Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁRSKA PRÁCA

Miroslav Dibák

Odhady distribučných a kvantilových funkcií

Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: Mgr. Luboš Prchal

Študijný program: matematika, finančná matematika

2007

Na tomto mieste by som chcel predovšetkým poďakovať vedúcemu bakalárskej práce Mgr. Lubošovi Prchalovi za vzácne rady a pripomienky pri vypracovaní tejto práce.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 9.8.2007

Obsah

Úvod **5**

1 Distribučná a kvantilová funkcia 6
1.1 Distribučná funkcia 6
1.2 Kvantilová funkcia 8
2 Empirické odhady 10
2.1 Empirická distribučná funkcia 10
2.2 Empirická kvantilová funkcia 14
3 Jadrové odhady 17
3.1 Jadrový odhad hustoty 17
3.2 Jadrový odhad distribučnej funkcie 19
3.3 Jadrový odhad kvantilovej funkcie 22
4 Simulácie 25

Záver **31**
Príloha **32**
Literatúra **34**

Názov práce: Odhady distribučních a kvantilových funkcí
Autor: Miroslav Dibák
Katedra: KPMS
Vedúci bakalárskej práce: Mgr. Luboš Prchal
e-mail vedúceho: prchal@karlin.mff.cuni.cz

Abstrakt: Odhadovanie distribučných a kvantilových funkcií nesie veľký význam v oblasti matematickej štatistiky. Po úvode tejto práce definujeme pojem náhodnej veličiny, distribučnej a kvantilovej funkcie. Ďalšia kapitola je venovaná empirickým odhadom distribučných a kvantilových funkcií a ich vlastnostiam. Následne si zavedieme jadrový odhad hustoty, distribučnej funkcie a kvantilovej funkcie, popíšeme strednú kvadratickú chybu týchto odhadov a voľbu šírky pásma. Na základe simulácií porovnáme odhady kvantilov pomocou rôznych jadrových odhadov.

Kľúčové slová: distribučná funkcia, kvantilová funkcia, empirický odhad, jadrový odhad, šírka pásma, simulácie

Title: Estimation of cumulative distribution and quantile functions
Author: Miroslav Dibák
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Luboš Prchal
Supervisor's e-mail address: prchal@karlin.mff.cuni.cz

Abstract: Cumulative distribution and quantile functions estimation has a great significance in a field of mathematical statistics. After the introduction we recall a notion of random variables, distribution and quantile functions. The next chapter is devoted to empirical distribution and quantile functions estimators and their properties. Consequently, we introduce the kernel density function, kernel distribution function and kernel quantile function estimators, we discuss the mean squared error of the estimators and methods to set up the corresponding bandwidths. We compare the different kernel quantile estimators by the means of simulations.

Keywords: distribution function, quantile function, empirical estimator, kernel estimator, bandwidth, simulations

Úvod

Určovanie pravdepodobnosti má veľa aplikácií v rôznych oblastiach ako je napríklad ekonómia, fyzika, medicína. Každý deň sa stretávame s názornými príkladmi z teórie pravdepodobnosti. Tak napríklad na autobusovej zastávke, kedy pozorujeme doby čakania na našu autobusovú linku, prídeme do obchodu a počet ľudí v obchode má opäť nejaké pravdepodobnostné rozdelenie. Stretávame sa teda s pojmom náhodnej veličiny.

Jendou z najzákladnejších charakteristík rozdelenia náhodnej veličniny je distribučná funkcia. Tá úplne popisuje rozdelenie pravdepodobnosti náhodnej veličiny. Avšak nie vždy, resp. málokedy túto distribučnú funkciu poznáme. Preto je potrebné nájsť spôsob, ako odhadnúť jej tvar a jej vlastnosti z napozorovaných dát. V tejto práci sa zaoberáme neparametrickými postupmi jej odhadovania. Vychádzame teda zo súboru napozorovaných dát. Jedným z možných spôsobov je odhadnúť túto funkciu empiricky. Takáto empirická distribučná funkcia má dobré vlastnosti, avšak jednou z nevýhod je jej nespojitosť. Preto sa v tejto práci zaoberáme aj spôsobom, akým je možné získať spojitú distribučnú funkciu. Prichádzame teda k pojmu jadrového vyhladzovania a jadrovej distribučnej funkcie ([3]). V tejto práci rozoberieme postup, akým sa jadrové vyhladzovanie prevádza.

Kým distribučnú funkciu odhadujeme pre určenie rozloženia pravdepodobnosti, zo štatistického hľadiska má veľký význam uvažovať obrátene, teda určitej konkrétnej pravdepodobnosti odhadnúť kvantil. Opäť si zavedieme pojem empirickej kvantilovej funkcie a jadrového vyhladzovania.

V simuláciách sa venujeme odhadom niektorých konkrétnych kvantilov. Na základe vygenerovaných dát odhadneme hodnoty týchto kvantilov štyrmi spôsobmi a porovnáme jednotlivé odhadnuté hodnoty so skutočnou hodnotou príslušného kvantilu.

Na záver ešte pripomeňme, že pomocou zdrojového kódu v prílohe tejto práce si čitateľ môže tieto odhady v programe Mathematica vyskúšať.

Kapitola 1

Distribučná a kvantilová funkcia

Na začiatok si zavedieme niekoľko základných pojmov. Predpokladajme teda, že $(\Omega, \mathcal{A})$ je merateľný priestor. Prvky množiny Ω nazývame *elementárnymi javmi* a značíme ich ω. Prvky $\sigma-$algebry $\mathcal{A}$ nazývame *javy* a značíme veľkými písmenami. *Pravdepodobnosť* P je definovaná ako miera na $\mathcal{A}$ s vlastnosťou $P(\Omega)=1$, teda P je množinová funkcia na $\mathcal{A}$ s vlastnosťami

(*i*) $P(A) \geq 0$, $A \in \mathcal{A}$

(*ii*) $P(\Omega)=1$, $P(\emptyset)=0$

(*iii*) $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

ak je $\{A_n\}$ postupnosť po dvoch disjunktných javov. Trojicu $(\Omega, \mathcal{A}, P)$ nazývame *pravdepodobnostný priestor.* Nech $\Omega = R$. $\sigma-$algebra generovaná systémom všetkých otvorených podmnožín v R sa značí $\mathcal{B}$ a jej prvky sa nazývajú *borelovské množiny v* R. Majme teda pevne daný pravdepodobnostný priestor $(\Omega, \mathcal{A}, P)$. Merateľnú reálnu funkciu $X : (\Omega, \mathcal{A}) \rightarrow (R, \mathcal{B})$ nazývame *náhodnou veličinou.*

1.1. Distribučná funkcia

Nech X je náhodná veličina. Potom jej *distribučnou funkciou* nazveme funkciu $F(x)$, ktorá je pre všetky reálne x definovaná vzťahom

$$F(x) = P(X \leq x), \quad \forall x \in \mathbb{R}$$

kde $P(X \leq x)$ značí pravdepodobnosť javu, že X nadobúda hodnoty menšie nanajvýš rovné x. Distribučná funkcia sa nazýva *diskrétna,* ak existuje konečná alebo spočetná postupnosť vcelku rôznych reálnych čísel $\{x_n\}_{n \in N_0}$, kde $N_0 \subseteq \mathbb{N}$, $\mathbb{N}$ je

množina prirodzených čísel, a odpovedajúca postupnosť kladných čísel $\{p_n\}_{n\in N_0}$ tak, že $\sum_{n\in N_O} p_n = 1$, $P(X = x_n) = p_n$ a

$$F(x) = \sum_{n\in N_0; x_n \le x} p_n \ , . \qquad \forall\, x \in \mathbb{R}. \tag{1.1}$$

Distribučná funkcia, pre ktorú platí rovnosť (1.1) teda zodpovedá diskrétnemu rozdeleniu. Ak existuje funkcia $f(x) \ge 0$ taká, že pre každé reálne x platí

$$F(x) = \int_{-\infty}^{x} f(t)\, dt\ , \qquad \forall\, x \in \mathbb{R}\ ,$$

potom hovoríme, že distribučná funkcia $F(x)$ zodpovedá spojitému rozdeleniu. Funkciu $f(x)$ nazývame *hustota* náhodnej veličiny X.

Pozrime sa na vlastnosti distribučnej funkcie. Distribučná funkcia $F(x)$ náhodnej veličiny X má tieto vlastnosti:

(*i*) $0 \le F(x) \le 1$ pre $x \in (-\infty\ , \infty)$,

(*ii*) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to+\infty} F(x) = 1$,

(*iii*) $F(x)$ je neklesajúca; pre $x_1 < x_2$ je $F(x_1) \le F(x_2)$,

(*iv*) nech $x_1 < x_2$, potom $P(x_1 < X \le x_2) = F(x_2) - F(x_1)$,

(*v*) $F(x)$ je sprava spojitá a má najviac spočetne mnoho bodov nespojitosti.

Graf každej distribučnej funkcie diskrétnej náhodnej veličiny má „schodovitý“ tvar a je nespojitý vo všetkých bodoch x_i. Príslušný „skok“ funkcie $F(x)$ v bode x_i je rovný pravdepodobnosti p_i.

Ak uvažujeme náhodnú veličinu X so spojitým rozdelením, tak jej distribučná funkcia je spojitá na intervale $(-\infty\ , \infty)$, teda na množine všetkých reálnych čísel.

