VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

ODHADY DISKRÉTNÍCH ROZDĚLENÍ PRAVDĚPODOBNOSTI PRO APLIKACE
ESTIMATES OF DISCRETE PROBABILITY DISTRIBUTIONS FOR APPLICATIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. Jakub Mašek

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. Zdeněk Karpíšek, CSc.

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadání diplomové práce

Ústav:	Ústav matematiky
Student:	**Bc. Jakub Mašek**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**doc. RNDr. Zdeněk Karpíšek, CSc.**
Akademický rok:	2015/16

Ředitel ústavu Vám v souladu se zákonem č. 111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma diplomové práce:

Odhady diskrétních rozdělení pravděpodobnosti pro aplikace

Stručná charakteristika problematiky úkolu:

Studium moderních efektivních metod odhadů diskrétních rozdělení pravděpodobnosti z pozorovaných hodnot náhodných veličin, s ohledem na aspekty jejich aplikací v technických a dalších oborech.

Cíle diplomové práce:

Popis, zhodnocení a rozvoj efektivních statistických metod odhadů rozdělení pravděpodobnosti respektujících omezení a neurčitost dat, jejich realizace na PC a aplikace na konkrétních datových souborech.

Seznam literatury:

MONTGOMERY, D. C. a RUNGER, G. Applied Statistics and Probability for Engineers. New York: John Wiley & Sons, 2010.

ANDĚL, J. Statistické metody. Praha: MATFYZPRESS, 2011.

ANDĚL, J. Základy matematické statistiky. Praha: MATFYZPRESS, 2007.

AGRESTI, A. Categorical Data Analysis. New York: John Wiley & Sons, 2002.

VAJDA, I. Theory of Statistical Inference and Information. London: Kluwer Academic Press, 1989.

SCOTT, D. W. Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley, 1992.

Články a materiály z odborných časopisů, sborníků konferencí a Internetu dle pokynů vedoucího diplomové práce.

Termín odevzdání diplomové práce je stanoven časovým plánem akademického roku 2015/16.

V Brně, dne 10. 12. 2015

L. S.

prof. RNDr. Josef Šlapal, CSc.
ředitel ústavu

doc. Ing. Jaroslav Katolický, Ph.D.
děkan

Abstrakt

Diplomová práce je zaměřena na řešení statistického problému nalezení rozdělení pravděpodobnosti diskrétní náhodné veličiny na základě pozorovaných dat. Tyto odhady jsou získány minimalizací pseudokvazinormy, která je zde zavedena. Práce se zabývá vlastnostmi této pseudokvazinormy. Obsahuje také praktickou aplikaci této metody.

Summary

Master's thesis is focused on solution of the statistical problem to find a probability distribution of a discrete random variable on the basis of the observed data. These estimates are obtained by minimizing pseudo-quasinorm which is introduced here.The thesis further focuses on atributes of this pseudo-quasinorm. It also contains practical application of these methods.

Klíčová slova

f-divergence, kvazinorma, odhad diskrétního rozdělení pravděpodobnosti, chí-kvadrát rozdělení

Keywords

f-divergence f-quasinorm, discrete probability distribution estimation, chi-squared distribution

MAŠEK, J.*Odhady diskrétních rozdělení pravděpodobnosti pro aplikace.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2016. 43 s. Vedoucí doc. RNDr. Zdeněk Karpíšek, CSc.

Prohlašuji, že jsem zadanou diplomovou práci zpracoval samostatně s přispěním vedoucího diplomové práce a používal jsem pouze literaturu v práci uvedenou.

Bc. Jakub Mašek

Rád bych poděkoval svému vedoucímu práce Doc. RNDr. Zdeňku Karpíškovi, CSc. za cenné rady a vedení.

Bc. Jakub Mašek

Obsah

1. Úvod

V matematické statistice je použití parametrických statistických testů i stanovení intervalových odhadů parametrů podmíněno předpoklady o rozdělení pravděpodobnosti pozorované náhodné veličiny nebo náhodného vektoru. Zanedbání předpokladu o tvaru rozdělení pravděpodobnosti může vést v praxi ke zcela zavádějícím výsledkům a proto je nutné věnovat tomuto problému patřičnou pozornost. Práce se zabývá fitováním (odhadům) diskrétních rozdělení pravděpodobnosti pomocí netradičních, resp. nových metod. Tyto metody jsou založeny na vyjádření jisté "normy"hledaného rozdělení, které je tak "nejblíže"rovnoměrnému rozdělení a současně splňuje předem dané požadavky. Poznamenejme, že v práci získané teoretické výsledky je možno také aplikovat na diskretizovaná pozorovaná spojitá rozdělení pravděpodobnosti.

