Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Tomáš Tomasy

Asymptotika při malém rozsahu výběru

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Radka Sabolová

Studijní program: matematika

Studijní obor: finanční matematika

Praha 2012

Na tomto mieste by som rád poďakoval vedúcej mojej bakalárskej práce Mgr. Radke Sabolovej za cenné rady, ochotu, s ktorou mi venovala svoj čas a za poskytnuté materiály. Veľká vďaka patrí aj mojej rodine a kamarátom, ktorí ma počas celého môjho štúdia podporovali a dôverovali mi.

Prehlasujem, že som túto bakalársku prácu vypracoval samostatne a výhradne s použitím citovaných prameňov, literatúry a ďalších odborných zdrojov.

Beriem na vedomie, že sa na moju prácu vzťahujú práva a povinnosti vyplývajúce zo zákona č. 121/2000 Sb., autorského zákona v platnom znení, hlavne skutočnosť, že Univerzita Karlova v Prahe má právo na uzavrenie licenčnej zmluvy o použití tejto práce ako školského diela podľa §60 odst. 1 autorského zákona.

V Prahe dňa 15.5.2012 Tomáš Tomasy

Název práce: Asymptotika při malém rozsahu výběru

Autor: Tomáš Tomasy

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Radka Sabolová, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V tejto práci budeme študovať správanie sa odhadov pre malý počet pozorovaní. Popíšeme si metódu sedlového bodu, ktorá je vhodná na riešenie tohto problému. Presnejšie sa budeme zaoberať aproximáciou hustoty daného odhadu, ktorú je súčasťou druhej a tretej kapitoly. V prvej kapitole uvedieme centrálnu limitnú vetu, M-odhady a ich asymptotické správanie medzi základnými pojmami.
V praktickej časti práce aplikujeme túto metódu na vybrané odhady pre niektoré rozdelenia a porovnáme ju s aplikovaním centrálnej limitnej vety. Výsledky predvedieme v grafoch a zhrnieme si ich v závere.

Klíčová slova: metóda sedlového bodu, centrálna limitná veta, M-odhad

Title: Small sample asymptotics

Author: Tomáš Tomasy

Department: Department of Probability and Mathematical Statistics

Supervisor: Mgr. Radka Sabolová, Department of Probability and Mathematical Statistics

Abstract: In this thesis we study the small sample asymptotics. We introduce the saddlepoint approximation which is important to approximate the density of estimator there. To derive this method we need some basic knowledge from probability and statistics, for example the central limit theorem and the M-estimators. They are presented in the first chapter.
In practical part of this work we apply the theoretical background on the given M-estimators and selected distribution. We also apply the central limit theorem on our estimators and compare it with small sample asymptotics. At the end we show and summarize the calculated results.

Keywords: saddlepoint approximation, central limit theorem, M-estimator

Obsah

Úvod **2**

1 Základné pojmy **3**

1.1 Centrálna limitná veta 3

1.2 M-odhady 3

1.2.1 M-odhad parametru posunutia 4

1.2.2 Asymptotické správanie M-odhadov 5

1.2.3 Asymptotické správanie MLE odhadov 5

1.2.4 Príklady 6

2 Aproximácia hustoty priemeru pomocou sedlového bodu **9**

2.1 Metóda najstrmšieho spádu 9

2.2 Aproximácia hustoty priemeru pomocou sedlového bodu 11

2.3 Súvislosť s metódou združených rozdelení 13

3 Všeobecné aproximácie pomocou sedlového bodu **16**

3.1 Jednorozmerné M-odhady 16

3.2 Viacrozmerné M-odhady 19

4 Simulácie **21**

4.1 Aproximácie hustoty pomocou sedlového bodu 21

4.1.1 Normálne rozdelenie 21

4.1.2 Gamma rozdelenie 23

4.2 Hustoty M-odhadov pre rôzne rozdelenia 26

4.2.1 Hustoty Huberovho odhadu 26

4.2.2 Hustoty maximálne vierohodných odhadov 28

Záver **30**

Zoznam použitej literatúry **31**

Prílohy **32**

Úvod

Mnohé odhady majú asymptoticky normálne rozdelenie, teda s ľahkými chvostami. Toto asymptotické rozdelenie aproximuje správanie odhadu veľmi dobre v strede, nie však na chvostoch. Vyvoláva to teda otázku, či je správne spoliehať sa na aproximáciu normálnym rozdelením v prípade, ak potrebujeme kvantily rozdelenia alebo tiež aj, aký je dostatočný počet pozorovaní, aby bola táto aproximácia dobrá.

Predložená práca sa zaoberá asymptotickým správaným odhadov pre malý rozsah výberu. Je rozdelená do štyroch kapitol, pričom prvé tri sa zaoberajú teóriou a posledná je venovaná praktickým ukážkam.

Pre vybudovanie samotnej teórie sú dôležité niektoré základné pojmy, poznatky a vzťahy týkajúce sa centrálnej limitnej vety a M-odhadov, ktoré si uvedieme v prvej kapitole. Nasleduje časť venovaná metóde sedlového bodu, ktorá rieši základný problém, a to ako aproximovať hustotu priemeru n nezávislých rovnako rozdelených veličín a časť, kde sa tejto metóde venujeme pre všeobecné aproximácie.

V praktickej časti práce sú aplikované popisované metódy, ktoré sú potom porovnávané s reálnymi hodnotami alebo dôsledkami centrálnej limitnej vety. Dosiahné výsledky si zhrnieme v závere.

Kapitola 1

Základné pojmy

V tejto kapitole uvedieme základné pojmy a tvrdenia, s ktorými budeme pracovať v ďalších kapitolách.

1.1 Centrálna limitná veta

Centrálna limitná veta označuje v teórii pravdepodobnosti tvrdenie, podľa ktorého sa rozdelenie výberového priemeru po vhodnej štandardizácii blíži k normálnemu rozdeleniu. Hovoríme, že náhodná veličina, ktorá má uvedené správanie, má asymptoticky normálne rozdelenie.

Veta 1. *Nech $X_1, X_2, ...$ je postupnosť nezávislých rovnako rozdelených náhodných veličín s konečnou strednou hodnotou μ, konečným rozptylom $\sigma^2 > 0$. Nech ďalej $E|X_i|^3 < \infty \quad \forall i = 1, 2, ...$ Označme*

$$S_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}.$$

Potom platí

$$S_n \xrightarrow{\mathcal{D}} S, \ n \to \infty,$$

kde

$$S \sim \mathcal{N}(0, 1).$$

1.2 M-odhady

M-odhady sú v prípade vhodne zvolenej funkcie ψ (definujeme neskôr) robustné odhady, to znamená, že v prípade, ak do istej miery pripúšťame v našich pozorovaných dátach zastúpenie aj odľahlých pozorovaní, odhad je navrhnutý tak, aby sa neprejavil ich nežiadúci vplyv vôbec, alebo len čiastočne.

P.J. Huber zavádza v [1] triedu M-odhadov, ktoré sú vlastne zovšeobecnením maximálne vierohodných odhadov. M-odhad T_n je definovaný ako riešenie minimalizačnej úlohy

$$\sum_{i=1}^{n} \rho(X_i, \theta) := min \quad \text{vzhľadom k } \theta \in \Theta \tag{1.1}$$

alebo

$$E_{P_n}[\rho(X, \theta)] = min, \quad \theta \in \Theta,$$

kde ρ je vhodne zvolená funkcia. P_n označuje empirické rozdelenie pravdepodobnosti vektoru pozorovaní $(X_1, ..., X_n)$. M-odhad T_n je riešením (môže byť aj jediným) rovnice

$$\sum_{i=1}^{n} \psi(X_i, \theta) = 0, \quad \theta \in \Theta \tag{1.2}$$

v prípade, ak ρ je diferencovateľná vzhľadom k θ so spojitou deriváciou

$$\psi(., \theta) = \frac{\partial}{\partial \theta} \rho(., \theta).$$

Definícia. *Hovoríme, že odhad T_n je M-odhad ak spĺňa rovnicu (1.1) alebo (1.2).*

1.2.1 M-odhad parametru posunutia

Dôležitým špeciálnym prípadom je model s parametrom posunutia θ, pri ktorom $X_1, ..., X_n$ je náhodný výber z rozdelenia s distribučnou funkciou $F(x - \theta)$, $\theta \in \mathbb{R}$, pričom distribučná funkcia F je všeobecne neznáma. M-odhad T_n potom definujeme ako riešenie minimalizačnej úlohy

$$\sum_{i=1}^{n} \rho(X_i - \theta) := min, \tag{1.3}$$

a pokiaľ je funkcia ρ navyše diferencovateľná s absolútne spojitou deriváciou ψ, povieme, že T_n je riešením rovnice