1.2. Kvantilová funkcia

Rozdelenie náhodnej veličiny je úplne popísané distribučnou funkciou. Niekedy je ale užitočné uvažovať obrátene, teda hľadať hodnoty x, ktoré spĺňajú $P(X \le x) = p$, pre predom stanovené $p \in (0,1)$. Zavedieme si preto nasledujúci pojem.

Definícia 1.1. *Nech X je náhodná veličina s distribučnou funkciou F. $P-tý$ kvantil (alebo $100p-tý$ percentil) tejto náhodnej veličiny je akékoľvek reálne číslo x_p, pre ktoré platí*

$$\lim_{h \to 0_+} F(x_p - h) \equiv F(x_p^-) \le p \quad \text{a} \quad F(x_p) \ge p \quad \forall p \in (0,1). \quad (1.2)$$

Ak je distribučná funkcia náhodnej veličiny X rýdzo rastúca, $p-$tý kvantil je jednoznačne určený zo vzťahu

$$x_p = F^{-1}(p) \equiv Q(p)$$

a platí

$$F(F^{-1}(p)) = p \quad \text{a} \quad F^{-1}(F(x)) = x, \quad \forall p \in (0,1) \ a \ \forall x \in \mathbb{R}$$

Takúto funkciu $Q(p)$, $p \in (0,1)$ nazývame *kvantilovou funkciou* náhodnej veličiny X. Pre krajné body 0 a 1 platí $Q(0) = \sup\{x \mid F(x) = 0\}$ a $Q(1) = \inf\{x \mid F(x) = 1\}$. Avšak inverznú funkciu je možné vytvoriť len k spojitej rastúcej distribučnej funkcii, teda len pre spojité rozdelenia. V prípade diskrétnych a zmiešaných rozdelení existujú intervaly, na ktorých je distribučná funkcia konštantná, a preto inverzná funkcia F^{-1} neexistuje. Pre tieto prípady je potrebné definíciu kvantilovej funkcie zobecniť

$$Q(p) = \inf \left\{ x \mid F(x) \ge p \right\} \quad \text{pre } p \in (0,1).$$

Kvantilová funkcia je spojitá zľava a je neklesajúca na svojom definičnom obore.

Na záver ešte spomeňme, že pre niektoré štatisticky významné kvantily sa používajú samostatné názvy. Napríklad $0{,}25-$ty kvantil nazývame *dolný kvartil*, $0{,}75-$ty kvantil sa nazýva *horný kvartil* a $0{,}5-$ty kvantil sa nazýva *medián.* Kvantily ako napríklad $0{,}95-$ty, $0{,}99-$ty alebo $0{,}995-$ty sú významné ako kritické hodnoty pre testové štatistiky.

Kapitola 2

Empirické odhady

2.1. Empirická distribučná funkcia

Nech $X_1,\ldots,X_n$ sú nezávislé rovnako rozdelené náhodné veličiny s distribučnou funkciou $F(x)$, teda pre $\forall i$, X_i má distribučnú funkciu $F(x)$. Pre náhodný výber z rozdelenia s distribučnou funkciou $F(x)$, *empirická distribučná funkcia* (EDF), ktorú označujeme $F_n(x)$, je jednoducho podiel pozorovaní menších alebo rovných konkrétnej hodnote x, teda

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

kde

$$I(X_i \le x) = \begin{cases} 1 & ak\ X_i \le x, \\ 0 & inak. \end{cases}$$

Usporiadajme hodnoty $X_1,\ldots,X_n$ do neklesajúcej postupnosti, teda $X_{1,n} \le X_{2,n} \le \cdots \le X_{n,n}$. Alternatívne môžme EDF $F_n(x)$ vyjadriť

$$F_n(x) = \begin{cases} 0 & \text{ak}\quad x < X_{1,n} \\ i/n & \text{ak}\quad X_{i,n} \le x < X_{i+1,n},\ i = 1,\ldots,n-1 \\ 1 & \text{ak}\quad x \ge X_{n,n} \end{cases}$$

$F_n(x)$ je po častiach konštantná. Pokiaľ sú všetky hodnoty $X_1,\ldots,X_n$ od seba rôzne, potom v každej z nich má $F_n(x)$ skok o veľkosti prevrátenej hodnoty veľkosti výberu, teda $1/n$. Ak sa však hodnota X_i v súbore $X_1,\ldots,X_n$ vyskytuje práve

$k-$krát, potom $F_n(x)$ má v bode $x = X_i$ skok o veľkosti k/n. Avšak pre spojité rozdelenia má tento jav nulovú pravdepodobnosť. Pred zavedením vlastností EDF ešte poznamenajme, že EDF má všetky vlastnosti distribučnej funkcie diskrétnej náhodnej veličiny (aj v prípade, že skutočné rozdelenie je spojité).

Vieme diskutovať niekoľko štatistických vlastností EDF. Zavedieme si preto náhodnú veličinu $T_n(x) = n\,F_n(x)$. Táto náhodná veličina nám reprezentuje počet hodnôt vo výbere, ktoré sú menšie alebo rovné konkrétnej hodnote x.

Veta 2.1. *Pre akékoľvek pevné reálne číslo x má náhodná veličina $T_n(x)$ binomické rozdelenie s parametrami n a $F(x)$, teda $T_n(x) \sim Bi(n, F(x))$.*

Dôkaz: Označme $\delta_i(x) = I(X_i \le x)$. Náhodné veličiny $\delta_1(x), \delta_2(x), \ldots, \delta_n(x)$ sú nezávislé a rovnako rozdelené, každá s alternatívnym (Bernoulliho) rozdelením s parametrom p, kde $p = P[\delta_i(x) = 1] = P(X_i \le x) = F(x)$. Pretože $T_n(x) = \sum_{i=1}^{n} \delta_i(x)$ je suma n nezávislých, rovnako (alternatívne) rozdelených náhodných veličín, môžme ľahko ukázať, že $T_n(x)$ má binomické rozdelenie s parametrami n a $p = F(x)$. □

Z Vety 2.1. a použitím vlastností binomického rozdelenia dostávame nasledujúce dôsledky.

Dôsledok 2.1. *Stredná hodnota a rozptyl $F_n(x)$ sú*

(*i*) $E[F_n(x)] = F(x)$

(*ii*) $Var[F_n(x)] = \dfrac{F(x)[1-F(x)]}{n}$

Vzťah (*i*) dôsledku ukazuje, že $F_n(x)$ je nestranný odhad $F(x)$. Vzťah (*ii*) ukazuje, že rozptyl $F_n(x)$ konverguje k nule pre $n \to \infty$. Teda použitím Čebyševovej nerovnosti môžme ukázať, že pre akékoľvek pevné reálne číslo x, $F_n(x)$ je konzistentný odhad $F(x)$, respektíve $F_n(x)$ konverguje v pravdepodobnosti k $F(x)$, teda $F_n(x) \xrightarrow{p} F(x)$ pre $n \to \infty$, resp.

$$P\left[\lim_{n\to\infty} F_n(x) = F(x)\right] = 1.$$

V nasledujúcej vete sa dozvieme, že z dostatočne veľkého náhodného výberu môžeme získať ľubovoľne podrobnú informáciu o distribučnej funkcii, keďže empirická distribučná funkcia konverguje s pravdepodobnosťou 1 rovnomerne na celej reálnej osi k distribučnej funkcii, keď rozsah výberu vzrastá do nekonečna.

Veta 2.2 (Glivenko-Cantelli). *Nech* $X_1, X_2, \ldots, X_n$ *sú nezávislé, rovnako rozdelené náhodné veličiny s distribučnou funkciou* $F(x)$. *Nech ďalej* $F_n(x)$ *je empirická distribučná funkcia náhodného výberu* $X_1, \ldots, X_n$. *Označme*

$$D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right|.$$

Potom platí

$$P\left[\lim_{n\to\infty} D_n = 0\right] = 1.$$

Dôkaz: ([2], strana 339) Uvažujme distribučnú funkciu $F(x)$, ktorá nie je degenerovaná, teda existujú aspoň dva body x, pre ktoré pre každé $h > 0$ platí $F(x) < F(x+h)$. Vieme, že $F(x)$ a $F_n(x)$ sú neklesajúce a zprava spojité obmedzené funkcie, ktoré nadobúdajú hodnoty medzi 0 a 1. Označme $F(x+0)$ a $F_n(x+0)$ limity funkcií v bode x zľava. Zvoľme ľubovoľné N prirodzené a nech $k = 0,1,\ldots,N$. Rozdelíme interval $[0,1]$ na N rovnakých častí dĺžky $\frac{1}{N}$. Pretože $F(x)$ nemusí byť v ľubovoľnom bode spojitá, zvoľme $x_{N,k}$ najmenšie x, ktoré vyhovuje nerovnosti

$$F(x) \le \frac{k}{N} \le F(x+0).$$

Zrejme platí $-\infty = x_{N,0} < x_{N,1} \le x_{N,2} \le \cdots \le x_{N,N} \le +\infty$, pričom pre dostatočne veľké N je aspoň jedna z nerovností medzi $x_{N,1}$ a $x_{N,N}$ ostrá vzhľadom k tomu, že $F(x)$ má aspoň dva body rastu. Vyšetríme teda rozdiel $\left|F_n(x) - F(x)\right|$ a taktiež limitu tohoto rozdielu zprava len v bodoch $x_{N,k}$. Voľme

$$D_n^{(1)} = \max_{1 \le k \le N} \left| F_n(x_{N,k}) - F(x_{N,k}) \right|, \quad D_n^{(2)} = \max_{1 \le k \le N} \left| F_n(x_{N,k} + 0) - F(x_{N,k} + 0) \right|$$

a označíme $D_{n,MAX} = \max(D_n^{(1)}, D_n^{(2)})$. Pre $0 \le k \le N-1$ za predpokladu $x_{N,k} < x_{N,k+1}$ zrejme platí

$$F(x_{N,k+1}) - F(x_{N,k} + 0) \le \frac{1}{N}.$$

Teraz využijeme vlastnosť, že $F_n(x)$ je neklesajúca. Ak $x_{N,k} < x \le x_{N,k+1}$, potom platí

$$F_n(x) \le F_n(x_{N,k+1}) \le F(x_{N,k+1}) + D_{n,MAX} \le F(x) + \frac{1}{N} + D_{n,MAX}$$

a

$$F_n(x) \ge F_n(x_{N,k} + 0) \ge F(x_{N,k} + 0) - D_{n,MAX} \ge F(x) - \frac{1}{N} - D_{n,MAX}.$$

