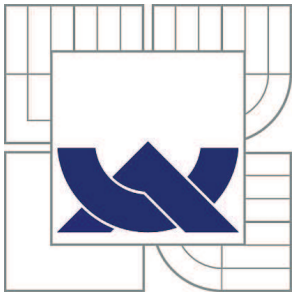


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
ÚSTAV MATEMATIKY

FACULTY OF MECHANICAL ENGINEERING
INSTITUTE OF MATHEMATICS

ODHADY DISKRÉTNÍHO ROZLOŽENÍ PRAVDĚPODOBNOSTI A BOOTSTRAP

ESTIMATION OF DISCRETE PROBABILITY DISTRIBUTION AND BOOTSTRAP

DIZERTAČNÍ PRÁCE
DOCTORAL THESIS

AUTOR PRÁCE — Ing. VERONIKA LACINOVÁ
AUTHOR

VEDOUCÍ PRÁCE — doc. RNDr. ZDENĚK KARPÍŠEK, CSc.
SUPERVISOR

BRNO 2014

Prohlašuji, že jsem dizertační práci vypracovala samostatně pod vedením doc. RNDr. Zdeňka Karpíška CSc. s použitím uvedené literatury.

Děkuji tímto doc. RNDr. Zdeňku Karpíškovi, CSc. za vedení během studia, za cenné rady, připomínky a konzultace při přípravě dizertační práce. Děkuji také rodině a všem svým blízkým za jejich podporu, kterou mi po dobu studia a přípravě této práce poskytovali.

Abstrakt

Dizertační práce se zabývá netradičními odhady diskrétního rozdělení pravděpodobnosti kategoriální veličiny z jejích pozorovaných hodnot. Při odhadech se využívá gradientu kvazinormy tohoto rozdělení pravděpodobnosti a tzv. přímkového odhadu. Zpřesnění gradientních odhadů bylo provedeno metodou bootstrap. Teoretické výsledky jsou pro vybrané kvazinormy ilustrovány na konkrétních příkladech.

Abstract

Doctoral thesis is focused on the unconventional methods of the discrete probability estimation of categorical quantity from its observed values. The gradient of quasinorm and so-called line estimation were emlopyed for these estimations. Bootstrap method was used for the improvement of accuracy. Theoretical results for selected quasinorms were illustrated on specific examples.

Klíčová slova

f-divergence, kvazinorma, diskrétní rozdělení pravděpodobnosti, gradientní odhad, bootstrap.

Keywords

f-divergence, quasinorm, discrete probability distribution, gradient estimation, bootstrap.

Obsah

1. Úvod a motivace

Základní praktickou úlohou při stochastickém modelování kategoriální veličiny X, která nabývá konečně mnoha různých hodnot x_j^*, $j=1,...,m$, kde $m \geq 2$, je odhad jejího rozdělení pravděpodobnosti z pozorovaných hodnot x_i, $i=1,...,n$, kde $n > m$. Předpokládáme, že pozorováním X získáme statistický soubor $(x_1,...,x_n)$ hodnot x_j^* a po jeho roztřídění dostaneme roztříděný statistický soubor $\left(\left(x_1^*, \frac{f_1}{n}\right),...,\left(x_m^*, \frac{f_m}{n}\right)\right)$, kde $\frac{f_j}{n} \neq 0$ je relativní četnost pozorované hodnoty x_j^*, $j=1,...,m$. Předpoklad nenulových relativních četností snadno zajistíme vynecháním jim odpovídajících hodnot x_j^*. Jestliže označíme odhadované rozdělení pravděpodobnosti $\mathbf{p}=(p_1,...,p_m)$, kde $p_j = P(X = x_j^*)$ je pravděpodobnost toho, že kategoriální veličina X nabude hodnotu x_j^*, jedná se vlastně o odhad parametrů $\mathbf{p}=(p_1,...,p_m)$ multinomického rozdělení pravděpodobnosti $M(n, p_1,..., p_m)$ při známém n. Jestliže byl statistický soubor $(x_1,...,x_n)$ získán výběrem s vracením a vzájemně nezávislými pozorováními X, je nestranným odhadem vektoru parametrů $\mathbf{p}=(p_1,...,p_m)$ vektor $\hat{\mathbf{p}} = \left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$.

V této práci předložíme jiný, v jistém smyslu pesimistický gradientní odhad $\mathbf{p}(t)$ diskrétního rozdělení náhodné veličiny, jeho vlastnosti a užití. Tento odhad je založen na pojmu *f*-divergence D_f, která slouží k měření „vzdálenosti“ dvou diskrétních rozdělení téže dimenze. Pomocí *f*-divergence je zaveden pojem pseudonormy, respektive kvazinormy rozdělení $\mathbf{p}$ jako „vzdálenost“ $D_f(\mathbf{p}, \mathbf{p}_0)$ od rovnoměrného rozdělení $\mathbf{p}_0$. Jde o analogii zavedení indukované normy na lineárním prostoru s metrikou pomocí neutrálního prvku. Rovnoměrné rozdělení $\mathbf{p}_0$ bylo zvoleno proto, že má na daném pravděpodobnostním prostoru největší neurčitost například ve smyslu Shannonovy entropie. Vhodný odhad rozdělení pravděpodobnosti je v práci realizován pomocí tzv. gradientního a přímkového odhadu.

Vyjdeme-li z empirického (pozorovaného) rozdělení četností $\frac{\mathbf{f}}{n} = \left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ a budeme-li se pohybovat po trajektorii největšího spádu směrem k rovnoměrnému rozdělení $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$, je gradientní odhad konstruován jako rozdělení nejbližší k rovnoměrnému rozdělení $\mathbf{p}_0$, kdy test dobré shody ještě rozdělení nezamítá. Úloha vede na řešení systému obyčejných diferenciálních rovnic 1. řádu, které jsou v práci odvozeny pro některé užívané kvazinormy. V případě kvadratické kvazinormy je tento systém lineární a dá se analyticky řešit.

V práci je popsán také tzv. přímkový odhad, kdy se nepohybujeme po trajektorii největšího spádu, ale přímo po přímce vycházející z empirického rozložení četností $\frac{\mathbf{f}}{n} = \left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ a procházející rozdělením $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$.

Gradientní i přímkové odhady jsou dostatečně vhodné pro aplikace a navíc lze zajistit vhodným postupem jejich asymptotickou nestrannost. V poslední kapitole je popsána metoda bootstrap, která je použita k zpřesnění gradientních odhadů.

Dizertační práce souvisí s pracemi [2], [31], které jsou zaměřeny na fitování rozdělení pravděpodobnosti za vedlejších momentových a obecných podmínek pomocí kvazinorem z pozorovaných hodnot diskrétních náhodných veličin. Tato práce se však zabývá odhady rozdělení pravděpodobnosti pozorovaných kategoriálních veličin [25] rovněž založených na kvazinormách. Původním impulzem studovat problematiku a aplikace kvazinorem je idea popsaná v práci [30].

Výsledky prezentované v práci jsou součástí řešení grantového projektu GAČR reg. č. P403/11/2085 „Konstrukce metod pro vícefaktorové měření komplexní podnikové výkonnosti ve vybraném odvětví“ a výzkumného projektu AKADEMIE STING IGA_AS_03 „Podpora řízení podniků“ . Dříve dosažené a publikované výsledky byly součástí řešení výzkumného projektu MŠMT České republiky čís. 1M06047 **„**Centrum pro jakost a spolehlivost výroby CQR“, grantového projektu GAČR reg. č. 103/05/0292 „Optimalizace navrhování progresivních betonových konstrukcí”, grantového projektu GAČR reg. čís. 103/08/1658 „Pokročilá optimalizace návrhu složených betonových konstrukcí“ a výzkumného záměru MŠMT České republiky čís. MSM0021630519 „Progresivní spolehlivé a trvanlivé nosné stavební konstrukce“. Projekt TAČR PID: TA02021449 „Systém inteligentních alarmů v energetickém provozu jaderných elektráren“.

2. F-divergence diskrétních rozdělení pravděpodobnosti

Předpokládejme, že je dán diskrétní pravděpodobnostní prostor (Ω, Σ, P). Až na výjimky však vystačíme s diskrétním pravděpodobnostním modelem $(\Omega, \mathbf{p})$, kde $\mathbf{p}$ je hustota pravděpodobnostní míry P. Omezujeme se tedy na konečný, případně spočetný, základní prostor Ω a σ–algebra Σ nehraje v našich úvahách podstatnou roli. Definice, věty a důsledky uvedené v tomto úvodním odstavci jsou převzaty z [1],[27].

V matematické statistice mají velký význam číselné míry podobnosti (vzdálenosti) dvojic hustot $\mathbf{p}$, $\mathbf{q}$ definované pomocí určité konvexní funkce $f:(0,\infty) \to \mathbb{R}$. Pro tyto míry se v literatuře vžil název *f*-divergence a hovoří se o *f*-divergenci pravděpodobnostních modelů $(\Omega, \mathbf{p})$, $(\Omega, \mathbf{q})$, resp. hustot $\mathbf{p}$, $\mathbf{q}$ nebo jim příslušných pravděpodobností P, Q. Existuje právě jedno spojité rozšíření $f(0)$, $f(\infty)$ takové, že rozšířená funkce je konvexní na $[0,\infty]$ a $f(0) > -\infty$. Bez újmy na obecnosti proto předpokládáme, že funkce f je definovaná na $[0,\infty]$, striktně konvexní v $u = 1$ a $f(0) > -\infty$.

Věta 2.1 Existuje limita

$$f(*) = \lim_{u \to \infty} \frac{f(u)}{u} \in \mathbb{R}^*,$$

kde $\mathbb{R}^*$ je množina reálných čísel rozšířená o nevlastní prvky $-\infty$ a ∞, přičemž platí

$$-\infty < f(1) < f(0) + f(*),$$

kde

$$\lim_{\substack{u \to 0^+ \\ v \to v_0}} uf(\frac{v}{u}) = v_0 f(*), \ \lim_{\substack{u \to 0^+ \\ v \to v_0}} vf(\frac{v}{u}) = v_0 f(0) \text{ pro každé } v_0 \in (0,\infty).$$

Důsledek 2.1 Jestliže je funkce $f(u)$ konvexní na $(0,\infty)$, pak funkce $\tilde{f}(u) = f(u) - f(1)$ je také konvexní na $(0,\infty)$ a $\tilde{f}(1) = 0$. Jestliže je funkce $f(u)$ striktně konvexní v $u = 1$, je také $\tilde{f}(u)$ striktně konvexní v $u = 1$ a $\tilde{f}(0) + \tilde{f}(*) > 0$.

Tento důsledek nám zajišťuje, že můžeme požadovat $f(1) = 0$ bez omezení obecnosti v následující definici.

Definice 2.1 Nechť funkce $f(u)$ je konvexní na $(0,\infty)$, striktně konvexní v $u = 1$ a $f(1) = 0$, ***f–divergencí*** pravděpodobnostních modelů $(\Omega, \mathbf{p})$, $(\Omega, \mathbf{q})$, resp. hustot $\mathbf{p}$, $\mathbf{q}$ na Ω rozumíme funkcionál

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\omega \in \Omega} q(\omega) f\left(\frac{p(\omega)}{q(\omega)}\right),$$

kde klademe $0 f\left(\frac{0}{0}\right) = 0$ a $0 f\left(\frac{p}{0}\right) = p f(*)$ pro všechna $p \in (0,1]$.

Definice 2.2 Pravděpodobnostní modely $(\Omega, \mathbf{p})$, $(\Omega, \mathbf{q})$, resp. jejich hustoty $\mathbf{p}$, $\mathbf{q}$, jsou ***ortogonální*** a píšeme $\mathbf{p} \perp \mathbf{q}$, jestliže existují takové disjunktní množiny $E, F \subset \Omega$, že

$$P(E)=\sum_{\omega\in E}p(\omega)=1 \text{ a } Q(F)=\sum_{\omega\in F}q(\omega)=1.$$

Následující věta vyjadřuje, jak *f*–divergence měří podobnost pravděpodobnostních modelů.

Věta 2.2 Pro libovolnou *f*-divergenci platí nerovnost

$$0\le D_f(\mathbf{p},\mathbf{q})\le f(0)+f(*),$$

přičemž obě dvě rovnosti nemohou nastat současně. Levá rovnost platí, právě když $\mathbf{p}=\mathbf{q}$ a pravá rovnost platí, právě když **p** a **q** jsou ortogonální a současně je $f(0)+f(*)<\infty$.

Z věty 2.2 plyne, že pravděpodobnostní modely $(\Omega,\mathbf{p})$ a $(\Omega,\mathbf{q})$ jsou si podobné, jestliže jejich *f*-divergence $D_f(\mathbf{p},\mathbf{q})$ je blízká 0. Maximální podobnost je vlastně shoda pravděpodobností *P* a *Q* na téže *σ*-algebře podmnožin množiny Ω, což nastává právě tehdy, když **p** a **q** jsou totožné na Ω. Naopak modely jsou tím více nepodobné, čím více se jejich *f*–divergence $D_f(\mathbf{p},\mathbf{q})$ blíží maximální hodnotě $f(0)+f(*)$. Maximálně divergentní jsou modely ortogonální. V tabulce 2.1 jsou uvedeny nejčastěji používané *f*-divergence.

Tab 2.1 Přehled nejpoužívanějších *f*-divergencí.

$f(u)$	Parametr	$f(0)+f(*)$	Označení $D_f(\mathbf{p},\mathbf{q})$ Název	Tvar $D_f(\mathbf{p},\mathbf{q})$
$u\ln u$	–	∞	$I(\mathbf{p},\mathbf{q})$ ***I-divergence***	$\sum_{x\in X}p(x)\ln\frac{p(x)}{q(x)}$
$\lvert u^{\beta}-1\rvert^{1/\beta}$	$\beta\in(0,1)$	2	$D_{\beta}(\mathbf{p},\mathbf{q})$ ***β-divergence***	$\sum_{x\in X}\lvert p(x)^{\beta}-q(x)^{\beta}\rvert^{1/\beta}$
	$\beta=1/2$	2	$D_{1/2}(\mathbf{p},\mathbf{q})$ ***Hellingerova vzdálenost***	$2\left(1-\sum_{x\in X}\left(p(x)q(x)\right)^{1/2}\right)$
$\lvert u-1\rvert^{\alpha}$	$\alpha=1$	2	$\chi^1(\mathbf{p},\mathbf{q})$, $V(\mathbf{p},\mathbf{q})$ ***Totální variace***	$\sum_{x\in X}\lvert p(x)-q(x)\rvert$
	$\alpha\in(1,\infty)$	∞	$\chi^{\alpha}(\mathbf{p},\mathbf{q})$ ***χ^{α} - divergence***	$\sum_{x\in X}\frac{\lvert p(x)-q(x)\rvert^{\alpha}}{q(x)^{\alpha-1}}$
	$\alpha=2$	∞	$\chi^2(\mathbf{p},\mathbf{q})$ ***χ^2 - divergence***	$\sum_{x\in X}\frac{(p(x)-q(x))^2}{q(x)}$
$sign(\alpha-1)(u^{\alpha}-1)$	$\alpha\in(0,1)$	1	$D^{\alpha}(\mathbf{p},\mathbf{q})$ ***α-divergence***	$1-\sum_{x\in X}p(x)^{\alpha}q(x)^{1-\alpha}$
	$\alpha\in(1,\infty)$	∞	$D^{\alpha}(\mathbf{p},\mathbf{q})$ ***α-divergence***	$\sum_{x\in X}\frac{p(x)^{\alpha}}{q(x)^{1-\alpha}}-1$

2.1 Duální *f*-divergence

Vzdálenost (metrika) dvou objektů bývá obvykle symetrickou funkcí těchto objektů. Definice *f*-divergence však tuto symetrii nevyžaduje, takže *f*-divergence nemusí být metrika. Obecně tedy platí

$$D_f(\mathbf{p},\mathbf{q}) \neq D_f(\mathbf{q},\mathbf{p}).$$

V tomto odstavci ukážeme, že vzdálenost **p** od **q** ve smyslu *f*-divergence může být rovna vzdálenosti **q** od **p** ve smyslu jiné *F*-divergence. Najdeme vztah mezi funkcemi f a F a ukážeme, že F splňuje požadavky na tzv. generující funkci *f*-divergence [2].

Věta 2.1.1 Nechť D_f je *f*-divergence a funkce $F(u)$ je dána vztahem

$$F(u) = uf\left(\frac{1}{u}\right),$$

kde generující funkce *f*(*u*) je dvakrát diferencovatelná v $(0,\infty)$. Pak D_F je opět *F*-divergencí a platí

$$D_F(\mathbf{p},\mathbf{q}) = D_f(\mathbf{q},\mathbf{p}).$$

Důkaz: Funkce $F(u)$ splňuje požadavky na funkci generující *F*-divergenci, neboť $F(1)=0$ a pro druhou derivaci platí

$$F''(u) = \frac{1}{u^3} f''\left(\frac{1}{u}\right) \geq 0,\ F''(1) > 0.$$

Po úpravách pak obdržíme $D_F(\mathbf{p},\mathbf{q}) = D_f(\mathbf{q},\mathbf{p})$. □

Poznámka 2.1.1 Vzhledem k tomu, že funkce *f*(*u*) je konvexní, striktně konvexní v $u=1$ a $f(1)=0$, je předpoklad existence druhé derivace ve větě 2.1.1. možné vypustit.

Definice 2.1.1 *F*-divergenci, která je generována funkcí

$$F(u) = uf\left(\frac{1}{u}\right),$$

budeme nazývat ***duální*** k *f*-divergenci generovanou funkcí f.

Poznámka 2.1.2 Duální *f*-divergence k duální *f*-divergenci je původní *f*-divergence.

Příklad 2.1.1 Najděte generující funkce duálních *F*-divergencí k používaným třídám divergencí z tabulky 2.1.

a) $f(u) = u\ln(u) \rightarrow F(u) = u\frac{1}{u}\ln\frac{1}{u} = -\ln u.$

b) $f(u) = \left|u^\beta - 1\right|^{1/\beta} \rightarrow F(u) = u\left|\frac{1}{u^\beta} - 1\right|^{1/\beta} = \left|u^\beta \frac{1}{u^\beta} - u^\beta\right|^{1/\beta} = \left|1-u^\beta\right|^{1/\beta} = f(u).$

c) $f(u) = |u-1|^\alpha \rightarrow F(u) = u\left|\frac{1}{u} - 1\right|^\alpha = u\frac{|1-u|^\alpha}{u^\alpha} = \frac{|u-1|^\alpha}{u^{\alpha-1}}.$

Poznámka 2.1.3 Z příkladu 2.1.1 a tabulky 2.1 vidíme, že *I*-divergence $I(\mathbf{p},\mathbf{q})$ je duální s logaritmickou divergencí $I_D(\mathbf{p},\mathbf{q})$, χ^2-divergence $\chi^2(\mathbf{p},\mathbf{q})$ je duální s Pearsonovou divergencí a Hellingerova vzdálenost $D_{1/2}(\mathbf{p},\mathbf{q})$ je duální sama se sebou.

Poznámka 2.1.4 Duální *F*-divergence k symetrické *f*-divergenci je původní *f*-divergence.

3. Kvazinormy a jejich vlastnosti

Dále vycházíme z následujících pojmů a vlastností popsaných ve [2], [3], [4].

Věta 3.1 Nechť (Ω, Σ, P) je konečný pravděpodobnostní prostor, kde P je libovolná pravděpodobnostní míra na Ω a $D_f(\mathbf{p},\mathbf{q})$ je f-divergence rozdělení (hustot) pravděpodobnosti $\mathbf{p} = (p_1,...,p_m)$, $\mathbf{q} = (q_1,...,q_m)$, $m > 1$, z tohoto prostoru. Označme

$$V(\mathbf{q}) = \int_S D_f(\mathbf{p},\mathbf{q})\, dS,$$

kde $S = \left\{ \mathbf{p} \in \mathbb{R}^m : \forall p_j \geq 0, \sum_{j=1}^{m} p_j = 1 \right\}$, integrál f-divergencí všech rozdělení $\mathbf{p}$ od nějakého pevně zvoleného rozdělení $\mathbf{q}$. Jestliže existuje funkce $V(\mathbf{q})$, $\mathbf{q} \in S$, a funkce

$$G(q_j) = \frac{\partial V(\mathbf{q})}{\partial q_j} = \int_S \frac{\partial D_f(\mathbf{p},\mathbf{q})}{\partial q_j} dS,$$

(tj. konvergují oba uvedené integrály), dále jestliže existuje $G'(q_j)$ v $[0,1]$ a funkce f má spojitou druhou derivaci v $(0,\infty)$, pak $V(\mathbf{q})$ nabývá absolutního minima na S v rozdělení pravděpodobnosti

$$\mathbf{p_0} = \left(\frac{1}{m},...,\frac{1}{m} \right).$$

(S je množina všech rozdělení pravděpodobnosti z (Ω, Σ, P) a oba integrály jsou integrály 1. druhu po nadploše S dimenze $m-1$ v $\mathbb{R}^m$.)

Důkaz: Označme

$$\Lambda(\mathbf{q},\lambda) = V(\mathbf{q}) + \lambda \left(\sum_{j=1}^{m} q_j - 1 \right)$$

Lagrangeovu funkci pro vázaný extrém funkce $V(\mathbf{q})$ za podmínky $\sum_{j=1}^{m} q_j = 1$, tj. extrém na nadploše S. Funkce f je konvexní (jde o f–divergenci), takže z předpokladu existence její druhé derivace je $f'' \geq 0$ v $(0,\infty)$. Protože

$$\frac{\partial D_f(\mathbf{p},\mathbf{q})}{\partial q_j} = f\left(\frac{p_j}{q_j} \right) - \frac{p_j}{q_j} f'\left(\frac{p_j}{q_j} \right),$$

pak

$$G'(q_j) = \int_S \frac{d}{dq_j} \left(f\left(\frac{p_j}{q_j} \right) - \frac{p_j}{q_j} f'\left(\frac{p_j}{q_j} \right) \right) dS = \int_S \frac{p_j^2}{q_j^3} f''\left(\frac{p_j}{q_j} \right) dS > 0$$

a odtud funkce $G(q_j)$ je rostoucí na $[0,1]$. Protože

$$\frac{\partial \Lambda(\mathbf{q},\lambda)}{\partial q_j} = G(q_j) + \lambda,$$

existuje na $[0,1]$ jediný kořen $q_j = G^{-1}(-\lambda)$ rovnice $G(q_j) + \lambda = 0$. Z podmínky $\sum_{j=1}^{m} q_j = 1$ pak dostaneme $q_j = \frac{1}{m}$, $j = 1,\ldots,m$. Dále je

$$\frac{\partial^2 V(\mathbf{q})}{\partial q_j \partial q_k} = \frac{\partial^2 \Lambda(\mathbf{q},\lambda)}{\partial q_j \partial q_k} = G'(q_j) > 0 \text{ pro } j = k,$$

$$= 0 \text{ pro } j \neq k,$$

$j = 1,\ldots,m$ a $k = 1,\ldots,m$. Z toho plyne, že Hessova matice funkce $V(\mathbf{q})$ je diagonální a pozitivně definitní. Jacobiho matice má pro jedinou podmínku $\sum_{j=1}^{m} q_j - 1 = 0$ hodnost 1, takže $V(\mathbf{q})$ má za dané podmínky v $\mathbf{q} = \mathbf{p_0} = \left(\frac{1}{m},...,\frac{1}{m}\right)$ absolutní minimum. □

Definice 3.1 Nechť $S = \left\{\mathbf{p} \in \mathbb{R}^m : \forall p_j \geq 0, \sum_{j=1}^{m} p_j = 1\right\}$ je množina všech diskrétních rozdělení pravděpodobnosti na Ω, $m > 1$. ***Kvazinormou*** rozdělení $\mathbf{p} \in S$ rozumíme *f*-divergenci $D_f(\mathbf{p},\mathbf{p}_0)$, kde $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$ a o funkci f říkáme, že ***generuje*** kvazinormu $D_f(\mathbf{p},\mathbf{p}_0)$ na S.

