jsou jednotlivé algoritmy metody „optimal sample fraction“ ověřovány. Zároveň je v některých případech studována i časová náročnost uvedené metody „optimal sample fraction“. Ze všech postupů uvedených v této kapitole je zřejmé, že je možné zlepšovat jak algoritmy pro aplikaci metody například pomocí nalezení jednodušších konzistentních odhadů parametru ρ (parametr druhého řádu), tak i použití metody „optimal sample fraction“ na jiné typy odhadů EVI (například na smíšené typy odhadů).

Cílem práce podle tezí bylo prostudovat neparametrické postupy založené na pořádkových statistikách a navrhnout vhodné varianty založené na bootstrapu. V tezích byla zmíněna také práce, která se zabývá modifikací robustního bootstrapu. Ukázalo se, že tato modifikace nebyla z hlediska výsledků dostatečně efektivní.

Proto jsme se v práci zabývali výše uvedenou metodou „optimal sample fraction“. Tuto metodu jsme dostatečně prostudovali ve 4. kapitole této práce. Věnovali jsme se všem základním odhadům EVI – Hillovu odhadu, Pickandsovu odhadu a momentovému odhadu.

U každého z těchto odhadů jsme hledali postup pro nalezení optimální hodnoty k (prostřední posloupnosti). Takto nalezené hodnoty jsme dále užili pro vytvoření bootstrapového algoritmu. Pro každý odhad je tento algoritmus jiný, i když lze nalézt podobné prvky, které se v těchto algoritmech objevovaly. Součástí těchto algoritmů bylo i

stanovení vhodných počátečních hodnot, pro co nejrychlejší konvergenci bootstrapové posloupnosti. Stanovení těchto hodnot jsme věnovali velké úsilí především v rozsáhlé simulační studii.

V poslední části uvedené kapitoly jsme se věnovali novým typům odhadů EVI, které byly vytvořeny vesměs v posledních pěti letech . Konkrétně šlo o odhady založené na technice PORT a metodice MVRB. Podobně jako u klasických odhadů jsme ověřili teoretické aspekty příslušných odhadů. Bootstrapové algoritmy jsou pro tyto odhady mnohem komplikovanější. Většinou je musíme provádět v několika ne příliš jednoduchých krocích. V závěrečných částech 4. kapitoly této práce je uvedena simulační studie, v níž jsme ověřili možnosti metody „optimal sample fraction“ pro všechny odhady – klasické, ale i nové.

V páté kapitole jsme využili předchozí výsledky na metodu kvantilové regrese. Dokázali jsme tvrzení o zprůměrovaných regresních kvantilech. Toto tvrzení bylo založeno na jistých větách z práce Dienstbier (2011) a umožňuje využít pro případ lineární regrese, v kterém jsou chyby náhodné veličiny z některé ze sfér přitažlivosti, postupy založené na bootstrapu pro odhad parametru EVI. Vzhledem k tomu, že využíváme lineární regresi, je zapotřebí mít typ odhadu, kterýje „location/scale invariant“. Takovým odhadem je Pickandsův odhad. I když víme, že není nejpřesnější, přesto se nám podařilo velmi pěkně odhadnout neznámý parametr γ. Pokud bychom dokázali výše uvedená tvrzení z páté kapitoly(obdoby vět 5.8 – 5.10) například pro odhady typu PORT (jsou také location/scale invariant), získali bychom přesnější výsledky, ovšem výpočty by byly časově mnohem náročnější.

Literatura:

Anděl, J.: Matematická statistika, SNTL, Praha, (1978)

Angus, J. E.: Asymptotic theory for bootstraping the extremes. Commun. Stat. Theory methods **22**, 15 – 30, (1993)

Araújo, S. I., Fraga Alves, M. I., Gomes, M. I.: Peaks over random Treshold metodology for tail index and high quantile estimation**,** Statistical Journal, Volume 4, Number 3, pp. 227-247, 2006

Athreya, K.: Bootstrap of the mean in the infinite variance case, Ann. Stat. 15 (2), 724 – 731, (1987)

Balakrishnan, N., Chan, P. S.: Order statistics from extreme value distribution, II: best linear unbiased estimates and some other uses, Comm. Statist. Simulation Comput., 21, 1219 – 1246, (1992)

Beirlant, J., Goegebeur, Y.,Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications, Willey,New Jersey, (2004)

Beirlant, J., Dierckx, G., Guillou, A.: Estimation of the extreme – value index and generalized quantile plots, Bernouli 11 (6), 949 – 970, (2005)

Berry, A. C. :The Accuracy of the Gaussian Approximation to the Sum of Independent Variates, *Transactions of the American Mathematical Society* **49** (1), 122–136, (1941).