Predchádzajúce úvahy nás vedú k nerovnosti $D_n \le D_{n,MAX} - \frac{1}{N}$. Poznamenajme, že náhodné veličiny D_n a $D_{n,MAX}$ závisia na ω, teda predchádzajúce nerovnosti platia skoro určite. Ako sme už uviedli skôr, pre každé pevné x platí

$$P\left[\lim_{n \to \infty} F_n(x) = F(x)\right] = 1 \quad \text{a} \quad P\left[\lim_{n \to \infty} F_n(x+0) = F(x+0)\right] = 1.$$

Vidíme teda, že $D_{n,MAX}$ konverguje k 0 skoro určite, pre $n \to +\infty$, preto

$$P\left[\limsup_{n \to \infty} D_n > \frac{1}{N}\right] = 0$$

pre každé prirodzené N, odkiaľ plynie, že

$$P\left[\lim_{n \to \infty} D_n = 0\right] = 1.$$

K úplnosti dôkazu dodajme, že ak je distribučná funkcia $F(x)$ degenerovaná, potom platí $F_n(x) = F(x)$ pre každé $n > 1$ a každé reálne x. □

Ďalšou užitočnou vlastnosťou EDF je jej asymptotická normalita, daná v nasledujúcej vete.

Veta 2.3. *Pre* $n \to \infty$*, limitné pravdepodobnostné rozdelenie* $F_n(x)$ *je normované normálne rozdelenie, teda*

$$\lim_{n\to\infty} P\left\{ \frac{\sqrt{n}\,[F_n(x) - F(x)]}{\sqrt{F(x)[1-F(x)]}} \le t \right\} = \Phi(t)$$

Dôkaz: Použitím Vety 2.1., Dôsledku 2.1. a centrálnej limitnej vety dostávame, že rozdelenie veličiny

$$\frac{[\,n\,F_n(x) - n\,F(x)\,]}{\sqrt{n\,F(x)[\,1-F(x)]}} = \frac{\sqrt{n}\,[F_n(x) - F(x)]}{\sqrt{F(x)[\,1-F(x)]}}$$

sa približuje k normovanému normálnemu rozdeleniu pre $n \to \infty$. □

Empirická distribučná funkcia má teda dobré konvergenčné vlastnosti. Avšak jednou z nevýhod EDF je jej nespojitosť. V kapitole 3 si preto zavedieme jadrový odhad distribučnej funkcie, ktorým získame spojitú distribučnú funkciu.

2.2. Empirická kvantilová funkcia

Pretože kvantilová funkcia je inverzná k distribučnej funkcii a empirická distribučná funkcia je odhadom distribučnej funkcie, je prirodzené kvantilovú funkciu odhadovať invertovaním empirickej distribučnej funkcie. Nech teda $X_1, X_2, \ldots, X_n$ sú nezávislé rovnako rozdelené náhodné veličiny. Usporiadajme tieto hodnoty do postupnosti $X_{1,n} \le X_{2,n} \le \cdots \le X_{n,n}$. *Empirickú kvantilovú funkciu* (EQF) $\widetilde{Q}_n(p) := F_n^{-1}(p)$, $0 < p \le 1$, definujeme nasledovne:

$$\widetilde{Q}_n(p) = \begin{cases} X_{1,n} & \text{ak } 0 < p \le \frac{1}{n} \\ X_{2,n} & \text{ak } \frac{1}{n} < p \le \frac{2}{n} \\ X_{3,n} & \text{ak } \frac{2}{n} < p \le \frac{3}{n} \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ X_{n,n} & \text{ak } \frac{n-1}{n} < p \le 1 \end{cases}$$

Teda $\widetilde{Q}_n(p) = \inf \left\{ x \mid F_n(x) \ge p \right\}$. Alternatívna definícia EQF je teda

$$\widetilde{Q}_n(p) = X_{(\lfloor np \rfloor + 1)},$$

kde $\lfloor np \rfloor$ značí spodnú celú časť súčinu np. Z toho vyplýva, že empirické (tiež výberové) kvantily (ozn. $\widetilde{x}_p$) sú vlastne usporiadané hodnoty vo výbere. Napríklad, ak $n = 10$, odhad $0{,}3-$tieho kvantilu alebo $30-$teho percentilu je jednoducho $\widetilde{Q}_{10}(0{,}3) = X_{3,n}$, pretože $\frac{2}{10} < 0{,}3 \le \frac{3}{10}$. Avšak poznamenajme, že $0{,}25-$ty empirický kvantil, alebo $25-$ty percentil je taktiež rovný $X_{3,n}$, pretože $\frac{2}{10} < 0{,}25 \le \frac{3}{10}$.

Pozrime sa na konvergenčné vlastnosti empirických kvantilov. Ak je teda x_p určené vzťahom (1.2) jednoznačne, potom pri $n \to \infty$, $\widetilde{x}_p \to x_p$ s pravdepodobnosťou 1.

Tvrdenie 2.2. *Predpokladajme, že $F(x)$ má hustotu pravdepodobnosti $f(x)$, ktorá je spojitou funkciou x, že x_p je určené jednoznačne a $f(x_p) > 0$. Potom*

$$n^{1/2}(\widetilde{x}_p - x_p) \xrightarrow{d} X \sim N\left(0, \frac{p(1-p)}{[f(x_p)]^2}\right).$$

Poznamenajme, že užitím vzťahu

$$P(\widetilde{x}_p < x) = P(n F_n(x) \ge np),$$

kde položíme $x = x_p + t/\sqrt{n}$, dostávame

$$P\left\lfloor (\widetilde{x}_p - x_p)\sqrt{n} < t \right\rfloor =$$

$$P\left\{ \frac{\sqrt{n}}{\sqrt{p(1-p)}} \left[F_n(x_p + t/\sqrt{n}) - F(x_p + t/\sqrt{n}) \right] > \frac{\sqrt{n}}{\sqrt{p(1-p)}} \left[p - F(x_p + t/\sqrt{n}) \right] \right\}$$

$$= P\left\{ \left(\frac{n}{p(1-p)} \right)^{1/2} \left[F_n(x_p) - F(x_p) \right] \geq - \frac{t\, f(x_p)}{\sqrt{p(1-p)}} \right\}$$

$$\to P\left\{ X \geq - \frac{t\, f(x_p)}{\sqrt{p(1-p)}} \right\} = P\left\{ X < \frac{t\, f(x_p)}{\sqrt{p(1-p)}} \right\},$$

kde X je náhodná veličina s rozdelením $N(0,1)$. Tento výsledok plynie z centrálnej limitnej vety aplikovanej na veličinu $F_n(x_p)$. Pri odvodení vzťahu sme užili aj skutočnosť, že

$$\sqrt{n}\left\{ \left[F_n(x_p + t/\sqrt{n}) - F_n(x_p) \right] - \left[F(x_p + t/\sqrt{n}) - F(x_p) \right] \right\} \xrightarrow{p} 0,$$

pretože ľavá strana má nulovú strednú hodnotu a jej rozptyl konverguje pri $n \to \infty$ k nule. Z predpokladov o $F(x)$ ďalej plynie, že pre $n \to \infty$ platí

$$\sqrt{n}\left[p - F(x_p + t/\sqrt{n}) \right] \to -t\, f(x_p).$$

Kapitola 3

Jadrové odhady

V predošlej kapitole sme rozoberali empirické odhady distribučnej a kvantilovej funkcie. Vieme teda, že empirická distribučná funkcia F_n má dobré konvergenčné vlastnosti a podáva dobrý obraz o tom, ako skutočná distribučná funkcia približne vyzerá. Má ale niekoľko podstatných nevýhod. Jednou z takýchto nevýhod je napríklad nespojitosť. V tejto kapitole si preto zavedieme jadrový odhad distribučnej funkcie, ktorý vychádza z jadrového odhadu hustoty. U jadrových odhadov je potrebné vhodne zvoliť *jadro* a vyhladzovací parameter, ktorý tiež nazývame *šírka pásma.* Ukážeme, prečo je potrebné vhodne zvoliť šírku pásma a ukážeme si tiež jednu z možností jej voľby.