Věta 3.2 Uvažujme rovnoměrné rozdělení $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$, které minimalizuje [3] integrál všech *f*-divergencí $D_f(\mathbf{p},\mathbf{q})$ na S a má maximální entropii [26]. Platí, že

a) $D_f(\mathbf{p},\mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m} f(mp_j) = \frac{1}{m}\left(\sum_{j=1}^{m-1} f_j(mp_j) + f\left(m\left(1 - \sum_{j=1}^{m-1} p_j\right)\right)\right)$,

b) $c > 0 \Rightarrow D_{cf}(\mathbf{p},\mathbf{p}_0) = cD_f(\mathbf{p},\mathbf{p}_0)$,

c) $D_f(\mathbf{p},\mathbf{p}_0)$ je nezáporná konvexní funkce na S, která je symetrická vzhledem k proměnným p_j, $j = 1,...,m$,

d) $D_f(\mathbf{p},\mathbf{p}_0) = 0 \Leftrightarrow \mathbf{p} = \mathbf{p}_0$.

Důkaz plyne bezprostředně z definice 3.1. □

Věta 3.3 Nechť je funkce $f(u)$ konvexní na $(0,\infty)$, striktně konvexní v $u = 1$ a $f(1) = 0$, pak

$$0 \leq D_f(\mathbf{p},\mathbf{p}_0) \leq D_f(\mathbf{p}_1,\mathbf{p}_0) = \frac{1}{m}\left(f(m) + (m-1)f(0)\right),$$

kde $\mathbf{p}_1 = (1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)$.

Důkaz: Zavedeme-li substituci

$$mp_j = u_j$$

a označíme

$$\frac{1}{m} f\left(mp_j\right) = \tilde{f}\left(u_j\right),$$

potom máme dokázat, že

$$\max_{u_j>0} \sum_{j=1}^{m} \tilde{f}\left(u_j\right) = (m-1)\,\tilde{f}(0) + \tilde{f}\left(\sum_{j=1}^{m} u_j\right).$$

Důkaz provedeme indukcí.

1. Pro $m=2$ zafixujeme $u_1+u_2=m$ a máme

$$\max \tilde{f}\left(u_1\right) + \tilde{f}\left(u_2\right) = \max \tilde{f}\left(u_1\right) + \tilde{f}\left(m-u_1\right)$$

$\tilde{f}\left(u_1\right) + \tilde{f}\left(m-u_1\right)$ je konvexní funkce na $[0,m]$, nabývá tedy svého maxima v jednom z hraničních bodů, tj. Buď v bodě $u_1=0$ nebo v bodě $u_1=m$. V obou případech je maximum shodné a to $\tilde{f}(0)+\tilde{f}(m)=\tilde{f}(0)+\tilde{f}\left(u_1+u_2\right)$.

2. Předpokládejme, že rovnost

$$\max \tilde{f}\left(u_1\right) + \tilde{f}\left(u_2\right) = \max \tilde{f}\left(u_1\right) + \tilde{f}\left(m-u_1\right)$$

platí pro $m=1,\ldots,k-1$. Nyní pro $m=k$ máme $\sum_{j=1}^{k} u_j = m \Rightarrow \sum_{j=1}^{k-1} u_j = m-u_k$, takže

$$\max \sum_{j=1}^{k} \tilde{f}\left(u_j\right) = \max_{u_k\in[0,m]} \max_{\sum_{j=1}^{k-1} u_j = m-u_k} \left(\sum_{j=1}^{k-1} \tilde{f}\left(u_j\right) + \tilde{f}\left(u_k\right) \right) =$$

$$= \max_{u_k\in[0,m]} \left(\left(\max_{\sum_{j=1}^{k-1} u_j = m-u_k} \sum_{j=1}^{k-1} \tilde{f}\left(u_j\right) \right) + \tilde{f}\left(u_k\right) \right).$$

Využijeme indukčního předpokladu a z předešlého výrazu obdržíme

$$\max_{u_k\in[0,m]} \left(\left(\max_{\sum_{j=1}^{k-1} u_j = m-u_k} \sum_{j=1}^{k-1} \tilde{f}\left(u_j\right) \right) + \tilde{f}\left(u_k\right) \right) = \max_{u_k\in[0,m]} \left((k-2)\,\tilde{f}(0) + \tilde{f}\left(m-k\right) + \tilde{f}\left(u_k\right) \right) =$$

$$= (k-2)\,\tilde{f}(0) + \max_{u_k\in[0,m]} \left(\tilde{f}\left(m-k\right) + \tilde{f}\left(u_k\right) \right) = (k-2)\,\tilde{f}(0) + \tilde{f}(0) + \tilde{f}(m) =$$

$$= (k-1)\,\tilde{f}(0) + \tilde{f}\left(\sum_{j=1}^{k-1} u_j\right).$$

Platí tedy

$$\max \sum_{j=1}^{k} \tilde{f}\left(u_j\right) = (k-1)\,\tilde{f}(0) + \tilde{f}\left(\sum_{j=1}^{k} u_j\right). \square$$

Poznámka 3.1 Kvazinormu můžeme chápat jako míru neurčitosti rozdělení **p** a platí

a) $D_f\left(\mathbf{p},\mathbf{p}_0\right)=0$ pro největší neurčitost,

b) $D_f\left(\mathbf{p},\mathbf{p}_0\right)=\frac{1}{m}\left(f(m)+(m-1)f(0)\right)$ pro nejmenší neurčitost.

Věta 3.4 Dvě různé generující funkce $f(u)$ mohou generovat tutéž kvazinormu.

Důkaz: Například funkce tvaru $\frac{(u-1)^2}{u} = u - 2 + \frac{1}{u}$ generuje kvazinormu $\frac{1}{m}\sum_{j=1}^{m}\left(mp_j - 2 + \frac{1}{mp_j}\right) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{mp_j} - 1$. Stejnou kvazinormu $\frac{1}{m}\sum_{j=1}^{m}\frac{1}{mp_j} - 1$ však generuje i funkce $\frac{1}{u} - 1$. □

Poznámka 3.2 Pro odhady diskrétních rozdělení pravděpodobnosti nejčastěji volíme

a) Hellingerovu vzdálenost $D_{1/2}(\mathbf{p},\mathbf{q})$, z níž získáme ***Hellingerovu kvazinormu***

$$H(\mathbf{p},\mathbf{p}_0) = \sum_{j=1}^{m}\left(\sqrt{p_j} - \sqrt{\frac{1}{m}}\right)^2 = 2 - \frac{2}{\sqrt{m}}\sum_{j=1}^{m}\sqrt{p_j}\,,$$

b) I-divergenci $I(\mathbf{p},\mathbf{q})$, z níž získáme ***Shannonovu kvazinormu***

$$S(\mathbf{p},\mathbf{p_0}) = \sum_{j=1}^{m}\left(p_j \ln p_j - \frac{1}{m}\ln\left(\frac{1}{m}\right)\right) = \sum_{j=1}^{m} p_j \ln p_j + \ln m\,,$$

c) χ^2-divergenci $\chi^2(\mathbf{p},\mathbf{q})$, z níž získáme ***kvadratickou kvazinormu***

$$K(\mathbf{p},\mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m}\left(mp_j - 1\right)^2$$

a ***Pearsonovu kvazinormu***

$$P(\mathbf{p},\mathbf{p_0}) = \frac{1}{m^2}\sum_{j=1}^{m}\frac{1}{p_j} - 1\,,$$

d) *G* ***-kvazinormu***

$$G(\mathbf{p},\mathbf{p_0}) = \frac{2}{m(k-2)}\sum_{j=1}^{m}\left(mp_j\right)^{\frac{2-k}{2}} - \frac{2}{k-2} \text{ pro } \forall k \in \mathbb{R}^+ - \{2\}.$$

Poznámka 3.3 V případě *G*-kvazinormy se jedná o celou třídu kvazinorem. Generující funkce této kvazinormy má tvar

$$f(u) = \frac{2}{k - 2\left(\sqrt{u}\right)^{k-2}} + u - \frac{k}{k-2}\text{, pro } \forall k \in \mathbb{R}^+ - \{2\}.$$

Definice 3.2 Nechť $D_f(\mathbf{p},\mathbf{p}_0)$ je kvazinorma a *F* je funkce daná vztahem (viz. definice 2.1.1)

$$F(u) = uf\left(\frac{1}{u}\right),$$

pak kvazinormu $D_F(\mathbf{p},\mathbf{p}_0)$ budeme nazývat ***duální kvazinormou*** k $D_f(\mathbf{p},\mathbf{p}_0)$.

Věta 3.5 Pro kvazinormu $D_f(\mathbf{p},\mathbf{p}_0)$ a k ní kvazinormu duální $D_F(\mathbf{p},\mathbf{p}_0)$ platí

a) $D_F(\mathbf{p},\mathbf{p}_0) = D_f(\mathbf{p}_0,\mathbf{p})$, $D_f(\mathbf{p},\mathbf{p}_0) = D_F(\mathbf{p}_0,\mathbf{p})$,

b) $D_F(\mathbf{p},\mathbf{p}_0)=\frac{1}{m}\sum_{j=1}^{m}F(mp_j)=\sum_{j=1}^{m}p_j f\left(\frac{1}{mp_j}\right)$.

Důkaz plyne bezprostředně z věty 2.1.1 a definic 3.1, 3.2. □

Ukázka některých duálních kvazinorem používaných v této práci je v tabulce 3.1.

Tab.3.1 Tabulka vybraných duálních kvazinorem.

$f(u)$	$D_f(\mathbf{p},\mathbf{p}_0)$	$F(u)$	$D_F(\mathbf{p},\mathbf{p}_0)$
$u\ln(u)$	Shannonova kvazinorma $\sum_{j=1}^{m}p_j\ln p_j+\ln m$	$-\ln u$	Logaritmická kvazinorma $-\frac{1}{m}\sum_{j=1}^{m}\ln(mp_j)$
$\left(u^{1/2}-1\right)^2$	Hellingerova kvazinorma $2-\frac{2}{\sqrt{m}}\sum_{j=1}^{m}\sqrt{p_j}$	$\left(u^{1/2}-1\right)^2$	Hellingerova kvazinorma $2-\frac{2}{\sqrt{m}}\sum_{j=1}^{m}\sqrt{p_j}$
$(u-1)^2$	kvadratická kvazinorma $\frac{1}{m}\sum_{j=1}^{m}(mp_j-1)^2$	$\frac{(u-1)^2}{u}$	Pearsonova kvazinorma $\frac{1}{m^2}\sum_{j=1}^{m}\frac{1}{p_j}-1$

3.1 Grafická interpretace kvazinorem

Na obrázcích 3.1.1, 3.1.2 a 3.1.3 jsou grafy isočar některých kvazinorem pro *m=3*, tj. pro $p_1\in[0,1]$, $p_2\in[0,1-p_1]$ a $p_3=1-p_1-p_2$. Jde o izočáry:

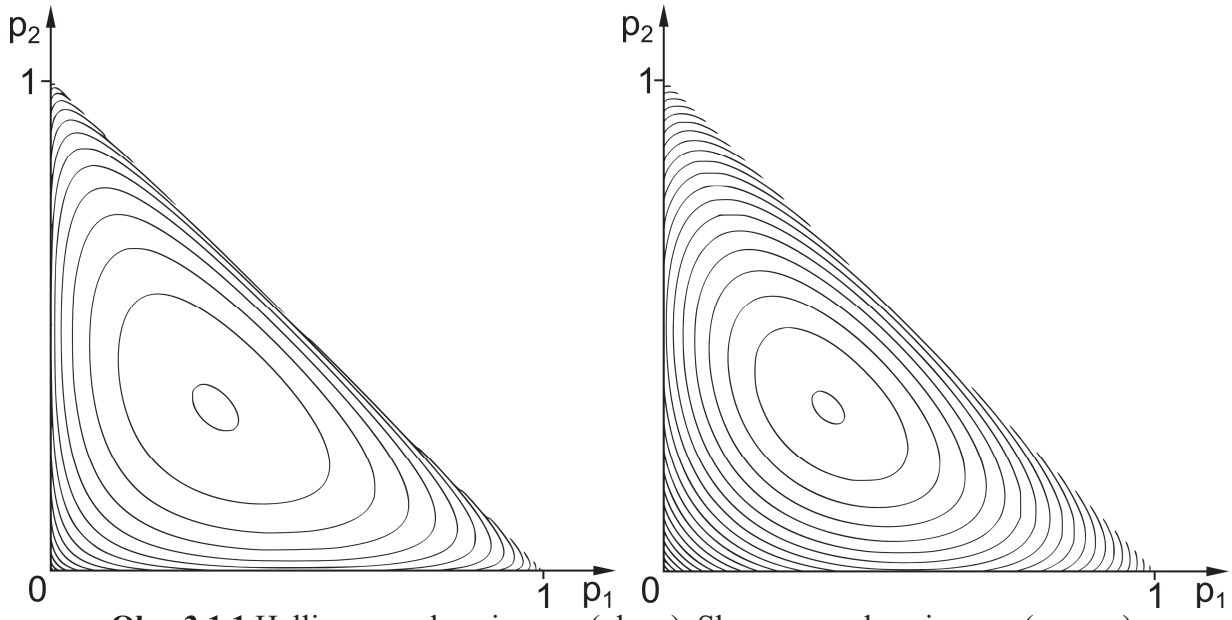
a) $2-\frac{2}{\sqrt{3}}\sum_{j=1}^{3}\sqrt{p_j}=c$ pro Hellingerovu kvazinormu,

b) $\sum_{j=1}^{3}p_j\ln p_j+\ln 3=c$ pro Shannonovu kvazinormu,

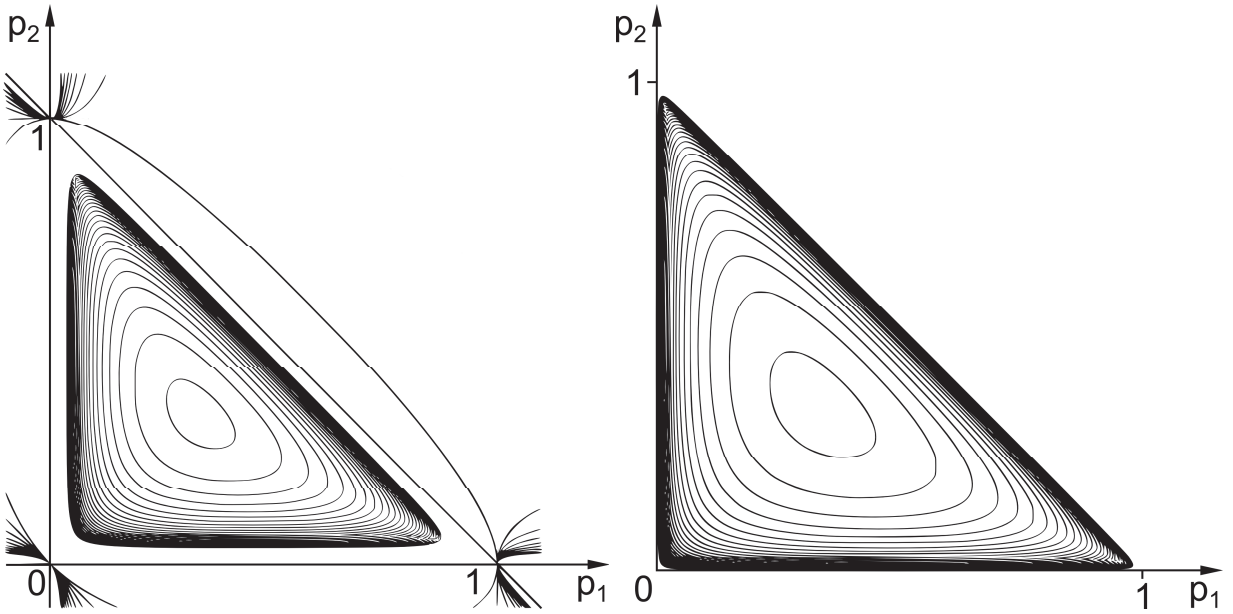
c) $\frac{1}{9}\sum_{j=1}^{3}\frac{1}{p_j}-1=c$ pro Pearsonovu kvazinormu,

d) $\frac{\sqrt{3}}{9}\sum_{j=1}^{3}\frac{1}{\sqrt{p_j}}-2=c$ pro G-kvazinormu ($k=3$),

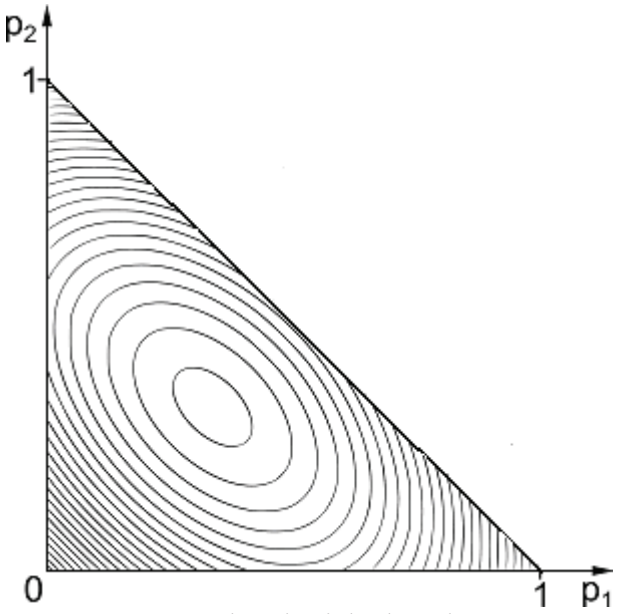
e) $\frac{1}{3}\sum_{j=1}^{3}(3p_j-1)^2=c$ pro kvadratickou kvazinormu.



Obr. 3.1.1 Hellingerova kvazinorma (vlevo), Shannonova kvazinorma (vpravo).



Obr. 3.1.2 Pearsonova kvazinorma (vlevo), G-kvazinorma (vpravo).



Obr. 3.1.3 kvadratická kvazinorma

4 Gradientní odhad

Idea gradientního odhadu rozdělení pravděpodobnosti $\mathbf{p}$ z pozorovaných hodnot náhodné veličiny X vychází z požadavku najít takové rozdělení v S, které je nejblíže rovnoměrnému rozdělení $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$ a k němuž se dostaneme od empirického rozdělení $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ v jistém smyslu co nejrychleji. Tomu odpovídá vhodná minimalizace zvolené kvazinormy $D_f(\mathbf{p},\mathbf{p}_0)$ a hledání rozdělení $\mathbf{p}$ na křivce největšího spádu v S, tj. křivce, jejíž tečný vektor je kolineární s gradientem této kvazinormy. Je zřejmé, že jde o úlohu, kde jsme vázáni podmínkou $\mathbf{p} \in S$. To nás spolu s konvexností generující funkce *f* opravňuje k následující definici [6].

Definice 4.1 Nechť $D_f(\mathbf{p},\mathbf{p}_0)$ je kvazinorma na S. ***Gradientním odhadem*** rozdělení pravděpodobnosti $\mathbf{p} \in S$ z empirického rozdělení $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right) \neq \left(\frac{1}{m},...,\frac{1}{m}\right)$ rozumíme takové rozdělení pravděpodobnosti $\mathbf{p}(t) \in S$, že

$$\frac{d}{dt}\mathbf{p}(t) = -gradD_f\left(\mathbf{p}(t),\mathbf{p}_0\right) \text{ pro } \forall t \in [0;\infty),$$

$$\mathbf{p}(0) = \frac{\mathbf{f}}{n} = \left(\frac{f_1}{n},...,\frac{f_m}{n}\right).$$

V případě, že $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right) = \left(\frac{1}{m},...,\frac{1}{m}\right)$, rozumíme gradientním odhadem rozdělení $\left(\frac{1}{m},...,\frac{1}{m}\right)$.

Věta 4.1 Jestliže funkce $f(u)$, která generuje kvazinormu $D_f(\mathbf{p},\mathbf{p}_0)$ na S, má vlastnosti uvedené v definici 2.1 a má spojitou derivaci $f'(u)$ pro $\forall u \in (0;\infty)$, pak existuje jediný gradientní odhad $\mathbf{p}(t) = \left(p_1(t),...,p_m(t)\right)$ rozdělení pravděpodobnosti $\mathbf{p} \in S$. Jeho složky $p_1(t),...,p_{m-1}(t)$ pro $\forall t \in [0;\infty)$ jsou partikulárním řešením soustavy obyčejných diferenciálních rovnic prvního řádu (SODR1)

$$p_1'(t) = -f'\left(mp_1(t)\right) + f'\left(m\left[1-\sum_{j=1}^{m-1}p_j(t)\right]\right),$$

$$p_2'(t) = -f'\left(mp_2(t)\right) + f'\left(m\left[1-\sum_{j=1}^{m-1}p_j(t)\right]\right),$$

...

$$p_{m-1}'(t) = -f'\left(mp_{m-1}(t)\right) + f'\left(m\left[1-\sum_{j=1}^{m-1}p_j(t)\right]\right),$$

s počátečními podmínkami

$$p_1(0) = \frac{f_1}{n},\ p_2(0) = \frac{f_2}{n},\ldots,\ p_{m-1}(0) = \frac{f_{m-1}}{n}$$

a složkou

$$p_m(t) = 1-\sum_{j=1}^{m-1}p_j(t) \text{ pro } \forall t \in [0;\infty).$$

Důkaz: SODR1 spolu s podmínkou $p_m(t) = 1 - \sum_{j=1}^{m-1} p_j(t)$ pro $\forall t \in [0;\infty)$ je zřejmě ekvivalentní s vektorovou rovnicí $\frac{d}{dt}\mathbf{p}(t) = -gradD_f(\mathbf{p}(t), \mathbf{p}_0)$ z definice 4.1. Protože generující funkce $f(u)$ má spojitou derivaci $f'(u)$ ve všech bodech $u \in (0;\infty)$, je vektorová funkce daná pravými stranami soustavy obyčejných diferenciálních rovnic ve větě 4.1 spojitá na podprostoru $T = \left\{ \mathbf{p} \in \mathbb{R}^{m-1} : \forall p_j > 0, \sum_{j=1}^{m-1} p_j = 1 \right\}$ prostoru *S*. Z toho plyne, že tato soustava má v *T* jediné partikulární řešení, které splňuje dané počáteční podmínky. □

Důsledek 4.1 Jestliže má navíc generující funkce $f(u)$ spojitou druhou derivaci pro $\forall u \in (0;\infty)$, pak má SODR1 jediný stacionární (asymptoticky stabilní) bod [5],[28]

$$\lim_{t\to\infty} \mathbf{p}(t) = \mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$$

a jde o bod typu uzel.

Důkaz: Stacionární bod množiny všech řešení uvedené SODR1 je řešení $\mathbf{p}(t)$, pro které je $p_j'(t) = 0$, $j = 1,\ldots,m-1$. Protože $f''(u) > 0$ v okolí bodu $u = 1$, má soustava

$$-f'\left(mp_j(t)\right) + f'\left(m\left[1 - \sum_{j=1}^{m-1} p_j(t)\right]\right) = 0 \quad j = 1,\ldots,m-1, \text{ pro } t \to \infty$$

jediné řešení $\mathbf{p} = (p_1,\ldots p_{m-1})$. Jde o řešení soustavy lineárních algebraických rovnic

$$\begin{aligned} 2p_1 + p_2 + \ldots + p_{m-1} &= 1, \\ p_1 + 2p_2 + \ldots + p_{m-1} &= 1, \\ &\ldots \\ p_1 + p_2 + \ldots + 2p_{m-1} &= 1, \end{aligned}$$

jejíž matice je regulární (má determinant rovný m). Snadno nahlédneme, že toto řešení je $\mathbf{p} = \left(\frac{1}{m},\ldots,\frac{1}{m}\right)$, což je současně stacionární bod SODR1 z věty 4.1. Z vlastností Jacobiho matice této SODR1 [5],[28] plyne, že jde o stacionární bod typu uzel. □

Obecně jsou složky gradientního odhadu uvedeny ve větě 4.1. Nyní uvedeme soustavy diferenciálních rovnic získané pro konkrétní generující funkce.