Bickel, P. J., Freedman, D.: Some asymptotic theory for the bootstrap, Ann. Stat. , 9(6), 1196 – 1217, (1981)

Billingsley, P.: Probability & Measure , (3rd ed.). New York: Wiley, *(*1995)*:*

Brunetti, M., L. Buffoni, F. Mangianti, M. Maugeri, and T. Nanni, Temperature,precipitation and extreme events during the last century in Italy. Global and Planetary Change, 40, 141–149, (2004)

Buchinsky, M.: The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach. Journal of Applied Econometrics, Vol. 13, 1 - 30, (1998)

Caeiro, F., Gomes, M. I.: A new class of estimators of a „scale“ sekond order parameters, Extremes **9**, 193 – 211,(2006)

Caeiro, F., Gomes, M. I.: On the Bootstrap Methodology for the Estimation of the Tail Sample Fraction, in Proceedings of COMPSTAT 2014, 21th International Conference on Computional Statistics, 545 – 552, (2014)

Caeiro, F., Gomes, M. I., Pestana, D.: Direct reduction of bias of the classical Hill estimator, Revstat **3,** 2, 111 – 136, (2005)

Čebyšev, P. L.: Sur deus Théorémes relativ aux probabilités, Acta Math. **14**, 305 - 315, (1890)

Christopeit, N.: Estimating parameters of an extreme value distribution by the method of moments, J. Statist. Plann. Inference, 41, 173 – 186, (1994)

Ciuperca, G., Mercadier, C.: Semi-parametric estimation for heavy toled distributions, Extremes **13**:1, 55 – 87, (2010)

Clauset, A., Shalizi, C. R., Newman, M. E .J., Power-Law Distributions in Empirical Data, SIAM Rev. 51(4), 661 – 703, (2009)

Colombi, S., Extreme value statistics of smooth random Gaussian fields and applications to kosmology, Extreme Value Statistics in Mathematics, Physics and Beyond,(2011)

Cooil,B.: Limiting multivariate distributions of intermediate order statistics. Ann. Probability **13**, 469-477, (1985)

Cooligan, H., Research Methods and Statistics in Psychology, Routledge, London and New York, (2009)

Danielsson, J., de Haan, L., Peng, L., de Vries, C.G.: Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation, J. Multivariate Anal. **76**, 226 – 248, (2001)

Davino, C., Furno, M., Vistocco, D.: Quantile Regression: Theory and Application, Wiley, (2014)

Davison, A. C., Hinkley, D. V.: Bootstrap Methods and their Application, Cambridge Series in Statistical and Probabilistic Mathematics, (1997)

Dekkers, A. L. M., de Haan, L.: On the Estimation of the Extreme-Value Index and Large Quantile Estimation. Ann. Stat. 17 (4), 1796 – 1832, (1989)

Dekkers, A. L. M., Einmahl, J. H. J., de Haan, L.: A Moment Estimator for the index of an Extreme – Value Distribution, Ann. Stat 17 (4), 1833 – 1855, (1989)

Dekkers, A. L. M., de Haan, L.: Optimal choice of sample fraction extreme-value estimation, J. Multivariate Anal. **47**, 173 – 195, (1993)

de Haan, L.: On regular Variation and its Applications to the Weak Convergence of Sample Extremes, Mathematical Centre Tract, 32, Amsterodam, (1970)

de Haan, L., Ferreira, A.: Extreme Value Theory. An Introduction, Springer, New York, (2006)

Dienstbier, J.: Stochastical inference in the model of extreme events, Ph.D Thesis, MFF UK, Praha, (2011)