3.1. Jadrový odhad hustoty

Jadrový odhad distribučnej funkcie vychádza z jadrového odhadu hustoty. Nech $X_1,\ldots,X_n$ sú nezávislé, rovnako rozdelené náhodné veličiny s distribučnou funkciou F. Jadrový odhad hustoty $f(x)$ je definovaný ([3])

$$\hat{f}_h(x) = \frac{1}{nh_f}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h_f}\right), \qquad (3.1)$$

kde *jadro* $k(t)$ je nejaká nezáporná obmedzená funkcia s $k(t) = k(-t)$ pre všetky reálne t a predpokladáme, že $\int_{-\infty}^{\infty} k(t)\,\mathrm{d}t = 1$ a $\int_{-\infty}^{\infty} t^2\, k(t)\,\mathrm{d}t < +\infty$. Pri odhadovaní funkčnej hodnoty v konkrétnom bode x umiestníme stred jadra do bodu x a vplyv každého referenčného bodu závisí od jeho blízkosti k bodu x. Príspevky všetkých bodov sú sčítané do celkového odhadu.

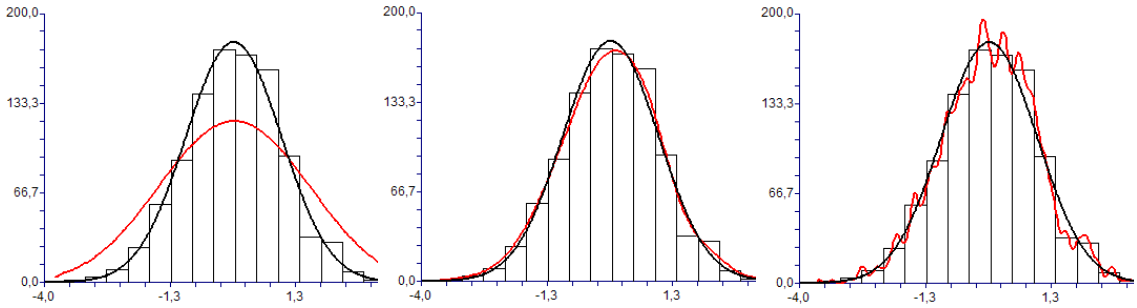
Poznamenajme, že ak za jadro k zvolíme indikátor $I(-1,1)$, dostávame histogram. Medzi najčastejšie používané jadrá patria jadrá uvedené v nasledujúcej tabuľke.

Jadro	Rovnica	Definičný obor
Rovnomerné	$\frac{1}{2}$	[-1,1]
Epanechnikov	$\frac{3}{4}(1-t^2)$	[-1,1]
Dvojváhové	$\frac{15}{16}(1-t^2)^2$	[-1,1]
Normálne	$\frac{1}{\sqrt{2\pi}}\mathrm{Exp}(-x^2/2)$	$\mathbb{R}$
Trojúholníkové	$1-\lvert t\rvert$	[-1,1]

Tabuľka 1. Niektoré bežne používané jadrá

Voľba jadra nie je z hľadiska vyhladzovania až tak veľmi podstatná. Podľa ([3]) je najoptimálnejšia voľba Epanechnikovho jadra.

Dôležitejšiu úlohu zohráva voľba vyhladzovacieho parametra h_f. Takzvaná *šírka pásma* h_f určuje, nakoľko hladká odhadovaná hustota bude: Väčšia hodnota h_f vedie k hladšej funkcii. Optimálna šírka závisí na počte pozorovaní n, platí $h_f \to 0$ pre $n \to \infty$, hustote f a jadre k. Bližšie sa voľbe optimálnej šírky pásma budeme venovať v ďalšom odstavci, kde odvodíme optimálnu šírku pásma pre jadrovú distribučnú funkciu. Pri voľbe šírky pásma h pre jadrový odhad hustoty sa napríklad pre normálne jadro používa tzv. „normal reference rule“ ([3]), kde položíme $h \approx 1{,}06\,\hat{\sigma}\, n^{-1/5}$. Na nasledujúcich grafoch vidíme, ako závisí vyhladená funkcia na šírke pásma h.



Graf 1. Jadrový odhad hustoty s voľbou neprimerane veľkej šírky ($5h_{opt}$) (vľavo), s voľbou optimálnej šírky h_{opt} (vstrede) a s voľbou neprimerane malej šírky (h_{opt} /5)(vpravo).

3.2. Jadrový odhad distribučnej funkcie

Obecne, integrovaním hustoty získavame distribučnú funkciu. Jadrový odhad distribučnej funkcie $F(x)$ odpovedajúci jadrovému odhadu hustoty je definovaný nasledovne:

$$\hat{F}(x) = \int_{-\infty}^{x} \hat{f}(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{nh_f} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_f}\right) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_f}\right), \qquad (3.2)$$

kde $K(x) = \int_{-\infty}^{x} k(t)\mathrm{d}t$. My si ukážeme, že vyhladzovací parameter pre jadrový odhad hustoty nie je vhodný pre jadrový odhad distribučnej funkcie, resp. pre jeho optimalitu. (ďalej namiesto h_f píšeme h).

Tradičnou mierou presnosti jadrového odhadu $\hat{F}(x)$ je *stredná kvadratická chyba* MSE definovaná

$$\begin{aligned} MSE(\hat{F}(x)) &= E[\hat{F}(x) - F(x)]^2 \\ &= E[\hat{F}(x) - E\hat{F}(x) + E\hat{F}(x) - F(x)]^2 \\ &= \left(E[\hat{F}(x)] - F(x)\right)^2 + Var[\hat{F}(x)] \\ &= Bias^2[\hat{F}(x)] + Var[\hat{F}(x)] \end{aligned}$$

Z toho vyplýva, že MSE je sumou umocnených vychýlok a rozptylu.

Strednú hodnotu vypočítame nasledovne

$$E[\hat{F}(x)] = \frac{1}{n} \sum_{i=1}^{n} E\left[K\left(\frac{x - X_i}{h}\right)\right] = \int K\left(\frac{x - y}{h}\right) f(y)\mathrm{d}y$$

Použijeme substitúciu $y = x+ht$ a následne v ďalšom kroku metódu per partes

$$K(-t)\,F(x + ht)\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} k(t)\,F(x + ht)\,\mathrm{d}t$$

Vzhľadom k limitným vlastnostiam $K(t)$ a $F(t)$ môžeme prvý člen zanedbať. Použitím Taylorovho rozvoja funkcie $F(x+ht)$ dostávame

$$\int_{-\infty}^{\infty} k(t)\left[F(x) + htF'(x) + \frac{h^2t^2}{2}F''(x) + \mathrm{O}(h^2) \right] \mathrm{d}t$$
$$= \int_{-\infty}^{\infty} k(t)\,F(x)\,\mathrm{d}t + \int_{-\infty}^{\infty} k(t)\,ht\,F'(x)\,\mathrm{d}t + \int_{-\infty}^{\infty} k(t)\frac{h^2t^2}{2}F''(x)\,\mathrm{d}t + \mathrm{O}(h^2)$$

Z vlastnosti jadra $\int t\,k(t)\,\mathrm{d}t = 0$ vyplýva, že druhý člen je rovný nule a dostávame

$$K(t)\,F(x)\Big|_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty} t^2 k(t)\,\mathrm{d}t\; f'(x)\,h^2 + \mathrm{O}(h^2)$$

Označíme $\mu_{2,K} = \int_{-\infty}^{\infty} t^2\,k(t)\,\mathrm{d}t$ a dostávame

$$E[\hat{F}(x)] = F(x) + \frac{\mu_{2,K}}{2} f'(x)\,h^2 + \mathrm{O}(h^2)$$

Rozptyl jadrovej distribučnej funkcie vypočítame podobným spôsobom a dostávame

$$Var[\hat{F}(x)] = E[\hat{F}(x)]^2 - \left\{E[\hat{F}(x)]\right\}^2$$
$$= \frac{F(x)[\,1-F(x)]}{n} - \frac{h}{n}C_{1,K}\,f(x) + \mathrm{O}(h/n)$$

kde $C_{1,K} = \int K(t)(1-K(t)) > 0$.

Môžme si teda porovnať strednú hodnotu a rozptyl empirickej distribučnej funkcie a jadrovej distribučnej funkcie. Zatiaľ čo stredná hodnota EDF je skutočná distribučná funkcia, u strednej hodnoty jadrovej distribučnej funkcie nám pribudlo vychýlenie. Môžme teda vidieť, že pre zvyšujúce sa h sa stredná hodnota zvyšuje (resp. znižuje, to závisí na znamienku $f'(x)$). Rozptyl sa pre zvyšujúce sa h zmenší a naopak pre znižujúce sa h sa rozptyl zväčší, ako je znázornené aj v grafe 1.