1. Generující funkce Hellingerovy kvazinormy a její derivace

$$f(u) = \left(u^{\frac{1}{2}} - 1\right)^2,$$

$$f'(u) = \left(1 - u^{-\frac{1}{2}}\right).$$

SODR1 pro generující funkci Hellingerovy kvazinormy

$$p_1'(t) = \left(mp_1(t)\right)^{-\frac{1}{2}} - \left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right)^{-\frac{1}{2}},$$

$$p_2^{'}(t) = \left(mp_2(t)\right)^{-\frac{1}{2}} - \left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right)^{-\frac{1}{2}},$$

...

$$p_{m-1}^{'}(t) = \left(mp_{m-1}(t)\right)^{-\frac{1}{2}} - \left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right)^{-\frac{1}{2}},$$

s počátečními podmínkami

$$p_1(0) = \frac{f_1}{n},\ p_2(0) = \frac{f_2}{n}, \ldots,\ p_{m-1}(0) = \frac{f_{m-1}}{n}$$

a složkou

$$p_m(t) = 1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t).$$

2. Generující funkce Shannonovy kvazinormy a její derivace

$$f(u) = u \ln u,$$

$$f^{'}(u) = \ln u + 1.$$

SODR1 pro generující funkci Shannonovy kvazinormy

$$p_1^{'}(t) = \ln\left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right) - \ln\left(mp_1(t)\right),$$

$$p_2^{'}(t) = \ln\left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right) - \ln\left(mp_2(t)\right),$$

...

$$p_{m-1}^{'}(t) = \ln\left(m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right) - \ln\left(mp_{m-1}(t)\right),$$

s počátečními podmínkami

$$p_1(0) = \frac{f_1}{n},\ p_2(0) = \frac{f_2}{n}, \ldots,\ p_{m-1}(0) = \frac{f_{m-1}}{n}$$

a složkou

$$p_m(t) = 1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t).$$

3. Generující funkce Pearsonovy kvazinormy a její derivace

$$f(u) = \frac{(u-1)^2}{u},$$

$$f'(u) = 1 - \frac{1}{u^2}.$$

SODR1 pro generující funkci Pearsonovy kvazinormy

$$p_1^{'}(t) = \left(mp_1(t)\right)^{-2} - \left[m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right]^{-2},$$

$$p_2^{'}(t) = \left(mp_2(t)\right)^{-2} - \left[m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right]^{-2},$$

...

$$p_{m-1}^{'}(t) = \left(mp_{m-1}(t)\right)^{-2} - \left[m\left(1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t)\right)\right]^{-2},$$

s počátečními podmínkami

$$p_1(0) = \frac{f_1}{n},\ p_2(0) = \frac{f_2}{n}, \ldots,\ p_{m-1}(0) = \frac{f_{m-1}}{n}$$

a složkou

$$p_m(t) = 1 - p_1(t) - p_2(t) - \cdots - p_{m-1}(t).$$

4. Generující funkce G-kvazinormy

$$f(u)=\frac{2}{k-2\left(\sqrt{u}\right)^{k-2}}+u-\frac{k}{k-2},\ \text{pro}\ \forall k\in\mathbb{R}^{+}-\{2\}$$

Generující funkce G-kvazinormy speciálně pro $k=3$ a její derivace

$$f(u)=\frac{2}{\left(\sqrt{u}\right)}+u-3,$$

$$f'(u)=-(u)^{-\frac{3}{2}}+1.$$

SODR1 pro generující funkci G-kvazinormy ($k=3$)

$$p_1'(t)=\left(mp_1(t)\right)^{-\frac{3}{2}}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\frac{3}{2}},$$

$$p_2'(t)=\left(mp_2(t)\right)^{-\frac{3}{2}}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\frac{3}{2}},$$

...

$$p_{m-1}'(t)=\left(mp_{m-1}(t)\right)^{-\frac{3}{2}}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\frac{3}{2}},$$

s počátečními podmínkami

$$p_1(0)=\frac{f_1}{n},\ p_2(0)=\frac{f_2}{n},\ldots,\ p_{m-1}(0)=\frac{f_{m-1}}{n}$$

a složkou

$$p_m(t)=1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t).$$

Poznámka 4.1 Gradientní odhad $\mathbf{p}(t)$ z věty 4.1 je odhad s parametrem $t\in[0;\infty)$ rozdělení pravděpodobnosti $\mathbf{p}$. Vhodnou hodnotu parametru $t_0\in[0;\infty)$ pro gradientní odhad můžeme najít pomocí testu dobré shody. Při použití Pearsonova testu, resp. Pitmanova – Hellingerova testu [7], [14], [32], je t_0 kořenem nelineární rovnice

$$\frac{1}{n}\sum_{j=1}^{m}\frac{f_j^{\,2}}{p_j(t)}-n=\chi^2_{1-\alpha},$$

resp.

$$8n\left(1-\sum_{j=1}^{m}\sqrt{p_j(t)\frac{f_j}{n}}\right)=\chi^2_{1-\alpha},$$

jejíž levá strana je zvolené testové kritérium a $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil chí-kvadrát rozdělení s $m-1$ stupni volnosti. Jde vlastně o nalezení takové hodnoty *t*, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\mathbf{p}(t)$ na hladině významnosti α. Oba testy jsou pouze asymptotické a pro praktické použití požadujeme, aby $np_j(t_0)>5$ pro $\forall j=1,\ldots,m$ [14].

Poznámka 4.2 Gradientní odhad $\mathbf{p}(t)$ je spojitá vektorová funkce pro $\forall t\in[0;\infty)$ a jejím grafem je křivka největšího spádu kvazinormy $D_f(\mathbf{p},\mathbf{p}_0)$ v S. Kvazinorma $D_f(\mathbf{p}(t),\mathbf{p}_0)$ je nerostoucí pro $t\in[0;\infty)$ a

$$D_f(\mathbf{p}(0),\mathbf{p}_0)=\frac{1}{m}\sum_{j=1}^{m}f\left(m\frac{f_j}{n}\right)\geq D_f(\mathbf{p}(t),\mathbf{p}_0)\geq\lim_{t\to\infty}D_f(\mathbf{p}(t),\mathbf{p}_0)=0,$$

takže se gradientní odhad $\mathbf{p}(t)$ pro rostoucí parametr $t \in [0;\infty)$ vzdaluje po křivce největšího spádu v S od empirického rozdělení $\frac{\mathbf{f}}{n} = \left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ směrem k rozdělení $\mathbf{p}_0 = \left(\frac{1}{m},...,\frac{1}{m}\right)$. Všechny gradientní odhady $\mathbf{p}(t)$ pro $\forall t \in [0;t_0]$ splňují zvolené testové kritérium z poznámky 4.1 na hladině významnosti alespoň α a gradientní odhad $\mathbf{p}(t_0)$ je „nejhorší“ z těchto odhadů, takže jej můžeme označit jako tzv. pesimistický gradientní odhad. Gradientní odhady $\mathbf{p}(t)$ tvoří vzhledem ke zvolené kvazinormě $D_f(\mathbf{p},\mathbf{p}_0)$ jednoparametrickou třídu konečných diskrétních rozdělení pravděpodobnosti s parametrem $t \in [0;\infty)$.

4.1 Gradientní odhad pomocí kvadratické kvazinormy

SODR1 ve větě 4.1 je obecně nelineární. K nalezení jejího řešení je nutno až na výjimky aplikovat některou numerickou metodu a současně hledat takovou hodnotu parametru *t*, který vyhovuje zvolené nelineární rovnici z poznámky 4.1. K nalezení řešení nelineární SODR1 byl vytvořen pomocí softwaru MATLAB program Pesfit 1.0. [8]. V programu se soustava diferenciálních rovnic řeší pomocí Rungeho-Kuttovy metody a k nalezení hodnoty *t*, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\mathbf{p}(t)$, byla užita metoda bisekce. Výjimkou, kdy získáme explicitně řešení dané SODR1, je kvazinorma použitá v následující větě, která vede na lineární SODR1 [6].

Věta 4.1.1 Nechť $f(u) = (u-1)^2$, takže $D_f(\mathbf{p},\mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m}(mp_j - 1)^2$ je tzv. kvadratická kvazinorma. Potom složky gradientního odhadu $\mathbf{p}(t) = (p_1(t),...,p_m(t))$ z empirického rozdělení $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ jsou pro $\forall t \in [0;\infty)$ partikulárním řešením nehomogenní lineární soustavy obyčejných diferenciálních rovnic prvního řádu (LSODR1) s konstantními koeficienty a konstantními pravými stranami

$$p_1'(t) = -4mp_1(t) - 2mp_2(t) - \cdots - 2mp_{m-1}(t) + 2m,$$
$$p_2'(t) = -2mp_1(t) - 4mp_2(t) - \cdots - 2mp_{m-1}(t) + 2m,$$
$$\cdots$$
$$p_{m-1}'(t) = -2mp_1(t) - 2mp_2(t) - \cdots - 4mp_{m-1}(t) + 2m,$$

s počátečními podmínkami

$$p_1(0) = \frac{f_1}{n},\ p_2(0) = \frac{f_2}{n},\ldots,\ p_{m-1}(0) = \frac{f_{m-1}}{n}$$

a složkou

$$p_m(t) = 1 - \sum_{j=1}^{m-1} p_j(t) \text{ pro } \forall t \in [0;\infty).$$

Gradientní odhad $\mathbf{p}(t)$ má složky

$$
\begin{aligned}
p_1(t) &= c_1 e^{-2m^2 t} + c_2 e^{-2mt} + \frac{1}{m},\\
p_2(t) &= c_1 e^{-2m^2 t} + c_3 e^{-2mt} + \frac{1}{m},\\
&\cdots\\
p_{m-2}(t) &= c_1 e^{-2m^2 t} + c_{m-1} e^{-2mt} + \frac{1}{m},\\
p_{m-1}(t) &= c_1 e^{-2m^2 t} - c_2 e^{-2mt} - \cdots - c_{m-1} e^{-2mt} + \frac{1}{m},\\
p_m(t) &= -(m-1) c_1 e^{-2m^2 t} + \frac{1}{m},
\end{aligned}
$$

kde

$$
\begin{aligned}
c_1 &= \frac{\frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_{m-1}}{n}}{m-1} - \frac{1}{m},\\
c_2 &= \frac{\frac{(m-2) f_1}{n} - \frac{f_2}{n} - \cdots - \frac{f_{m-2}}{n} - \frac{f_{m-1}}{n}}{m-1},\\
c_3 &= \frac{-\frac{f_1}{n} + \frac{(m-2) f_2}{n} - \cdots - \frac{f_{m-2}}{n} - \frac{f_{m-1}}{n}}{m-1},\\
&\cdots\\
c_{m-1} &= \frac{-\frac{f_1}{n} - \frac{f_2}{n} - \cdots + \frac{(m-2) f_{m-2}}{n} - \frac{f_{m-1}}{n}}{m-1}.
\end{aligned}
$$

Důkaz: Zabývejme se nejprve přidruženou homogenní soustavou k nehomogenní LSODR1

$$
\begin{aligned}
p_1'(t) &= -4mp_1(t) - 2mp_2(t) - \cdots - 2mp_{m-1}(t),\\
p_2'(t) &= -2mp_1(t) - 4mp_2(t) - \cdots - 2mp_{m-1}(t),\\
&\cdots\\
p_{m-1}'(t) &= -2mp_1(t) - 2mp_2(t) - \cdots - 4mp_{m-1}(t).
\end{aligned}
$$

Po substituci $s = -2mt$ dostaneme homogenní soustavu, jejíž maticový zápis je

$$\mathbf{q}'(s) = \mathbf{A}\mathbf{q}(s),$$

kde

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix},$$

a $p_j(t) = q_j(-mt)$, $j = 1,...,m-1$. Po elementárních úpravách dostaneme charakteristický polynom matice **A**

$$\det\left(\lambda\mathbf{E}-\mathbf{A}\right)=\left(\lambda-1\right)^{m-2}\left(\lambda-m\right)$$

a Jordanovu normální formu matice **A**

$$\mathbf{J}=\begin{pmatrix} m & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Odtud obdržíme matici vlastních vektorů matice **A,**

$$\begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & \cdots & -1 \end{pmatrix},$$

kde první sloupec odpovídá jednoduchému charakteristickému číslu $\lambda=m$ a ostatní sloupce $\left(m-2\right)$ – násobnému charakteristickému číslu $\lambda=1$. Fundamentální matice řešení homogenní soustavy $\mathbf{q}'\left(s\right)=\mathbf{A}\mathbf{q}\left(s\right)$ pak je

$$\mathbf{Q}\left(s\right)=\begin{pmatrix} e^{ms} & e^{s} & 0 & \cdots & 0 \\ e^{ms} & 0 & e^{s} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ e^{ms} & 0 & 0 & \cdots & e^{s} \\ e^{ms} & -e^{s} & -e^{s} & \cdots & -e^{s} \end{pmatrix},$$

takže obecné řešení této soustavy je

$$\mathbf{q}\left(s\right)=\mathbf{Q}\left(s\right)\mathbf{c},$$

kde $\mathbf{c}=\left(c_1,c_2,...,c_{m-1}\right)^{\mathrm{T}}$ je matice libovolných reálných konstant. Po substituci $s=-2mt$ dostaneme obecné řešení přidružené homogenní LSODR1 a snadno nahlédneme, že $\left(\frac{1}{m},\frac{1}{m},...,\frac{1}{m}\right)^{\mathrm{T}}$ je partikulární řešení dané nehomogenní LSODR1. Tato nehomogenní soustava má vzhledem k počátečním podmínkám vždy jediné partikulární řešení, takže konstanty $c_1,c_2,...,c_{m-1}$ pro počáteční podmínky

$$p_1\left(0\right)=\frac{f_1}{n},\ p_2\left(0\right)=\frac{f_2}{n},\ldots,\ p_{m-1}\left(0\right)=\frac{f_{m-1}}{n}$$

jsou řešením soustavy lineárních algebraických rovnic

$$\mathbf{Q}\left(0\right)\mathbf{c}+\left(\frac{1}{m},\frac{1}{m},...,\frac{1}{m}\right)^{\mathrm{T}}=\left(\frac{f_1}{n},\frac{f_2}{n},...,\frac{f_{m-1}}{n}\right)^{\mathrm{T}},$$

což lze lehce ověřit dosazením za $c_1,c_2,...,c_{m-1}$ z tvrzení věty. Tak získáme hledané složky $p_1\left(t\right),p_2\left(t\right),...,p_{m-1}\left(t\right)$ gradientního odhadu $\mathbf{p}\left(t\right)$ a zbývající složku $p_m\left(t\right)$ určíme ze vztahu $p_m\left(t\right)=1-\sum_{j=1}^{m-1}p_j\left(t\right)$. Složky řešení $\mathbf{q}\left(s\right)$ jsou spojité pro $\forall s\in\mathbb{R}$ a počáteční podmínky jsou

kladné, takže také všechny složky $p_1(t), p_2(t), ..., p_{m-1}(t)$ jsou spojité a kladné pro $\forall t \in [0;\infty)$. Pro $\forall t \in [0;\infty)$ je $1 \geq e^{-m^2 t}$, takže

$$p_m(t) \geq -(m-1)\left(\frac{\frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_{m-1}}{n}}{m-1} - \frac{1}{m} \right) e^{-2m^2 t} + \frac{1}{m} e^{-2m^2 t} = \frac{f_m}{n} e^{-2m^2 t} > 0 . \square$$

Důsledek 4.1.1 Jestliže volíme za hodnoty parametru t takovou posloupnost t_n v závislosti na rozsahu výběru n, že $\lim_{n\to\infty} t_n = 0$, pak složky gradientního odhadu $\mathbf{p}(t_n) = (p_1(t_n), ..., p_m(t_n))$ z věty 4.1.1 získané z empirického rozdělení $\frac{\mathbf{f}}{n} = \left(\frac{f_1}{n}, ..., \frac{f_m}{n} \right)$ jsou asymptoticky nestranné odhady složek pozorovaného rozdělení pravděpodobnosti $\mathbf{p} = (p_1, ..., p_m)$.

Důkaz: Složky empirického rozdělení $\frac{\mathbf{f}}{n}$ jsou nestrannými odhady složek rozdělení **p**, neboť střední hodnota $E\left(\frac{f_j}{n}\right) = p_j$, $j = 1, ..., m$. Odtud jsou střední hodnoty

$$E(c_1) = \frac{p_1 + p_2 + \cdots + p_{m-1}}{m-1} - \frac{1}{m},$$

$$E(c_2) = \frac{(m-2)p_1 - p_2 - \cdots - p_{m-2} - p_{m-1}}{m-1},$$

$$\cdots$$

$$E(c_{m-1}) = \frac{-p_1 - p_2 - \cdots + (m-2)p_{m-2} - p_{m-1}}{m-1},$$

takže pro $j = 1, ..., m-2$

$$\lim_{n\to\infty} E\left[p_j(t_n) \right] = \lim_{n\to\infty} \left[E(c_1) e^{-2m^2 t_n} + E(c_{j+1}) e^{-2mt_n} + \frac{1}{m} \right] = \cdots = p_j,$$

pro $j = m-1$

$$\lim_{n\to\infty} E\left[p_{m-1}(t_n) \right] = \lim_{n\to\infty} \left[E(c_1) e^{-2m^2 t_n} - E(c_2) e^{-2mt_n} - \cdots - E(c_{m-1}) e^{-2mt_n} + \frac{1}{m} \right] = \cdots = p_{m-1},$$

a pro $j = m$

$$\lim_{n\to\infty} E\left[p_m(t_n) \right] = \lim_{n\to\infty} \left[-(m-1) E(c_1) e^{-2m^2 t_n} + \frac{1}{m} \right] = \cdots = p_m . \square$$

4.2 Aplikace gradientních odhadů

Příklad 4.2.1 Počítačovou simulací pozorování rozdělení pravděpodobnosti

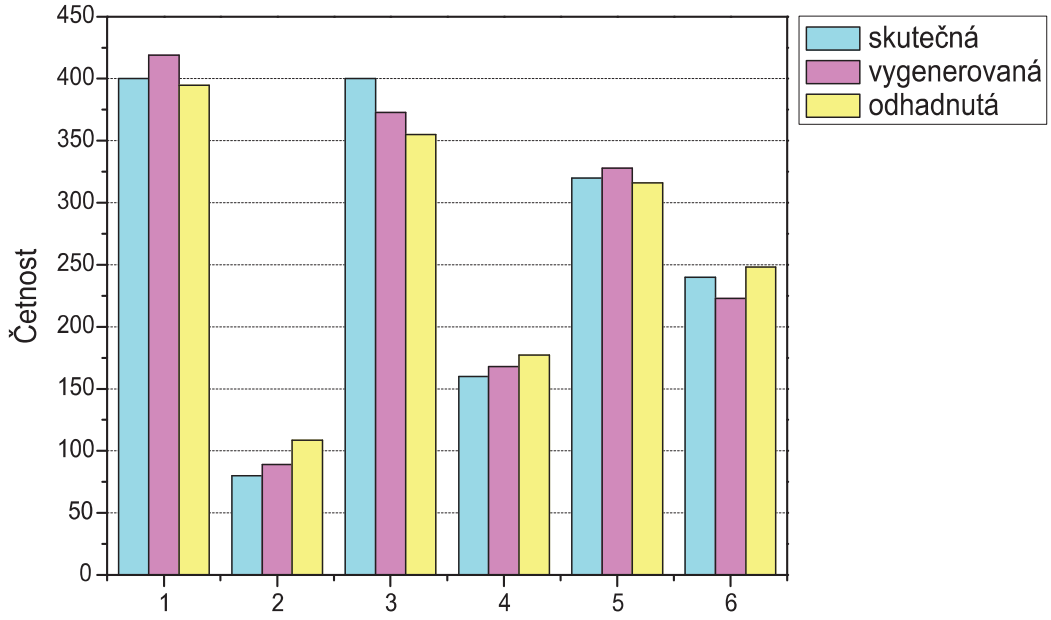
$$\mathbf{p} = (0,25; 0,05; 0,25; 0,10; 0,20; 0,15)$$

(např. na falešné šestistěnné hrací kostce) byl získán statistický soubor o rozsahu 1600. Po roztřídění pozorovaných hodnot byl v MS Excelu vypočten gradientní odhad $\mathbf{p}(t_0)$ rozdělení **p**

pomocí kvadratické kvazinormy. Hladina významnosti byla zvolena $\alpha = 0,05$. Výsledky uvedené v tabulce 4.2.1 a graf 4.2.1 ilustrují dobrou aproximaci rozdělení **p** [9].

Tab. 4.2.1 Nasimulované a odhadnuté četnosti pomocí kvadratické kvazinormy. Parametr $t_0 = 0,01191148$.

j	f_j	f_j/n	$p_j(t_0)$	$np_j(t_0)$	p_j
1	419	0,261875	0,246778	394,845	0,25
2	89	0,055625	0,067999	108,798	0,05
3	373	0,233125	0,221857	354,972	0,25
4	168	0,105000	0,110797	177,276	0,10
5	328	0,205000	0,197478	315,965	0,20
6	223	0,139375	0,155090	248,145	0,15



Obr. 4.2.1 Skutečné, vygenerované a odhadnuté četnosti.

Příklad 4.2.2 Kovový upínací přípravek byl lisován ve formě, potažen lepidlem a umístěn do lisu, aby byla vstříknuta pryž k vytvoření konečného výrobku. Při sledování neshodnosti dávky výrobků byl zjištěn celkový počet 314 vad v této dávce. Jednotlivé druhy vad jsou uvedeny v tabulce 4.2.2. Pro zjištění druhů vad, které nejvíce ovlivňují neshodnost výrobků byla provedena Paretova analýza a to z dat získaných jak ze záznamů výstupní kontroly jakosti, tak z dat odhadnutých pomocí jednotlivých kvazinorem. Počty vad získané z výstupní kontroly a odhadnuté počty vad vypočítané gradientním odhadem jsou pro jednotlivé kvazinormy uvedeny také v tabulce 4.2.2. Výpočet byl proveden pomocí vytvořeného softwaru Pesfit 1.0. V tabulce 4.2.3 jsou původní a odhadnuté procentuální kumulativní četnosti pro Paretovu analýzu [10].

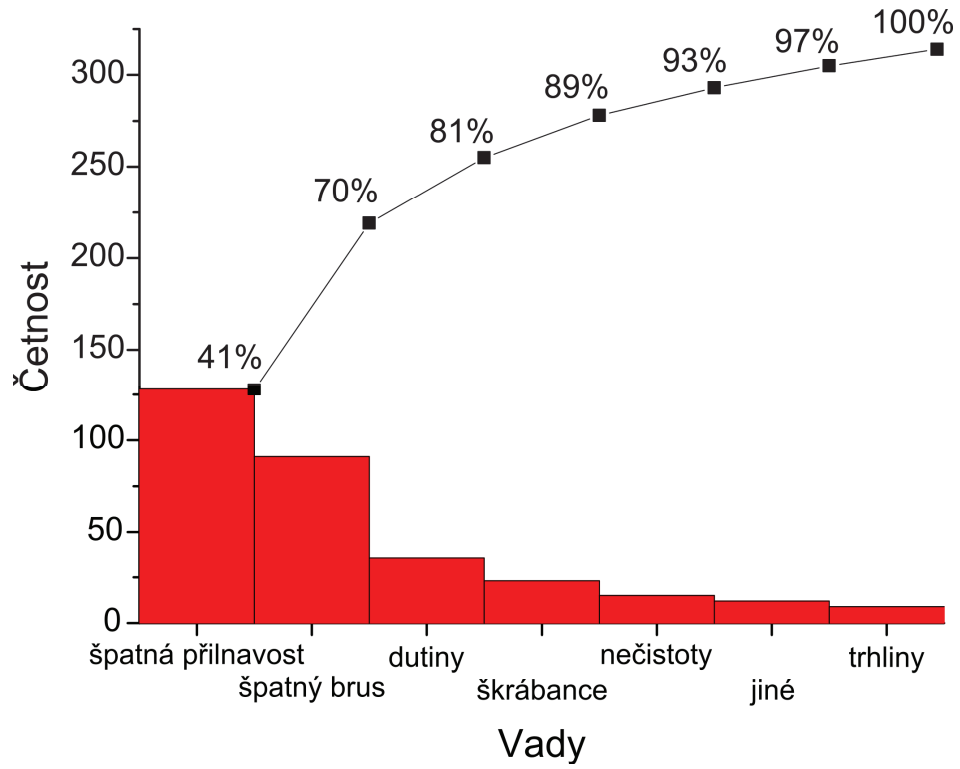
Tab. 4.2.2 Gradientní dhad četností vad pro jednotlivé kvazinormy.