Dietrich, D., Hüsler, J.: Minimum distance estimators in extreme value distributions, Comm. Statist. Theory Methods, 25, 695 – 703, (1996)

Dodge,Y.: The Oxford Dictionary of Statistical Terms, OUP, ISBN 0-19-920613-9,(2003)

Draisma, G., de Haan, L., Peng, L., Pereira, T. T.: A Bootstrap-based Method to Achieve Optimality in Estimating the Extreme-value Index, Extremes **2** (4), 367 – 404, (1999)

Drees, H.: Refined Pickands estimators for extreme value index, Ann. Stat. 23 (6), 2059 – 2080, (1995)

Drees, H., Kaufmann, E.: Selecting the optimal sample fraction in univariate extreme value estimation, Stochastic Process and thein Applications **75**, 149 – 172, (1998)

Drees, H: On smooth statistical tail functionals, Scandinavian Journal of Statistics 25 (1), 434-448, (1998)

Edgeworth, F. Y.:The asymmetrical probability curve, Philos. Mag. 5th Ser. **41**, 90 – 99, (1896)

Edgeworth, F. Y.: The law error, proc. Cambridge Philos. Sooc. **20**, 36 – 65, (1905)

Efron, B.: Bootstrap methods: another look at the jacknife, Ann. Statistics **7**, 1-26, (1979)

Efron, B.: The Jacknife, the Bootstrap, and Other Resampling Plants, SIAM, Philadelphia, (1982)

Embrechts, P., C. Klüppelberg, and T. Mikosch, Modelling Extremal Events for Insurance and Finance. Springer-Verlag, 645 pp, (1997)

Esseen, C. G.: A moment inequality with an application to the central limit tvorem, *Skand. Aktuarietidskr.* **39**, 160–170, (1956).

Farmer, J. D., Physicists attempt to scale the ivory towers of finance, Computing in Science and Engineering, (1999)

Ferreira, A., de Haan, L., Peng, L.: On optimizing the estimation of high quantiles of probability distribution, Statistics 37 (5), 401 – 434, (2003)

Fisher, R. A., Tippet, L. H. C.: Limiting forms of the frequency distribution of the largem or smallest member of sample, Proc. Camb. Phil. Soc. 24, 180 – 190, (1928)

Fisher, R. A., Yates, F. : Statistical tables for biological, agricultural and medical research, Edinburg, Oliver and Boyd, (1945)

Fraga Alves, M. I.: A location invariant Hill – type estimator, Extremes 4, 199 – 217, (2001)

Fraga Alves, M. I.; de Haan, L., Lin, T.: Estimation of the parameter controlling the speed of convergence in extreme value theory, Math. Meth. Statist., 12, 155–176, (2003)

Fraga Alves, M. I., Gomes M. I., de Haan, L., Neves, C.: The mixed moment estimator and location invariant alternatives, Extremes 12,14 – 185, (2009)

Fraga Alves, M. I.; de Haan, L., Lin, T.: Third order extended regular variation, Pub. l'Institut Math., 80, 109–120, (2006)

Frèchet, M.: Sur le loi de probabilitè de l' ècart maximum, Ann. Soc. Polonaise Math., 6, 93–116, (1927)

Galambos, J.: The asymptotic theory of extreme order statistics, 2nd ed., Krieger, Melbourne, Florida, (1987)

Geluk, J., de Haan, L.:Regular Variation, Extensions and Tauberian Theorems, CWI Tract 40, Amsterdam, (1987)

Geluk, J., de Haan, L.: On bootstrap sample size extreme value theory, Publ. Inst. Math. (Beograd) (N. S.) **71** (85), 21 – 25, (2002)

Gnedenko, B. V.: Sur la distribution limite du terme maximum dune série alétoire, Ann. Math. 44, 423 – 453, (1943)

Goegebeuer, Y., Beirlant, J., de Wet, T.: Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal treshold and sekond order estimation, Revstat **6**:1, 51 – 69, (2008)

Gomes, M. I., Oliveira, O.:The Bootstrap Methodology in Statistics of Extremes-Choice of the Optimal Sample Fraction, Extremes **4**(4), 331 – 358, (2001)