$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0. \tag{1.4}$$

Z podmienok (1.3) a (1.4) ihneď vyplýva, že odhad T_n je ekvivariantný vzhľadom k posunutiu, to znamená, že spĺňa

$$T_n(X_1 + c, ..., X_n + c) = T_n(X_1, ..., X_n) + c \quad \forall c \in \mathbb{R}.$$

Na druhej strane však T_n nie je všeobecne ekvivariantný k meradlu, tj. všeobecne neplatí

$$T_n(cX_1, ..., cX_n) = cT_n(X_1, ..., X_n) \text{ pre } c > 0.$$

1.2.2 Asymptotické správanie M-odhadov

Nech $\{X_i, i = 1, 2, ...,\}$ je postupnosť náhodných veličín s rovnakou distribučnou funkciou $F(x, \theta),\ \theta \in \Theta$, kde Θ je otvorený interval $\mathbb{R}^1$. M-odhad parametru θ je riešením minimalizácie

$$\sum_{i=1}^{n} \rho(X_i, \theta) := min\ \ \theta \in \Theta.$$

Predpokladajme, že $\rho(x, \theta)$ je absolútne spojitá v θ s deriváciou $\psi(x, \theta) = \frac{\partial}{\partial\theta}\rho(x, \theta)$. Ak je $\psi(x, \theta)$ spojitá v θ, potom hľadáme M-odhad T_n medzi koreňmi rovnice (1.2).

Ak funkcia $E_\theta\rho(X, t)$ má jediné minimum v bode $t = \theta$ a sú splnené ďalšie podmienky buď na hladkosť $\psi(x, \theta)$, alebo $F(x, \theta)$, potom existuje postupnosť $\{T_n\}$ koreňov rovnice (1.2) taká, že pri $n \to \infty$

$$\sqrt{n}(T_n - \theta) = \mathcal{O}_p(1),$$

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}\gamma(\theta)} \sum_{i=1}^{n} \psi(X_i, \theta) + \mathcal{O}_p(n^{-1/2}),$$

kde $\gamma(\theta) = E_\theta\psi_0(X, \theta),\ \psi_0(x, \theta) = \frac{\partial}{\partial\theta}\psi(x, \theta)$. Odtiaľ ďalej vyplýva, že $\sqrt{n}(T_n - \theta)$ má asymptoticy normálne rozdelenie

$$\mathcal{N}(0, \sigma^2(\psi, F)),\ \text{kde } \sigma^2(\psi, F) = \frac{E_\theta\psi(X, \theta)}{\gamma^2(\theta)}.$$

1.2.3 Asymptotické správanie MLE odhadov

Máme nasledujúce predpoklady.

P1: Nech Ω je parametrický priestor, ktorý obsahuje taký neprázdny otvorený interval ω, že skutočná hodnota parametra θ_0 patrí do ω.

P2: Nech $\mathbf{X} = (X_1, ..., X_n)'$, kde X_i sú nezávislé rovnako rozdelené náhodné veličiny s hustotou $f(x, \theta)$ vzhľadom k nejakej σ-konečnej miere μ.

P3: Nech $M = \{x : f(x, \theta) > 0\}$ nezávisí na θ.

P4: Nech $\theta_1, \theta_2 \in \Omega$. Potom $f(x, \theta_1) = f(x, \theta_2)$ skoro všade platí práve vtedy, keď $\theta_1 = \theta_2$.

Veta 2. *Nech $\{f(x, \theta), \theta \in \Omega\}$ je regulárny systém hustôt s Fisherovou mierou informácie $J(\theta)$ a nech platia predpoklady P1 až P4. Nech $\theta_0 \in \omega$ je skutočná hodnota parametru a nech sú splnené nasledujúce predpoklady.*

(a) Pre všetky $\theta \in \omega$ a skoro všetky $x \in M$ existuje derivácia $f'''(x, \theta) = \frac{\partial^3 f(x,\theta)}{\partial \theta^3}$.

(b) Pre všetky $\theta \in \omega$ platí

$$\int_M f''(x, \theta) d\mu(x).$$

(c) Existuje taká nezáporná merateľná funkcia $H(x)$, že $E_{\theta_0} H(x) < \infty$ a pritom pre skoro všetky $x \in M$ a pre všetky θ také, že $|\theta - \theta_0| < \varepsilon$ pre nejaké dostatočne malé $\varepsilon > 0$ platí

$$\left| \frac{\partial^3 f(x, \theta)}{\partial \theta^3} \right| \leq H(x).$$

Potom platí nasledujúce tvrdenie.

Ak pre každé dostatočne veľké n a pre každú hodnotu $\boldsymbol{X}$ existuje taký koreň T_n vierohodnostnej rovnice, že T_n je konzistentným odhadom parametru θ_0, potom

$$\sqrt{n}(T_n - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left[0, \frac{1}{J(\theta)}\right].$$

Dôkaz. Dôkaz je môžné nájsť v [5]. □

1.2.4 Príklady

(1) Priemer: funkcia $\psi(x, \theta) = x - \theta$ spĺňa rovnicu (1.2), a teda nám dáva M-odhad $T_n = \bar{X}_n$ (funkcia $\rho(x, \theta) = \frac{(x-\theta)^2}{2}$).

(2) Maximálne vierohodný odhad: $\psi(x, \theta) = -\frac{\partial}{\partial \theta} \log f(x, \theta)$ pre triedu funkcií hustoty $f(x, \theta)$ nám dáva, že odhad T_n je riešením vierohodnostnej rovnice

$$\frac{\partial}{\partial \theta} \log \left(\prod_{i=1}^{n} f(X_i, \theta) \right) = 0.$$

(3) Huberov odhad: funkcia

$$\psi_H(x) = \begin{cases} x & \dots |x| \leq k \\ k \; sgn(x) & \dots |x| > k \end{cases}$$

je lineárna v ohraničenom intervale $[-k, k]$ a je konštantná mimo intervalu. Navrhol a zdôvodnil ju P.J.Huber v [1]. Funkcia pomáha nájsť M-odhad parametru polohy rozdelenia pravdepodobností vhodný pre okolie normálneho rozdelenia. Príslušný M-odhad sa teda nazýva Huberov odhad. Funkciu $\psi_H(x)$ môžeme vidieť na obrázku (1.2.4).

Limitné prípady Huberovho odhadu:

(i): Pre $k = \infty$ je Huberovým odhadom aritmetický priemer.

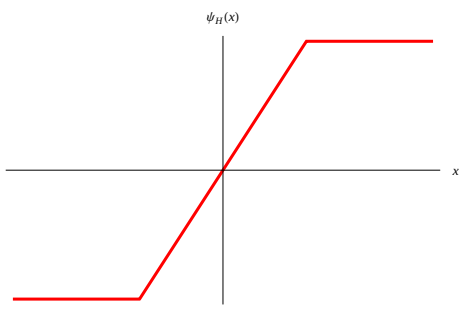
(ii): Pre $k = 0$ dostávame výberový medián.

Ako je uvedené v [1], najrobustnejší odhad je jednoznačne určený a zodpovedá nasledujúcej funkcii ρ:

$$\rho(t) = \begin{cases} \frac{1}{2}t^2 & \dots |t| \leq k \\ k|t| - \frac{1}{2}k^2 & \dots |t| > k \end{cases}$$

V [1] je ďalej odvodený aj asymptotický rozptyl tohto odhadu, ktorý využijeme neskôr pri asymptotickom správaní Huberovho odhadu.

$$\sigma^2 = \frac{\int_{-k}^{k} t^2 F(dt) + k^2 \int_{-\infty}^{-k} F(dt) + k^2 \int_{k}^{\infty} F(dt)}{\left(\int_{-k}^{k} F(dt)\right)^2}. \tag{1.5}$$



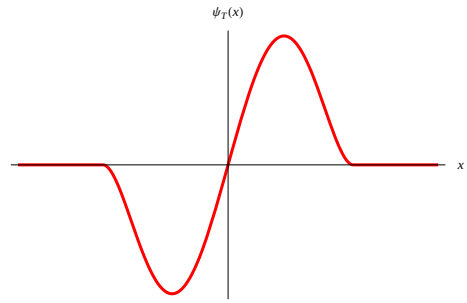
Obr. 1.1: Huberova funkcia

(4) Medián: $X = F^{-1}(\frac{1}{2})$. Funkcia

$$\psi(x, \theta) = sgn(x - \theta).$$

(5) Tukeyho *biweight* funkcia:

$$\psi_T(x) = \begin{cases} x[1-(\frac{x}{k})^2] & \dots |x| \leq k \\ 0 & \dots |x| > k \end{cases}$$



Obr. 1.2: Tukeyho „biweight“ funkcia

Kapitola 2

Aproximácia hustoty priemeru pomocou sedlového bodu

V tejto časti uvedieme techniku sedlového bodu pre základný problém, a to aproximáciu hustoty priemeru n nezávislých rovnako rozdelených náhodných veličín.