Název vady	Četnosti vad	Odhadnuté četnosti vad				
		Kvadratická kvazinorma	Pearsonova kvazinorma	Hellingerova kvazinorma	Shannonova kvazinorma	G-kvazinorma $k=3$
Špatná přilnavost	128	111	126	119	116	125
Špatný brus	91	81	89	84	82	88
Dutiny	36	43	24	29	33	25
Škrábance	23	26	24	25	26	24
Nečistoty	15	20	18	20	21	19
Jiné	12	18	17	19	19	17
Trhliny	9	15	16	18	17	16

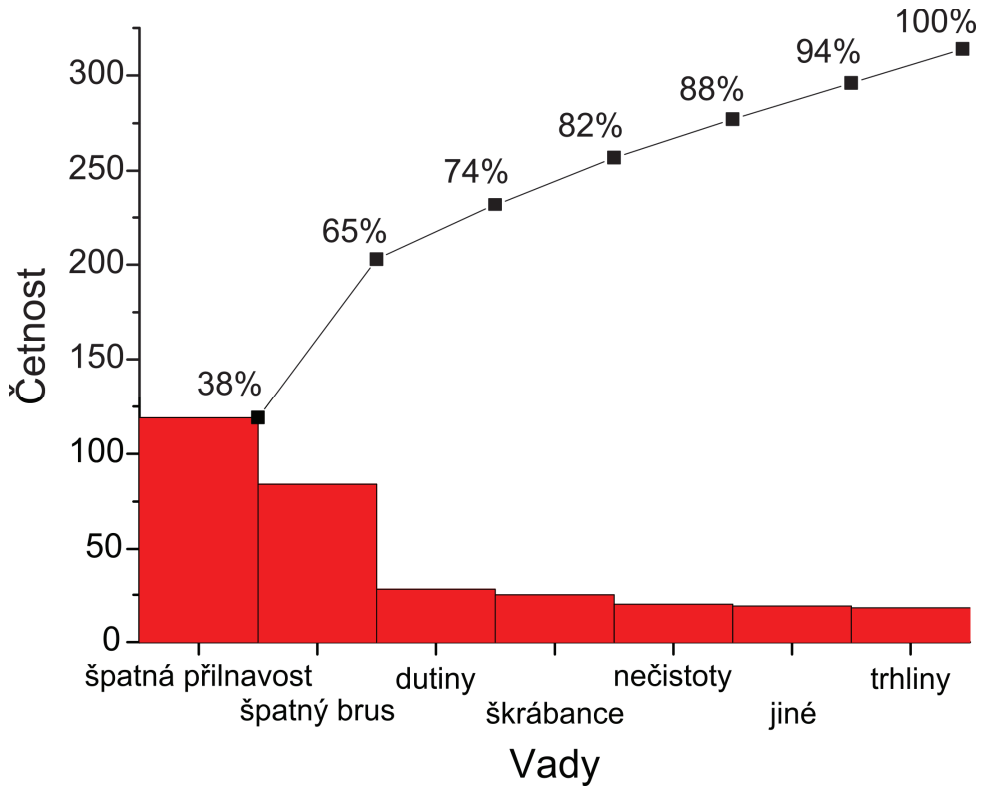
Tab. 4.2.3 Kumulativní četnosti gradientního odhadu vad.

Název vady	Kumulativní četnosti vad	Odhadnuté kumulativní četnosti vad				
		Kvadratická kvazinorma	Pearsonova kvazinorma	Hellingerova kvazinorma	Shannonova kvazinorma	G-kvazinorma $k=3$
Špatná přilnavost	41%	35%	40%	38%	37%	40%
Špatný brus	70%	61%	68%	65%	63%	68%
Dutiny	81%	75%	76%	74%	74%	76%
Škrábance	89%	83%	84%	82%	82%	83%
Nečistoty	93%	89%	89%	88%	89%	89%
Jiné	97%	95%	95%	94%	95%	95%
Trhliny	100%	100%	100%	100%	100%	100%

Na obr. 4.2.2 je Paretův diagram pro počty vad získané z výstupní kontroly a na obr. 4.2.3 je Paretův diagram pro odhadnuté počty vad vypočítané gradientním odhadem pomocí Hellingerovy kvazinormy.



Obr. 4.2.2 Paretův diagram pro data získaná z výstupní kontroly.



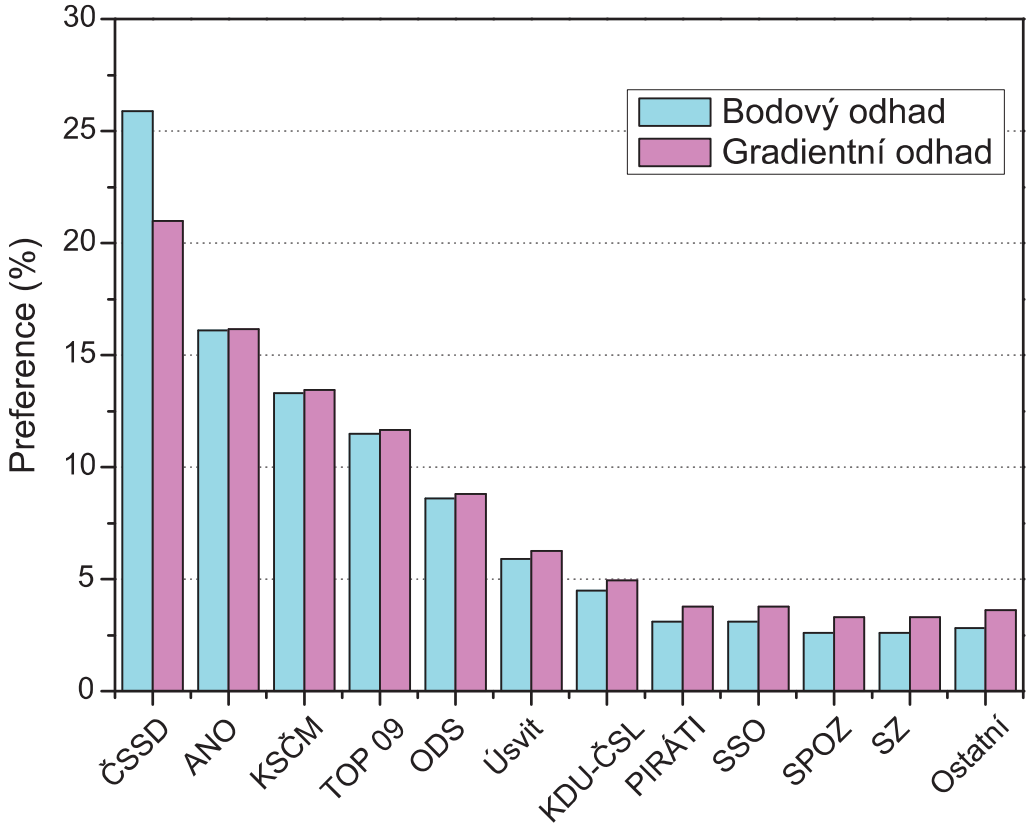
Obr. 4.2.3 Paretův diagram pro odhadnuté počty vad vypočítané pomocí gradientního odhadu Hellingerovou kvazinormou.

Pokud bychom za rozhodovací kritérium zvolili kolik příčin ze všech analyzovaných způsobuje 80% neshodných výrobků, pak by se měla firma pro zvýšení jakosti zaměřit na odstranění závad typu špatná přilnavost, špatný brus a výrobky s dutinami. Tyto tři vady způsobují celkem 81% všech závad. Pokud bychom vycházeli z dat odhadnutých pomocí kvazinorem, museli bychom být ještě opatrnější a snížit také počet vad typu škrábance.

Příklad 4.2.3
Agentura STEM provedla ve dnech 8. – 15. října 2013 na vybrané reprezentativní skupině průzkum volebních preferencí [29]. Průzkumu se zúčastnilo 1052 obyvatel ČR ve věku 18-69 let. V tabulce 4.2.4 je zaznamenán výsledek průzkumu, tj. pozorované četnosti a bodové odhady preferencí určené pomocí relativních četností. V téže tabulce jsou také odhady četností a odhady volebních preferencí vypočítané pomocí gradientního odhadu Hellingerovou kvazinormou. Výsledky dokládá připojený sloupcový graf.

Tab. 4.2.4 Odhady četností a odhady volebních preferencí vypočítané pomocí gradientního odhadu Hellingerovou kvazinormou z dat získaných předvolebním průzkumem, $t_0 = 0,0463$.

Strany	Veřejné mínění	Bodový odhad preferencí (%)	Gradientní odhad četností	Gradientní odhad preferencí (%)
ČSSD	273	25,9	221	20,98
ANO	169	16,1	170	16,15
KSČM	140	13,3	141	13,44
TOP 09	121	11,5	123	11,67
ODS	91	8,6	93	8,81
Úsvit	62	5,9	66	6,27
KDU-ČSL	47	4,5	52	4,95
PIRÁTI	33	3,1	40	3,77
SSO	33	3,1	40	3,77
SPOZ	27	2,6	35	3,29
SZ	27	2,6	35	3,29
OSTATNÍ	29	2,8	38	3,61



Obr. 4.2.4 Sloupcový graf preferencí

Tab. 4.2.5 Srovnání odhadů se skutečným výsledkem voleb.

	Bodový odhad preferencí (%)	Gradientní odhad preferencí (%)	Skutečný výsledek voleb (%)
ČSSD	25,9	20,98	20,45
ANO	16,1	16,15	18,65
KSČM	13,3	13,44	14,91
TOP 09	11,5	11,67	11,99
ODS	8,6	8,81	7,72
Úsvit	5,9	6,27	6,88
KDU-ČSL	4,5	4,95	6,78
PIRÁTI	3,1	3,77	2,66
SSO	3,1	3,77	2,46
SPOZ	2,6	3,29	1,51
SZ	2,6	3,29	3,19
OSTATNÍ	2,8	3,61	2,80

5. Přímkový odhad

Vhodný odhad $\mathbf{p}(t)=\left(p_1(t),...,p_m(t)\right)$ rozdělení pravděpodobnosti $\mathbf{p}$ z pozorovaných hodnot náhodné veličiny X můžeme v $S=\left\{\mathbf{p}\in\mathbb{R}^m : \forall p_j \geq 0, \sum_{j=1}^{m} p_j = 1\right\}$ najít tak, že se nebudeme pohybovat po křivce největšího spádu, ale po přímce vycházející z empirického rozložení četností $\frac{\mathbf{f}}{n}=\left(\frac{f_1}{n},...,\frac{f_m}{n}\right)$ procházející rovnoměrným rozdělením $\mathbf{p}_0=\left(\frac{1}{m},...,\frac{1}{m}\right)$ [11].

Definice 5.1 ***Přímkovým odhadem*** rozdělení pravděpodobnosti $\mathbf{p}\in S$ z empirického rozdělení $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right)\neq\left(\frac{1}{m},...,\frac{1}{m}\right)$ rozumíme takové rozdělení $\mathbf{p}(t)=\left(p_1(t),...,p_m(t)\right)\in S$, že

$$p_j(t)=\frac{f_j}{n}+\left(\frac{1}{m}-\frac{f_j}{n}\right)t,\ t\in T,\ j=1,...,m,$$

kde T je taková podmnožina $\mathbb{R}$, aby $p_j(t)\in[0,1]$ pro $\forall j=1,...,m.$

V případě, že $\left(\frac{f_1}{n},...,\frac{f_m}{n}\right)=\left(\frac{1}{m},...,\frac{1}{m}\right)$, rozumíme přímkovým odhadem rozdělení $\left(\frac{1}{m},...,\frac{1}{m}\right)$.

Poznámka 5.1 Složky $p_j(t)$ odhadu $\mathbf{p}(t)$ rozdělení $\mathbf{p}$ jsou konvexní kombinace odpovídajících složek $\frac{\mathbf{f}}{n}$ a $\mathbf{p}_0$, speciálně $p_j(0)=\frac{f_j}{n}$ a $p_j(1)=\frac{1}{m}$ pro $\forall j=1,...,m.$

Poznámka 5.2 Přímkový odhad je stejně jako gradientní odhad odhad s jedním parametrem. Vhodnou hodnotu parametru t_0 můžeme najít pomocí testu dobré shody. Při použití Pearsonova testu, resp. Pitmanova – Hellingerova testu [7], je t_0 kořenem nelineární rovnice

$$\frac{1}{n}\sum_{j=1}^{m}\frac{f_j^2}{p_j(t)}-n=\chi^2_{1-\alpha},$$

resp.

$$8n\left(1-\sum_{j=1}^{m}\sqrt{p_j(t)\frac{f_j}{n}}\right)=\chi^2_{1-\alpha},$$

jejíž levá strana je zvolené testové kritérium a $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil chí-kvadrát rozdělení s $m-1$ stupni volnosti. Jde vlastně o nalezení takové hodnoty *t*, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\mathbf{p}(t)$ na hladině významnosti α. Oba testy jsou pouze asymptotické a pro praktické použití požadujeme, aby $np_j(t_0)>5$ pro $\forall j=1,\ldots,m$.

Věta 5.2 Přímkový odhad je totožný s tzv. diskrétním jádrovým odhadem s mocninnými jádry [12].

$$\hat{p}_j(x)=\frac{f_j}{n}\frac{1}{1+cm}+\frac{c}{1+cm},\ \text{přičemž}\ c\in[0,\infty)\ \text{a}\ j=1,...,m.$$

Důkaz: Uvažujme rovnici přímky $p_j(t)=\frac{f_j}{n}+\left(\frac{1}{m}-\frac{f_j}{n}\right)t$, pro $t\in T$, kde T je taková podmnožina $\mathbb{R}$, aby $p_j(t)\geq 0$ pro $\forall j=1,...,m.$ Rovnici přepišme do tvaru $p_j(t)=\frac{f_j}{n}(1-t)+\frac{1}{m}t$. Porovnáme-li diskrétní jádrový odhad $\hat{p}_j(x)=\frac{f_j}{n}\frac{1}{1+cm}+\frac{c}{1+cm}$ pro $c\in[0,\infty)$ a $j=1,...,m$ s přímkovým odhadem, vidíme, že pro složky vektoru platí $\frac{1}{1+cm}=1-t\Rightarrow t=1-\frac{1}{1+cm}=\frac{cm}{1+cm}$ a $\frac{c}{1+cm}=\frac{t}{m}\Rightarrow t=\frac{cm}{1+cm}$. □

Poznámka 5.3 U přímkového odhadu si můžeme vybrat, zda budeme z pozorovaných hodnot náhodné veličiny X hledat tzv. ***pesimistický odhad*** $\mathbf{p}(t_0)=\left(p_1(t_0),...,p_m(t_0)\right)$, nebo naopak odhad nejvíce vzdálený od rovnoměrného rozdělení $\mathbf{p}_0$ ležící na hranici zamítnutí testu dobré shody. Takovému odhadu odpovídají záporné hodnoty parametru t a budeme ho nazývat ***optimistický odhad***. Podobně můžeme postupovat při gradientním odhadu z kapitoly 4.

Příklad 5.1
Ilustrační příklad přímkového odhadu pro $m=2$.

$$p_1(t)=\frac{f_1}{n}+\left(\frac{1}{2}-\frac{f_1}{n}\right)t,$$

$$p_2(t)=1-\left[\frac{f_1}{n}+\left(\frac{1}{2}-\frac{f_1}{n}\right)t\right].$$

K nalezení vhodné hodnoty parametru t použijeme například Pearsonův test

$$\frac{1}{n}\left(\frac{f_1^2}{p_1(t)}-\frac{(n-f_1)^2}{p_2(t)}\right)-n-\chi^2_{1-\alpha}=0.$$

Po dosazení pravděpodobností $p_1(t), p_2(t)$ do Pearsonova testu dostaneme rovnici

$$\frac{1}{n}\left(\frac{f_1^2}{\frac{f_1}{n}+\left(\frac{1}{2}-\frac{f_1}{n}\right)t}-\frac{(n-f_1)^2}{1-\left[\frac{f_1}{n}+\left(\frac{1}{2}-\frac{f_1}{n}\right)t\right]}\right)-n-\chi^2_{1-\alpha}=0.$$

Odtud po úpravě

$$\frac{2f_1^2}{2f_1+nt-2f_1t}+\frac{2n^2-4nf_1+2f_1^2}{2n-2f_1-nt+2f_1t}-n-\chi^2_{1-\alpha}=0.$$

Z předchozího vztahu dostaneme kvadratickou rovnici vzhledem k hledanému parametru t

$$t^2\left(n^3-4n^2f_1+4nf_1^2+\chi^2_{1-\alpha}n^2-4\chi^2_{1-\alpha}nf_1+4\chi^2_{1-\alpha}f_1^2\right)+t\left(8\chi^2_{1-\alpha}nf_1-8\chi^2_{1-\alpha}f_1^2-2\chi^2_{1-\alpha}n^2\right)+$$
$$+4\chi^2_{1-\alpha}f_1^2-4\chi^2_{1-\alpha}nf_1=0.$$

Nyní dosadíme do rovnice konkrétní hodnoty. Uvažujme například falešnou kostku, na které nám ze sta pokusů padlo pouze třicetkrát sudé číslo. Pak

$$n = 100,$$
$$f_1 = 30,$$
$$\chi^2_{1-\alpha} \doteq 3,8415.$$

Po dosazení má kvadratická rovnice tvar

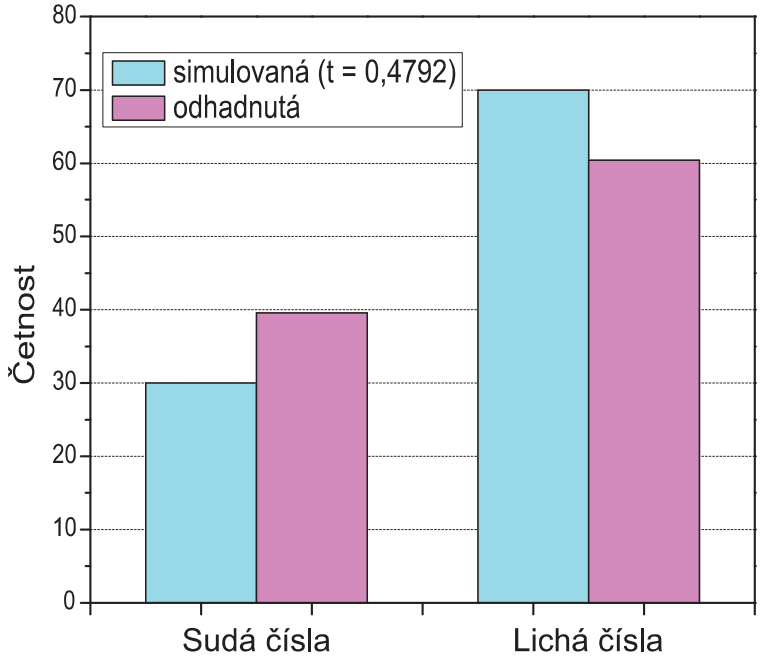
$$166146,3285t^2 - 12292,6571t - 32268,22484 = 0.$$

Řešení kvadratické rovnice

$$t_1 = 0,4792,$$
$$t_2 = -0,4053.$$

Odhad vypočítaný pro parametr t_1

$$p_1 = 0,3958,$$
$$p_2 = 0,6042.$$

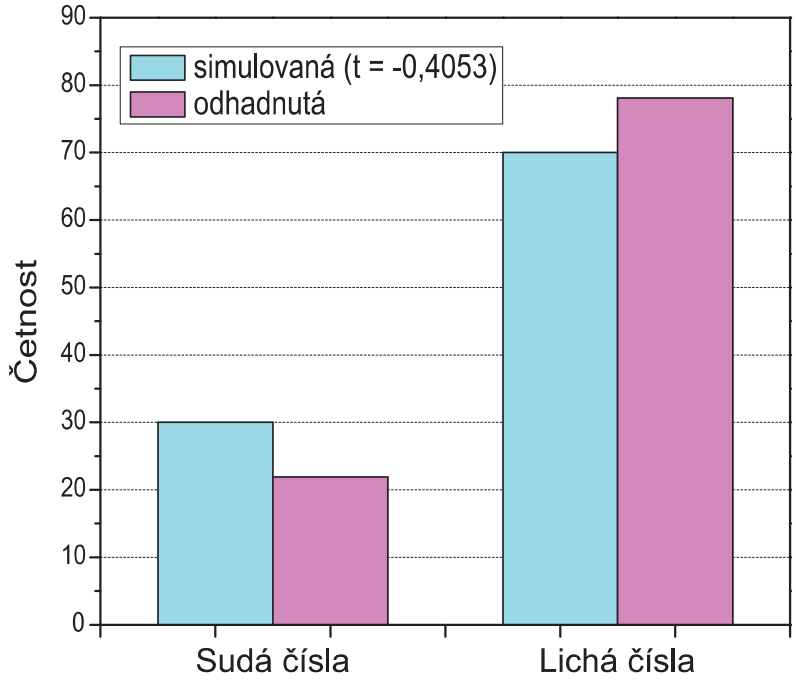


Obr. 5.1 Sloupcový graf četností pro falešnou kostku

Vypočítaný parametr $t_2 < 0$. Jde o optimistický přímkový odhad, kdy hledáme odhad nejvzdálenější od rovnoměrného rozdělení $\mathbf{p}_0$ ležící na hranici zamítnutí testu dobré shody.

Odhad vypočítaný pro parametr t_2

$$p_1 = 0,2189,$$
$$p_2 = 0,7811.$$



Obr. 5.2 Sloupcový graf četností pro falešnou kostku, pohybujeme-li se směrem od rovnoměrného rozdělení

Příklad 5.2 V tabulce 5.1 je uveden počet manželství v České republice, která nevydržela ani jeden rok [19]. Vypočítaný je odhad četností rozvodů pomocí jednotlivých kavazinorem a přímkového odhadu. U přímkového odhadu jsou uvedeny oba dva typy. Jak odhad pesimistický blížící se k rovnoměrnému rozdělení, tak odhad optimistický. Uvedena je také kritická hodnota parametru t_0, kdy ještě nezamítáme vhodnost daného rozdělení na hladině významnosti $\alpha = 0{,}05$.

Tab 5.1 Skutečné a odhadnuté počty manželství v ČR, která nevydržela ani jeden rok.

Rok	**2012**	**2011**	**2010**	**2009**	**2008**	**2007**	**2006**	t_0
Skutečné četnosti	204	235	254	158	277	341	307	0
Kvadratická kvazinorma	220	244	259	183	277	328	264	0,0379
Pearsonova kvazinorma	217	241	257	185	277	336	263	0,0383
Hellingerova kvazinorma	218	243	258	184	277	332	263	0,0675
Shannonova kvazinorma	219	243	259	184	277	331	263	0,0541
G-kvazinorma ($k = 4/3$)	218	243	258	184	277	333	263	0,0535
Přímkový odhad optimistický	188	229	254	127	285	369	324	0,3263
Přímkový odhad pesimistický	222	242	254	193	269	309	288	0,3609

6. Gradientní a přímkové odhady na kontingenčních tabulkách

Pojmy uvedené v definici 6.1 a tabulky 6.1, 6.2 jsou převzaty z [13].

Definice 6.1 Nechť náhodný vektor $\mathbf{Z} = (X, Y)$ má diskrétní rozdělení, přičemž veličina X nabývá hodnot $1,\ldots,r$ a veličina Y nabývá hodnot $1,\ldots,s$, $r,s \geq 2$. Označme

$$f_{i.} = \sum_{j=1}^{s} f_{ij}\,,\ f_{.j} = \sum_{i=1}^{r} f_{ij}$$

Předpokládejme, že se uskutečnil náhodný výběr o rozsahu n z tohoto rozdělení. Nechť f_{ij} je počet těch případů, kdy se ve výběru vyskytla dvojice (i, j). Matici $\left(f_{ij}\right)$ pak nazveme ***kontingenční tabulka*** a čísla $f_{i.}$ a $f_{.j}$ ***marginální četnosti*** (viz. Tabulka 6.1).

Tab. 6.1 Kontingenční tabulka.