Práce je rozdělena do 8 kapitol. První kapitola po úvodu pojednává o už zjištěných možnostech fitování rozdělení. Další dvě kapitoly zavádí hojně používaný pojem chí--kvadrát rozdělení a kvazinorma. Také se zde popisují jejich vlastnosti využívané později v hlavní přínosné kapitole této práce. Následuje kapitola Hledání minima f-kvazinorem, kde jsou uvedeny nutné podmínky minima pro nelineární optimalizaci využívané později.

Hlavní těžiště práce je v kapitole F-T pseudokvazinorma, kde je předložena nová tzv. Freeman-Tukyho pseudokvazinorma. Tato charakteristika vychází z dříve studovaného pojmu kvazinormy a jsou zde prezentovány její základní vlastnosti pro fitování pozorovaného diskrétního rozdělení pravděpodobnosti za vedlejších momentových podmínek.

Kapitola Aplikace pseudokvazinorem je zaměřena formou příkladů na aplikace fitování pomocí uvedené pseudokvazinormy.

2. Odhady rozdělení pravděpodobnosti

Odhady (fitování) rozdělení pravděpodobnosti pozorovaných náhodných veličin a náhodných vektorů mají zásadní význam pro odhady parametrů i testování statistických hypotéz. Podle způsobu realizace odhadů rozdělení je můžeme rozdělit na empirické a inferenční.

1. **Empirické odhady** rozdělení pravděpodobnosti jsou založeny na:
   - teoretických a zkušenostních předpokladech a informacích o tvaru rozdělení,
   - grafických vyjádření statistických souborů (histogramy, polygony aj.).
2. Mezi **inferenční (indukční) metody odhadů** rozdělení pravděpodobnosti patří zejména:
   - Pearsonovy křivky,
   - Gramovy – Charlierovy řady
   - Johnsonovy křivky,
   - jádrové odhady,
   - odhady pomocí kvazinorem.

Postup odhadování a verifikace rozdělení pravděpodobnosti na základě získaného statistického souboru probíhá v krocích:

1. Grafické znázornění statistického souboru.
2. Vlastní odhad rozdělení.
3. Testování shody rozdělení.

Při empirických odhadech (ale i při odhadech inferenčních) se musíme vyrovnat s řadou problémů, zejména

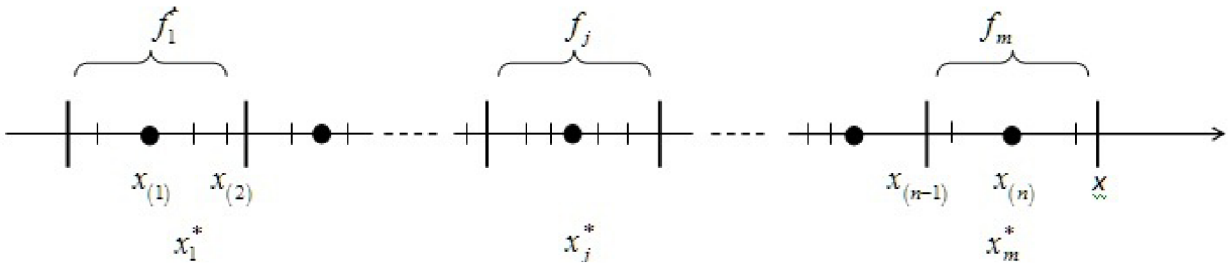
- rozhodnout, zda se jedná spojitou anebo diskrétní náhodnou veličinu s ohledem na přesnost pozorování (měření) a možný obor jejích hodnot,
- posoudit šikmost a špičatost rozdělení pozorované náhodné veličiny,
- posoudit vícemodalitu rozdělení pozorované náhodné veličiny,
- odfiltrovat extrémně odchýlené pozorované hodnoty,
- zvážit nutnost respektování dimenze vícerozměrného statistického souboru.