Gomes, M. I., de Haan, L., Peng, L.: Semi – parametric estimation of the second order parametr – asymptotic and finite sample behaviour, Extremes 5 (4), 387 – 414, (2002)

Gomes, M. I., Martins, M. J.: „Asymptoticaly unbiased" estimators of the tail index based on external estimation of the second order parameter, Extremes, **5**:1, 5 – 31, (2002)

Gomes, M. I., de Haan, L., Henriques Rodrigues, L.: Tail Index estimation for heavy-tailed models : accomodation of bias in weighted log-excesses, J. Royal Statistical Society B, DOI: 10.1111/ j.1467-9869.2007.00620.x, (2004)

Gomes, M. I., Fraga Alves, M. I., Araujo Santos, P.:PORT Hill and Moment Estimators for Heavy-Tailed Models. Tech. Rep. 15/07, CEAUL, (2007)

Gomes, M. I., Martins, M .J., Neves, M. : Improving second order reduction bias extreme value index estimation, Revstat **5**, 2, 177 – 207, (2007)

Gomes, M. I., Fraga Alves, M. I., Araújo Santos, P.: PORT Hill and Moment Estimators for Heavy – Tailed Models, Journal of Statistical Computation and Simulation, **37**(7), 1281 – 1306, (2008)

Gomes, M. I., Henriques – Rodrigues, L.: Adaptive PORT-MVRB Estimation: an Empirical Comparison of two Heuristic Algorithms, Journal of Statistical Computation and Simulation, **83**(6), 1129 – 1144, (2011)

Gomes, M. I., Figueiredo, F., Neves, M. M.: Adaptive Estimation of Heavy Right Tails: resampling – based Methods in Action, Extremes **15(**4**)**, 463 – 489, (2012)

Greene, W. H.: Econometric Analysis, sixth edition, New Jersey, (2008)

Greenwood, J. A., Landwehr, J. M., Matalas, N. C.,Wallis, J.R. : Probability weighted moments: definitionand relation to parameters of several distributions expressable in inverse form, Water ressource research,15, 1049 – 1054, (1979)

Gurney, M.: The Variance of the Replication Method for Estimating Variances for the CPS "SampleDesign. Unpublished memorandum, U."S.Bureau of the Census, (1962)

Chernic, M. R.: Bootstrap methods: A guide for practicioners and researchers, Wiley & Sons, New Jersey, (2008)

Hall, P.: On the number of bootstrap simulations required to construct a confidence interval, Ann. Stat., **14**(4), 1453–1462, (1986)

Hall, P.: Edgeworth expansions for Student's t – statistic under minimal moment conditions, Ann. Prob., **15**(3), 920 – 931, (1987)

Hall, P.: Rate of konvergence in bootstrap approximations, Ann. Prob., **16**(4), 1665-1684, (1988)

Hall, P.: On efficient bootstrap simulation, Biometrika, **76**(3), 613–617, (1989)

Hall, P.: Asymptotic properties of the bootstrap for heavy – tailed distributions, Ann. Prob., **18**(3), 1342 – 1360, (1990a)

Hall, P.: Using the Bootstrap to Estimate Mean Squared Error and Select Smothing Parametr in Nonparametric Problems, Journal of multivariate analysis **32**, 177-203 (1990b)

Hall, P.: The Bootstrap and Edgeworth Expansion, Springer-Verlag, New York, (1992)

Hall, P., Jing, B. Y.: Comparison of bootstrap and asymptotic approximation to the distribution of a heavy-tailed mean, Statistica Sinica 8, 887-906, (1998)

Hall, P., Welsh, A. H.: Adaptive estimate of parameters of regular variation, Ann. Statist. 13, 331 – 341, (1985)

Hill, B. M.: A simple general Approach to Interfernce about the tail of a Distribution, Ann. Statistics 3, 1163 – 1174, (1975)

Hosking, J. R. M.: Algorithm AS 215: Maximum likelyhood estimation of the parameters of the generalized extreme value distribution, J. R. Stat. Soc. Ser. C. Appl. Stat., 34, 301 – 310, (1985)