X	Y $1\ldots s$			$\sum$
1	f_{11}	$\cdots$	f_{1s}	$f_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	f_{r1}	$\cdots$	f_{rs}	$f_{r.}$
$\sum$	$f_{.1}$	$\cdots$	$f_{.s}$	n

Nyní označme $p_{ij} = \mathrm{P}(X = i, Y = j)$, $p_{i.} = \sum_{j=1}^{s} p_{ij}$, $p_{.j} = \sum_{i=1}^{r} p_{ij}$. Čísla $p_{i.}$ a $p_{.j}$ se nazývají ***marginální pravděpodobnosti***, rozdělení pravděpodobnosti vektoru $\mathbf{Z}$ zapisujeme maticí (viz. Tabulka 6.2).

Tab. 6.2 Matice pravděpodobností.

X	Y $1\ldots s$			$\sum$
1	p_{11}	$\cdots$	p_{1s}	$p_{1.}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	p_{r1}	$\cdots$	p_{rs}	$p_{r.}$
$\sum$	$p_{.1}$	$\cdots$	$p_{.s}$	1

Poznámka 6.1 Náhodné veličiny f_{ij} mají sdružené multinomické rozdělení s parametrem n a s pravděpodobnostmi p_{ij}.

Poznámka 6.2 Náhodný vektor $\mathbf{Z}$ reprezentuje náhodný výběr při n pozorováních dvourozměrné kategoriální veličiny $(\mathbf{A}, \mathbf{B})$, kde $\mathbf{A}$ nabývá r různých kvalitativních hodnot a $\mathbf{B}$

nabývá s různých kvalitativních hodnot. $(\mathbf{A},\mathbf{B})$ má simultání rozdělení (matici) pravděpodobností $\mathbf{p}\in\mathbf{S}=\left\{p_{i,j}\geq 0,(i,j)\in\{1,\ldots r\}\times\{1,\ldots s\},\sum_{i=1}^{r}\sum_{j=1}^{s}p_{ij}=1\right\}$. Odhady pravděpodobností p_{ij} z pozorovaných četností f_{ij} můžeme najít pomocí následujícího gradientního odhadu (analogie definice 4.1).

Definice 6.2 Nechť $D_f(\mathbf{p},\mathbf{p_0})$ je kvazinorma na *S,* kde $\mathbf{p}_0=\left\{\frac{1}{rs},\ldots,\frac{1}{rs}\right\}$. Gradientním odhadem rozdělení pravděpodobnosti $\mathbf{p}\in S$ z empirického rozdělení četností f_{ij} z tabulky 6.1 rozumíme takové rozdělení pravděpodobnosti $\mathbf{p}(t)\in S$, že

$$\frac{d}{dt}\mathbf{p}(t)=-gradD_f\left(\mathbf{p}(t),\mathbf{p}_0\right)\text{ pro }\forall t\in[0;\infty),$$

$$\mathbf{p}(0)=\left(\frac{f_{11}}{n},\ldots,\frac{f_{ij}}{n},\ldots,\frac{f_{rs}}{n}\right).$$

V případě, že $\frac{f_{ij}}{n}=\frac{1}{rs}$ pro $\forall$ *i,j*, klademe $\mathbf{p}(t)=\left(\frac{1}{rs},\ldots,\frac{1}{rs}\right)$ pro $\forall\ t\in[0;\infty)$.

Věta 6.1 Jestliže funkce $f(u)$, která generuje kvazinormu $D_f(\mathbf{p},\mathbf{p}_0)$ na S, má vlastnosti uvedené v definici 2.1 a má spojitou derivaci $f'(u)$ pro $\forall u\in(0;\infty)$, pak existuje jediný gradientní odhad $\mathbf{p}(t)=\left(p_{11}(t),\ldots,p_{ij}(t),\ldots,p_{rs}(t)\right)$ rozdělení pravděpodobnosti $\mathbf{p}\in S$. Jeho složky $p_{ij}(t)$, $(i,j)\in\{1,\ldots,r\}\times\{1,\ldots,s\}-\{(r,s)\}$, jsou pro $\forall t\in[0;\infty)$ partikulárním řešením soustavy obyčejných diferenciálních rovnic prvního řádu (SODR1)

$$p_{ij}(t)=-f'\left(rsp_{ij}(t)\right)+f'\left(rs\left[1-\sum_{\substack{i=1\\(i,j)\neq(r,s)}}^{r}\sum_{j=1}^{s}p_{ij}(t)\right]\right),$$

s počátečními podmínkami

$$p_{ij}(0)=\frac{f_{ij}}{n}$$

a složkou

$$p_{rs}(t)=1-\sum_{\substack{i=1\\(i,j)\neq(r,s)}}^{r}\sum_{j=1}^{s}p_{ij}(t).$$

Důkaz je totožný s důkazem věty 4.1. □

Poznámka 6.3 Pro kvadratickou kvazinormu dostaneme explicitní řešení SODR z věty 6.1 ve tvaru

$$p_{11}(t) = c_{11}e^{-2r^2s^2t} + c_{12}e^{-2rst} + \frac{1}{rs},$$
$$p_{12}(t) = c_{11}e^{-2r^2s^2t} + c_{13}e^{-2rst} + \frac{1}{rs},$$
$$\dots$$
$$p_{r,s-2}(t) = c_{11}e^{-2r^2s^2t} + c_{r,s-1}e^{-2rst} + \frac{1}{rs},$$
$$p_{r,s-1}(t) = c_{11}e^{-2r^2s^2t} - c_{12}e^{-2rst} - \dots - c_{r,s-1}e^{-2rst} + \frac{1}{rs},$$

kde

$$c_{11} = \frac{\frac{f_{11}}{n} + \frac{f_{12}}{n} + \dots + \frac{n_{r,s-1}}{n}}{rs-1} - \frac{1}{rs},$$
$$c_{12} = \frac{\frac{(rs-2)f_{11}}{n} - \frac{f_{12}}{n} - \dots - \frac{f_{r,s-1}}{n}}{rs-1},$$
$$\dots$$
$$c_{r,s-1} = \frac{-\frac{f_{11}}{n} - \frac{f_{12}}{n} - \dots + \frac{(rs-2)f_{r,s-2}}{n} - \frac{f_{r,s-1}}{n}}{rs-1},$$

a

$$p_{rs}(t) = -(rs-1)c_{11}e^{-2r^2s^2t}.$$

Poznámka 6.4 Přímkový odhad $\mathbf{p} \in S$ z empirického rozdělení z tabulky 6.1 definujeme podobně jako přímkový odhad z definice 5.1:

$$p_{ij}(t) = \frac{f_{ij}}{n} + \left(\frac{1}{rs} - \frac{f_{ij}}{n}\right)t,\ t \in T, i = 1,\dots,r,\ j = 1,\dots,s,$$

kde T je taková podmnožina $\mathbb{R}$, aby $p_{ij}(t) \in [0,1]$ pro $\forall\ i = 1,\dots,r,\ j = 1,\dots,s$.

V případě, že $\frac{f_{ij}}{n} = \frac{1}{rs}$ pro $\forall\ i,j$, klademe $p_{ij}(t) = \left(\frac{1}{rs},\dots,\frac{1}{rs}\right)$ pro $\forall\ t \in [0;\infty)$.

Poznámka 6.5 Vhodné hodnoty parametrů t pro gradientní i přímkový odhad stanovíme pomocí Pearsonova nebo Pittmanova-Hellingerova testového kritéria stejně jako v kapitolách 4 a 5.

Například při použití Pearsonova testu je t_0 kořenem nelineární rovnice

$$n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{f_{i.}f_{.j}} - n = \chi^2,$$

jejíž levá strana je zvolené testové kritérium a $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil chí-kvadrát rozdělení s $rs-1$ stupni volnosti. Oba testy jsou pouze asymptotické a pro praktické použití požadujeme, aby $np_{ij}(t_0) > 5$ pro $\forall\ i = 1,\dots,r,\ j = 1,\dots,s$.

Při zpracování dat se nejčastěji setkáváme s testováním statistické hypotézy, že náhodné veličiny X a Y jsou na sobě nezávislé. Využijeme-li gradientního odhadu, můžeme hledat odhad pravděpodobností p_{ij}, respektive četností n_{ij} za určitých vstupních podmínek. Zjistíme-li, například, že pro náhodné veličiny X, Y hypotézu o nezávislosti nezamítáme, budeme hledat odhad takový, abychom nezamítli hypotézu o nezávislosti a zároveň vhodnosti daného rozdělení. Další možností je hledat rozdělení, které je svázáno s původním rozdělením přes marginální řádkové či sloupcové pravděpodobnosti.

Poznámka 6.6 Gradientním odhadem se přiblížíme po křivce největšího spádu k rovnoměrnému rozdělení, pro které máme nezávislost zaručenou.

6.1 Příklady užití gradientních odhadů na kontingenční tabulky

Příklad 6.1.1 U 125 studentů VUT byl hodnocen dojem, jaký mají studenti po napsání závěrečného testu z předmětu Aplikovaná statistika a plánování experimentu. Výsledky experimentu jsou zaznamenány v tabulce 6.1.1 i s výsledným hodnocením zkoušky. Pomocí kvadratické kvazinormy a přímkového odhadu najděte „pesimistický" odhad četností. Na asymptotické hladině významnosti 0,05 otestujte hypotézu, že výsledek testu (zkoušky) nezávisí na dojmu studentů po napsání testu a ověřte, že odhadnuté četnosti splňují původní předpoklad nezávislosti (závislosti). Gradientní odhad kvadratickou kvazinormou i přímkový odhad byly vypočteny pomocí řešitele v MS Excelu.

Tab. 6.1.1 Tabulka skutečných četností.

zkouška	dojem		$f_{j.}$
	dobrý	špatný	
ano	19	13	32
ne	37	56	93
$f_{.k}$	56	69	125

Tab. 6.1.2 Odhad kvadratickou kvazinormou, parametr $t_0 = 0,02523$.

zkouška	dojem		$f_{j.}$
	dobrý	špatný	
ano	24	20	44
ne	39	42	81
$f_{.k}$	63	62	125

Tab. 6.1.3 Přímkový odhad, parametr $t_0 = 0,46387$.

zkouška	dojem		$f_{j.}$
	dobrý	špatný	
ano	25	21	46
ne	34	45	79
$f_{.k}$	59	66	125

Jedná se o čtyřpolní tabulku, můžeme tedy k testování hypotézy o nezávislosti užít charakteristiku OR, tzv. podíl šancí. V prvním případě obdržíme interval spolehlivosti $(0,976;5,016)$, pro odhad kvadratickou kvazinormou dostaneme interval $(0,649;2,842)$ a pro přímkový odhad $(0,719;3,094)$. Ve všech případech zahrne interval spolehlivosti jedničku, nezamítáme tedy na hladině významnosti 0,05 hypotézu o nezávislosti výsledku zkoušky na dojmu studenta. Hypotézu na hladině významnosti 0,05 nezamítáme ani v případě, použijeme-li χ^2-test. V tomto případě je pro skutečné četnosti hodnota $\chi^2 = 3,79 \leq 3,84$, pro kvadratickou kvazinormu je hodnota $\chi^2 = 0,66 \leq 3,84$ a pro přímkový odhad $\chi^2 = 1,18 \leq 3,84$.

Příklad 6.1.2 V severozápadním Skotsku byla provedena studie, která měla prokázat, zda je procentuální zastoupení krevních skupin na celém území homogenní či nikoli. Výsledky pozorování převzaté ze sbírky Osborn (1979), která je citována v [13], jsou uvedeny v tabulce 6.1.4.

Tab. 6.1.4 Tabulka četností zastoupení krevních skupin v jednotlivých oblastech.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	33	6	56	5	100
Annandale	54	14	52	5	125
Nithsdale	98	35	115	5	253
Celkem	185	55	223	15	478

Budeme testovat hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné. Testová statistika má tvar $\chi^2 = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{f_{i.}f_{.j}} - n$. Hypotézu zamítáme na asymptotické hladině významnosti α, když $\chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha)$. Neboť $\chi^2 = 10,45$ a $\chi^2_6(0,05) \doteq 12,59$, hypotézu nezamítáme [13].

Nyní vypočítáme „pesimistický" odhad pomocí kvadratické kvazinormy (výsledky jsou uvedeny v tabulce 6.1.5) a otestujeme nezávislost. Výpočty byly provedeny pomocí Řešitele v MS Excelu.

Tab. 6.1.5 Odhad pomocí kvadratické kvazinormy, parametr $t_0 = 0,0034$.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	32	7	53	6	98
Annandale	51	14	49	6	120
Nithsdale	92	33	108	27	260
Celkem	175	54	210	39	478

Hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné na hladině významnosti $\alpha = 0,05$ nezamítáme, neboť $\chi^2 = 10,305 \leq 12,59$ a zároveň nezamítáme hypotézu o vhodnosti nalezeného rozdělení.

Nyní provedeme opět odhad pomocí kvadratické kvazinormy a přidáme požadavek na marginální četnosti. Počet stupňů volnosti bude tedy $rs-2$. Odhadnuté četnosti jsou uvedeny v tabulkách 6.1.6.- 6.1.9.

Tab. 6.1.6 Odhad pomocí kvadratické kvazinormy za požadavku, aby v oblasti Eskade bylo vybráno 100 osob. Parametr $t = 0,000281$.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	33	6	56	5	100
Annandale	53	14	52	5	124
Nithsdale	97	35	114	8	254
Celkem	183	55	222	18	478

Hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné na hladině významnosti $\alpha = 0,05$ nezamítáme, neboť $\chi^2 = 9,937 \leq 12,59$. V tomto případě již nejde o odhad pesimistický, ale našli jsme rozdělení, pro které nezamítáme hypotézu o vhodnosti nalezeného rozdělení ($\chi^2 = 1,16295 \leq 18,30703$) a které splňuje požadavek na marginální četnost. Skutečnost, že nezamítáme hypotézu o vhodnosti nalezeného rozdělení, snadno poznáme také ze skutečnosti, že parametr $t = 0,000281$ je výrazně menší, než jeho kritická hodnota t_0 uvedená v tabulce 6.1.5.

Tab. 6.1.7 Odhad pomocí kvadratické kvazinormy za požadavku, aby v oblasti Annandale bylo vybráno právě 125 osob. Parametr $t = 0,000207$.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	33	6	55	5	99
Annandale	54	14	52	5	125
Nithsdale	98	35	114	7	254
Celkem	184	55	222	17	478

Hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné na hladině významnosti $\alpha = 0,05$ nezamítáme, neboť $\chi^2 = 6,161 \leq 12,59$. Hypotézu o vhodnosti nalezeného rozdělení nezamítáme, $\chi^2 = 0,59838 \leq 18,30703$.

Pokud provedeme odhad pomocí kvadratické kvazinormy za požadavku, aby v oblasti Nithsdale bylo vybráno právě 253 osob, pak je parametr *t* roven téměř nule a dostaneme původní četnosti. V případě, že bychom připustili řešení, kdy na hladině významnosti $\alpha = 0,05$ zamítáme hypotézu o vhodnosti rozdělení, dostaneme četnosti uvedené v tabulce 6.1.8.

Tab. 6.1.8 Odhad pomocí kvadratické kvazinormy za požadavku, aby v oblasti Nithsdale bylo vybráno právě 253 osob. Parametr $t = 0,0092$, $\chi^2 = 41,9102 \geq 18,30703$.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	32	10	50	10	102
Annandale	49	17	47	10	123
Nithsdale	84	34	98	37	253
Celkem	165	61	195	57	478

Stejně tak, jak jsme kladli podmínky na marginální četnosti pro jednotlivé oblasti, můžeme klást požadavky na zkoumaný počet lidí s určitou krevní skupinou. Například budeme hledat rozdělení takové, aby bylo vybráno právě 55 lidí s krevní skupinou B. Výsledky jsou uvedeny v tabulce 6.1.9.

Tab. 6.1.9 Odhad pomocí kvadratické kvazinormy za požadavku, že lidí s krevní skupinou B bylo vybráno právě 55. Parametr $t = 0,0027$.

Oblast	Krevní skupina				
	A	B	0	AB	celkem
Eskade	32	7	53	6	98
Annandale	52	14	50	5	121
Nithsdale	93	34	109	23	259
Celkem	177	55	212	34	478

Hypotézu, že rozdělení krevních skupin je ve všech třech oblastech stejné na hladině významnosti $\alpha = 0,05$ nezamítáme, neboť $\chi^2 = 10,416 \leq 12,59$. Stejně tak nezamítáme hypotézu o vhodnosti nalezeného rozdělení, $\chi^2 = 15,3830 \leq 18,30703$.

Příklad 6.1.3
U 130 uchazečů o zaměstnání bylo sledováno, zda byli přijati do zaměstnání či nikoliv v závislosti na jejich pohlaví. Výsledky průzkumu jsou uvedeny v tabulce 6.1.10. Pomocí přímkových odhadů (pesimistický, optimistický) byl nalezen závislý i nezávislý odhad tohoto rozdělení pravděpodobnosti. Výpočty byly provedeny pomocí Řešitele v MS Excelu.

Tab. 6.1.10 Tabulka původních četností.

zaměstnavatel A	muž	žena
přijati	20	13
nepřijati	39	58

Jedná se o čtyřpolní tabulku, je tedy možné k testování hypotézy o nezávislosti použít opět tzv. podíl šancí. Hypotézu o nezávislosti na asymptotické hladině $0,05$ zamítáme, interval spolehlivosti $(1,0202;5,1312)$ neobsahuje jedničku. Stejně tak hypotézu zamítáme i v případě, že použijeme χ^2-test, protože $\chi^2 = 4,23 > 3,84$.

Nyní najdeme odhad, kdy hypotézu o nezávislosti nezamítneme. K výpočtu použijeme přímkový odhad. Výsledky jsou uvedeny v tabulce 6.1.11.

Tab. 6.1.11 Četnosti vypočítané pomocí přímkového odhadu, $t_0 = 0,44928$.

zaměstnavatel A	muž	žena
přijati	26	22
nepřijati	36	47

Pro přímkový odhad dostaneme interval spolehlivosti $(0,7406;3,1135)$ a pro χ^2-test, platí $\chi^2 = 1,33 \leq 3,84$. Hypotézu o nezávislosti tedy na asymptotické hladině 0,05 nezamítáme.

Pomocí modifikovaného přímkového odhadu je možné najít řešení, které splňuje původní předpoklad, tedy řešení, kdy hypotézu o nezávislosti zamítneme. Odhadnuté četnosti jsou

uvedeny v tabulce 6.1.12. Interval spolehlivosti je v tomto případě $(1,4292;10,1259)$ a pro χ^2-test platí $\chi^2 = 7,96 > 3,84$.

Tab. 6.1.12 Četnosti vypočítané pomocí modifikovaného přímkového odhadu, $t_0 = -0,319136$.

zaměstnavatel A	muž	žena
přijati	16	7
nepřijati	41	66

7. Nová třída kvazinorem

V této kapitole odvodíme z tzv. Cressie-Readovy statistiky [15] celou třídu *f*-divergencí vedoucí na asymptotický chí-kvadrát test. Pro odvozenou kvazinormu ukážeme jistou analogii s G-kvazinormou a zbývajícími dosud používanými kvazinormami a odvodíme kvazinormu k ní duální.

Věta 7.1 Jestliže $\lambda \in \mathbb{R}$, pak Cressie-Readova statistika

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} f_j \left\{ \left(\frac{f_j}{np_j} \right)^{\lambda} - 1 \right\},$$

speciálně pro $\lambda = 0$ a $\lambda = -1$

$$2nI^{-1} = \lim_{\lambda \to -1} \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} f_j \left\{ \left(\frac{f_j}{np_j} \right)^{\lambda} - 1 \right\} = -2n \sum_{j=1}^{m} p_j \ln \left(\frac{f_j}{np_j} \right),$$

$$2nI^{0} = \lim_{\lambda \to 0} \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} f_j \left\{ \left(\frac{f_j}{np_j} \right)^{\lambda} - 1 \right\} = 2 \sum_{j=1}^{m} f_j \ln \left(\frac{f_j}{np_j} \right),$$

má asymptoticky chí-kvadrát rozdělení s $m-1$ stupni volnosti pro všechna λ.

Důkaz: Pro $\lambda \neq 0, \lambda \neq -1$ přepišme $2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} f_j \left\{ \left(\frac{f_j}{np_j} \right)^{\lambda} - 1 \right\}$ ve tvaru

$$2nI^{\lambda} = \frac{2n}{\lambda(\lambda+1)} \sum_{j=1}^{m} p_j \left\{ \left(1 + \frac{f_j - np_j}{np_j} \right)^{\lambda+1} - 1 \right\}.$$

Označíme-li $V_j = \frac{f_j - np_j}{np_j}$ a výraz $\left(1+V_j\right)^{\lambda+1}$ rozvineme v Taylorovu řadu, dostaneme

$$2nI^{\lambda} = \frac{2n}{\lambda(\lambda+1)} \sum_{j=1}^{m} p_j \left(1 + (\lambda+1)V_j + \frac{\lambda(\lambda+1)}{2} V_j^2 + o_p(1) - 1 \right) = \frac{2n}{\lambda} \sum_{j=1}^{m} p_j \left(V_j + \frac{\lambda V_j^2}{2} \right) +$$

$$+ o_p(1) = \sum_{j=1}^{m} p_j \left(n^{1/2} V_j \right)^2 + o_p(1),$$ přičemž $o_p(1)$ konverguje k nule pro $n \to \infty$. Pak

$$2nI^{\lambda} = \sum_{j=1}^{m} \frac{\left(f_j - np_j \right)^2}{np_j} \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\}.$$

Speciálně pro $\lambda = -1$ dostaneme

$2nI^{-1} = -2n \sum_{j=1}^{m} p_j \ln \left(\frac{f_j}{np_j} \right) = -2n \sum_{j=1}^{m} p_j \ln \left(1 + \frac{f_j - np_j}{np_j} \right)$ a funkci $\ln\left(1+V_j\right)$ rozvineme v Taylorovu řadu, pak $2nI^{-1} = -2n \sum_{j=1}^{m} p_j \left(V_j - \frac{V_j^2}{2} + o_p(1) \right)$, přičemž $o_p(1)$ konverguje k nule pro $n \to \infty$. Odtud

$$2nI^{-1} = \sum_{j=1}^{m} \frac{\left(f_j - np_j \right)^2}{np_j}.$$

Pro $\lambda = 0$ dostaneme

$$2nI^0 = 2\sum_{j=1}^{m} f_j \ln\left(\frac{f_j}{np_j}\right) = -2\sum_{j=1}^{m} f_j \ln\left(\frac{np_j}{f_j}\right) = -2\sum_{j=1}^{m} f_j \ln\left(1 + \frac{np_j - f_j}{f_j}\right).$$

Nyní položíme $V_j = \dfrac{np_j - f_j}{f_j}$ a funkci $\ln(1+V_j)$ opět rozvineme v Taylorovu řadu. Pak

$$2nI^0 = -2\sum_{j=1}^{m} f_j \left(V_j - \frac{V_j^2}{2} + o_p(1)\right) = \sum_{j=1}^{m} \frac{\left(np_j - f_j\right)^2}{f_j} + o_p(1) = \sum_{j=1}^{m} \frac{\left(f_j - np_j\right)^2}{f_j} \frac{np_j}{np_j} + o_p(1).$$

Protože podíl $\dfrac{np_j}{f_j}$ konverguje k jedné pro $n \to \infty$ a $o_p(1)$ konverguje k nule pro $n \to \infty$, pak

$$2nI^0 = \sum_{j=1}^{m} \frac{\left(f_j - np_j\right)^2}{np_j}. \square$$

Poznámka 7.1 Z třídy Cressie-Readových statistik je možné vzhledem k λ odvodit třídu *f*-divergencí a následně kvazinorem vedoucí na asymptotický chí-kvadrát test.