Aby sme odhadli celkovú presnosť odhadu pre všetky x, kvadratickú chybu integrujeme cez reálnu os a odpovedajúca miera nazývaná ako *stredná integrovaná kvadratická chyba* MISE, je definovaná ako

$$MISE(\hat{F}(x)) = \int E[\hat{F}(x) - F(x)]^2\,\mathrm{d}x$$
$$= \int (\,E[\hat{F}(x)] - F(x))^2\,\mathrm{d}x + \int Var[\hat{F}(x)]\mathrm{d}x.$$

Po dosadení strednej hodnoty a rozptylu teda dostávame, že

$$MISE(\hat{F}(x)) = \frac{1}{n}\int F(x)[1-F(x)]\,\mathrm{d}x - \frac{h}{n}\int K(t)[1-K(t)]\,\mathrm{d}t + \frac{h^4}{4}\left\{\int t^2 k(t)\,\mathrm{d}t\right\}^2 \int (f'(x))^2\,\mathrm{d}x + \mathrm{O}(h/n + h^4)\,.$$

MISE je funkciou šírky h. Pre optimálne vyhladenie je teda potrebné nájsť optimálnu šírku h tak, aby MISE vyjadrujúca chybu odhadu nadobudla čo možno najmenšie hodnoty. Je teda potrebné nájsť

$$h_{opt} = \arg\min_h MISE(h)\,.$$

Položíme teda prvú deriváciu MISE rovnú nule, dostávame

$$-\frac{1}{n}\int K(t)[1-K(t)]\,\mathrm{d}t + h^3\left\{\int t^2 k(t)\,\mathrm{d}t\right\}^2 \int [f'(x)]^2\,\mathrm{d}x = 0,$$

z tohto vzťahu vyjadríme h a dostávame vzťah pre optimálnu šírku pásma, teda

$$h_{opt} = \left\{\frac{\int K(t)[1-K(t)]\,\mathrm{d}t}{n\left\{\int t^2 k(t)\,\mathrm{d}t\right\}^2 \int [f'(x)]^2\,\mathrm{d}x}\right\}^{1/3} = \left\{\frac{C_{1,K}}{\mu_{2,K}^2 \int [f'(x)]^2\,\mathrm{d}x}\right\}^{1/3} n^{-1/3}. \quad (3.3)$$

Túto hodnotu ale v praxi nájsť nevieme, pretože nepoznáme hustotu. Tú ale môžme odhadnúť jadrovým odhadom. Označme si preto $R(f') = \int [f'(x)]^2\,\mathrm{d}x$. Tento vzťah môžme upraviť nasledovne

$$R(f') = \int [f'(x)]^2\,\mathrm{d}x = \int f'(x) f'(x)\,\mathrm{d}x\,,$$

použitím metódy per partes dostávame

$$\int f'(x) f'(x)\,\mathrm{d}x = f(x) f'(x)\Big|_{-\infty}^{\infty} - \int f''(x) f(x)\,\mathrm{d}x$$

Na základe limitných vlastností $f(x)$ a $f'(x)$ môžme prvý člen zanedbať a druhý člen je strednou hodnotou $f''(x)$, dostávame

$$\int f''(x) f(x)\,\mathrm{d}x = E[f''(x)] = R(f').$$

Dostávame teda upravený vzťah pre odhad optimálnej šírky pásma

$$\hat{h}_{opt} = \left(\frac{C_{1,K}}{\mu_{2,K}^2 \hat{R}(f')} \right)^{1/3} n^{-1/3}, \qquad (3.4)$$

kde $R(f')$ odhadneme ako

$$\hat{R}(f') = -\frac{1}{n^2 \beta_1^3} \sum_{i=1}^{n} \sum_{j \neq i} \phi^{(2)} \left(\frac{X_i - X_j}{\beta_1} \right) - \frac{1}{(2\pi)^{1/2} n \beta_1^3},$$

$$\phi^{(2)}(t) = (2\pi)^{-1/2} (t^2 - 1) \exp(-t^2/2).$$

Jednou z alternatív voľby β_1 je $n^{-0.3}$. Odhad (3.4) nazývame *plug-in* odhadom.

Ďalším z možných spôsobov voľby optimálnej šírky je použiť takzvané *krosvalidačné* kritérium, ktoré je definované nasledovne ([5])

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \int [I(x - X_i) - \hat{F}_{n;-i}(X_i)]^2 \, \mathrm{d}x,$$

kde $\hat{F}_{n;-i}$ je jadrový odhad určený z pozorovaní X_i, ale konštruovaný z dát vynechaním pozorovaní X_i. Existuje niekoľko ďalších odhadov optimálnej šírky (viď [4] a [5]). Tieto odhady ale nepracujú správne. Zvolia veľmi malú hodnotu šírky pásma, a preto odhady nie sú vhodne vyhladené.

3.3. Jadrový odhad kvantilovej funkcie

Jedným z možných spôsobov, ako odhadovať kvantily je použiť jadrový odhad distribučnej funkcie s dosadením do obecného tvaru kvantilovej funkcie, teda

$$\hat{Q}_1(p) = \inf \left\{ x \mid \hat{F}(x) \geq p \right\}.$$

Ako sa presvedčíme v simuláciach, tento odhad často nemusí byť dostatočne presný. Preto si zavedieme jadrový odhad kvantilovej funkcie. Nech teda $X_1, X_2, \ldots X_n$ sú

nezávislé, rovnako rozdelené náhodné veličiny. Usporiadajme tieto hodnoty do postupnosti $X_{1,n} \le X_{2,n} \le \cdots \le X_{n,n}$. Jadrový odhad kvantilovej funkcie je definovaný

$$\hat{Q}_2(p) = \int_0^1 \widetilde{Q}_n(t)\,\mathrm{d}_t K\left(\frac{t-p}{h_n}\right). \qquad (3.5)$$

Tento odhad môžme vyjadriť nasledovne

$$\begin{aligned}\hat{Q}_2(p) &= \int_0^1 \frac{1}{h_n} k\left(\frac{t-p}{h_n}\right)\widetilde{Q}_n(t)\,\mathrm{d}t \\ &= \sum_{i=1}^{n} X_{i,n} \frac{1}{h_n}\int_{(i-1)/n}^{i/n} k\left(\frac{t-p}{h_n}\right)\mathrm{d}t\,.\end{aligned}$$

Podobne ako pri jadrovom odhade distribučnej funkcie vieme diskutovať o strednej kvadratickej chybe $\mathrm{MSE}(\hat{Q}(p))$. Nech teda Q'' je spojitá v okolí bodu p a nech je jadro $k = K'$ symetrické okolo 0. Potom pre každé pevné $p \in (0,1)$,

$$\begin{aligned}MSE(\hat{Q}(p)) &= Var(\hat{Q}(p)) + Bias^2(\hat{Q}(p)) \\ &= \frac{p(1-p)}{n}\left[Q'(p)\right]^2 - 2\frac{h}{n}\left[Q'(p)\right]^2 \int_{-\infty}^{\infty} u\,k(u)K(u)\,\mathrm{d}u \\ &\quad + \frac{h^4}{4}\left[Q''(p)\right]^2\left[\int_{-\infty}^{\infty} u^2 k(u)\,\mathrm{d}u\right]^2 + \mathrm{O}(h/n) + \mathrm{O}(h^4).\end{aligned}$$

Z toho vyplýva, že pre každé $p \in (0,1)$, asymptoticky optimálna šírka pásma h_{opt} je

$$h_{opt}(p) = \alpha(k)\,\beta(Q)\,n^{-1/3},$$

kde

$$\alpha(k) = \left[\ 2\int_{-\infty}^{\infty} u\,k(u)\,K(u)\,\mathrm{d}u \Big/ \left\{\int_{-\infty}^{\infty} u^2 k(u)\,\mathrm{d}u\right\}^2\right]^{1/3}$$

a $\beta(Q) = \left[Q'(p)/Q''(p)\right]^{2/3}$.

Asymptoticky optimálna šírka $h_{opt}(p)$ teda závisí na prvej a druhej derivácii kvantilovej funkcie, ktorú nepoznáme. Avšak tieto derivácie je možné opäť odhadnúť. Problémom je, že tieto odhady opäť závisia na ich optimálnej šírke pásma.

Alternatívnou a podstatne jednoduchšou voľbou šírky h je použitie vzťahu ([8])

$$h = \left[\frac{p(1-p)}{n+1} \right]^{1/2}. \qquad (3.6)$$

Ako jadrový odhad pomocou takejto voľby šírky h pracuje sa presvedčíme v našich simuláciach.

Na záver ešte spomeňme, že v praxi sa často používa aproximácia odhadu (3.5) vyjadrená vzťahom

$$\hat{Q}_3(p) = \sum_{i=1}^{n} k\left(\frac{i-\frac{1}{2}}{n} - p \right) X_{i,n} \Big/ \sum_{j=1}^{n} k\left(\frac{j-\frac{1}{2}}{n} - p \right).$$

Tento odhad tiež porovnáme s ostatnými odhadmi v našich simuláciách.