Hubback, J. A.: Sampling for Rice Yield in Bihar and Orissa, Sankhyā: The Indian Journal of Statistic., Vol. 7, No. 3, Apr., 1946, (1923)

Hušek, R., Pelikán, J: Aplikovaná ekonometrie, Teorie a praxe, Praha, Professional Publishing, (2003)

Cheng, S., Pan, J.: Asymptotic Expansions for Distribution Function of Moment Estimator for the Extreme Value Index, Science of China, **43**(11), 1131 – 1143, (2000)

Cheng, S., Pan, J.: Asymptotic Expansions of Estimators for the Tail Index with Applications, Scand, J. Statist., 717 – 728, (1998)

Cheng, S., Peng, L.: Confidents Intervals for the Tail Index, Bernoulli, **7**(5), 751 – 760, (2001)

Jarrett, R. G., Maritz, J.**,** S.: A Note on Estimating the Variance of the Sample Median, Journal of the American Statistical Association, Vol. **73 – 361,** 194 – 196, (1978)

Joyce, P., Rokyta, D. R., Beisel, C. J., Orr, H. A., A General Extreme Value Theory Model for the Adaptation of DNA Sequences Under Strong Selection and Weak Mutation, Genetics 180, 1627 – 1643, (2008)

Jurečková, J., Picek, J., Averaged regression quantiles. In: Contemporary Developments in Statistical Theory (S. Lahiri et al., eds.), Springer Proc. in Math. & Statistics 68, 203–216 (2014)

Kaiser, O., Horenko, I., On inference of statistical regression models for extreme events based on incomplete observation data, University of Lugano, (2013)

Katz, R. W., Parlange R. W., Naveau P., Statistics of extremes of hydrology, Advances in Water Resources 25, 1287 - 1304 , (2002)

Koenker, R.: Quantile regression, Number 38, Cambridge university press, (2005)

Koenker, R.: quantreg: Quantile Regression, R package version 5.11, (2015)

Koenker, R, R., Bassett, G.: Regression Quantiles. Econometrica, Vol. 46, 33 - 50, (1978)

Koenker, R., Bassett, G.: Robust test for heteroscedasticity based on regression quantiles. Econometrica, Vol. 50, str. 43 - 61, (1982)

Kohout, V., Picek, J.: Boostrap and the Moment of Tail Index, 29th European Meeting of Statisticians, Budapest, (2013a)

Kohout, V., Picek, J.: Bootstrap in estimating the Extreme Value Index, Int. conf. Precipitation Extremes in a Changing Climate, hejnice, (2013b)

Kohout, V.: Využití bootstrapu pro nalezení shape parametru gama rozdělení extrémních hodnot – simulační studie, XXXI. Inter. Colloq. on the Manag. of educational Process, 97 – 107,(2013)

Landwehr, J., Matalas, N, Wallis, J.: Probability Weighted Moments compared with some traditional techniquesin estimating Gumbel parameters and quantiles, Water Resources Research, 15, 1055 – 1064, (1979)

Li, J., Peng., Z., Nadarajah, S.: A class of unbiased location invariant Hill-type estimators for heavy tailed distributionsm Electronic Journal of Statistics, 829 – 847, (2008)

Ling, C., Peng, Z., Nadarajah, S.: A location invariant Moment-type estimator I. Theory Probab. Math. Stat. **76**, 23–31 (2007a)

Ling, C., Peng, Z., Nadarajah, S.: A location invariant Moment-type estimator II. Theory Probab. Math. Stat. **77**, 177–189 (2007b)

MacLeod, A. J.: A remark on the algorithm 215: Maximum likelyhood estimation of the parameters of the generalized extreme value distribution, Applied Statistics, 38, 198 – 199, (1989)

Mahalanobis, P. C.: A revision of Risley`s anthropometric data relating Chittagong Hill tribes, Sankhyā: The Indian Journal of Statistic., Vol. 1, 267 - 276, (1931)

Manly, B. F. J.: Randomization, Bootstrap and Monte Carlo Methods in Biology, Cambridge Series in Statistical and Probabilistic Mathematics, (2006)