Základná idea sa používa v zložitejších modeloch na odvodenie priblíženia pomocou sedlového bodu pre všeobecnú štatistiku. Pri dobrom pochopení techniky v tomto základnom prípade umožnuje priamu aplikáciu pre zložitejšie a dôležitejšie situácie. Metódu vyvinunul H.E.Daniels a predstavil ju v základnom dokumente [2]. Uvedené metódy sú prevzaté z [3].

2.1 Metóda najstrmšieho spádu

Predstavíme tu všeobecnú techniku, pomocou ktorej môžeme vypočítať asymptotický rozvoj integrálov formy

$$\int_P e^{v.w(z)}\xi(z)dz. \tag{2.1}$$

Hodnota reálneho parameteru v je veľká a kladná. Ďalej w a ξ sú analytické funkcie s parametrom z v oblasti komplexnej roviny, ktorá obsahuje integračnú cestu[1] P. Technika sa nazýva *metóda najstrmšieho spádu* a používa sa na odvodenie aproximácie pomocou sedlového bodu k hustote priemeru.

Aby sme vypočítali hore uvedený integrál, môžeme ľubovoľne upraviť integračnú cestu P, za predpokladu, že sa stále nachádzame v oblasti, kde sú w a ξ analytické funkcie. Zmeníme P tak, že:

(i) nová integračná cesta prechádza bodom, kde derivácia $w'(z)$ funkcie w je

[1] preklad z angl. "path"z [3]

nulová;

(ii) pokiaľ imaginárna časť funkcie w, $\Im w(z)$, leží na novej ceste, tak je konštantná.

Ak napíšeme:

$$z = x + iy, \quad z_0 = x_0 + iy_0$$

$$w(z) = u(x, y) + iv(x, y), \quad w'(z_0) = 0$$

a označme symbolom S zobrazenie $(x, y) \longmapsto u(x, y)$, potom podľa Cauchy-Riemannových diferenciálnych rovníc

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \; a \; \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

to znamená, že bod (x_0, y_0) nemôže byť maximum alebo minimum, ale musí byť sedlovým bodom oblasti[2] S. Naviac, spádnice sú k vrstevniciam $u(x, y) =$ konšt. dané (opäť podľa Cauchy-Riemanových diferenciálnych rovníc) podľa kriviek $\Im w(z) = v(x, y) =$ konšt. Pretože cesty ležiace na oblasti S zodpovedajú spádniciam a vrstevniciam, sú cestami najstrmšieho spádu (prípadne výstupu), podmienka (ii) spomenutá vyššie znamená, že integrovanie pozdĺž cesty, kde $\Im w(z)$ je konštantná, implikuje, že sa pohybujeme pozdĺž ciest s najstrmším spádom zo sedlového bodu (x_0, y_0) na oblasti S.

Preto, keď prechádzame najstrmšou cestou cez sedlový bod, máme:

$$w(z) = u(x, y) + iv(x_0, y_0)$$

$$= u(x_0, y_0) + iv(x_0, y_0) - (u(x_0, y_0) - u(x, y)) = w(z_0) - \gamma(x, y) \tag{2.2}$$

kde γ je reálna funkcia

$$\gamma(x, y) = u(x_0, y_0) - u(x, y)$$

V prípade, že môžeme zmeniť integračnú cestu a vyjadriť integrál ako sumu integrálov pozdĺž cesty najstrmšieho spádu zo sedlového bodu, vyplýva z (2.1) a (2.2), že môžeme uvažovať len integrály tvaru

$$\int_P e^{v.w(z)} \xi(z) dz = \int_0^\infty \exp^{-v.\gamma} \xi(z) \frac{dz}{d\gamma} d\gamma.$$

Typickým príkladom, kde sa využívajú integrály v takomto tvare, je odvodenie aproximácie hustoty priemeru pomocou sedlového bodu.

[2]preklad z angl. "surface"z [3]

2.2 Aproximácia hustoty priemeru pomocou sedlového bodu

V tejto podkapitole ukážeme aplikáciu metódy najstrmšieho spádu a tiež aproximáciu k rozdeleniu priemeru n nezávislých rovnako rozdelených veličín.

Uvažujme náhodný výber $\boldsymbol{X} = (X_1, ..., X_n)$ z rozdelenia so známou distibučnou funkciu F a hustotou f. Naším cieľom je aproximovať hustotu f_n aritmetického priemeru

$$T_n(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Označme $M(\alpha)$ ako momentovú vytvárajúcu funkciu

$$M(\alpha) = \int_{-\infty}^{+\infty} e^{\alpha x} f(x)dx$$

a $K(\alpha)$ ako kumulatívnu vytvárajúcu funkciu

$$K(\alpha) = \log M(\alpha),$$

kde predpokladáme, že existujú pre α reálne z nejakého intervalu (c_1, c_2). Potom podľa Fourierovej inverznej transformácie môžeme hustotu $f_n(t)$ napísať ako

$$\begin{aligned} f_n(t) &= \frac{n}{2\pi}\int_{-\infty}^{+\infty} M^n(ir)e^{-inrt}dr \\ &= \frac{n}{2\pi i}\int_{\Im} M^n(z)e^{-ntz}dz \end{aligned} \tag{2.3}$$

kde $\Im$ je imaginárna os komplexnej roviny a vďaka tomu, že môžeme integrovať cez ľubovoľnú priamku rovnobežnú s imaginárnou osou, môžeme hustotu ďalej prepísať ako

$$\begin{aligned} f_n(t) &= \frac{n}{2\pi i}\int_{r-i\infty}^{r+i\infty} M^n(z)e^{-ntz}dz \\ &= \frac{n}{2\pi i}\int_{r-i\infty}^{r+i\infty} \exp(n[K(z) - zt])dz. \end{aligned} \tag{2.4}$$

V tomto bode máme integrál (2.4) vo forme základného tvaru (2.1), kde $v = n$, $w(z) = K(z) - z.t$, t je pevné, $\xi(z) = \frac{n}{2\pi i}$ a cesta P je rovnobežná s imaginárnou

osou prechádzajúca bodom r.

Z (i) (na str.7) vyplýva, že

$$w'(z) = K'(z) - t = 0$$

a nová cesta bude prechádzať sedlovým bod z_0, ktorý je riešením

$$K'(z_0) = t.$$

Daniels v [2] ukazuje, že pri všeobecných podmienkach je sedlový bod z_0 jediný a reálny na otvorenom intervale (c_1, c_2) a $K''(z_0) > 0$. Odteraz budeme označovať $z_0 = \alpha_0 \in \mathbb{R}$.

Podmienka (ii) požaduje aby $\Im w(z) \equiv$ konšt. cez novú cestu, takže $\Im w(z) \equiv \Im w(\alpha_0) = 0$, kde $w(\alpha_0)$ je reálne. Toto nám umožnuje upraviť P. Skonštruujeme malý kruh okolo sedlového bodu α_0 a nasledujeme cestu najstrmšieho spádu od sedlového bodu vo vnútri kruhu. Je potrebné rozdeliť pôvodný integrál (2.3) na súčet dvoch integrálov

$$f_n(t) = I_0 + I_1,$$

kde I_0 integrujeme vo vnútri kruhu a I_1 zas mimo kruhu. Podrobnejší postup je uvedený v [3].

Nakoniec prichádzame k nasledujúcemu asymptotickému rozvoju

$$f_n(t) = \left[\frac{n}{2\pi K''(\alpha_0)}\right]^{1/2} \exp\{n[K(\alpha_0) - \alpha_0 t]\} \times \left\{1 + \frac{1}{n}\left[\frac{1}{8}\lambda_4(\alpha_0) - \frac{5}{24}\lambda_3^2(\alpha_0)\right] + ...\right\},$$

kde α_0 je stanovený na základe rovnice pre výpočet sedlového bodu

$$K'(\alpha_0) = t,$$

a

$$\lambda_3(\alpha_0) = K'''(\alpha_0)/[K''(\alpha_0)]^{3/2}$$

$$\lambda_4(\alpha_0) = K^{(IV)}(\alpha_0)/[K''(\alpha_0)]^2$$

sú štandardné kritériá šikmosti a špicatosti. A týmto sa dostávame k hlavnému výrazu

$$g_n(t) = \left[\frac{n}{2\pi K''(\alpha_0)}\right]^{1/2} \exp\{n[K(\alpha_0) - \alpha_0 t]\}, \tag{2.5}$$

ktorý nazývame *aproximáciou (hustoty) pomocou sedlového bodu.*

2.3 Súvislosť s metódou združených rozdelení

Aproximáciu hustotu priemeru $f_n(t)$ sme odvodili pomocou metódy najstrmšieho spádu a aproximácie pomocou sedlového bodu. V tejto časti uvedieme aproximáciu pomocou idey združených hustôt a normálneho rozdelenia a ukážeme, že týmto spôsobom docielime rovnaký výsledok.