Označíme-li $2nI^\lambda = \dfrac{2n}{\lambda(\lambda+1)} \sum_{j=1}^{m} q_j \left\{\left(\dfrac{q_j}{p_j}\right)^\lambda - 1\right\}$, pak platí pro odvozenou třídu kvazinorem

$$I^\lambda = \frac{1}{\lambda(\lambda+1)m} \sum_{j=1}^{m} \left\{\left(\frac{1}{mp_j}\right)^\lambda - 1\right\}, \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\}.$$

Speciálně pro $\lambda = 0$ a $\lambda = -1$

$$I^0 = \sum_{j=1}^{m} q_j \ln\left(\frac{q_j}{p_j}\right) = \sum_{j=1}^{m} \frac{1}{m} \ln\left(\frac{1}{mp_j}\right) = \frac{1}{m}\sum_{j=1}^{m} \ln\frac{1}{m} - \ln p_j,$$

$$I^{-1} = \sum_{j=1}^{m} p_j \ln\left(\frac{p_j}{q_j}\right) = \sum_{j=1}^{m} p_j \ln\left(\frac{p_j}{m}\right) = \sum_{j=1}^{m} p_j \ln p_j + \ln m.$$

Definice 7.1 Kvazinormu

$$CR_f(\mathbf{p}, \mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m} \sum_{j=1}^{m} \left\{\left(\frac{1}{mp_j}\right)^\lambda - 1\right\} \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\},$$

speciálně pro $\lambda = 0$

$$CR_f(\mathbf{p}, \mathbf{p_0}) = \frac{1}{m}\sum_{j=1}^{m} \ln\frac{1}{m} - \ln p_j$$

a pro $\lambda = -1$

$$CR_f(\mathbf{p}, \mathbf{p_0}) = \sum_{j=1}^{m} p_j \ln p_j + \ln m,$$

budeme nazývat ***Cressie-Readova kvazinorma***.

Poznámka 7.2 Generující funkce Cressie-Readovy kvazinormy má tvar

$$f(u)=\frac{1}{\lambda(\lambda+1)}\left(u^{-\lambda}-1\right) \text{ pro } \lambda\in\mathbb{R}-\{-1,0\},$$

speciálně pro $\lambda=0$

$$f(u)=-\ln(u)$$

a pro $\lambda=-1$

$$f(u)=u\ln(u).$$

Poznámka 7.3 Složky gradientního odhadu $\mathbf{p}(t)=\left(p_1(t),...,p_{m-1}(t)\right)$ pro $\lambda\in\mathbb{R}-\{-1,0\}$ získané pomocí Cressie-Readovy kvazinormy jsou podle věty 4.1 partikulárním řešením soustavy obyčejných diferenciálních rovnic

$$\begin{aligned}
p_1'(t)&=\frac{1}{\lambda+1}\left(mp_1(t)\right)^{-\lambda-1}-\frac{1}{\lambda+1}\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\lambda-1},\\
p_2'(t)&=\frac{1}{\lambda+1}\left(mp_2(t)\right)^{-\lambda-1}-\frac{1}{\lambda+1}\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\lambda-1},\\
&\ldots\\
p_{m-1}'(t)&=\frac{1}{\lambda+1}\left(mp_{m-1}(t)\right)^{-\lambda-1}-\frac{1}{\lambda+1}\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-\lambda-1},
\end{aligned}$$

speciálně pro $\lambda=0$

$$\begin{aligned}
p_1'(t)&=\left(mp_1(t)\right)^{-1}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-1},\\
p_2'(t)&=\left(mp_2(t)\right)^{-1}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-1},\\
&\ldots\\
p_{m-1}'(t)&=\left(mp_{m-1}(t)\right)^{-1}-\left[m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right]^{-1}
\end{aligned}$$

a pro $\lambda=-1$

$$\begin{aligned}
p_1'(t)&=\ln\left(m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right)-\ln\left(mp_1(t)\right),\\
p_2'(t)&=\ln\left(m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right)-\ln\left(mp_2(t)\right),\\
&\ldots\\
p_{m-1}'(t)&=\ln\left(m\left(1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t)\right)\right)-\ln\left(mp_{m-1}(t)\right),
\end{aligned}$$

s počátečními podmínkami

$$p_1(0)=\frac{f_1}{n},\ p_2(0)=\frac{f_2}{n},\ldots,\ p_{m-1}(0)=\frac{f_{m-1}}{n}$$

a složkou

$$p_m(t)=1-p_1(t)-p_2(t)-\cdots-p_{m-1}(t).$$

Poznámka 7.4 Cressie-Readova kvazinorma zahrnuje všechny naše popsané kvazinormy. Konkrétně pro $\lambda=1$ dostaneme Pearsonovu kvazinormu, pro $\lambda=-\frac{1}{2}$ Hellingerovu

kvazinormu a pro $\lambda = -2$ kvadratickou kvazinormu. V případě, že $\lambda = -1$ obdržíme Shannonovu kvazinormu a pro $\lambda = 0$ duální kvazinormu k Shannonově kvazinormě.

Věta 7.2 Cressie-Readova kvazinorma je až na násobek totožná s G-kvazinormou.

Důkaz: Vyjdeme z generující funkce G-kvazinormy $f(u) = \frac{2}{k-2}\frac{1}{u^{(k-2)/2}} + u - \frac{k}{k-2}$ pro $k \in \mathbb{R}^+ - \{2\}$. Pak $G(\mathbf{p}, \mathbf{p_0}) = \frac{1}{m}\left(\sum_{j=1}^{m} \frac{2}{k-2}\left(\frac{1}{mp_j}\right)^{\frac{k-2}{2}} + mp_j - \frac{k}{k-2}\right) =$

$$= \frac{1}{m}\sum_{j=1}^{m}\frac{2}{k-2}\left(\frac{1}{mp_j}\right)^{\frac{k-2}{2}} - \frac{2}{k-2} = \frac{2}{k-2}\frac{1}{m}\sum_{j=1}^{m}\left\{\left(\frac{1}{mp_j}\right)^{\frac{k-2}{2}} - 1\right\} \text{ pro } k \in \mathbb{R}^+ - \{2\}.$$

Odtud dostaneme novou generující funkci G-kvazinormy

$$f(u) = \frac{2}{k-2}\left\{\left(\frac{1}{u}\right)^{\frac{k-2}{2}} - 1\right\}, \text{ pro } k \in \mathbb{R}^+ - \{2\}.$$

Zavedeme-li substituci $\lambda = \frac{k-2}{2}$, má generující funkce G-kvazinormy tvar

$$f(u) = \frac{1}{\lambda}\left\{\frac{1}{u^{\lambda}} - 1\right\}, \text{ pro } \lambda \in \mathbb{R}^+,$$

který je až na násobek totožný s generující funkcí Cressie-Readovy kvazinormy

$$f(u) = \frac{1}{\lambda(\lambda+1)}\left(\frac{1}{u^{\lambda}} - 1\right) \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\}. \square$$

Poznámka 7.5 Využijeme-li definici 2.1.1, pak duální kvazinorma ke Cressie-Readově kvazinormě $CR_f(\mathbf{p}, \mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(\frac{1}{mp_j}\right)^{\lambda} - 1\right\}$ je

$$CR_F(\mathbf{p}, \mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(mp_j\right)^{\lambda+1} - 1\right\} \text{ pro } \lambda \in \mathbb{R} - \{0, -1\},$$

speciálně pro $\lambda = 0$

$$F(u) = u\ln(u)$$

a pro $\lambda = -1$

$$F(u) = -\ln(u).$$

Věta 7.3 Cressie-Readova kvazinorma tvoří uzavřenou třídu kvazinorem vzhledem k dualitě.

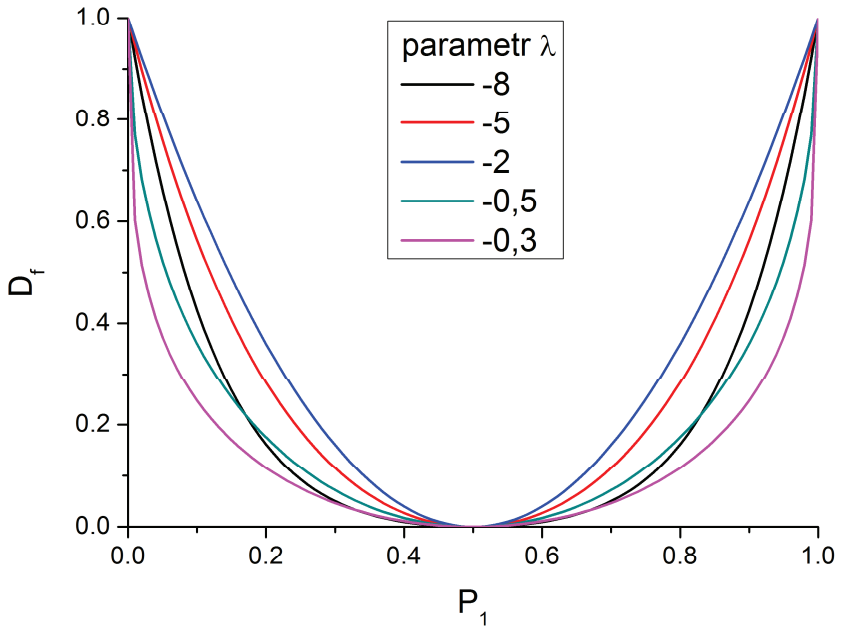
Důkaz: Duální Cressie-Readovu kvazinormu $CR_F(\mathbf{p},\mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(mp_j\right)^{\lambda+1}-1\right\}$ pro $\lambda \in \mathbb{R}-\{0,-1\}$ dostaneme z $CR_f(\mathbf{p},\mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(\frac{1}{mp_j}\right)^{\lambda}-1\right\}$ substitucí $\lambda \to -1-\lambda$.

speciálně pro $\lambda = 0$ je $CR_f(\mathbf{p},\mathbf{p_0}) = -\ln(u) \Rightarrow CR_F(\mathbf{p},\mathbf{p_0}) = u\ln u$, což je Cressie-Readova kvazinorma pro $\lambda = -1$. □

8. Porovnání kvazinorem

Pro praxi je užitečné najít kritéria pro výběr nejvhodnější kvazinormy, respektive generující funkce $f(u)$. Jednou z možností je vybrat kvazinormu, která nejrychleji konverguje ke hledanému rozdělení pravděpodobnosti, tedy tu, která je „nejstrmější“. Musíme však respektovat tvrzení b) z věty 3.2, tj. že generující funkce $f(u)$ a $cf(u)$, kde $c>0$, $c\neq 1$ je konstanta, definují dvě kvazinormy lišící se pouze násobkem c. Aby byly kvazinormy mezi sebou porovnatelné, podělíme vždy kvazinormu D_f jejím maximem $N=\max_{\mathbf{p}\in S} D_f$, pokud její maximum existuje. Maximum N nabývá kvazinorma D_f podle věty 3.2 v bodech $\mathbf{p}_1 = \\ =(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)$. Takto normovaná kvazinorma má maximum rovno jedné. Vzhledem k tomu, že Cressie-Readova kvazinorma $CR_f(\mathbf{p},\mathbf{p}_0)$ zahrnuje všechny dosud uvedené kvazinormy, použijeme ji k jejich porovnání. V úvahu budeme brát pouze $\lambda<0$, protože v opačném případě se jedná o neohraničenou funkci.

Speciálně pro $m=2$ si můžeme vykreslit hodnoty příslušné normované kvazinormy v závislosti na p_1. Jak je vidět z obrázku 8.1, ze zobrazených kvazinorem je nejstrmější kvadratická kvazinorma ($\lambda=-2$) a nejméně strmá kvazinorma pro $\lambda=-0{,}3$. V některých případech si však nemusíme být jisti, například porovnáváme-li $\lambda=-8$ a $\lambda=-0{,}5$. V takovém případě je vhodné spočítat plochu pod kvazinormou. Čím bude plocha větší, tím je kvazinorma strmější. Pro vyšší dimenze je nutné odvodit objem V_D m-rozměrného tělesa ohraničeného kvazinormou D_f a simplexem S.



Obr.8.1 Graf závislosti hodnoty Cressie-Readovy kvazinormy na p_1 pro vybraná λ.

Věta 8.1 Pro Cressie-Readovu kvazinormu, kde $\lambda\in\mathbb{R}^{-}-\{-1\}$, je

$$\mathrm{V_D} = \frac{1}{\lambda(\lambda+1)}\left(m^{-\lambda}\prod_{i=0}^{m-2}\frac{1}{(1-\lambda+i)} - \frac{1}{(m-1)!}\right)$$

a

$$N = \frac{m^{-(\lambda+1)}-1}{\lambda(\lambda+1)}.$$

Důkaz: Uvažujme Cressie-Readovu kvazinormu vedoucí na asymptotický chí-kvadrát test $CR_f(\mathbf{p},\mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(\frac{1}{mp_j}\right)^{\lambda}-1\right\}$ s generující funkcí $f(u) = \frac{1}{\lambda(\lambda+1)}\left(u^{-\lambda}-1\right)$ pro $\lambda \in \mathbb{R}^{-} - \{-1\}$.

Označme

$$\mathrm{V_D} = \frac{1}{m\lambda(\lambda+1)}\int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-2}}\left[\left(mp_1\right)^{-\lambda}+\left(mp_2\right)^{-\lambda}+\cdots+\right.$$
$$\left.+\left(m\left(1-p_1-p_2-\cdots-p_{m-1}\right)\right)^{-\lambda}-m\right]dp_{m-1}\ldots dp_1.$$

Tento integrál vypočítáme pomocí následujících integrálů

$$\mathrm{I_1} = \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-2}} 1\,dp_{m-1}\ldots dp_1 =$$
$$= \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-3}}\left(1-p_1-p_2-\cdots-p_{m-2}\right)dp_{m-2}\ldots dp_1 =$$

$$= \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-4}}(-1)^2\frac{1}{2}\left(1-p_1-p_2-\cdots-p_{m-3}\right)^2 dp_{m-3}\ldots dp_1 =$$
$$= \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-5}}(-1)^3\frac{1}{2.3}\left(1-p_1-p_2-\cdots-p_{m-4}\right)^3 dp_{m-4}\ldots dp_1 = \cdots =$$
$$= \int_0^1(-1)^{m-1}\frac{1}{2.3....(m-2)}\left(1-p_1\right)^{m-2}dp_1 = \frac{1}{(m-1)!}$$

$$\mathrm{I_2} = \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-2}} {p_1}^{-\lambda}dp_{m-1}\ldots dp_1 =$$
$$= \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-3}}\left(1-p_1-p_2-\cdots-p_{m-2}\right){p_1}^{-\lambda}dp_{m-2}\ldots dp_1 =$$
$$= \int_0^1\int_0^{1-p_1}\ldots\int_0^{1-p_1-\cdots-p_{m-4}}-\frac{1}{2}\left(1-p_1-p_2-\cdots-p_{m-3}\right)^2{p_1}^{-\lambda}dp_{m-3}\ldots dp_1 = \cdots =$$

$$= \int_0^1 (-1)^{m-1} \frac{1}{2.3...(m-2)} (1-p_1)^{m-2} p_1^{-\lambda} =$$

$$= (-1)^{m-2} \frac{1}{2.3...(m-2)} \int_0^1 \left[\binom{m-2}{0} p_1^{m-2-\lambda} - \binom{m-2}{1} p_1^{m-3-\lambda} + \cdots \binom{m-2}{m-2} p_1^{-\lambda} \right] dp_1 =$$

$$= (-1)^{m-2} \frac{1}{2.3...(m-2)} \left[\frac{\binom{m-2}{0}}{m-1-\lambda} p_1^{m-1-\lambda} - \frac{\binom{m-2}{1}}{m-2-\lambda} p_1^{m-2-\lambda} + \cdots \frac{\binom{m-2}{m-2}}{1-\lambda} p_1^{1-\lambda} \right]_0^1 =$$

$$= (-1)^{m-2} \frac{1}{2.3...(m-2)} \left(\frac{\binom{m-2}{0}}{m-1-\lambda} - \frac{\binom{m-2}{1}}{m-2-\lambda} + \frac{\binom{m-2}{2}}{m-3-\lambda} - \cdots \frac{\binom{m-2}{m-2}}{1-\lambda} \right) =$$

$$= \frac{(-1)^{m-2}}{2.3...(m-2)} \sum_{i=0}^{m-2} (-1)^i \frac{\binom{m-2}{i}}{m-1-\lambda-i} = \prod_{i=0}^{m-2} \frac{1}{1-\lambda+i},$$

$$\mathrm{I}_3 = \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-2}} (1-p_1-p_2-\cdots p_{m-1})^{-\lambda} \, dp_{m-1} \ldots dp_1 =$$

$$= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-3}} \frac{1}{(1-\lambda)} (1-p_1-p_2-\cdots p_{m-2})^{1-\lambda} \, dp_{m-2} \ldots dp_1 =$$

$$= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-4}} \frac{1}{(1-\lambda)} \frac{1}{(2-\lambda)} (1-p_1-p_2-\cdots p_{m-3})^{2-\lambda} \, dp_{m-3} \ldots dp_1 =$$

$$= \int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-p_1-\cdots-p_{m-5}} \frac{1}{(1-\lambda)} \frac{1}{(2-\lambda)} \frac{1}{(3-\lambda)} (1-p_1-p_2-\cdots p_{m-4})^{3-\lambda} \, dp_{m-4} \ldots dp_1 = \cdots =$$

$$= \int_0^1 \frac{1}{1-\lambda} \frac{1}{2-\lambda} \cdots \frac{1}{m-2-\lambda} (1-p_1)^{m-2-\lambda} \, dp_1 = \frac{1}{1-\lambda} \frac{1}{2-\lambda} \cdots \frac{1}{m-1-\lambda} = \prod_{i=0}^{m-2} \frac{1}{1-\lambda+i}.$$

Pak

$$\mathrm{V_D} = \frac{1}{\lambda(\lambda+1)} \left(m^{-\lambda} \prod_{i=0}^{m-2} \frac{1}{(1-\lambda+i)} - \frac{1}{(m-1)!} \right).$$

$$N = \max_{\mathbf{p} \in S} CR_f(\mathbf{p}, \mathbf{p_0}) = \max_{\mathbf{p} \in S} \frac{1}{\lambda(\lambda+1)m} \sum_{j=1}^{m} \left\{ \left(\frac{1}{mp_j} \right)^{\lambda} - 1 \right\} = \frac{1}{\lambda(\lambda+1)m} \sum_{j=1}^{m} \left\{ \left(\frac{1}{mp_j} \right)^{\lambda} - 1 \right\} =$$

$$= \frac{m^{-(\lambda+1)}-1}{\lambda(\lambda+1)}. \square$$

Poznámka 8.1 Cressie-Readova kvazinorma je pro $\lambda \in \mathbb{R}^- - \{-1\}$ tím strmější, čím větší je velikost podílu $\frac{V_D}{N}$. Pro $\lambda > 0$ je Cressie-Readova kvazinorma neohraničená funkce a porovnání přes objemy není možné.

Definice 8.1 ***Mírou strmosti kvazinormy*** D_f rozumíme číslo $\frac{V_D}{N}$, kde V_D je objem *m*-rozměrného tělesa ohraničeného kvazinormou D_f a simplexem

$$S = \left\{ \mathbf{p} \in \mathbb{R}^m : \forall p_j \geq 0, \sum_{j=1}^{m} p_j = 1 \right\} \text{ a } N = \max_{\mathbf{p} \in S} D_f .$$

Poznámka 8.2 Konkrétně pro Cressie-Readovu kvazinormu, kde $\lambda = -\frac{1}{2}$ (Hellingerova kvazinorma), je

$$\mathrm{V_D} = \frac{4}{(m-1)!} - \frac{4}{\sqrt{m}} \left(\frac{2^{m-1}}{(2m-1)!!} \right)$$

a

$$N = 4\left(1 - m^{-1/2}\right).$$

Pro $\lambda = -2$ (kvadratická kvazinorma) je

$$\mathrm{V_D} = \frac{m^2}{(m+1)!} - \frac{1}{2(m-1)!}$$

a

$$N = \frac{1}{2}(m-1).$$

V tabulce 8.1 je vyčíslený podíl $\frac{\mathrm{V_D}}{N}$ pro vybraná λ Cressie-Readovy kvazinormy a dimenzi *m*. Z tabulky je zřejmé, že pro $m < 6$ je vhodné volit $\lambda = -2$, tedy kvadratickou kvazinormu a pro $m \geq 6$ je již vhodnější volit $\lambda = -0,5$, tedy Hellingerovu kvazinormu. Největší hodnoty podílu jsou pro jednotlivá *m* v tabulce zvýrazněny modrou barvou, naopak nejmenší hodnoty barvou červenou.

Tab 8.1 Hodnoty podílu $\frac{V_D}{N}$ pro vybraná λ Cressie-Readovy kvazinormy a dimenzi *m*.

	m						
lambda	2	3	4	5	6	9	12
-0,2	0,10044	0,04811	0,01563	3,841E-03	7,585E-04	2,204E-06	2,196E-09
-0,5	0,19526	0,09019	0,02857	6,887E-03	1,340E-03	3,772E-06	3,680E-09
-2	0,33333	0,12500	0,03333	6,944E-03	1,190E-03	2,480E-06	1,927E-09
-5	0,28889	0,06607	0,01130	1,589E-03	1,921E-04	1,697E-07	6,762E-11
-10	0,18022	0,02270	0,00233	2,081E-04	1,665E-05	5,101E-09	8,523E-13

9. Bootstrapové intervalové odhady

Pokud neznáme rozdělení pravděpodobnosti náhodné veličiny *X*, není k dispozici intervalový odhad jejího parametru θ, nechceme nebo nemůžeme aplikovat asymptotické odhady, například pro malý rozsah výběru, pak je možné nahradit statistický soubor pozorovaných hodnot novým bootstrapovým souborem. Ten získáme ze souboru původního náhodným výběrem s opakováním (s vracením). Prezentované pojmy a postupy jsou převzaty z [16], [17].

Definice 9.1 Ze statistického souboru pozorovaných hodnot $(x_1,\ldots,x_n)$ náhodného výběru $(X_1,\ldots,X_n)$ vytvoříme nový statistický soubor $(x_1^*,\ldots,x_n^*)$ náhodným výběrem hodnot x_i s opakováním (s vracením). Takto získaný náhodný výběr se nazývá ***bootstrapový výběr***, resp. ***bootstrapový soubor.***

Poznámka 9.1 Při odhadu střední kvadratické chyby, rozptylu, směrodatné odchylky a vychýlení odhadu $\hat{\theta}$ parametru θ rozdělení pravděpodobnosti náhodné veličiny *X* postupujeme následovně [17].

1. Z pozorovaných hodnot $(x_1,\ldots,x_n)$ náhodného výběru $(X_1,\ldots,X_n)$ vypočítáme odhad $\hat{\theta}$ parametru θ.
2. Realizujeme *B* náhodných bootstrapových výběrů o rozsahu *n* z pozorovaných hodnot $(x_1,\ldots,x_n)$.
3. Pro každý bootstrapový výběr vypočítáme odhad parametru θ a označíme jej $\hat{\theta}_{b,i}$, kde $i=1,\ldots,B$.

Poznámka 9.2 Počet všech různých bootstrapových výběrů je $\binom{n+B-1}{B}$.

Poznámka 9.3 Bootstrapovým odhadem střední kvadratické chyby MSE odhadu $\hat{\theta}$ je

$$\widehat{\mathrm{MSE}}_b = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_{b,i}-\hat{\theta}\right)^2,$$

bootstrapovým odhadem rozptylu $D\left(\hat{\theta}\right)$ je

$$\hat{D}\left(\hat{\theta}\right)_b = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_{b,i}-\frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_{b,j}\right)^2,$$

bootstrapovým odhadem směrodatné odchylky $\sigma\left(\hat{\theta}\right)$ je

$$\hat{\sigma}\left(\hat{\theta}\right)_b = \sqrt{\hat{D}\left(\hat{\theta}\right)_b},$$

bootstrapovým odhadem vychýlení $\mathrm{Bias}\left(\hat{\theta}\right)$ odhadu $\hat{\theta}$ je

$$\hat{B}\left(\hat{\theta}\right)_b = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i}-\hat{\theta}.$$

Poznámka 9.4 Odhad $\hat{\theta}$ získaný z původního statistického souboru $(x_1,\ldots,x_n)$ je bodovým

odhadem parametru θ, ale můžeme jej dle potřeby také nahradit aritmetickým průměrem

$$\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i}\,.$$

Pomocí bootstrapových výběrů získáme bootstrapové intervalové odhady se spolehlivostí $1-\alpha$ střední hodnoty, rozptylu a směrodatné odchylky náhodné veličiny *X*.