2.1. Empirické odhady rozdělení

Neroztříděný statistický soubor $(x_1, \ldots, x_n)$ nebo uspořádaný statistický soubor $\big(x_{(1)}, \ldots, x_{(n)}\big)$, $x_{(i)} \leq x_{(i+1)}$, $i = 1, \ldots, n$ s rozsahem n převedeme na roztříděný statistický soubor (variační řadu): Přitom x_j^* je střed třídy, f_j je četnost hodnot $x_{(i)}$ v j-té třídě, $j = 1, \ldots, m$;

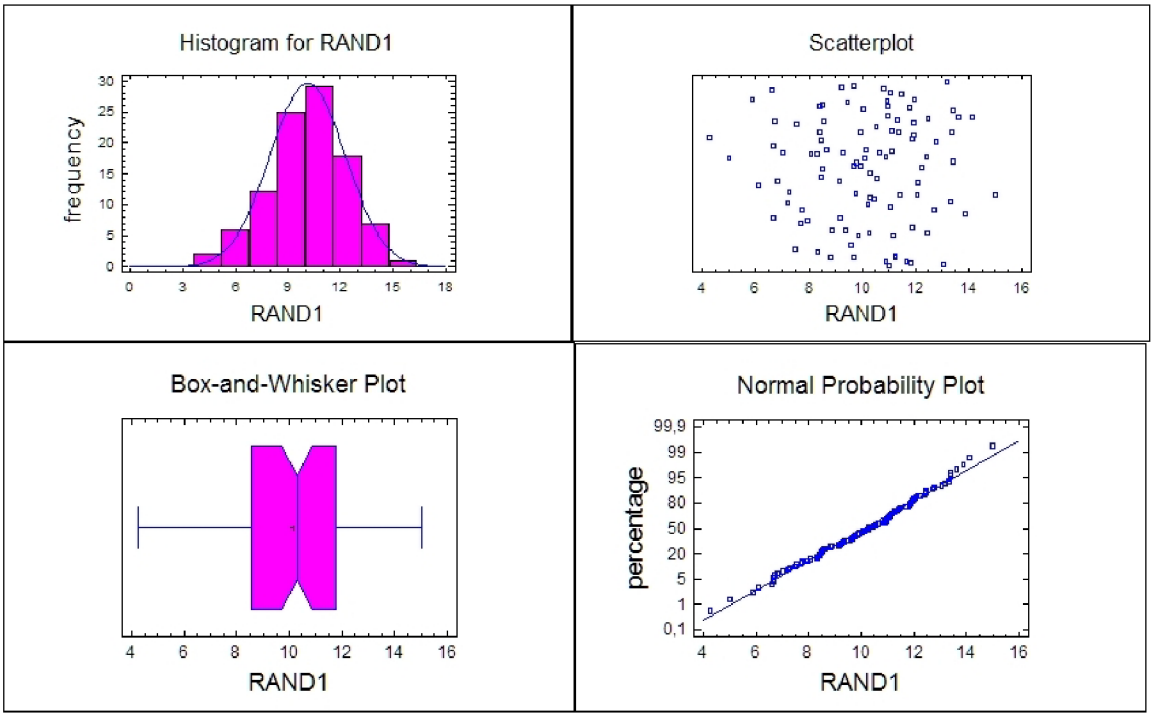
x_j^*	$x_1^*, \ldots, x_m^*$
f_j	$f_1, \ldots, f_m$

třídy jsou zleva otevřené a zprava uzavřené intervaly. Grafické vyjádření roztřídění:



Počet tříd je $m < n$ a obvykle jej volíme přibližně $1 + 3,3 \log n$ pro soubory symetrického charakteru nebo $\sqrt{n}$ až $2\sqrt{n}$ pro soubory nesymetrického charakteru.

Příklad: Příklad ilustruje základní grafické zpracování statistického souboru získaného generátorem náhodných čísel pro normální rozdělení se střední hodnotou $\mu = 2$ a rozptylem $\sigma^2 = 10$, odhady parametrů a následnou verifikaci pomocí testů shody. Grafy a výpočty byly realizovány pomocí statistického softwaru Statgraphics.



Data variable: RAND1
100 values ranging from 4,25663 to 14,9898

Fitted normal distribution:
mean = 10,131
standard deviation = 2,15925

Tests for Normality for RAND1
Computed Chi-Square goodness-of-fit statistic = 19,52
P-Value = 0,55182

Shapiro-Wilks W statistic = 0,981384
P-Value = 0,600628
Z score for skewness = 0,837386
P-Value = 0,402374
Z score for kurtosis = -0,439584
P-Value = 0,660235
Goodness-of-Fit Tests for RAND1