Asi najjednoduchšia a najčastejšie používaná aproximácia hustoty priemeru je aproximácia normálnym rozdelením. Funguje veľmi dobre okolo stredu rozdelenia, no na chvostoch už nie. Hlavnou myšlienkou teda bude precentrovať hustotu na náš bod záujmu t pomocou priemerov združenej hustoty a potom použiť aproximáciu normálneho rozdelenia na nami upravený problém. Uvedenú metódu a jej popis sme prevzali z [3]

Teda zavedieme konjugovanú (združenú) hustotu

$$h_t(x) = c(t)\exp(\alpha(t)(x-t))f(x),$$

kde $c(t)$ je zvolená tak, aby h_t bola hustota a $\alpha(t)$ je zvolená tak, že

$$\int (x-t)\exp(\alpha(t)(x-t))f(x)dx = 0 \quad \text{tj. } E_{h_t} = t. \tag{2.6}$$

Rozptyl označíme ako $\sigma^2(t)$, teda

$$\sigma^2(t) = c(t)\int (x-t)^2\exp(\alpha(t)(x-t))f(x)dx$$

Dôležitým bodom v tejto časti je zápis konjugovanej hustoty, v ktorom využijeme kumulatívnu vytvárajúcu funkciu

$$K(\alpha) = \log\int e^{\alpha x}f(x)dx$$

a sedlový bod, ktorý je riešením rovnice

$$K'(\alpha_0) = t.$$

Potom môžeme napísať

$$\frac{\int x\exp(\alpha_0 x)f(x)dx}{\int \exp(\alpha_0 x))f(x)dx} = t$$

alebo

$$\frac{\int (x-t)\exp(\alpha_0 x)f(x)dx}{\int \exp(\alpha_0 x))f(x)dx} = 0.$$

Vynásobením čitateľa aj menovateľa výrazom $\exp(-\alpha_0 t)$ dostaneme

$$\frac{\int (x-t)\exp(\alpha_0(x-t))f(x)dx}{\int \exp(\alpha_0(x-t)))f(x)dx} = 0$$

alebo

$$\int (x-t)\exp(\alpha_0(x-t))f(x)dx = 0.$$

Pri porovnaní s (2.6) vidíme, že $\alpha_0 = \alpha(t)$, to znamená, že $\alpha(t)$ je sedlovým bodom pre (pevný bod) t. Navyše

$$c^{-1}(t) = \int \exp(\alpha_0(t)(x-t))f(x)dx = \exp(K(\alpha(t)) - \alpha(t)t),$$

a preto

$$-\log c(t) = K(\alpha(t)) - \alpha(t)t = K(\alpha_0) - \alpha_0 t.$$

Podobne

$$\begin{aligned} K''(\alpha_0) &= K''(\alpha(t)) \\ &= \frac{\int \exp(\alpha(t)x)f(x)dx \int x^2 \exp(\alpha(t)x)f(x)dx - (\int x\exp(\alpha(t)x)f(x)dx)^2}{(\int \exp(\alpha(t)x)f(x)dx)^2} \end{aligned}$$

a navyše, keď vynásobíme čitateľa aj menovateľa výrazom $(e^{-\alpha(t)t})^2$ dostávame

$$K''(\alpha_0) = E_{h_t}X^2 - (E_{h_t}X)^2 = \sigma^2(t).$$

Keď to všetko zhrnieme, máme

$$\alpha(t) = \alpha_0,\ -\log c(t) = K(\alpha_0) - \alpha_0 t,\ \sigma^2(t) = K''(\alpha_0). \tag{2.7}$$

V ďalšom kroku uvážme $h_{t,n}$ ako hustotu priemeru pre h_t a to, ako súvisí s hustotou priemeru $f_n(t)$ pre f. Ako je odvodené v [3], môžeme usúdiť nasledovné:

$$f_n(t) = c^{-n}(t)h_{t,n}(t).$$

Táto upravená rovnica nám teda určuje vzťah aproximácie hustoty $h_{t,n}(t)$ k nami požadovanej aproximácii hustoty $f_n(t)$.

Posledným krokom je aproximovať hustotu $h_{t,n}$. [3] uvádza, že túto hustotu môžeme vyjadriť ako

$$h_{t,n} = \sqrt{\frac{n}{2\pi}}\frac{1}{\sigma(t)}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

a teraz nám už len stačí dosadiť do rovnice pre f_n vyššie

$$f_n = \sqrt{\frac{n}{2\pi}} \frac{c^{-n}(t)}{\sigma(t)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Dovedie nás to až k aproximácii pomocou sedlového bodu pre hustotu priemeru (small sample aproximation)

$$g_n(t) = \sqrt{\frac{n}{2\pi}} \frac{c^{-n}(t)}{\sigma(t)}.$$

V prípade že využijeme (2.7), vidíme, že dostávame úplne rovnaký výraz (2.5).

Kapitola 3

Všeobecné aproximácie pomocou sedlového bodu

V tejto kapitole použijeme aproximácie získané pre priemer, aby sme dosiahli aproximácie pre zložitejšie štatisitky. Najprv budeme uvažovať jednorozmerné M-odhady, tj. T_n je riešenie $\sum_{i=1}^{n} \psi(x_i, t) = 0$.

Naším cieľom je aproximovať hustotu odhadu. Hoci asymptotické výsledky sú dostupné vo viacerých prípadoch, zvyčajne nevieme, či tieto rozdelenia sú dobré aproximácie pre malý alebo stredný počet pozorovaní.

3.1 Jednorozmerné M-odhady

Ako spomíname už v prvej kapitole, M-odhad je definovaný ako riešenie T_n rovnice

$$\sum_{i=1}^{n} \psi(x_i, t) = 0, \tag{3.1}$$

pre pozorovania $x_1,\ x_2, ..., x_n$.

Označme teda hustotu $f_n(t)$ ako hustotu priemeru T_n, kde pozorovania sú nezávislé s hustotou f. Aby sme aproximovali $f_n(t)$ použijeme aproximáciu hustoty priemeru pomocou sedlového bodu, ako je odvodené v druhej kapitole.

Pre dosiahnutie aproximácie hrá veľmi dôležitú úlohu konjugovaná hustota. V našom prípade, keď uvažujeme M-odhady, konjugovaná hustota pre pevný bod t je definovaná ako

$$h_t(x, \theta) = c(t) \exp(\alpha(t)\psi(x, t)) f(x, \theta).$$

Vhodnou voľbou $\alpha(t)$ použijeme konjugovanú hustotu na centrovanie priemeru

T_n k t v zmysle toho, že $E_{h_t}(T_n) = t$. Umožnuje nám to použiť prvú časť Edgeworthovho rozvoja na aproximáciu hustoty f_n k t.

Nasledujúce predpoklady pre ψ a $f(x, \theta)$ kladieme za účelom rozvinutia aproximácie. D^n označuje n-tú deriváciu funkcie $\psi(x, \theta)$ podľa θ.

A.1 Rovnica $\sum_{i=1}^{n} \psi(X_i, t) = 0$ má jediné riešenie T_n a rovnica

$$\int \psi(x, t) e^{\alpha(t)\psi(x,t)} f(x, \theta) dx = 0$$

má jediné riešenie $\alpha(t)$.

A.2 Existuje otvorená podmnožina $U \subset R$ taká, že

(i) $F_\theta(U) = 1$ pre každú $\theta \in \Theta$

(ii) $D\psi(x, \theta)$, $D^2\psi(x, \theta)$, $D^3\psi(x, \theta)$ existujú.