Kapitola 4

Simulácie

Naše simulácie sú zamerané na porovnanie efektívnosti fungovania štyroch odhadov kvantilových funkcií (resp. kvantilov). Prvým odhadom je empirická kvantilová funkcia, teda

$$(1) \quad \widetilde{Q}_n(p) = X_{(\lfloor np \rfloor + 1)}.$$

Druhým odhadom je invertovanie jadrovej distribučnej funkcie, teda

$$(2) \quad \hat{Q}_1(p) = \inf \left\{ t \;\middle|\; \hat{F}(t) > p \right\}.$$

Tieto dva odhady porovnáme s odhadmi

$$(3) \quad \hat{Q}_2(p) = \int_0^1 \widetilde{Q}_n(t)\, \mathrm{d}_t\, K_n(p,t) \quad \text{a}$$

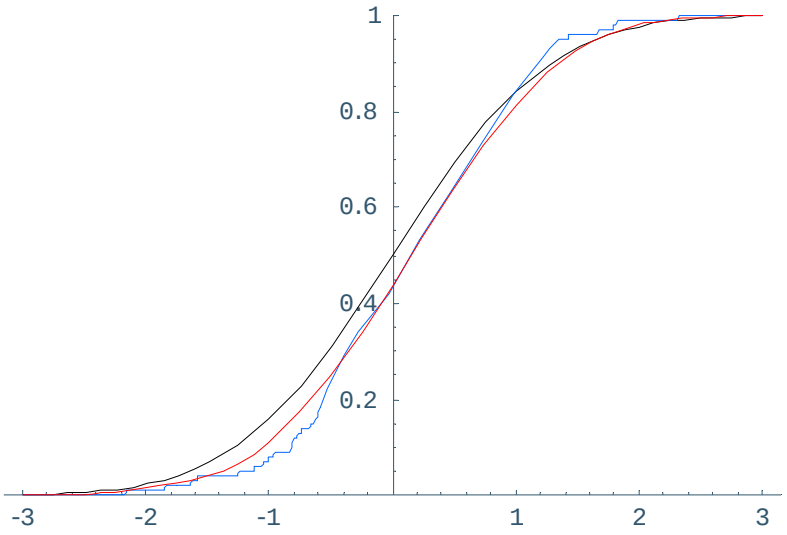
$$(4) \quad \hat{Q}_3(p) = \sum_{i=1}^{n} k\left(\frac{i-\frac{1}{2}}{n} - p \right) X_{i,n} \Big/ \sum_{j=1}^{n} k\left(\frac{j-\frac{1}{2}}{n} - p \right).$$

V druhom odhade sme vychádzali z jadrového odhadu distribučnej funkcie. Pri tomto odhade sme volili Epanechnikovo jadro, $\frac{3}{4}(1-t^2)$ pre $|t| \le 1$. Vzhľadom k odhadu distribučnej funkcie použijeme jeho integrál

$$K(t) := \begin{cases} 0 & pre\ t < -1 \\ \frac{1}{2} + \frac{3}{4}t - \frac{1}{4}t^3 & pre\ |t| \le 1 \\ 1 & pre\ t > 1 \end{cases}$$

Z hľadiska voľby vyhladzovacieho parametra h pre KDFE sme vychádzali z [4], [5] a plug-in odhadu zavedeného v kapitole 3.2. Najspoľahlivejšie pracoval plug-in odhad. Pre porovnanie, v nasledujúcom grafe vidíme skutočnú distribučnú funkciu, KDFE so šírkou h zvolenou kritériom LNO (viď Sarda, 1993) a KDFE so šírkou h

zvolenou plug-in odhadom uvedeným v kapitole 3. KDFE sú konštruované z výberu o veľkosti 100 z N(0,1) rozdelenia.



Graf 2.: Skutočná distribučná funkcia rozdelenia N(0,1) (čierna), KDFE so šírkou *h* zvolenou kritériom LNO (modrá) a KDFE so šírkou *h* zvolenou plug-in odhadom (červená).

Pri našom treťom odhade sme opäť použili Epanechnikovo jadro. Pri voľbe vyhladzovacieho parametra h sme vychádzali z [8], za vyhladzovací parameter sme teda dosadili vzťah (3.6) uvedený v kapitole 3. Poznamenajme, že za podmienky $hn^{5/6} \to \infty$ pre pevné $p \in (0,1)$ sú odhady (3) a (4) asymptoticky ekvivalentné ([8]). Táto podmienka je pri voľbe šírky h podľa vzťahu (3.6) splnená.

V našich simuláciach vychádzame zo štyroch rozdelení uvedených v nasledujúcej tabuľke.

Tabuľka 2. Distribučné funkcie použité v simuláciách

Rozdelenie	**Distribučná funkcia**
Normované normálne	N(0,1)
Exponenciálne	Exp(1)
Logaritmicko-normálne	LN(0,1)
t-studentovo (2 stupne voľnosti)	t_2

Pre každé z týchto štyroch rozdelení sme vygenerovali 500 výberov veľkosti 25, 50, 100 a 250. Z každého výberu sme našimi štyrma odhadmi vypočítali hodnoty

kvantilov 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95 a 0.99 a pre každý kvantil sme vypočítali priemernú hodnotu z vypočítaných 500 hodnôt. Tieto hodnoty nájdeme v nasledujúcich tabuľkách. Tmavo vyznačené hodnoty sú najpresnejšie odhady hodnoty konkrétnych kvantilov.

Tabuľka 3. Odhady kvantilov N(0,1)

N(0,1)									
n	**kvantil**								
	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
	skutočná hodnota								
	-2,326	-1,644	-1,281	-0,674	0	0,674	1,281	1,644	2,326
25	-1,938	-1,561	**-1,286**	**-0,657**	-0,027	0,629	1,239	1,505	1,938
	-2,525	-1,903	-1,502	-0,813	**-0,021**	0,770	1,458	1,852	**2,521**
	-1,938	-1,561	**-1,286**	**-0,657**	-0,027	0,629	1,239	1,506	1,938
	-1,938	**-1,712**	-1,308	-0,704	-0,024	**0,670**	**1,258**	**1,657**	1,938
50	-2,246	-1,625	-1,228	**-0,678**	0,015	0,662	1,316	1,619	**2,259**
	-2,543	-1,838	-1,439	-0,764	-0,008	0,749	1,420	1,811	2,527
	-2,249	**-1,656**	-1,243	-0,680	0,017	0,669	1,329	1,620	2,251
	-2,249	-1,673	**-1,307**	-0,681	**-0,008**	**0,670**	**1,277**	**1,637**	2,252
100	-2,144	-1,584	-1,224	-0,655	0,011	0,693	1,307	1,696	2,483
	-2,513	-1,793	-1,402	-0,742	**-0,004**	0,736	1,396	1,786	2,492
	-2,144	-1,601	-1,262	-0,662	0,009	0,693	1,308	1,701	2,471
	-2,334	**-1,655**	**-1,293**	**-0,679**	**-0,004**	**0,676**	**1,282**	**1,655**	**2,304**
250	-2,297	-1,635	-1,269	-0,668	0,005	0,676	1,287	1,648	2,310
	-2,408	-1,722	-1,343	-0,706	0,002	0,711	1,350	1,729	2,429
	-2,297	-1,632	-1,264	-0,668	0,007	**0,675**	1,297	**1,646**	2,304
	-2,311	**-1,636**	**-1,277**	**-0,669**	**0,001**	**0,675**	**1,286**	1,647	**2,323**

Tabuľka 4. Odhady kvantilov Exp(1)

Exp(1)									
n	**kvantil**								
	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
	skutočná hodnota								
	0,010	0,051	0,105	0,287	0,693	1,386	2,302	2,995	4,605
25	**0,038**	0,084	0,127	0,326	0,712	**1,365**	2,329	**2,836**	3,904
	-0,740	-0,379	-0,168	0,249	0,793	1,517	2,443	3,220	**4,771**
	0,038	0,078	**0,122**	0,320	**0,703**	1,348	**2,291**	2,832	3,795
	0,038	**0,064**	**0,122**	**0,305**	0,708	1,411	2,346	3,168	3,795
50	**0,019**	**0,060**	0,124	**0,295**	0,721	1,397	2,419	3,020	4,524
	-0,382	-0,160	-0,007	**0,295**	0,736	1,433	2,360	3,065	4,745
	0,019	0,061	0,124	**0,295**	0,716	**1,383**	2,420	**2,992**	**4,543**
	0,019	**0,060**	**0,113**	0,296	**0,697**	**1,389**	**2,330**	3,035	**4,543**
100	0,019	0,060	0,114	0,301	0,710	1,421	2,373	3,123	5,185
	-0,216	-0,052	0,063	0,301	0,703	1,396	2,313	3,012	**4,713**
	0,019	0,060	0,114	0,294	**0,695**	1,399	2,350	3,095	5,232
	0,014	**0,055**	**0,109**	**0,287**	0,687	**1,380**	**2,306**	**3,007**	4,718
250	**0,011**	**0,053**	0,109	0,290	0,699	1,392	2,331	**2,994**	4,590
	-0,113	0,009	0,097	0,296	0,700	1,395	2,310	3,007	4,705
	0,011	**0,053**	0,109	**0,289**	0,697	**1,387**	2,321	**2,996**	4,568
	0,011	**0,053**	**0,107**	**0,289**	**0,693**	1,388	**2,306**	3,005	**4,610**

Tabuľka 5. Odhady kvantilov LN(0,1)