Poznámka 9.5 Nechť $\overline{x}$ je aritmetický průměr a s^2 je rozptyl původního statistického souboru $(x_1,\ldots x_n)$, $\overline{x}_{b,i}$ je aritmetický průměr a $s_{b,i}^2$ je rozptyl statistického souboru z *i*-tého bootstrapového výběru, $i=1,\ldots,B$, pak

1. Bootstrapový intervalový odhad střední hodnoty $E(X)$ se spolehlivostí $1-\alpha$ je

$$\left[\overline{x}-t_{b,1-\alpha/2}\frac{s}{\sqrt{n-1}},\overline{x}-t_{b,\alpha/2}\frac{s}{\sqrt{n-1}}\right],$$

kde $t_{b,P}$ je *P*-kvantil statistického souboru $(t_{b,1},...,t_{b,B})$ a $t_{b,i}=\dfrac{\overline{x}_{b,i}-\overline{x}}{s_{b,i}}\sqrt{n-1}$, $i=1,\ldots,B$.

2. Bootstrapový intervalový odhad rozptylu $D(X)$ se spolehlivostí $1-\alpha$ je

$$\left[\frac{ns^2}{\chi^2_{b,1-\alpha/2}},\frac{ns^2}{\chi^2_{b,\alpha/2}}\right],$$

kde $\chi^2_{b,P}$ je *P*-kvantil statistického souboru $(\chi^2_{b,1},...,\chi^2_{b,B})$ a $\chi^2_{b,i}=\dfrac{ns^2_{b,i}}{s^2}$, $i=1,\ldots,B$.

3. Bootstrapový intervalový odhad směrodatné odchylky $\sigma(X)$ se spolehlivostí $1-\alpha$ obdržíme z odhadu $D(X)$ pomocí odmocniny.

Existuje řada dalších způsobů stanovení intervalových odhadů parametrů založených na bootstrapových výběrech [16], [17]. Mimo uvedených oboustranných intervalových odhadů se používají dle potřeby také jednostranné bootstrapové intervalové odhady.

Předchozí kapitoly byly věnovány speciálním typům bodových odhadů diskrétního rozdělení pravděpodobnosti (gradientní a přímkový odhad). Jednou z možností, jak získat odhady intervalové, je použití metody bootstrap. V programovacím jazyce C++ byl vytvořen program Shine bootstrap [20], který umožňuje vygenerovat bootstrapové výběry a následně spočítat odhad rozdělení pravděpodobnosti z těchto bootstrapových souborů pomocí kvadratické kvazinormy. Nyní uvedeme dva příklady kategoriálních veličin, u nichž předvedeme získání intervalových odhadů jejich rozdělení pravděpodobnosti s využitím programu Shine bootstrap [20]. Bootstrapové odhady rozdělení pravděpodobnosti byly částečně publikovány v [18].

Příklad 9.1 Kovový upínací přípravek byl lisován ve formě, potažen lepidlem a umístěn do lisu, kde byla vstříknuta pryž k vytvoření konečného výrobku. Při sledování neshodnosti dávky výrobků byl zjištěn celkový počet 314 vad v této dávce. Odhadnuté počty vad pro jednotlivé kvazinormy i výsledky Paretovy analýzy jsou již uvedeny v příkladu 4.2.2. Z důvodu nedostatečného počtu pozorování bylo pomocí programu Shine Bootstrap [20] vygenerováno tisíc bootstrapových výběrů a z každého výběru spočítán odhad pomocí kvadratické kvazinormy (tisíc odhadů pro každý typ vady). Každý bootstrapový výběr byl stejného rozsahu, jako původní soubor, tedy $n=314$. V tabulce 9.1 jsou spočítány charakteristiky odhadnutých parametrů pro

jednotlivé typy vad, které byly vypočítány z bootstrapových souborů. V tabulce 9.2 je zapsán bootstrapový odhad střední hodnoty a rozptylu pravděpodobnosti $p_j \quad j = 1,...,m$ dle vztahů uvedených v poznámce 9.2. Spolehlivost byla zvolena $0,95$. Jako průměrná hodnota $\bar{x}$ byl volen odhad pravděpodobností p_j kvadratickou kvazinormou skutečného (původního) statistického souboru. Podobně za s^2 byl dosazen rozptyl tohoto odhadnutého souboru.

V tabulce 9.3 jsou zapsány skutečné četnosti jednotlivých vad, průměrný odhad četností kvadratickou kvazinormou a bootstrapový intervalový odhad četností.

Tab. 9.1 Charakteristiky odhadnutých parametrů kvadratickou kvazinormou, které byly vypočítány z tisíce bootstrapových výběrů.

Název vady	Průměr	Medián	Minimum	Maximum	Směrodatná odchylka	Rozptyl
Špatná přilnavost	0,371504	0,372358	0,372358	0,291666	0,000761	0,027584
Špatný brus	0,274932	0,274172	0,274172	0,199755	0,000559	0,023639
Dutiny	0,109539	0,109735	0,109735	0,059762	0,000269	0,016407
Škrábance	0,072207	0,071751	0,071751	0,023652	0,000167	0,012909
Nečistoty	0,048943	0,048402	0,048402	0,018613	0,000129	0,011343
Trhliny	0,082311	0,040082	0,082949	0,044394	0,000155	0,012431
Jiné	0,040566	0,082949	0,040082	0,013143	0,000092	0,009577

Tab. 9.2 Bootstrapový odhad střední hodnoty a rozptylu.

	Odhad kvadratickou kvazinormou	Bootstrapový odhad střední hodnoty		Rozptyl	Bootstrapový odhad rozptylu	
Špatná přilnavost	0,37100	0,319937	0,427853	0,233359	0,223085	0,251829
Špatný brus	0,27520	0,227991	0,327762	0,199465	0,190684	0,215252
Dutiny	0,10970	0,076666	0,146480	0,097666	0,093366	0,105396
Škrábance	0,07201	0,044685	0,102434	0,066825	0,063883	0,072114
Nečistoty	0,04878	0,02601	0,074131	0,046401	0,044358	0,050073
Trhliny	0,08313	0,053947	0,115622	0,076219	0,072864	0,082252
Jiné	0,04008	0,019346	0,063165	0,038474	0,036780	0,041519

Tab. 9.3 Četnosti skutečných naměřených vad, průměrný odhad kvadratickou kvazinormou a bootstrapový intervalový odhad četností.

Název vady	Skutečné četnosti naměřených vad	Průměrný odhad četností kvadratickou kvazinormou.	Bootstrapový odhad četností	
Špatná přilnavost	128	117	100	134
Špatný brus	91	86	72	103
Dutiny	36	34	24	46
Škrábance	23	23	14	32
Nečistoty	15	15	8	23
Trhliny	9	13	17	36
Jiné	12	26	6	20

Příklad 9.2 Ve dnech 13.1. 2013 až 23.1. 2013 byl v obci Brno-Jehnice proveden průzkum týkající se druhého kola přímé volby prezidenta. Tázáno bylo celkem 50 občanů. Všichni z dotázaných byli starší osmnácti let a uvedli trvalé bydliště v Jehnicích. Celkem 32 dotázaných uvedlo, že budou volit Karla Schwarzenberga a 18 občanů uvedlo, že budou volit Miloše Zemana. Vzhledem k malému počtu dotázaných bylo pomocí programu Shine bootstrap [20] vygenerováno sto bootstrapových náhodných výběrů stejného rozsahu, jako byl původní statistický soubor, tedy $n = 50$. U všech bootstrapových výběrů byl spočítán v programu Shine bootstrap odhad pravděpodobnostní funkce pomocí kvadratické kvazinormy a následně v programu Excel spočítán bootstrapový intervalový odhad střední hodnoty p_1 a p_2.Výsledky jsou uvedeny v tabulce 9.4.

Tab. 9.4 Tabulka pravděpodobností, bootstrapový intervalový odhad střední hodnoty a rozptylu, $n = 50$.

	p_1	p_2
Původní pravděpodobnostní funkce	0,64	0,36
Odhad kvadratickou kvazinormou	0,5014	0,4986
Průměr ze sta odhadů kvadratickou kvazinormou	0,5577	0,4423
Bootstrapový intervalový odhad $E(X)$ se spolehlivostí 0,95.	$[0{,}3793;0{,}5228]$	$[0{,}4772;0{,}6207]$
Bootstrapový intervalový odhad $D(X)$ se spolehlivostí 0,95.	$[0{,}2500;0{,}2653]$	

Jelikož je bootstrapový intervalový odhad parametru p_1 resp. p_2 poměrně široký, zvětšíme rozsah statistického souboru. Vezmeme v úvahu původní pravděpodobnostní funkci $\mathbf{p} = (0{,}64;0{,}36)$ a rozsah souboru zvětšíme na $n = 600$. Rozsah souboru byl zvolen s ohledem na to, že v prvním kole se zúčastnilo voleb v Jehnicích 611 občanů a předpokládáme, že ve druhém kole bude účast podobná. V programu Shine bootstrap bylo vygenerováno sto bootstrapových souborů, přičemž každý měl rozsah $n = 600$. Z těchto souborů byl vypočítán odhad pravděpodobnostní funkce pomocí kvadratické kvazinormy a následně v programu Excel spočítán bootstrapový intervalový odhad střední hodnoty p_1 a p_2.Výsledky jsou uvedeny v tabulce 9.5.

Tab. 9.5 Tabulka pravděpodobností, bootstrapový intervalový odhad střední hodnoty a rozptylu, $n = 600$.

	p_1	p_2
Původní pravděpodobnostní funkce	0,64	0,36
Odhad kvadratickou kvazinormou	0,6008	0,3992
Průměr ze sta odhadů kvadratickou kvazinormou	0,5989	0,4011
Bootstrapový intervalový odhad $E(X)$ se spolehlivostí 0,95.	$[0{,}5609;0{,}63254]$	$[0{,}36746;0{,}4391]$
Bootstrapový intervalový odhad $D(X)$ se spolehlivostí 0,95.	$[0,2345;0,2496]$	

Druhého kola voleb se zúčastnilo 595 voličů, z toho platných hlasů bylo 593. Skutečné výsledky voleb jsou uvedeny v tabulce 9.6.

Tab. 9.6 Skutečné výsledky voleb obce Brno-Jehnice.

	p_1	p_2
Skutečné výsledky voleb v Jehnicích	0,6087	0,3912

Bootstrapový intervalový odhad parametru p_1 resp. parametru p_2 pro $n = 50$ nezahrne skutečný výsledek voleb na rozdíl od výsledků získaných pro rozsah souboru $n = 600$.

10. Závěr

Cíle dizertační práce byly zaměřeny na vytvoření a popis vlastností nových metod odhadů rozdělení pravděpodobnosti kategoriální náhodné veličiny z pozorovaných četností. Na rozdíl od klasických bodových odhadů jednotlivých pravděpodobností daného rozdělení umožňují v práci popsané metody respektovat případné další podmínky kladené na rozdělení a to ve smyslu maximální neurčitosti tohoto rozdělení. Neurčitost je vyjádřena tzv. kvazinormou, která vychází z *f*-divergence (tj. vzdálenosti) hledaného rozdělení od maximálně neurčitého rozdělení.

V práci jsou předloženy a studovány nové typy odhadů rozdělení pravděpodobnosti kategoriální náhodné veličiny, konkrétně gradientní a přímkové odhady. Gradientní odhady $\mathbf{p}(t)$ jsou vzhledem k variabilitě volby parametru t do jisté míry blízké mocninným diskrétním jádrovým odhadům [12]. Umožňují libovolnou volbu parametru t z intervalu $[0;t_0]$ a tím odhad přiblížit k empirickému rozdělení. Snaha přiblížit odhad co nejvíce k rovnoměrnému rozdělení, tzv. pesimistický odhad, však může vést k navýšení podílu složek s malou četností na úkor těch vyšších. Z tohoto důvodu je vhodné u gradientních odhadů zkoumat i modifikovanou úlohu, kdy hledáme odhad nejvzdálenější od rovnoměrného rozdělení ležící na hranici zamítnutí testu dobré shody. Tento typ odhadu je v práci uveden pro přímkový odhad.

Gradientní odhad je závislý nejen na zvoleném parametru t, ale také na generující funkci f, respektive kvazinormě. Odvozena byla nová třída kvazinorem, tzv. Cressie-Readova kvazinorma $CR_f(\mathbf{p},\mathbf{p_0}) = \frac{1}{\lambda(\lambda+1)m}\sum_{j=1}^{m}\left\{\left(\frac{1}{mp_j}\right)^{\lambda}-1\right\}$ zahrnující mimo jiné všechny kvazinormy uvedené v dizertační práci. Vzhledem k tomu, že každá konkrétní Cressie-Readova kvazinorma vede na asymptotický chí-kvadrát test, je možné hledat vhodnou hodnotu parametru $t_0 \in [0;\infty)$ pro gradientní odhad pomocí testu dobré shody, kde bude parametr λ určen tak, aby odpovídal příslušné zvolené kvazinormě. V budoucnu bychom chtěli využít i jiné míry vzdálenosti než *f*-divergence, například míry vzdálenosti vycházející z transformace stabilizující rozptyl.

Gradientní odhady vyžadují numerické řešení získaných nelineárních soustav diferenciálních rovnic. Za tímto účelem byl pomocí softwaru MATLAB vytvořen program Pesfit. Jediná kvazinorma, která vede na lineární soustavu diferenciálních rovnic s konstantními koeficienty, je kvadratická kvazinorma $D_f(\mathbf{p},\mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m}(mp_j-1)^2$. Při řešení nelineárních soustav diferenciálních rovnic se ukázalo, že výpočet hodnoty t_0 je citlivý na „strmost“ zvolené kvazinormy $D_f(\mathbf{p},\mathbf{p}_0)$, respektive na volbu generující funkce. V dizertační práci jsou proto kvazinormy mezi sebou porovnány právě s ohledem na jejich „strmost“. Čím je kvazinorma „strmější“, tím rychleji dospějeme ke hledanému rozdělení pravděpodobnosti. Ukázalo se, že pro $m<6$ je nejvhodnější volit kvadratickou kvazinormu, pro $m \geq 6$ je již vhodnější volit Hellingerovu kvazinormu. Předložené odhady založené na kvazinormách mají sice poněkud samoúčelný charakter, ale na řadě úloh se ukazuje jejich praktická použitelnost. Gradientní odhady byly například aplikovány na kontingenční tabulky. Přidány byly vstupní podmínky kladené na marginální četnosti hledaného rozdělení pravděpodobnosti, nebo požadavek nezávislosti hledaného rozdělení.

K zpřesnění gradientních odhadů byla využita metoda bootstrapových odhadů, kterou je vhodné požít zejména, pokud není rozsah souboru dostatečně velký. S využitím metody bootstrap byly bodové gradientní odhady rozšířeny na intervalové odhady. Vytvořen byl

program Shine Bootstrap, pomocí kterého můžeme vygenerovat potřebný počet bootstrapových výběrů a navíc z každého výběru spočítat gradientní odhad kvadratickou kvazinormou. V současné době se vytváří nový software, který by měl být jednodušší pro uživatele. Software Shine Bootstrap je programován v jazyce Python [21] za použití knihovny NumPy [22]. Grafické rozhraní je vytvořeno pomoci knihovny wxWidgets [23].

Ukázalo se, že přímkové odhady založené původně na gradientním přístupu jsou ekvivalentní s mocninnými diskrétními jádrovými odhady [12] a můžeme zajistit jejich asymptotickou nestrannost. V našem případě se však opět nabízí možnost testovat jejich vhodnost pomocí Cressie-Readovy statistiky.

Mimo výše uvedených dalších možných směrů bádání se nabízí také zabývat se gradientními a přímkovými odhady vedenými směrem na množinu všech rozdělení simultánních pravděpodobností ve čtyřpolní kontingenční tabulce, které respektují nezávislost obou dichotomických kategorií.

Vlastní dosažené výsledky v rámci dizertace byly povětšinou publikovány. Mimo osmnáct číslovaných poznámek, grafické prezentace, příklady a textové komentáře jde zejména o:

- Definice 3.2, 4.1, 5.1, 6.2, 7.1 a 8.1.
- Věta 3.3, 3.4, 3.5, 4.1, 4.1.1, 5.2, 6.1, 7.1 (pouze důkaz), 7.2, 7.3 a 8.1.
- Důsledek 4.1 a 4.1.1.
- Spoluautorství autorizovaného softwaru Pesfit a Shine Bootstrap.

11. Dodatky

11.1 Dodatek A

Pesfit 1.0
Program Pesfit 1.0 slouží k nalezení gradientního odhadu rozdělení prověpodobnosti ***p*** z pozorovaných četností *f* pomocí zvolené kvazinormy. Složky gradientního odhadu jsou partikulárním řešením soustavy obyčejných diferenciálních rovnic prvního řádu (SODR1). Soustava diferenciálních rovnic se v programu řeší pomocí Rungeho-Kuttovy metody a k nalezení hodnoty *t*, kdy ještě nezamítáme hypotézu o vhodnosti rozdělení $\mathbf{p}(t)$, byla užita metoda bisekce. Program Pesfit 1.0. byl napsán v prostředí MATLAB 7.1.0.246 (R14).

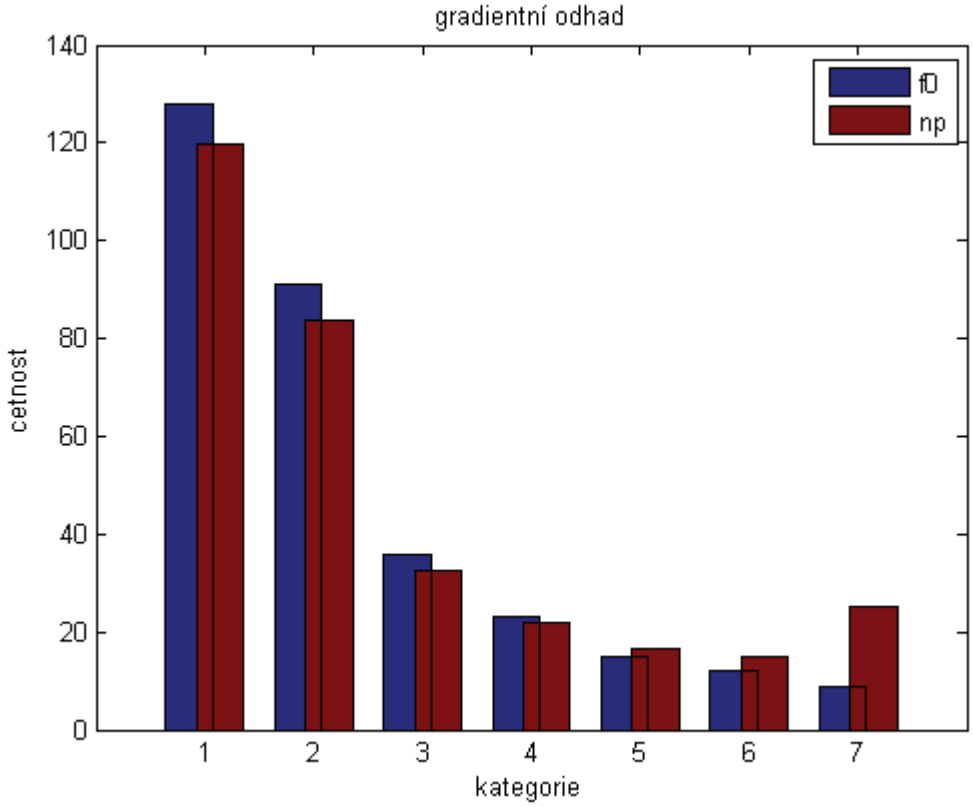
Ukázka zdrojového textu pro Hellingerovu kvazinormu

```
% Nalezení gradientního odhadu pomocí Hellingerovy kvazinormy,
% původní četnosti {128,91,36,23,15,12,9}.

function kvazi3(m,n,v)
%m - pocet kategorii
%n - rozsah souboru
m=7
n=314
format short;
tol=0.001; %tolerance na reseni nelinearni rce
t_end=0.5; %prvni odhad t0
tspan=[0 t_end];
f0=[128 91 36 23 15 12]; %cetnosti

p0=f0/n; %pocatecni podminka

[t,p]=ode45(@f,tspan,p0); %reseni pomoci Runge-Kutta
pm=1-(p(:,1)+p(:,2)+p(:,3)+p(:,4)+p(:,5)+p(:,6));
p=[p pm];

Tl=length(p); %pocet kroku
tau=t_end/Tl; %delka kroku
f0(m)=n-sum(f0);

chi=12.59157742; %kvantil chi-kvadrat rozdeleni
krit=zeros(Tl,1);
tfin=0;
for i=1:Tl
  for j=1:m
     krit(i)=krit(i)+f0(j)*f0(j)/p(i,j);
  end
  krit(i)=(krit(i)/n)-n-chi;
  if krit(i)<=0 %urceni posledniho t, kdy nezamitame hypotezu
      tfin=i;
  end
end
krit;
a=tfin;
b=tfin+1;
p=p(tfin,:); %pocatecni podminka pro 1.bisekci
l=1;
krit=krit(tfin);
```

```
if -krit<tol
    t0=tfin*tau
    krit
    p
else %nulovy bod funkce krit hledame v intervalu (a,b) bisekci
    while (-krit>tol)|(krit>0)
        c=(a+b)/2;
        t0=c*tau;
        [t,p]=ode45(@f,[a*tau t0 b*tau],p(1,1:m-1)); %reseni dif.soustavy pro
t0
        pm=1-(p(:,1)+p(:,2)+p(:,3)+p(:,4)+p(:,5)+p(:,6));
        p=[p pm];
        krit=zeros(3,1);
        for i=1:3
            for j=1:m
                krit(i)=krit(i)+f0(j)*f0(j)/p(i,j);
            end
            krit(i)=(krit(i)/n)-n-chi;
        end
        if krit(2)==0
            a=c;
            b=c;
        elseif krit(2)*krit(3)>0
            b=c;
            l=1;
            else
            a=c;
            l=2;
        end
        krit=krit(2);
    end
    t0  %nalezeny nulovy bod
    krit %hodnota kriteria v t0
    p=p(2,:) %vysledne pravdepodobnosti
end

    D=0;
    for j=1:m
    D=D+sqrt(p(j));
    end
    D=2-2/sqrt(m)*D %hodnota Hellingerovy kvazinormy

np=n*p
cetnost=[f0;np]';  %vykresli sloupcovy graf
bar(cetnost,1.5);
title('gradientní odhad');
xlabel('kategorie');
ylabel('cetnost');
legend('f0','np');

function dpdt = f(t,p) %Hellingerova kvazinorma
dpdt=[(m*p(1))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)
      (m*p(2))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)
      (m*p(3))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)
      (m*p(4))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)
      (m*p(5))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)
      (m*p(6))^(-1/2)-(m*(1-p(1)-p(2)-p(3)-p(4)-p(5)-p(6)))^(-1/2)];
end

end
```

Výstup

```
m = 7

n = 314

t0 = 0.1250

krit = -7.5291e-004

p = 0.3808   0.2663   0.1035   0.0700   0.0526   0.0472   0.0797

D = 0.1494

np = 119.5699  83.6097  32.5116  21.9790  16.5102  14.8063  25.0132
```



11.2 Dodatek B

Shine Bootstrap

Program Shine Bootstrap umožňuje generovat potřebný počet bootstrapových výběrů a navíc z každého výběru spočítat gradientní odhad kvadratickou kvazinormou. Další možnosti programu jsou uvedeny v manuálu společně se způsobem zadávání příkazů. Programovacím jazykem bylo C++. Příkazová řádka byla vytvořena s pomocí knihovny wxWidgets [23]. Využito bylo také řešiče NOMAD [24].