A.3 Pre každý kompakt $K \subset \Theta$

(i) $\sup_{\theta_0 \in K} E_{\theta_0} \mid D^2\psi(X, \theta) \mid^4 < \infty$

(ii) existuje $\varepsilon > 0$ a také $sup_{\theta_0 \in K} E_{\theta_0}(max_{|\theta - \theta_0| \leq \varepsilon} \mid D^3\psi(X, \theta) \mid^3) < \infty$

A.4 $E_{\theta_0}\psi(X, \theta_0)$ a $A(\theta_0) = E_{\theta_0} D\psi(X, \theta_0) \neq 0$ pre každú $\theta_0 \in \Theta$

A.5 Funkcie $A(\theta)$ a $E_{\theta_0}[D^2\psi(X, \theta))]^2$ sú spojité na Θ.

A.6 $|D\psi(x, \theta)|$ je ohraničená hodnotou k.

Veta 3. *Ak platia predpoklady A.1-A.6, potom*

$$f_n(t_0) = c^{-n}(t_0) h_{t_0,n}(t_0),$$

kde $h_{t_0,n}(t)$ je hustota odhadu T_n pri združenej hustote $h_{t_0}(x, \theta)$ a

$$c^{-1}(t_0) = \int \exp(\alpha(t_0) f(x, \theta)) dx.$$

Dôkaz. Uvedený v [3] na strane 44. □

Pri využití tejto vety, aproximáciu $f_n(t_0)$ môžeme vypočítať priamo pomocou aproximácie $h_{t_0,n}(t_0)$. Je dôležité teda podotknúť, že pre každý bod t_0, používame rôznu združenú hustotu.

Veta 4. *(Battacharya a Ghosh,1978) Predpokladajme, že A.2-A.5 platia. Potom existuje taká postupnosť štatistík $\{T_n\}$, že pre každý kompakt $K \subset \Theta$*

$$\begin{aligned} \inf_{\theta_0 \in K} P_{\theta_0}(|T_n - \theta_0| &< d_0 n^{-1/2}(\log n)^{-1/2}, \text{ kde } T_n \text{ rieši (3.1))} \\ &= 1 - \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

kde d_0 môže závisieť na hodnote k.

Zhrnutie:

Nasledujúca väzba nám umožnuje napísať T_n ako odhadnutý priemer. Uvažujme teraz Taylorov rozvoj druhého stupňa z (3.1) okolo bodu θ_0:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i, T_n)$$

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} \psi(x_i, T_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\psi(x_i, \theta_0) + D\psi(x_i, \theta_0)(T_n - \theta_0) + \frac{1}{2}D^2\psi(x_i, \theta_0)(T_n - \theta_0)^2\right] + R_n(T_n), \end{aligned} \tag{3.2}$$

kde $R_n(T_n) = \mathcal{O}_p|T_n - \theta_0|^3 = \mathcal{O}_p(1/n)$.

V závere s využitím Viet 3 a 4 a tiež predpokladov A.1-A.5 zhrnieme výsledky v nasledujúcej vete.

Veta 5. *Ak T_n reprezentuje riešenie $\sum_{i=1}^{n} \psi(x_i, t) = 0$ a podmienky A.1-A.5 platia, potom hustota z T_n pripúšťa nasledujúce asymptotické rozšírenie:*

$$f_n(t_0) = \sqrt{\frac{n}{2\pi}} c^{-n}(t_0)\left[\frac{A(t_0)}{\sigma(t_0)} + \mathcal{O}(n^{-1})\right],$$

kde $\alpha(t_0)$ je riešenie rovnice

$$\int \psi(x, t_0) e^{\alpha\psi(x,t_0)} f(x, \theta) dx = 0,$$

$$c^{-1}(t) = \int \exp(\alpha\psi(x, t_0)) f(x), \tag{3.3}$$

$$\sigma^2(t_0) = E_{t_0}\psi^2(x, t_0), \tag{3.4}$$

$$A(t_0) = E_{t_0}[D\psi(x, t_0)], \tag{3.5}$$

kde E_{t_0} je stredná hodnota vzhľadom ku konjugovanej hustote

$$h_{t_0}(x) = c(t_0)\exp(\alpha(t_0)\psi(x,t_0))f(x).$$

Teda Veta 5 nám dáva nasledujúcu aproximáciu hustoty f_n:

$$g_n(t_0) = \sqrt{\frac{n}{2\pi}}c^{-n}(t_0)\frac{A(t_0)}{\sigma(t_0)}. \tag{3.6}$$

3.2 Viacrozmerné M-odhady

Vrátime sa opäť k M-odhadom, ale teraz, pre úplnosť tejto kapitoly, zvážime viacrozmerný prípad. M-odhad $\mathbf{T}_n$ pre $\theta = (\theta_1, \theta_2, ..., \theta_p)$ je riešením nasledujúceho systému (sústavy) rovníc:

$$\sum_{i=1}^{n} \psi_j(x, \mathbf{t}) = 0 \text{ pre } j = 1, ..., p. \tag{3.7}$$

Naším cieľom je, ako v predchádzajúcej podkapitole, aproximovať hustotu $f_n(\mathbf{t})$ priemeru $\mathbf{T}_n$. Odvodenie je podobne ako predtým, ale $\mathbf{T}_n$ je rozšírený ako viacrozmerný priemer. Parameter $\alpha(\mathbf{t})$, v tomto prípadne ako p-rozmerný vektor, získame ako riešenie

$$\int \psi_j(\mathbf{x},\mathbf{t}) \exp\left\{\sum_{j=1}^{p} \alpha_j \psi_j(\mathbf{x},\mathbf{t})\right\} f(x)dx = 0 \quad j = 1, ..., p.$$

Nasledujúce predpoklady sú viacrozmerná verzia predpokladov z podkapitoly 3.1 a ešte poznačme, že D_j znamená deriváciu podľa θ_j.

A.1 Systém rovníc (3.7) má jediné riešenie

A.2 Existuje otvorená podmožina $U \subseteq R^n$ taká, že

(i) pre každú $\theta \in \Theta$ platí $F_\theta(U) = 1$ a

(ii) derivácie $D_j\psi_r(x,\theta), D_kD_j\psi_r(x,\theta), D_lD_kD_j\psi_r(x,\theta)$ existujú pre $1 \leq r, j, k, l \leq p$.

A.3 Pre každý kompakt $K \subset \Theta$,

(i) pre $0 \leq j,\ k \leq p,\ 1 \leq r \leq p$,

$$\sup_{\theta_0 \in K} E_{\theta_0}|D_kD_j\psi_r(X,\theta_0)|^4 < \infty.$$

(ii) existuje také $\epsilon > 0$, že pre $1 \leq r,\ j, k, l \leq p$,

$$\sup_{\theta_0 \in K} E_{\theta_0}(\max_{|\theta-\theta_0|<\epsilon} |D_l D_k D_j \psi_r(X, \theta)|)^3 < \infty.$$

A.4 Pre každé $\theta_0 \in \Theta$

$$E_{\theta_0} \psi_r(X, \theta_0) = 0$$

a matice

$$A(\theta_0) = E_{\theta_0}\left[\frac{\partial \psi}{\partial \theta}(X, \theta_0)\right],$$

$$C(\theta_0) = E_{\theta_0}[\psi(X, \theta_0)\psi^T(X, \theta_0)]$$

nie sú singulárne.

A.5 Funkcie $A(\theta)$ a $E_\theta[(D_{k_1}, D_{j_1}, \psi_{r_1})(D_{k_2}, D_{j_2}, \psi_{r_2})]$, kde $0 \leq j_1, j_2, k_1, k_2 \leq p,\ k_1 + j_1 \geq 1,\ k_1 + j_1 \geq 1,\ 1 \leq r_1,\ r_2 \leq p$ sú spojité na Θ.

Veta 6. *Ak $\boldsymbol{T}_n$ reprezentuje riešenie sústavy rovníc (3.7) a podmienky A.1 až A.5 sú splnené, potom asymptotický rozvoj pre hustotu odhadu $\boldsymbol{T}_n$ je*

$$f_n(\boldsymbol{t}_0) = (n/2\pi)^{p/2} c^{-n}(\boldsymbol{t}_0)|A||\Sigma|^{-1/2}\{1 + \mathcal{O}(1/n)\},$$

kde $\alpha(t_0)$ je riešením rovnice

$$\int \psi_r(x, t_0) \exp\left\{\sum_{j=1}^{p} \alpha_j \psi_j(x, t_0)\right\} f(x)dx = 0 \text{ pre } r = 1, ..., p,$$

$$c^{-1}(t_0) = \int \exp\left\{\sum_{j=1}^{p} \alpha_j \psi_j(x, t_0)\right\} f(x)dx,$$

$$A = \left\{E\frac{\partial \psi(x, t)}{\partial t_r}\ |_{t=t_0}\right\}_{1 \leq r, j \leq p}, \quad \Sigma = \{E\psi_j(x, t_0)\psi_r(x, t_0)\}_{1 \leq r,\ j \leq p}$$

a všetky stredné hodnoty sú vypočítané vzhľadom ku konjugovanej hustote

$$h_{t_0}(x) = c(t_0) \exp\left\{\sum_{j=1}^{p} \alpha_j(\boldsymbol{t}_0)\psi_j(\boldsymbol{x}, \boldsymbol{t}_0)\right\} f(x).$$

Kapitola 4

Simulácie

4.1 Aproximácie hustoty pomocou sedlového bodu

V tejto podkapitole bude naším cieľom aproximovať hustotu priemeru vybraných rozdelení $f_n(t)$, kde využijeme poznatky z druhej kapitoly o sedlovom bode.