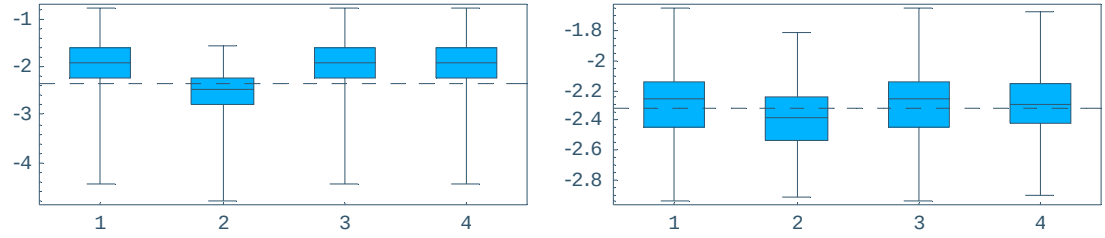
LN(0,1)									
n	**kvantil**								
	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
	skutočná hodnota								
	0,097	0,193	0,277	0,509	1	1,963	3,602	5,180	10,240
25	**0,161**	0,230	0,301	0,561	1,038	**1,957**	**3,674**	4,977	**8,070**
	-0,740	-0,342	-0,080	0,443	1,157	2,235	3,976	5,592	8,059
	0,161	0,233	0,299	0,541	**1,027**	1,944	3,748	**4,995**	8,063
	0,161	**0,208**	**0,297**	**0,522**	1,035	2,040	3,890	6,066	8,063
50	0,115	0,205	0,302	0,517	1,029	1,974	3,674	4,977	**11,093**
	-0,408	-0,109	0,093	0,496	1,073	2,106	3,766	5,393	8,327
	0,111	**0,196**	0,301	**0,515**	1,021	**1,963**	3,842	**5,245**	**11,093**
	0,111	0,197	**0,283**	0,516	**0,999**	1,973	**3,669**	5,369	**11,093**
100	0,124	0,210	0,294	0,527	1,026	2,023	3,790	5,604	13,963
	-0,171	0,056	0,209	0,523	1,057	2,100	3,836	5,486	9,061
	0,118	0,209	0,289	0,522	1,025	2,024	3,790	5,633	13,855
	0,102	**0,198**	**0,281**	**0,515**	**1,014**	**2,000**	**3,714**	**5,416**	**11,405**
250	0,103	**0,196**	0,283	**0,513**	1,010	1,967	3,688	5,266	10,488
	-0,031	0,141	0,260	0,519	1,012	2,184	4,019	5,563	9,958
	0,103	**0,196**	0,283	0,514	1,006	**1,964**	3,651	**5,212**	**10,433**
	0,101	**0,196**	**0,280**	0,514	**1,001**	**1,964**	**3,617**	5,232	10,703

Tabuľka 6. Odhady kvantilov t_2

t_2									
n	**kvantil**								
	0,01	0,05	0,1	0,25	0,5	0,75	0,9	0,95	0,99
	skutočná hodnota								
	-6,964	-2,919	-1,885	-0,816	0	0,816	1,885	2,919	6,964
25	-5,778	-2,794	**-1,964**	-0,805	-0,004	0,785	2,007	2,840	**7,340**
	-5,010	-3,220	-2,167	-0,988	-0,010	0,965	2,132	3,117	5,709
	-6,290	**-2,837**	-1,973	**-0,816**	**-0,003**	**0,793**	**2,002**	**2,844**	7,342
	-6,290	-4,043	-2,085	-0,889	-0,006	0,858	2,103	4,415	7,342
50	-8,463	**-3,103**	**-1,849**	-0,845	0,028	0,835	2,100	3,089	7,972
	-5,979	-3,168	-2,072	-0,923	**0,009**	0,931	2,066	3,120	5,800
	-8,463	-3,111	-1,810	**-0,818**	0,042	**0,820**	2,067	**3,016**	**7,948**
	-8,463	-3,255	-1,967	-0,819	0,010	0,826	**1,958**	3,187	**7,948**
100	-6,205	-2,787	-1,833	-0,794	0,019	0,850	1,968	3,162	11,525
	-6,540	-3,068	-2,001	-0,891	**-0,0002**	0,886	1,978	**3,038**	**6,454**
	-6,275	**-2,851**	**-1,871**	**-0,809**	0,010	0,841	1,979	3,182	12,106
	-9,076	-3,071	-1,952	-0,832	-0,002	**0,825**	**1,925**	3,041	9,088
250	**-7,128**	**-2,948**	-1,858	-0,809	**0,007**	**0,820**	1,926	2,975	7,343
	-7,386	-3,014	-1,945	-0,852	0,010	0,871	1,955	3,007	7,772
	-7,361	-2,974	-1,867	-0,810	0,017	0,832	1,932	**2,959**	**7,220**
	-7,718	-2,994	**-1,902**	**-0,811**	0,011	0,833	**1,912**	2,985	7,580

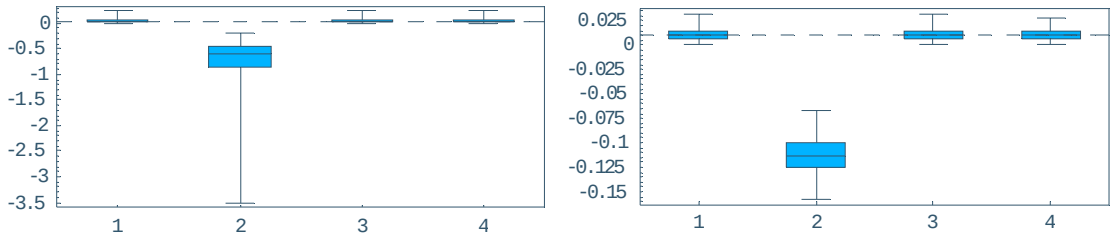
Na základe týchto pozorovaní vieme konštatovať, že odhad (2) až na niekoľko výnimočných prípadov pracuje podstatne horšie ako ostatné odhady. Pri odhadovaní kvantilov normovaného normálneho rozdelenia N(0,1) pracuje odhad (4) lepšie ako odhad (3) a (1). Avšak ako sme už spomínali skôr, za podmienky $hn^{5/6} \to \infty$ pre pevné $p \in (0,1)$ sú odhady (3) a (4) asymptoticky ekvivalentné, čo

je pri voľbe parametra h podľa vzťahu (3.6) splnené. Nasledujúce grafy porovnávajú 500 hodnôt kvantilu 0,01 rozdelenia N(0,1) vypočítaných odhadom (1),(2),(3) a (4).



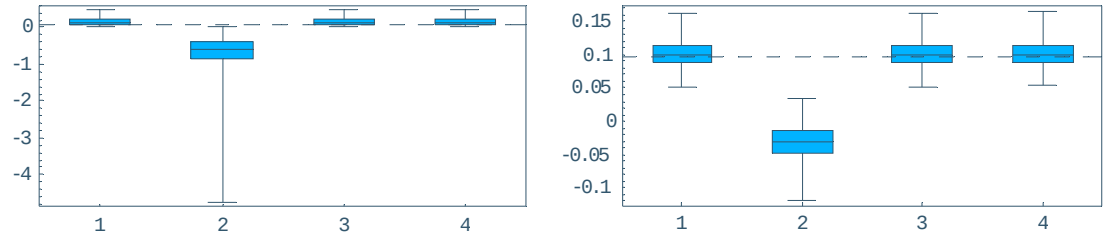
Graf 3.: Simulované hodnoty vypočítané odhadom (1),(2),(3) a (4) rozdelenia N(0,1) veľkosti výberu n=25 (vľavo) a veľkosti výberu n=250 (vpravo) kvantilu p=0,01.

V prípade veľkosti výberu n=25 bola hodnota kvantilu 0,01 odhadnutá presnejšie odhadom (2). Na základe grafu vidíme, že rozptyl simulovaných hodnôt je u odhadu (2) v prípade veľkosti výberu n=250 nižší. V nasledujúcom grafe vidíme hodnoty kvantilu 0,01 exponenciálneho rozdelenia Exp(1) odhadovaného štyrma spôsobmi pre veľkosť výberu n=25 a n=250.



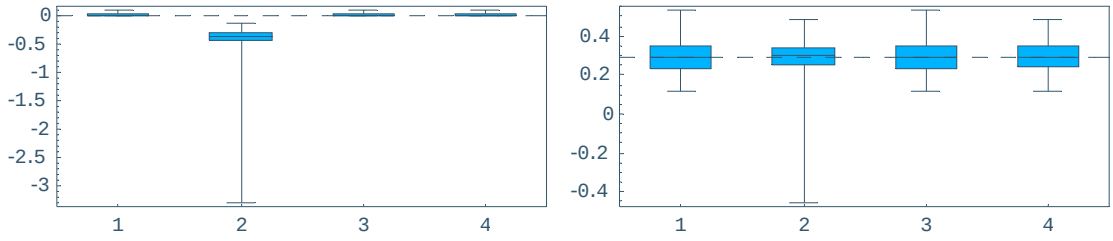
Graf 4.: Simulované hodnoty vypočítané odhadom (1),(2),(3) a (4) rozdelenia Exp(1) veľkosti výberu n=25 (vľavo) a veľkosti výberu n=250 (vpravo) kvantilu p=0,01.

Pri exponenciálnom rozdelení Exp(1) dokonca vidíme, že hodnoty kvantilu 0,01 nadobúdajú záporné hodnoty. Pri veľkosti výberu n=250 je rozptyl nasimulovaných hodnôt podstatne menší, no simulované hodnoty sú stále záporné. Podobne je to u logaritmicko-normálneho rozdelenia, ako ukazujú nasledujúce grafy.



Graf 5.: Simulované hodnoty vypočítané odhadom (1),(2),(3) a (4) rozdelenia LN(0,1) veľkosti výberu n=25 (vľavo) a veľkosti výberu n=250 (vpravo) kvantilu p=0,01.