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	6,57937	4	5,00	0,20
6,57937	7,36383	8	5,00	1,80
7,36383	7,8931	4	5,00	0,20
7,8931	8,31375	3	5,00	0,80
8,31375	8,67463	9	5,00	3,20
8,67463	8,99871	2	5,00	1,80
8,99871	9,29902	3	5,00	0,80
9,29902	9,58398	3	5,00	0,80
9,58398	9,85968	6	5,00	0,20
9,85968	10,131	6	5,00	0,20
10,131	10,4024	4	5,00	0,20
10,4024	10,6781	4	5,00	0,20
10,6781	10,963	4	5,00	0,20
10,963	11,2633	10	5,00	5,00
11,2633	11,5874	4	5,00	0,20
11,5874	11,9483	6	5,00	0,20
11,9483	12,3689	5	5,00	0,00
12,3689	12,8982	5	5,00	0,00
12,8982	13,6827	7	5,00	0,80
above 13,6827		3	5,00	0,80

Chi-Square = 17,6 with 17 d.f. P-Value = 0,414482

Estimated Kolmogorov statistic DPLUS = 0,0348125
Estimated Kolmogorov statistic DMINUS = 0,0698938
Estimated overall statistic DN = 0,0698938
Approximate P-Value = 0,713002

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	**0,0698938**	**0,70418**	**>=0.10***
Kuiper V	0,104706	1,06088	>=0.10*
Cramer-Von Mises W^2	0,0539431	0,0542128	0,4510*
Watson U^2	0,0474934	0,0477309	0,5026*
Anderson-Darling A^2	0,312734	0,315149	0,5434*

*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the currently selected distribution. Other P-values are based on general tables and may be very conservative.

2.2. Pearsonovy křivky

Odhad rozdělení pravděpodobnosti pomocí vychází z diferenciální rovnice pro hustotu pravděpodobnosti$f(x)$ pozorované náhodné veličiny X ve tvaru

$$\frac{df}{f} = \frac{x+d}{c_0 + c_1 x + c_2 x^2} dx$$

. K. Pearson zavedl celkem 7 typů rozdělení značených I až VII, přičemž některé z nich mají ještě výrazně odlišené podtypy značené indexy. Celkem se tedy uvádí 12 typů a podtypů. Obecné řešení diferenciální rovnice je v

$$f = f_0 \mathrm{e}^{v(x)},$$

kde

$$v\,(x) = \int \frac{x+d}{c_0 + c_1 x + c_2 x^2} dx$$

Typově závisí integrál na hodnotách koeficientů kvadratického polynomu ve jmenovateli integrandu. Konstanty c_0, c_1, c_2, d lze vyjádřit pomocí normovaných momentů r_1, r_2, r_3, r_4 hustoty pravděpodobnosti $f\,(x)$ ve tvaru

$$c_0 = -\sigma^2 \frac{s+1}{s-2}, \quad c_1 = -d = -\frac{\sigma r_3}{2} \frac{s+2}{s-2}, \quad c_2 = \frac{1}{s-2},$$

kde

$$s = \frac{6(r_4 - r_3^2 - 1)}{3r_3^2 - 2r_4 + 6},$$

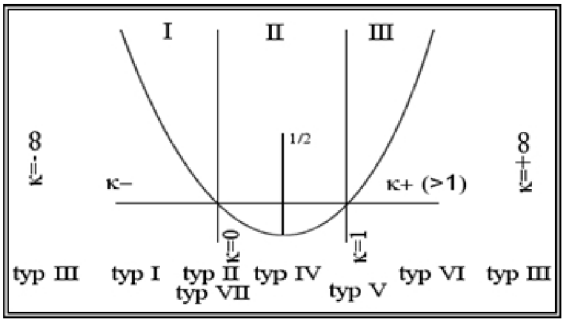
přičemž $r_1 = 0$, $r_2 = 1$ a σ^2 je rozptyl pozorované náhodné veličiny X. Z diskriminantu kvadratického polynomu $c_0 + c_1 x + c_2 x^2$ určíme typ pozorovaného rozdělení, konkrétně podle hodnoty charakteristiky

$$\kappa = \frac{c_1^2}{4 c_0 c_2} = \frac{r_3^2 (r_4 + 3)^2}{4(4r_4 - 3r_3^2)(2r_4 - 3r_3^2 - 6)}.$$

Následující obrázek popisuje rozdělení Pearsonových křivek pomocí grafu paraboly $y = \kappa(\kappa - 1)$

I. pro $\kappa < 0$
II. pro $\kappa = 0 a r_4 < 3$
III. pro $\kappa = \pm\infty$
IV. pro $0 < \kappa < 1$
V. pro $\kappa = 1$
VI. pro $\kappa > 1$
VII. pro $\kappa = 0$ a $r_4 > 3$

Je-li $\kappa = 0$ a $r_4 = 3$, jedná se o normální rozdělení.