4.1.1 Normálne rozdelenie

Máme náhodný výber $X_1, .., X_n$ z rozdelenia $\mathcal{N}(\mu, \sigma^2)$. Vieme teda, že

$$f_n(t) = \sqrt{\frac{n}{2\pi\sigma^2}} \exp^{\frac{-n(t-\mu)^2}{2\sigma^2}}.$$

Momentová vytvárajúca funkcia je definovaná ako

$$M_x(z) = \int_{-\infty}^{\infty} e^{zx} f(x) dx,$$

a pomocou tohto vzťahu odvodíme momentovú vytvárajúcu funkciu pre normálne rozdelenie.

$$M_x(z) = \int_{-\infty}^{\infty} e^{zx} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \exp\left(\frac{-\mu^2}{2\sigma^2}\right) \int_{-\infty}^{\infty} e^{zx} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2+2x\mu}{2\sigma^2}}.$$

Argumenty exponenciál upravíme do jedného, zjednodušíme ho, doplníme na štvorec a využijeme vlastnosť hustoty

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) = 1.$$

Takže

$$\begin{aligned} zx + \frac{-x^2+2x\mu}{2\sigma^2} &= \frac{-x^2+2x(z\sigma^2+\mu)}{2\sigma^2} = \\ &= \frac{-(x^2-2x(z\sigma^2+\mu)+(z\sigma^2+\mu)^2)+(z\sigma^2+\mu)^2}{2\sigma^2} \end{aligned}$$

a

$$\begin{aligned} M_x(z) &= \exp\left(\frac{-\mu^2}{2\sigma^2}\right)\exp\left(\frac{(z\sigma^2+\mu)^2}{2\sigma^2}\right) = \\ &= \exp\left(\mu z + \frac{1}{2}\sigma^2 z^2\right). \end{aligned}$$

Kumulatívna vytvárajúca funkcia $K_x(z)$ má tvar

$$K_x(z) = \log(M_x(z)) = \mu z + \frac{1}{2}\sigma^2 z^2$$

a jej 1. a 2. deriváciu odvodíme jednoducho

$$K'_x(z) = \mu + \sigma^2 x,\ K''_x(z) = \sigma^2.$$

V nasledujúcom kroku využijme aproximáciu pomocou sedlového bodu, ktorú sme si odvodili v druhej kapitole:

$$g_n(t) = \left[\frac{n}{2\pi K''(\alpha_0)}\right]^{1/2} \exp\{n[K(\alpha_0) - \alpha_0 t]\}. \tag{4.1}$$

Pre $\mathcal{N}(0,1)$ hľadáme sedlový bod α_0 ako riešenie rovnice $K'(\alpha_0) = t$

$$=>\ \alpha_0(t) = t$$

ďalej

$$K(\alpha_0) = \mu\alpha_0 + \frac{1}{2}\sigma^2\alpha_0^2 = \frac{1}{2}\alpha_0^2 = K(t) = \frac{1}{2}t^2$$

a

$$K''(\alpha_0) = \sigma^2 = 1,$$

takže pre $\mathcal{N}(0,1)$ po dosadení do (4.1) máme:

$$g_n(t) = \sqrt{\frac{n}{2\pi}} e^{\frac{-nt^2}{2}} = f_n(t).$$

Všeobecne pre $\mathcal{N}(\mu, \sigma^2)$ hľadáme sedlový bod α_0 podľa rovnice $K'(\alpha_0) = t$

$$\mu + \sigma^2 \alpha_0 = t$$

$$=> \ \alpha_0 = \frac{t-\mu}{\sigma^2}$$

ďalej

$$K(\alpha_0) = \mu\alpha_0 + \frac{1}{2}\sigma^2\alpha_0^2 = K(\frac{t-\mu}{\sigma^2}) = .. = \frac{t^2-\mu^2}{2\sigma^2}$$

a

$$K''(\alpha_0) = \sigma^2,$$

takže pre $\mathcal{N}(\mu, \sigma^2)$ po dosadení do (4.1) máme:

$$g_n(t) = \sqrt{\frac{n}{2\pi\sigma^2}} e^{\frac{-n(t-\mu)^2}{2\sigma^2}}.$$

Z našich vypočtov vidíme, že pre normálne rozdelenie $\mathcal{N}(\mu, \sigma^2)$ je aproximácia hustoty priemeru rovná samotnej hustote priemeru.

4.1.2 Gamma rozdelenie

V nasledujúcom príklade aproximujeme hustotu priemeru gamma rozdelenia $f_n(t)$, kde výpočet je založený na rovnakom princípe ako v predchádzajúcom príklade.

Máme náhodný výber $X_1, .., X_n$ z rozdelenia $\Gamma(a, b)$, kde parametry $a \in \mathbb{R}^+$, $b \in \mathbb{R}^+$. Potom hustota gamma rozdelenia je definovaná ako $f(x) = \frac{b^{-a}}{\Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$, pre $x > 0$. Vieme, že pre náhodnú veličinu $Y = \sum_{i=1}^n X_i$ platí

$$Y \sim \Gamma(na, b)$$

a pre aritmetický priemer potom dostávame

$$\frac{1}{n} Y \sim \Gamma(na, b/n).$$

Hustotu priemeru teda vyjadríme ako

$$f_n(x) = \frac{(b/n)^{-na}}{\Gamma(na)} x^{na-1} e^{-\frac{nx}{b}}.$$

V ďalšom kroku si odvodíme momentovú vytvárajúcu funkciu, ktorá nám pomôže pri ďalšom výpočte aproximácie.

$$\begin{aligned} M_x(z) &= \frac{b^{-a}}{\Gamma(a)} \int_0^\infty e^{zx} e^{\frac{-x}{b}} x^{a-1} dx = \frac{b^{-a}}{\Gamma(a)} \int_0^\infty e^{(z-\frac{1}{b})x} x^{a-1} dx \\ &= \frac{b^{-a}}{\Gamma(a)} \frac{\Gamma(a)}{(\frac{1}{b}-z)^a} = (1-bz)^{-a}. \end{aligned}$$

Kumulatívna vytvárajúca funkcia $K_x(z)$ má potom tvar

$$K_x(z) = \log(M_x(z)) = -a\log(1-bz)$$

a jej 1. a 2. deriváciu vyjadríme nasledovne

$$\begin{aligned} K'_x(z) &= \frac{ab}{1-bz}, \\ K''_x(z) &= \frac{ab^2}{(1-bz)^2}. \end{aligned}$$

Pre $\Gamma(a,b)$ hľadáme sedlový bod α_0 ako riešenie rovnice $K'_x(\alpha_0) = t$

$$\frac{ab}{1-b\alpha_0} = t$$

$$=> \ \alpha_0(t) = \frac{t-ab}{bt}$$

ďalej

$$K_x(\alpha_0) = K_x\left(\frac{t-ab}{bt}\right) = -a\log\left(1-\frac{t-ab}{t}\right) = -a\log\left(\frac{ab}{t}\right)$$

a

$$K''_x(\alpha_0) = \frac{ab^2t^2}{a^2b^2} = \frac{t^2}{a},$$

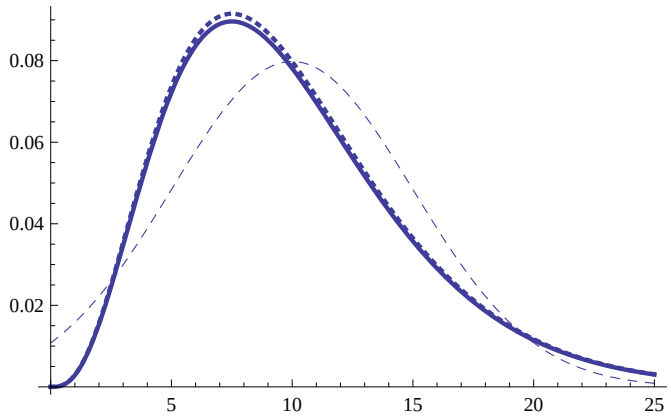
takže pre $\Gamma(a,b)$ po dosadení do (4.1) máme:

$$g_n(t) = \sqrt{\frac{na}{2\pi t^2}} \exp\left[n\left(-a\log\left(\frac{ab}{t}\right) - \frac{t-ab}{b}\right)\right],$$