Marcinkiewicz, J., Zygmund, A.: Quelques inégalités pour les opérations linéaires, Fund. math. 32, 113 – 121, (1939)

Mc Carthy, P. G.: Pseudo replication – half samples, Rewiev of the international statistical institut, **37:3**, 239 – 264, (1969a)

Mc Carthy, P. G. and many others: A REPORT of the National Center of Health Statistics (Series2, No. 14), (1969b)

Newey, W., Powell, J.: Eficient estimation of linear and type I censored regression models under conditional quantile restrictions. Econometric Theory, 6, 295 - 317, (1990)

Peng, L.: Asymptotically unbiased estimators for extreme value index, Stat. Probab. Lett. 38, 107 – 114, (1998)

Peng,L., Qi, Y.: Asymptotic normality of Hill estimator in sekond-order submodel of regular variation, Chinese Ann. Math. Ser. A, **18**(5), 539 – 544, (1997)

Petrov, V. V.: Sum of Independent random Variables, Springer – Verlag, Berlin, (1975)

Pickands III, J.: Statistical inference using extréme order statistics, Ann. Stat. 3, 1163-1174, (1975)

Pickands, J. III: The continuous and differentiable domains of attraction of the extreme value distributions, Ann. Probab. **14**, 996 – 1004, (1986)

Potter, H. S. A.: The Mean Value of Certain Dirichlet Series, Proc. London Math. Soc. 47, 1 – 19, (1942)

Powell, J.: Least absolute deviation estimation for the censored regression model, Journal of Econometrics, Vol. 25, 303 – 325, (1984)

Prášková, Z.: Metoda bootstrap, Robust 2004, JČMF, 299 – 314, (2004)

Prescott. P, Walden, A. T.: Maximum likelyhood estimation of the parametersof the generalized extreme – value distribution, Biometrika, 67, 723 – 724, (1980)

Prescott. P, Walden, A. T.:Maximum likelyhood estimationof the parameters of the free – parameter generalizedextreme – value distribution from censored samples, J. Statist. Comput. Simulation, 16, 241 – 250, (1983)

Ralston, A.: Základy numerické matematiky, Academia, Praha, (1978)

Reiss, R. - D., Thomas, M.: Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields, 3rd Edition, Birkhäuser Verlag, Basel – Boston – Berlin, (2007)

Resnick, I. S.: Extreme Values, Regular Variation, and point Processes, Springer, New York, (1987)

Segers, J.: Generalized Pickands Estimators for Extreme Value Index, J. Stat. Plan. Infer. 128 (2), 381 – 396, (2005)

Shao, J., Tu, D.: The Jacknife and the Bootstrap, Springer, (1995)

Shorack, G., Wellner, J.: Empirical Processes with Applications to Statistics, John Wiley & Sons, (1986)

Silbergleit, V. M., Forecast of the most geomagnetically disturbed days, Earth Planets Space, 51, 19–22, (1999)

Simon, J. L., Basic Research Methods in Social Science, Random House, New York, (1969)

Simon. J. L., Weidenfeld, D.: SIMPLE: Computer Program for Monte Carlo Statistics Teaching, American Statistician, Nov., (1974), (letter)

Singh, K.: On the asymptotic accuracy of Efron`s bootstrap, Ann. Stat., 9(6), 1187 – 1195, (1981)

Smith P. J.: A Recursive Formulation of the Old Problem of Obtaining Moments from Cumulants and Vice Versa, The American Statistician, Volume 49, Issue 2, pages 217-218, (1995)

Smith, R. L.: Maximum likelyhood estimation in a class of nonregular cases, Biometrika, 72, 67 – 90, (1985)

Štěpán, J.: Teorie pravděpodobnosti, Academia, Praha, (1987)

Tasche, D. J.: Unbiasedness in Least Quantile Regression,in Conf. Developments in Robust Statistics, (2001)

Vanderbei, R. J.: Linear programming: Foundation and extension, Springer, (2013)

von Mises, R.: La distribution de la plus grande de n valeurs. Reprinted in Selected Papers Volumen II, Amer. Math. Soc. , Providence, R. I., 271 – 294, (1936)