Příklad: Hodnoty v následující tabulce udávají věk X vědeckých pracovníků v SSSR v roce 1928. Relativní četnosti n_j jsou v ‰ a statistický soubor je roztříděný.

Třídy	n_j	Třídy	n_j
20 - 24	11	55 - 59	67
25 - 29	93	60 - 64	40
30 - 34	163	65 - 69	24
35 - 39	178	70 - 74	12
40 - 44	176	75 - 79	3
45 - 49	132	80 - 84	1
50 - 54	100	Σ	1000

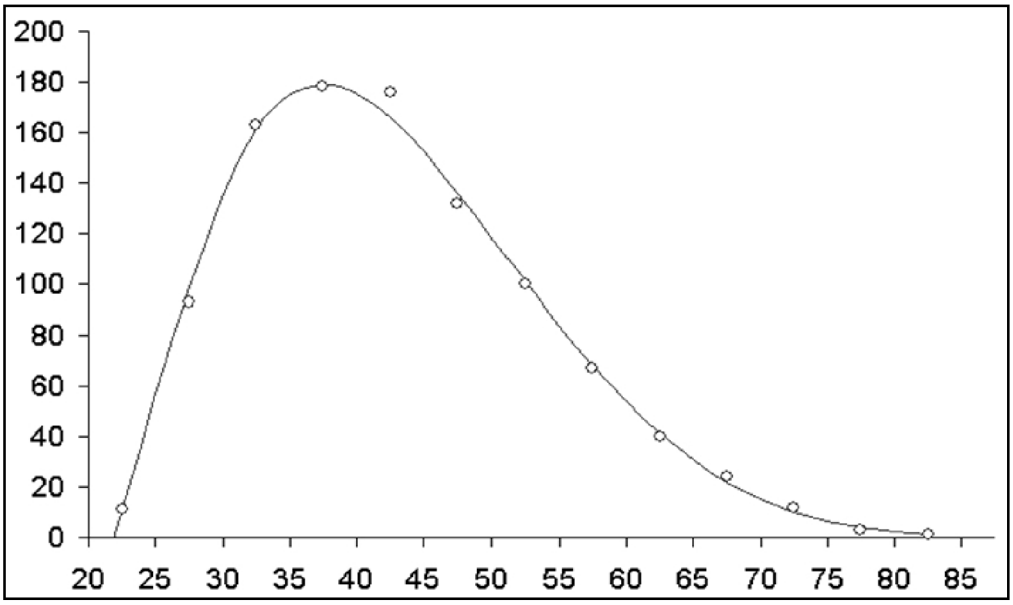
Z tabulky vypočteme empirické charakteristiky $\bar{x} = 42,935$; $\sigma = 2,203$; $r_3 = 0,605$; $r_4 = 2,968$. Pak je $\kappa = -0,26 < 0$, takže použijeme Pearsonovu křivku typu I. Dalším výpočtem pak obdržíme odhady četností ve tvaru

$$\tilde{n}_j = 179,43 \left(1 + \frac{x_j^*}{3,017}\right)^{1,256} \left(1 - \frac{x_j^*}{12,043}\right)^{5,016}$$

Vypočtené četnosti jsou v následující tabulce. Vykreslením původních četností n_j a křivky

j	Třídy	Střed třídy x_j^*	Pozorovaná četnost n_j	Odhadnutá četnost $\tilde{n}_j$
1	20 - 24	22,5	11	12
2	25 - 29	27,5	93	99
3	30 - 34	32,5	163	161
4	35 - 39	37,5	178	179
5	40 - 44	42,5	176	166
6	45 - 49	47,5	132	136
7	50 - 54	52,5	100	101
8	55 - 59	57,5	67	68
9	60 - 64	62,5	40	41
10	65 - 69	57,5	24	22
11	70 - 74	72,5	12	10
12	75 - 79	77,5	3	4
13	80 - 84	82,5	1	1
Σ	-	-	1000	1000

prokládající vypočtené četnosti $\tilde{n}_j$ do grafu dostáváme výsledný tvar rozdělení



Z tabulky i grafu je patrná dobrá aproximace neznámého rozdělení a potvrdil by ji asi i chí-kvadrát test rozdělení pravděpodobnosti, avšak není znám celkový počet pracovníků ve výběru.