čo môžeme ešte upraviť na tento výsledný tvar aproximácie:

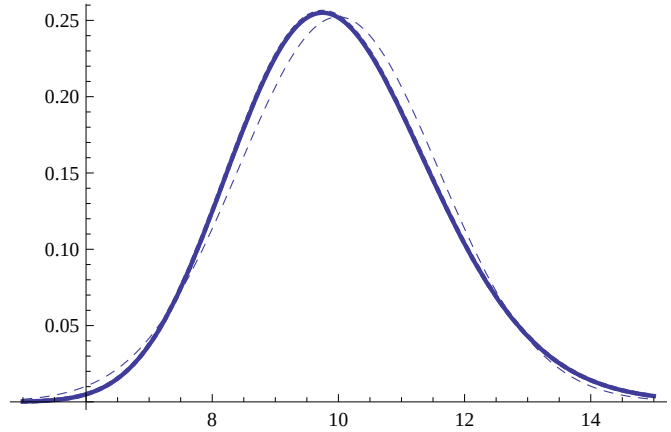
$$g_n(t) = \sqrt{\frac{na}{2\pi t^2}} \left(\frac{ab}{t}\right)^{-na} \exp\left[\frac{-n(t-ab)}{b}\right].$$

Na obrázku (4.1) môžeme vidieť ako vyzerá skutočná hustota priemeru gamma



Obr. 4.1: $n = 2$

rozdelenia (plná čiara), g_n ako aproximácia hustoty priemeru pomocou metódy sedlového bodu (bodkovane) a nakoniec aj hustota priemeru spočítaná pomocou centrálnej limitnej vety (čiarkovane; kap. 1.1). V tejto simulácii pracujeme s gamma rozdelením s parametrami $\alpha = 2$ a $\beta = 5$ a počet $n = 2$. Teda môžeme usúdiť že SSA[1] je omnoho presnejšia ako CLV[2] pre malé n, Totiž so zvyšujúcim sa počtom pozorovaní dokáže CLV presnejšie kopírovať skutočnú hustotu ako môžeme vidieť na obrázku (4.2), kde už $n = 20$.



Obr. 4.2: $n = 20$

[1]Small sample aproximation
[2]Centrálna limitná veta

4.2 Hustoty M-odhadov pre rôzne rozdelenia

V tejto podkapitole sa bude zaoberať hustotami vybraných M-odhadov pre rozdelenie $\mathcal{N}(0,1)$, gamma rozdelenie a studentovo rozdelenie. Všetky simulácie sú vypočítané v programe Wolfram Mathematica a je ich možné nájsť v prílohe.

4.2.1 Hustoty Huberovho odhadu

Z prvej kapitoly už vieme, že Huber navrhol funkciu ψ pre odhad ako

$$\psi(x,t) = \begin{cases} x - t & \dots |x-t| \leq k \\ k\ sgn(x-t) & \dots |x-t| > k. \end{cases}$$

Ďalej si vyjadríme deriváciu funkcie ψ, ktorú využijeme neskôr v našom výpočte.

$$\frac{\partial \psi(x,t)}{\partial x} = \begin{cases} 1 & \dots |x-t| \leq k \\ 0 & \dots |x-t| > k. \end{cases}$$

Hustota štandardizovaného normálneho rozdelenia $\mathcal{N}(0,1)$ je definovaná ako

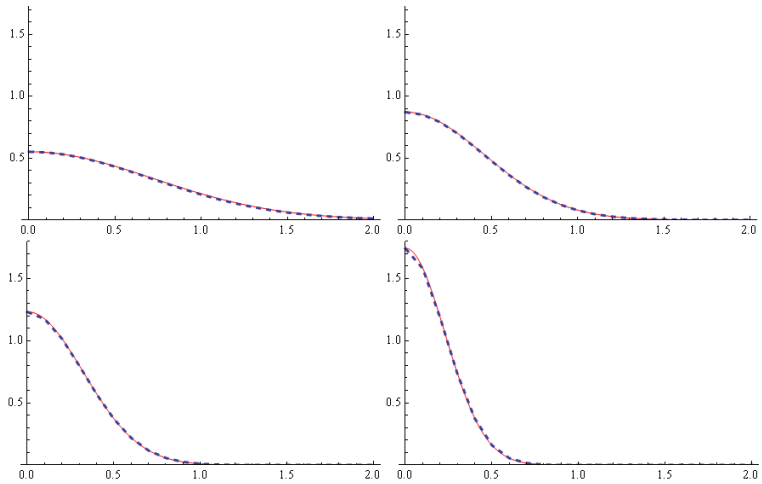
$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Jednotlivé kroky a oblasti numerickej integrácie závisia na tom ako rozdelenie vyzerá, keď uvážime rozdelenie pre konkrétne hustoty. Keď chceme aproximovať hustotu T_n podľa vzorca (3.6), musíme najprv nájsť $\alpha(t)$ ako riešenie rovnice

$$\int \psi(x,t) e^{\alpha(t)\psi(x,t)} f(x,\theta) dx = 0.$$

Pre všetky rozdelenia a odhady je možné využiť fakt, že $\alpha(t) = 0$ vždy pre $t = \theta$. To nám umožnuje dobrý počiatočný bod pre Newton-Rapsonovu metódu pre ostatné body t. Potom už len c, σ a A sú aproximované numerickou integráciou pri použití vzorcov (3.3), (3.4) a (3.5). Na záver stačí jednotlivé podvýpočty vložiť do (3.6).

Naším výstupom bude teda aproximácia hustoty Huberovho odhadu, ktorý pre $n = 2, 5, 10, 20$ a $k = 1.4$ vidíme na obrázku (4.3). Túto aproximáciu tu porovnáme s asymptotickým správaním Huberovho odhadu, ktoré odvodíme s využitím CLV a asymptotického rozptylu Huberovho odhadu (1.5). V prípade, že požadujeme hustotu Huberovho odhadu pre gamma rozdelenie alebo t-rozdelenie,



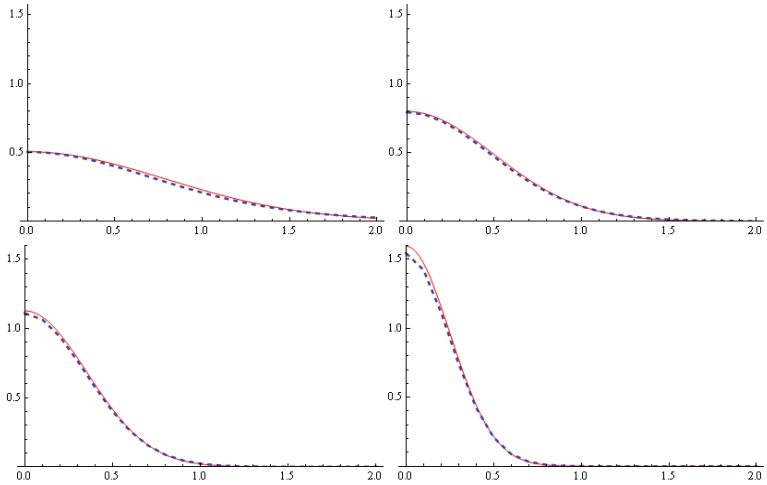
Obr. 4.3: hustoty Huberovho odhadu pre N(0,1)

postupujeme podobne. Teraz si vyjadríme hustotu gamma a t-rozdelenia.

$$f_2(x) = \frac{b^{-a}}{\Gamma(a)} x^{a-1} e^{-\frac{x}{b}}, \text{ pre } x > 0 \tag{4.2}$$