Shine Bootstrap - Manual

1 Literals

Identifiers: Begin with a letter, can contain alphanumeric characters and an underscore `_`.

```
x, A, x1, my_variable, c_142
```

Numbers: In decimal notation, with dot `.` as separator.

```
42, 12.34
```

Strings: Sequence of arbitrary characters enclosed in double quotes `"`.

```
"Whatever."
```

2 Datatypes

`v`	vector
`A`	matrix
`i`	integer
`d`	double
`.`	arbitrary of the above
`s`	string

3 Execute, Import, Save

`import(s)` (File to be opened and scanned for data)
Returns vector or matrix of the numeric data contained in the file.

```
> x = import("Data.txt")
```

`save(s, .)` (File to save to, Data to save)
`save(s, ., s)` (File to save to, Data to save, Description)

Saves the data, overwrites the file.
Saves the data with the description, overwrites the file.

```
> save("Data.txt", x)
> save("Data.txt", x, "The big sample.")
```

`execute(s)` (File name)
Opens the given script file and executes the commands.

```
> execute("Script.txt")
```

4 Indices

Starting from 0.

```
> v[0] = 5
> A[0,1] = 42
```

`> A[:,0]` First column of matrix A
`> A[0,:]` First row of matrix A

5 Info commands

`what` Prints all defined variables.

`size(.)` Returns (number of rows) x (number of columns).
`rows(.)` Returns number of rows.
`cols(.)` Returns number of columns.

6 Plotting commands

`plot(v, v)` Plots x coordinates vs. y coordinates as a curve.
`hist(v)` Displays histogram of v.
`hist(v, i)` Displays histogram of v with given number of clases.

7 Basic operations on vectors and matrices

`.'` Transpose.

```
> y = x'
```

`sum(.)` Sum of all entries.
`rsum(.)` Column vector of row sums.
`csum(.)` Row vector of column sums.

`abs(.)` Absolute value.

`max(.)` Maximum of all entries.
`rmax(.)` Column vector of row maxima.

`cmax(.)` Row vector of column maxima.

`min(.)` Minimum of all entries.
`rmin(.)` Column vector of row minima.
`cmin(.)` Row vector of column minima.

8 Useful matrices and vectors

`ones(i)` Row vector of 1's of given length.
`ones(i, i)` Matrix of 1's of given dimensions.

`zeros(i)` Row vector of 0's of given length.
`zeros(i, i)` Matrix of 0's of given dimensions.

`eye(i)` Identity matrix (square) of given size.

9 Functions

`cos(.)` Cosine
`exp(.)` Exp
`log(.)` Logarithm
`sin(.)` Sine
`tg(.)` Tangens

10 Bootstrap

`bootstrap(v, i, i)` (Sample to select from, Number of bootstraps, Size of one bootstrap)

Matrix (Number of bootstraps) x (Size of one bootstrap), each row is one bootstrap sample.

```
> y = bootstrap(x, 100, 10)
```

11 Generators of distributions

`binomial(i, d, i)` (Binomial N, Probability, Count)
`binomial(i, d, i, i)` (Binomial N, Probability, Rows, Columns)

Returns row vector of length (Count)
Returns matrix (Rows x Columns)

```
> bi = binomial(10, 1/2, 100)
> bi = binomial(10, 1/2, 10, 100)
```

`weibull(d, d, d, i)` (b (shape), delta (scale), c (shift), Count)
`weibull(d, d, d, i, i)` (b (shape), delta (scale), c (shift), Rows, Columns)

Returns row vector of length (Count)
Returns matrix (Rows x Columns)

```
> bi = weibull(1, 1, 0, 100)
> bi = weibull(1, 1, 0, 10, 100)
```

`uniform(d, d, i)` (left, right, count)
`uniform(d, d, i, i)` (left, right, rows, columns)

Uniform distribution between 'left' and 'right'. Returns row vector (length Count).
Uniform distribution between 'left and 'right'. Returns matrix (Rows x Columns).

```
> uni = uniform(0, 1, 10)
> uni = uniform(-1, 1, 5, 5)
```

`discrete(v, v, i)` (Probabilities, Values, Count)
`discrete(v, v, i, i)` (Probabilities, Values, Rows, Columns)

Probabilities and Values must have the same dimension.
Performs selection with replacement from given values with corresponding pro-babilities, returns row vector of (Count) values.

Returns matrix (Rows x Columns).

```
> p = ones(6)
> x = 1:1:6
> dis = discrete(p, x, 100)
> dis = discrete(p, x, 10, 100)
```

`frechet(d, d, d, i)` (Gamma, Sigma, Mu, Count)
`frechet(d, d, d, i, i)` (Gamma, Sigma, Mu, Rows, Cols)

12 Fitting

`fitweibull(v)` (Weibull sample to fit)
`fitweibull(A)` (Matrix with rows as Weibull samples to fit)

Returns b (shape), delta (scale), c (shift) as a row vector.
Returns matrix with rows as the fitted parameters: b (shape), delta (scale), c (shift).

`fitdiscrete(v, v)` (Sample to fit, Underlying values)
`fitdiscrete(A, v)` (Matrix with rows as samples to fit, Underlying values)

Returns a row vector: frequencies of each value from (Underlying values) in (Sample to fit)

Returns a matrix with rows: frequencies of each value from (Underlying values) in (Sample to fit)

```
> fit = fitdiscrete(dis, x)
```

`fitqnorm(v, v)` Analogous to fitdiscrete, quasinorm used for fitting.
`fitqnorm(A, v)` Analogous to fitdiscrete, quasinorm used for fitting.

`fitfrechet(v)` Gamma, Sigma, Mu

13 Command line

`clc` Clears the editor window.
`clear` Clears the editor window.

Command history on arrows (up, down).

12 Použitá literatura

[1] VAJDA, I. Teória informácie a štatistického rozhodovania. Bratislava: Alfa, 1982.

[2] ŠÁCHA, J. Kvazinormy diskrétních rozdělení pravděpodobnosti a jejich aplikace. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2012. 12-13s. Vedoucí dizertační práce doc. RNDr. Zdeněk Karpíšek, CSc..

[3] KARPÍŠEK, Z., JURÁK, P., ŠÁCHA, J. Divergences for Discrete Probability Distribution Estimations. In Summer School DATASTAT ´06, Proceedings, Masaryk Univeristy, Brno, 2007, pp. 109-120, ISBN 978-80-210-4493-7.

[4] NERADOVÁ, V. Progresivní metody odhadu rozdělení pravděpodobnosti. Diplomová práce (vedoucí Karpíšek, Z.), FSI VUT v Brně, Brno 2007.

[5] HALANAY, A. Diferential Equations, stability, oscilations, time lags. New York, Academic Press 1966.

[6] KARPÍŠEK, Z., NERADOVÁ, V. Estimation of Categorical Variable Probability Distribution (Odhad rozdělení pravděpodobnosti kategoriální veličiny). In 7th International Conference APLIMAT 2008. Bratislava, 5. – 8. 2. 2008, Book of abstracts p. 101, ISBN 978-80-89313-02-0, Proceedings pp. 1145-1154, ISBN 978-80-89313-03-7.

[7] KARPÍŠEK, Z., SADOVSKÝ, Z., ŠÁCHA, J. Pitman – Hellinger Test of Fit. In 4th International Conference APLIMAT 2005 (part II). Bratislava, 2005, pp. 471-478, ISBN 80-969264-2-X.

[8] ŽAMPACHOVÁ, E., LACINOVÁ, V., KARPÍŠEK, Z. Pesfit 1.0. Autorizovaný software, Brno, CQR VUT v Brně, 2008.

[9] KARPÍŠEK, Z., NERADOVÁ, V., ŽAMPACHOVÁ, E. A Contribution to the Estimation of Discrete Probability Distribution. In MENDEL 2008. 14th International Conference on Soft Computing. Brno, 18. – 20. 6. 2008, pp. 287-292, ISBN 978-80-214-3675-6.

[10] KARPÍŠEK, Z., LACINOVÁ, V. Aplikace odhadů rozdělení pravděpodobnosti kategoriální veličiny v průzkumových šetřeních. In Sborník konference Daně – teorie a praxe 2010. Brno: Akademie Sting, 2010, pp. 41-47. ISBN 978-80-86342-91-7.

[11] KARPÍŠEK, Z., LACINOVÁ, V. Pesimistické odhady rozdělení pravděpodobnosti kategoriální veličiny. In Informační bulletin České statistické společnosti. Praha, 2011, 10 stran. ISSN 1210-8022.

[12] VÁVRA, F. et al. Discrete Kernels. Austrian Journal of Statistics, Volume 35, 2006, Number 2, 3, pp. 365-370.

[13] ANDĚL, J. Základy matematické statistiky. Praha: MATFYZPRESS, 2005. ISBN 80-86732-40-1.

[14] ANDĚL, J. Statistické metody, Praha: MATFYZPRESS, 2007, ISBN 978-80-7378-003-6.

[15] CRESSIE, N., READ, T. R. C. Multinomial Goodness of Fit Test. J. R. Statist. Soc. B, 46(3), pp. 440-464 (1984), MR 0790631.

[16] DAVISON, A. C., HINKLEY, D.V. Bootstrap Methods and Their Applications.

Cambridge: Cambridge University Press, 2003. ISBN 0-521-57471-4.

[17] EFRON, B.,TIBSHIRANI, R. J. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57, Chapman & Hall/CRC, Boca Raton, USA, 1994, ISBN 0-412-04231-2.

[18] PAVLÍČKOVÁ, L. Metoda bootstrap a její aplikace. Diplomová práce (vedoucí Karpíšek, Z.). ÚM FSI VUT Brno, 2009.

[19] Český statistický úřad-rozvodovost, 15.9.2013, dostupné z <http://www.czso.cz/csu/2013edicniplan.nsf/p/4019-13>, <http://www.czso.cz/csu/2012edicniplan.nsf/t/F00036926F/$File/1413121109.pdf>

[20] NOVOTNÝ, J., LACINOVÁ, V., KARPÍŠEK, Z. Shine Bootstrap. Autorizovaný software, Brno: CQR VUT v Brně, 2011.

[21] Python Programming Language – Official Website, 1.8.2013, dostupné z <http://www.python.org/>.

[22] NumPy, package for scientific computing with Python, 1.8.2013, dostupné z <http://www.numpy.org/>.

[23] WxWidgets, Cross-Platform GUI Library, 23.7.2013, dostupné z <http://www.wxwidgets.org/>.

[24] NOMAD a blackbox optimization software, 23.7.2013, dostupné z <http://www.gerad.ca/nomad/Project/Home.html>.

[25] AGRESTI, A. Categorical Data Analysis. Second Edition (2002), New York, John Wiley & Sons, ISBN 0-471-36093-7.

[26] BUCK, B., MACAULAY, V.A. Maximum Entropy in Action. New York, 1992, Oxford University Press. ISBN 0-19-853941-X.

[27] VAJDA, I. Theory of statistical inference and information. London, Kluver Academic Press, 1989, ISBN 9027727813.

[28] KALAS, J., RÁB, M. Obyčejné diferenciální rovnice. Brno, 2012, ISBN 978-80-210-5815-6.

[29] Průzkum agentury STEM, 20.10.2013, dostupné z <http://aktualne.centrum.cz/domaci/volby/clanek.phtml?id=793173>

[30] PAULOV, J. Teória informacií a regionálna analýza. Pokroky matematiky, fyziky a astronomie, Praha: JČMF, (27) 4, 1982.

[31] JURÁK, P. Odhady diskrétních rozdělení pravděpodobnosti pomocí kvazinorem. Disertační práce, školitel Karpíšek, Z. Ústav matematiky FSI VUT v Brně, 2008, 93 s.

[32] PITMAN, E.J.G. Some Basic Theory for Statistical Inference. New York: John Wiley,1979.

13. Seznam použitých zkratek a symbolů

$\mathbb{R}$	množina reálných čísel
$\mathbb{R}^*$	množina reálných čísel rozšířená o nevlastní prvky ∞ a $-\infty$
Ω	základní prostor
Σ	jevové pole
P	pravděpodobnostní míra
f_j	absolutní četnost
n	rozsah statistického souboru
m	mohutnost základního prostoru Ω
$\frac{f_j}{n}$	relativní četnost
$M(n, p_1, ..., p_m)$	multinomické rozdělení pravděpodobnosti
f	generující funkce *f*-divergence
$D_f(\mathbf{p},\mathbf{q})$	*f*–divergence hustot **p**, **q**
$D_F(\mathbf{p},\mathbf{q})$	duální *f*-divergence k $D_f(\mathbf{p},\mathbf{q})$
S	množina všech diskrétních rozdělení pravděpodobnosti na Ω
$\mathbf{p}$	rozdělení pravděpodobnosti
$\mathbf{p}_0$	rovnoměrné rozdělení pravděpodobnosti
$D_f(\mathbf{p},\mathbf{p}_0)$	kvazinorma diskrétního rozdělení pravděpodobnosti $\mathbf{p}$
$D_F(\mathbf{p},\mathbf{p}_0)$	duální kvazinorma k $D_f(\mathbf{p},\mathbf{p}_0)$
$H(\mathbf{p},\mathbf{p}_0)$	Hellingerova kvazinorma
$S(\mathbf{p},\mathbf{p}_0)$	Shannonova kvazinorma
$P(\mathbf{p},\mathbf{p}_0)$	Pearsonova kvazinorma
$G(\mathbf{p},\mathbf{p}_0)$	G-kvazinormu
$\mathbf{p}(t)$	gradientní odhad
$\mathbf{p}(t_0)$	pesimistický gradientní odhad
α	hladina významnosti
$\chi^2_{1-\alpha}$	$(1-\alpha)$-kvantil chí-kvadrát rozdělení
$\mathbf{J}$	Jordanova normální forma matice
$E(X)$	střední hodnota náhodné veličiny X
n_{ij}	počet případů, kdy se ve výběru vyskytla dvojice (i, j)
$n_{i.}n_{.j}$	marginální četnosti
$p_{i.}p_{.j}$	marginální pravděpodobnosti
$2nI^{\lambda}$	Cressie-Readova statistika
$CR_f(\mathbf{p},\mathbf{p}_0)$	Cressie-Readova kvazinorma
$\frac{V_D}{N}$	míra strmosti kvazinormy D_f

V_D	objem *m*-rozměrného tělesa ohraničeného kvazinormou D_f a S
N	maximální hodnota kvazinormy
B	počet bootstrapových výběrů
MSE	střední kvadratická chyba
$\widehat{\mathrm{MSE}}_b$	bootstrapový odhad střední kvadratické chyby
$\hat{D}\left(\hat{\theta}\right)_b$	bootstrapový odhad rozptylu
$\hat{B}\left(\hat{\theta}\right)_b$	bootstrapový odhad vychýlení
$\chi^2_{b,P}$	*P*-kvantil statistického souboru $\left(\chi^2_{b,1},...,\chi^2_{b,B}\right)$

14 Přehled publikací autorky dizertační práce

14.1 Přehled publikací týkajících se dizertační práce

[1] NERADOVÁ, V. Progresivní metody odhadu rozdělení pravděpodobnosti. Diplomová práce (vedoucí Karpíšek, Z.). FSI VUT v Brně, Brno 2007.

[2] KARPÍŠEK, Z., NERADOVÁ, V. Estimation of Categorical Variable Probability Distribution (Odhad rozdělení pravděpodobnosti kategoriální veličiny). In 7th International Conference APLIMAT 2008. Bratislava, 5. – 8. 2. 2008, Book of abstracts p. 101, ISBN 978-80-89313-02-0, Proceedings pp. 1145-1154, ISBN 978-80-89313-03-7.

[3] KARPÍŠEK, Z., JURÁK, P., NERADOVÁ, V. Divergences and Quasi-norms of Discrete Probability Distributions. In 6th International Conference APLIMAT 2007 (Part I). Bratislava, 6. - 9. 2. 2007, pp. 387-395, ISBN 978-80-969562-5-8.

[4] KARPÍŠEK, Z., NERADOVÁ, V., SADOVSKÝ, Z. Pesimistické odhady diskrétního rozdělení pravděpodobnosti. In Konference Inteligentní systémy pro praxi. Lázně Bohdaneč, 30. – 31. 1. 2008, anotace pp. 65-68, plný text 9 p. na CD ROM, ISBN 978-80-7399-354-2.

[5] KARPÍŠEK, Z., NERADOVÁ, V., ŽAMPACHOVÁ, E. A Contribution to the Estimation of Discrete Probability Distribution. In MENDEL 2008. 14th International Conference on Soft Computing. Brno, 18. – 20. 6. 2008, pp. 287-292, ISBN 978-80-214-3675-6.

[6] KARPÍŠEK, Z., NERADOVÁ, V., ŽAMPACHOVÁ, E. Odhad rozdělení pravděpodobnosti kategoriální veličiny. In REQUEST ´08. Sborník konference CQR. Brno, 25. – 28. 6. 2008, pp. 93-101, ISBN 978-80-214-3774-6.

[7] KARPÍŠEK, Z., LACINOVÁ, V., PAVLÍČKOVÁ, L. Odhady parametrů a rozdělení pravděpodobnosti. In Analýza dat 2010/I – Statistické metody pro technologii a výzkum. Pardubice: TriloByte, CQR, 2010, pp. 47-76, ISBN 978-80-904-053-2-5.

[8] KARPÍŠEK, Z., LACINOVÁ, V. Odhady diskrétního rozdělení pravděpodobnosti s použitím kvazinorem a bootstrapu. In Analýza dat 2010/II – Statistické metody pro technologii a výzkum. Pardubice: TriloByte, CQR, 2010, pp. 131-145, ISBN 978-80-904053-3-2.

[9] KARPÍŠEK, Z., LACINOVÁ, V. Aplikace odhadů rozdělení pravděpodobnosti kategoriální veličiny v průzkumových šetřeních. In Sborník konference Daně – teorie a praxe 2010. Brno: Akademie Sting, 2010, pp. 41-47, ISBN 978-80-86342-91-7.

[10] LACINOVÁ, V., KARPÍŠEK, Z., SADOVSKÝ, Z. Pesimistické odhady rozdělení pravděpodobnosti kategoriální veličiny. In Informační Bulletin České Statistické společnosti. Praha 2011, roč.22, č.2, pp. 138-145, ISSN:1210-8022.

[11] KARPÍŠEK, Z., LACINOVÁ, V., SADOVSKÝ, Z. Metody odhadů parametrů a rozdělení pravděpodobnosti. In Analýza dat 2011/I . Pardubice: TriloByte, 2011, pp. 105-134, ISBN 978-80-904053-4-9.

[12] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z ., LACINOVÁ, V. Bootstrap v indexové analýze. In Analýza dat 2012/II . Pardubice: TriloByte, 2012, pp. 151-160, ISBN 978-80-904053-6-3.

[13] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z ., LACINOVÁ, V. Užití metody bootstrap v indexové analýze. In Daně-teorie a praxe 2012. Brno: Akademie Sting, 2012, pp. 39-45, ISBN 978-80-87482-10-0.

[14] ŽAMPACHOVÁ, E., LACINOVÁ, V., KARPÍŠEK, Z. Pesfit 1.0. Autorizovaný software. Brno: CQR VUT v Brně, 2008.

[15] NOVOTNÝ, J., LACINOVÁ, V., KARPÍŠEK, Z. Shine Bootstrap. Autorizovaný software. Brno: CQR VUT v Brně, 2011.

[16] LACINOVÁ, V., KARPÍŠEK, Z. Line Estimate of Discrete Probability Distribution. Mendel Journal series. 2013, pp.412-415, ISSN 1803-3814.

[17] KARPÍŠEK, Z., NERADOVÁ, V., JURÁK, P., ŠÁCHA, J. Referát semináře Statistika, optimalizace a funkcionální modelování (SOFM), Divergences and quasi-norms for discrete probability distribution estimations, http://mat.fme.vutbr.cz/STATaOPT/ (autorizovaný přístup), 31.12.2007.

[18] LACINOVÁ, V., KARPÍŠEK, Z.. Referát semináře Statistika, optimalizace a funkcionální modelování (SOFM), Pesimistické odhady rozdělení pravděpodobnostikategoriální veličiny , http://mat.fme.vutbr.cz/STATaOPT/ (autorizovaný přístup), 31.12.2012.

[19] LACINOVÁ, V., KARPÍŠEK, Z. Line Estimate of Discrete Probability Distribution. In MENDEL 2013 - 19th International Conference on Soft Computing. June 19-21, 2013, Brno. pp. 412-415, ISSN 1803-3814 (*Mendel Journal Series on CD*).

[20] KARPÍŠEK, Z., LACINOVÁ, V. Použití metody bootstrap pro odhady statistik. Analýza dat 2013. Pardubice: TriloByte, 2013, s. 119-129, ISSN 1805-6903.

[21] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z. LACINOVÁ, V. Indexová analýza s bootstrapem. In REQUEST 2012. Informační bulletin České statistické společnosti, roč. 25, Praha, 2014, 9 s. (v tisku). ISSN 1210-8022.

[22] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z. LACINOVÁ, V. Užití metody bootstrap pro indexovou analýzu. *ACTA STING*, Brno, 2014, 9 s. (v tisku). ISSN 1805-1391 (Print), ISSN 1805-6873 (Online).

[23] NOVOTNÝ, J. KARPÍŠEK, Z., LACINOVÁ, V. Shine_Bootstrap New. Autorizovaný software. Nová verze programu ve Windows pro odhady rozdělení pravděpodobnosti kategoriální veličiny pomocí kvazinorem a bootstrapových odhadů. Brno: GAČR P403/11/2085, TAČR TA02021449, 2013.

[24] LACINOVÁ, V., KARPÍŠEK, Z. Gradient and Line Estimates Employed in Surveyes. In MENDEL 2014 - 20th International Conference on Soft Computing. June 25-27, 2014, Brno. 8 pp., ISSN 1803-3814 (Mendel Journal Series on CD) - přijato k publikaci.

14.2 Další publikace

[1] KARPÍŠEK, Z., LACINOVÁ, V. Intervalová spolehlivost systému. In Sborník konference REQUEST ´09. Liberec 2009, pp. 74-79, ISBN 978-80-7372-619-5.

[2] KARPÍŠEK, Z., LACINOVÁ, V. System Reliability Computed by Interval Arithmetic. In MENDEL 2010 - 16th International Conference on Soft Computing. Brno: FME BUT, 2010, pp. 565-570, ISBN 978-80-214-4120-0.

[3] KARPÍŠEK, Z., LACINOVÁ, V. System Reliability Computed by Interval Arithmetic. In Mendel Journal Series – MENDEL 2010. Brno, 2010, 6 pp. on CD, ISSN 1803-3814.

[4] KARPÍŠEK, Z., LACINOVÁ, V. IJK – Algorithm to Calculate the Interval Reliability. In Aplimat 2011 - 10th International Conference on Applied Mathematics. Bratislava: FME SUT, 2011. Book of abstracts p. 123, ISBN 978-80-89313-52-5, Proceedings on CD pp. 577-584, ISBN 978-80-89313-51-8.

[5] KARPÍŠEK, Z., LACINOVÁ, V. IJK – Algorithm to Calculate the Interval Reliability. In Journal of Applied Mathematics – Aplimat 2011. Bratislava, 2011, p.577-584 ISSN 1337-6365.

[6] KARPÍŠEK, Z., KOCMANOVÁ, A., KRÁL, D ., LACINOVÁ, V. Aplikace intervalové aritmetiky v indexové analýze. ACTA STING, 2012, roč. 1, č. 1/2012, pp. 13-24. ISSN: 1805-1391.

[7] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z., LACINOVÁ, V. Intervalová aritmetika ve spolehlivosti a indexové analýze. In Analýza dat 2011/II. Pardubice: TriloByte, 2011, pp. 121-130, ISBN 978-80-904053-5-6.

[8] KARPÍŠEK, Z., KOCMANOVÁ, A., KRÁL, D., Z. LACINOVÁ, V. Aplikace intervalové aritmetiky v indexové analýze. *ACTA STING*, Brno, 2012, č. 1, s. 13-24, ISSN 1805-1391 (Print), ISSN 1805-6873 (Online).

[9] KARPÍŠEK, Z., KOCMANOVÁ, A., SADOVSKÝ, Z. LACINOVÁ, V. Fuzzy aritmetický průměr statistického souboru. In Daně – Teorie a praxe 2013. Brno: Akademie Sting, 2013, pp. 40-46. ISBN 978-80-87482-10-0.

[10] KARPÍŠEK, Z., LACINOVÁ, V. Fuzzy funkční průměr. In REQUEST 2013. Informační bulletin České statistické společnosti, roč. 25, Praha, 2014, 10 pp. (v recenzním řízení). ISSN 1210-8022.