V nasledujúcich grafoch si ešte porovnáme odhady kvantilov 0,01 a 0,25 u exponenciálneho rozdelenia Exp(1) pre veľkosť výberu n=50.



Graf 6.: Simulované hodnoty vypočítané odhadom (1),(2),(3) a (4) rozdelenia Exp(1) veľkosti výberu n=50 pre kvantil 0,01 (vľavo) a kvantil 0,25 (vpravo).

Na základe nasimulovaných hodnôt vidíme, že odhad (2) pre rozdelenia Exp(1) a LN(0,1) nepracuje správne. Súvisí to s vyhladzovacím parametrom h. Hoci skutočná distribučná funkcia týchto rozdelení nie je pre záporné hodnoty x definovaná, hodnoty blízke nule majú pri jadrovom odhade distribučnej funkcie predsalen nejakú váhu aj pre záporné hodnoty x blízko nuly.

Na záver ešte poznamenajme, že u logaritmicko-normálneho rozdelenia odhad (2) ani raz neodhadol hodnotu príslušného kvantilu najpresnejšie z našich štyroch odhadov. Pri rozdeleniach s nosičom $\mathbb{R}$ pracuje tento odhad lepšie, avšak odhady (1),(3) a (4) sú stále presnejšie.

Záver

V tejto práci sme sa venovali odhadom distribučných a kvantilových funkcií. Popísali sme si empirické a jadrové odhady a ich vlastnosti.

V simuláciách sme sa venovali štyrom možným spôsobom odhadovania kvantilov a na základe týchto simulácií môžme povedať, že odhad kvantilov pomocou invertovania jadrovej distribučnej funkcie nepracoval správne, najmä u exponenciálneho a logaritmicko-normálneho rozdelenia. Je to spôsobené tým, že vyhladzovací parameter jadrovej distribučnej funkcie volíme globálne pre všetky hodnoty x. Naopak je to u jadrovej kvantilovej funkcie, kde vyhladzovací parameter závisí na hodnote p. V porovnaní empirického odhadu a jadrového odhadu kvantilov pracovali lepšie jadrové odhady, avšak empirický odhad je v porovnaní s jadrovými odhadmi tiež dostatočne presný, resp. hodnoty kvantilov odhadnuté empiricky sú tiež blízko hodnôt skutočných kvantilov. V porovnaní jadrových odhadov pracoval najlepšie odhad (4). Avšak rozdiely medzi odhadovanými kvantilmi odhadmi (3) a (4) sú minimálne. Vieme teda povedať, že odhady kvantilov pomocou jadrovej kvantilovej funkcie pracujú najlepšie.

Z praktického hľadiska je pri odhadovaní kvantilov najjednoduchšie použiť empirickú kvantilovú funkciu. Hoci jadrové odhady pracujú trochu lepšie, je potrebné uvažovať o voľbe jadra a vyhladzovacieho parametra, zatiaľ čo u empirickej kvantilovej funkcie potrebujeme len usporiadať dáta do neklesajúcej postupnosti.

Príloha

V prílohe uvádzame zdrojový kód v programe Mathematica použitý v simuláciách.

```
Needs["Statistics`  "]
Rozdelenie  := NormalDistribution  [0, 1]
EDF[x_, data_] := Total[UnitStep[x - data]] / Length[data]
EQF[p_, data_] := Sort[data][[Floor[p * Length[data]] + 1]]
```

definícia Epanechnikovho jadra:

```
K[t_] := Piecewise[{{0, t < -1}, {∫_-1^t 3/4 (1 - x^2) dx, Abs[t] ≤ 1}, {1, t > 1}}]
k[t_] := Piecewise[{{3/4 (1 - t^2), Abs[t] ≤ 1}, {0, Abs[t] > 1}}]
```

definícia jadrovej distribučnej funkcie a voľby šírky *h* plug-in odhadom:

```
KDFE[x_, data_, h_] := 1/Length[data] Sum_{i=1}^{Length[data]} K[(x - data[[i]])/h]
Hopt[data_] :=
 Abs[
  (∫_-1^1 K[t] (1 - K[t]) dt / (∫_-1^1 t^2 (∂t K[t]) dt)^2) /
   ((-1/((Length[data])^2 ((Length[data])^-0.3)^3)
     Sum_{i=1}^{Length[data]} ((Sum_{j=1}^{i-1} (2 π)^-1/2 ((data[[i]] - data[[j]])/(Length[data])^-0.3)^2 - 1) Exp[-((data[[i]] - data[[j]])/(Length[data])^-0.3)^2/2]) +
     Sum_{j=i+1}^{Length[data]} (2 π)^-1/2 ((data[[i]] - data[[j]])/(Length[data])^-0.3)^2 - 1) Exp[-((data[[i]] - data[[j]])/(Length[data])^-0.3)^2/2]) -
    1/((2 π)^1/2 Length[data] ((Length[data])^-0.3)^3)) Length[data])]^(1/3)
```

definícia jadrových odhadov kvantilových funkcií, kvantilov, simulácií, výpočty jednotlivých hodnôt kvantilov podľa jednotlivých odhadov a vykreslenie grafov:

$$KQFE1[p\_, data\_] := \int_0^1 EQF[p, data] \left(\partial_t K\left[\frac{t-p}{\left(\frac{p(1-p)}{(n+1)} \right)^{1/2}} \right] \right) dt$$

$$KQFE2[p\_, data\_] := \left(\sum_{i=1}^{Length[data]} k\left[\frac{\frac{i-\frac{1}{2}}{Length[data]} - p}{\sqrt{\left(\frac{p(1-p)}{Length[data]+1} \right)}} \right] Sort[data][[i]] \right) \Big/ \left(\sum_{j=1}^{Length[data]} k\left[\frac{\frac{j-\frac{1}{2}}{Length[data]} - p}{\sqrt{\left(\frac{p(1-p)}{Length[data]+1} \right)}} \right] \right)$$

```
Kvantily = {0.01 , 0.05 , 0.1 , 0.25 , 0.5 , 0.75 , 0.9 , 0.95 , 0.99 };

PocetSimulaci = 500 ;
VelkostVyberu = 25 ;
Pokusy = Table [Random [Rozdelenie ], {PocetSimulaci }, {VelkostVyberu }];

OptimH = Map [Hopt , Pokusy ];

Hodnoty1 = Table [EQF [Kvantily [[j]], Pokusy [[i]]],
   {i, 1, PocetSimulaci }, {j, 1, Length [Kvantily ]}];
Hodnoty2 = Table [x /. FindRoot [KDFE [x, Pokusy [[i]], OptimH [[i]]] == Kvantily [[j]], {x, 0}],
   {i, 1, PocetSimulaci }, {j, 1, Length [Kvantily ]}];
Hodnoty3 = Table [KQFE1 [Kvantily [[j]], Pokusy [[i]]],
   {i, 1, PocetSimulaci }, {j, 1, Length [Kvantily ]}];
Hodnoty4 = Table [KQFE2 [Kvantily [[j]], Pokusy [[i]]],
   {i, 1, PocetSimulaci }, {j, 1, Length [Kvantily ]}];

{Mean [Hodnoty1 ], Mean [Hodnoty2 ], Mean [Hodnoty3 ], Mean [Hodnoty4 ]} // MatrixForm
{Variance [Hodnoty1 ], Variance [Hodnoty2 ], Variance [Hodnoty3 ], Variance [Hodnoty4 ]}
 // MatrixForm

Table [DisplayTogether [BoxWhiskerPlot [Transpose [Hodnoty1 ][[j]],
    Transpose [Hodnoty2 ][[j]], Transpose [Hodnoty3 ][[j]], Transpose [Hodnoty4 ][[j]],
    BoxStyle → Hue [0.55 ]],
  Plot [x /. Flatten [NSolve [CDF [Rozdelenie , x] == Kvantily [[j]], x]], {k, 0, 5},
    PlotStyle → Dashing [{0.03 }]]], {j, 1, Length [Kvantily ]}]
```

Literatúra

[1] Dupač,V.; Hušková, M.: *Pravděpodobnost a matematická statistika,* Karolinum, Praha, 2001

[2] Rényi, A.: *Teorie pravděpodobnosti,* Academia, Praha, 1972

[3] Scott, D.W.: *Multivariate Density Estimation: theory, practice, and visualization,* John Wiley & Sons, Texas, 1992

[4] Altman, N.; Léger, Ch.: *Bandwidth selection for kernel distribution function estimation,* Journal of Statistical Planning and Inference **46**, s.195-214, 1995

[5] Bowman, A.; Hall, P.; Prvan, T.: *Bandwidth selection for the smoothing of distribution functions,* Biometrika **85**, 4, s.799-808, 1998

[6] Cheng, Ch.; Parzen, E.: *Unified estimators of smooth quantile and quantile density functions,* Journal of Statistical Planning and Inference **59**, s.291-307, 1997

[7] Yang, Sh.: *A Smooth Nonparametric Estimator of a Quantile Function,* Journal of the American Statistical Association, Vol.**80**, No.392, 1985

[8] Simon, J.S.; Marron, J.S.: *Kernel Quantile Estimators,* Journal of the American Statistical Association, Vol.**85**, No.410, 1990

[9] Rao, C.R.; *Lineární metody statistické indukce a jejich aplikace,* Academia, Praha, 1978