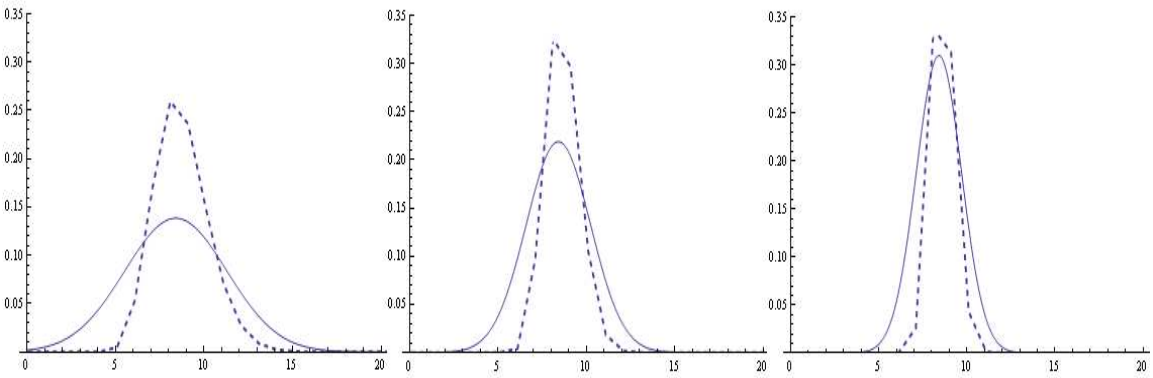
$$f_3(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \tag{4.3}$$

V naších simuláciách budeme aj naďalej pracovať s t-rozdelením so 7 stupňami volnosti a $\Gamma(2,5)$, a teda interpretované výsledky môžeme vidieť na obrázku (4.4) a (4.5). Pre gamma rozdelenie volíme $k = 1.345\sigma$ ako je uvedené v [7]. V grafe



Obr. 4.4: hustoty Huberovho odhadu pre t-rozdelnie so 7 stupňami voľnosti, $n = 2, 5, 10, 20$

(4.5) vidíme, že aplikovanie centrálnej limitnej vety je vhodné až pre veľký počet porozovaní.



Obr. 4.5: hustoty Huberovho odhadu pre $\Gamma(2,5)$, $n = 20, 50, 100$

4.2.2 Hustoty maximálne vierohodných odhadov

Pre maximálne vierohodný odhad, teda M-odhad platí, že

$$\psi(x) = -\frac{f'(x)}{f(x)}.$$

V prípade, že uvažujeme rozdelenie $\mathcal{N}(0,1)$, dostávame $\psi(x) = x$, čo je teda vhodne zvolená funkcia pre priemer ako M-odhad. A túto možnosť už máme vypočítanú.

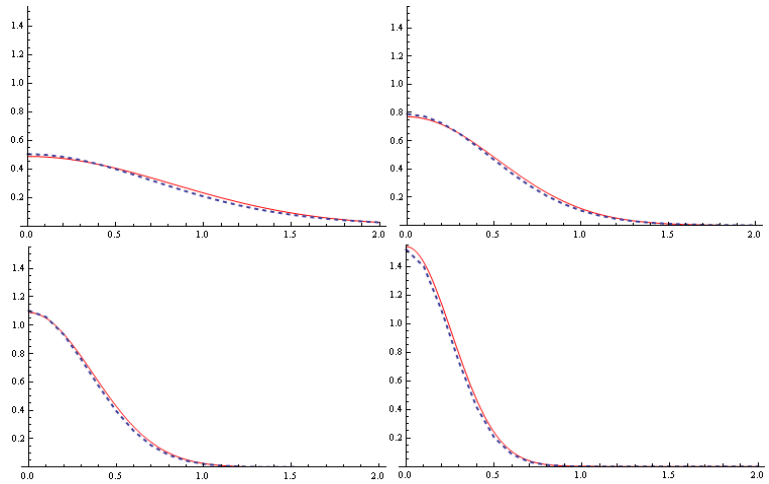
Máme teraz t-rozdelenie so 7 stupňami voľnosti a hustotou (4.3). Teda

$$\psi(x) = \frac{8x}{7 + x^2}$$

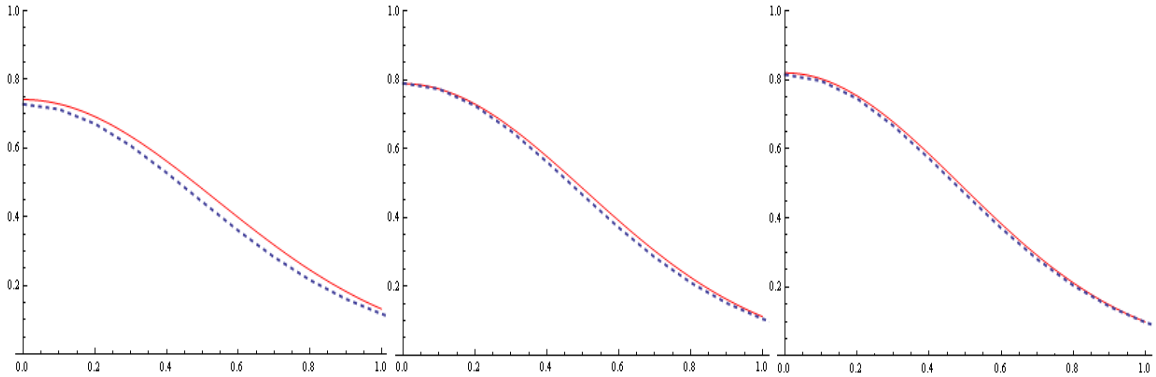
Ďalej môžeme vo výpočte pokračovať ako v predchádzajúcej podkapitole. Hustoty môžeme vidieť na obrázku (4.6), kde pre asymptické správanie maximálne vierohodného odhadu sme využili poznatky z podkapitoly (1.2.3). Hustoty vypočítané metódou sedlového bodu sú nakreslené prerušovanou čiarou a pomocou centrálnej limitnej vety červenou čiarou.

Pretože vieme, že hustota normálneho rozdelenia je dosť podobná t-rozdeleniu a navyše aproximácia sedlovým bodom a pomocou centrálnej limitnej vety je pre normálne rozdelenie rovnaká, ukážeme ešte ako sa bude správať napríklad hustota maximálne vierohodného odhadu pri rastúcom počte stupňov voľnosti v. Zvolíme $v = 4, 7, 10$.

Na obrázku (4.7) teda vidíme, že so zvyšujúcim sa počtom stupňov voľnosti sa asymptotické rozdelenie maximálne vierohodného odhadu (červenou čiarou) viac podobá hustote odhadu, ktorú sme spočítali pomocou metódy sedlového bodu (prerušovanou čiarou).



Obr. 4.6: hustoty maximálne vierohodného odhadu pre t-rozdelnie so 7 stupňami voľnosti, $n = 2, 5, 10, 20$



Obr. 4.7: hustoty maximálne vierohodného odhadu pre t-rozdelnie so 4,7 a 10 stupňami voľnosti, $n = 5$

Záver

V tejto práci sme sa venovali konkrétnemu problému, a to aproximácii hustoty odhadu pre rôzne rozdelenia s malým počtom pozorovaní. Popísali sme metódy a následne sme ich aplikovali na konkrétne, nami zvolené, odhady a rozdelenia.

Pri použití metódy sedlového bodu sme zistili, že aproximácia hustoty priemeru je dokonca presná pre normálne rozdelenie, prípadne sa mierne líši pre iné rozdelenia. Pre ďalšie odhady je táto metóda tiež veľmi vhodná, čo dokazujú aj naše výsledky.

Pri porovnaní použitia centrálnej limitnej vety a metódy sedlového bodu sme došli k záveru, že centrálna limitná veta má svoje využitie hlavne pri väčšom počte pozorovaní n.

Zoznam použitej literatúry

[1] P.J.Huber. *Robust estimation of a location parameter.* Ann.Math.Statist., 1964. 36,73-101.

[2] Daniels H.E. *Saddlepoint Approximations in Statistics* Ann.Math.Statist., Volume 25, Number 4, 631-650., 1954.

[3] Field C.A., Ronchetti E. *Small Sample Asymptotics.* Institute of Mathematical Statistics - Monograph Series, Hayward (CA), 1990. chapter 3.,4.

[4] Jurečková J., Sabolová R. *Finite-sample density and its small sample asymptotic approximation.* Statistics Probability Letters(2011), doi:10.1016/j.spl.2011.03.034

[5] Anděl J. *Základy matematické statistiky.* 2 vydání Praha:MATFYZPRESS, 2007. ISBN 80-7378-001-01

[6] Jurečkova J. *Robustní statistické metody* Univerzita Karlova v Praze - Nakladatelství Karolinum, Praha 2001, ISBN 80-246-0259-8

[7] John Fox *Robust Regression.* Appendix to An R and S-PLUS Companion to Applied Regression. January 2002

[8] Wolfram Mathematica 7: *Documentation Centre*

Prílohy

Na priloženom CD sa nachádzajú nasledujúce údaje

- Bakalárska práca vo formáte pdf
- Zdrojový kód programu Mathematica