2.3. Gramovy – Charlierovy řady

Třídy těchto rozdělení vychází z vyjádření funkce $\ln \varphi(t; n, p)$, kde

$$\varphi(t; n, p) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \mathrm{e}^{\mathrm{i}tx} = \left[p\left(\mathrm{e}^{it} - 1\right) + 1\right]^n$$

je charakteristická funkce binomického rozdělení Bi(n,p), pomocí ortogonálního systému funkcí tvořícího bázi založenou:

- ve spojitém případě (vzhledem k t) na hustotě normálního rozdělení (**typ A**),
- v diskrétním případě (vzhledem k p) na pravděpodobnostní funkci Poissonova rozdě--lení (**typ B**).

Hustota pravděpodobnosti pro rozdělení **typu A** má tvar

$$f_A(x) = f(x) - \frac{r_3}{6} f^{(3)}(x) + \frac{r_4 - 3}{24} f^{(4)}(x) - \frac{r_5 - 10r_3}{120} f^{(5)}(x) + \frac{r_6 - 15r_4 + 30}{720} f^{(6)}(x) \ + \cdots,$$

kde

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}$$

je hustota normovaného normálního rozdělení a $f^{(n)}(x)$ je n-tá derivace $f(x)$ podle x. Ve většině praktických případů postačuje zjednodušený tvar s pouze prvními třemi členy

$$f_A(x) = f(x) - \frac{r_3}{6} f^{(3)}(x) + \frac{r_4 - 3}{24} f^{(4)}(x).$$

První člen pravé části nám dává normální rozdělení, druhý člen reflektuje šikmost a třetí člen špičatost hledaného rozdělení. Četnost výskytu v případě roztříděného souboru určíme pro středy tříd x_j^* ze vztahu (n je rozsah pozorování)

$$\tilde{n}_j = \frac{n}{\sigma} f_A\left(x_j^*\right).$$

Příklad: Měřením mezi pevnosti $X\,(kg/cm^2)$ při stlačení vzorku z borového dřeva podél vláken byly po roztřídění získány hodnoty v následující tabulce

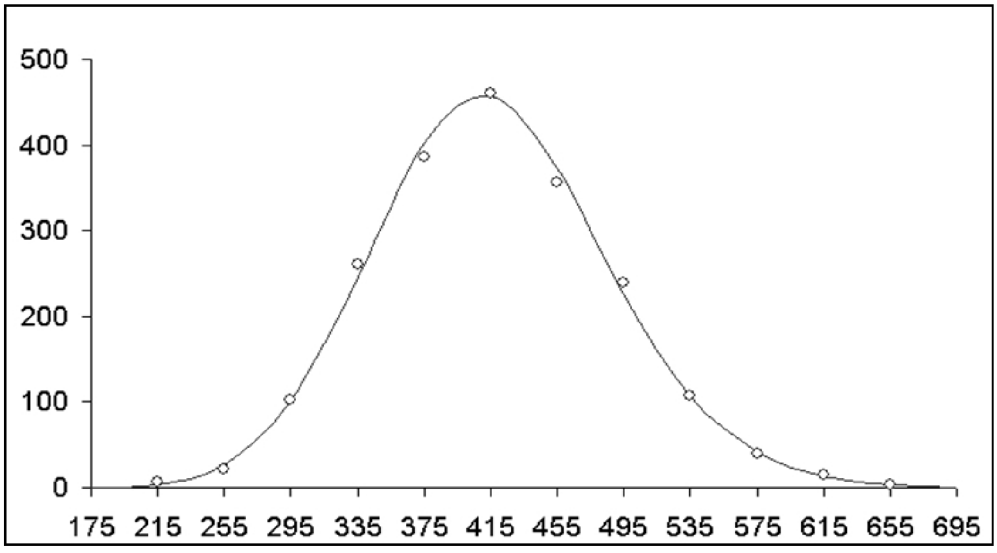
x_j^*	215	255	295	335	375	415	455
n_j	7	22	102	260	386	461	356
x_j^*	495	535	575	615	655	Σ	
n_j	239	108	40	15	4	2000	

Výpočtem s použitím známých Sheppardových korekcí pro třídění získáme z tabulky $\bar{x} = 416,64$; $\sigma = 1,752$; $r_3 = 0,205$; $r_4 = 3,021$ a obdržíme odhad hustoty rozdělení pravděpodobnosti

$$f_A(x) = f(x) - 0,034\, f^{(3)}(x) + 0,001\, f^{(4)}(x).$$

V následující tabulce jsou odhady četností

x_j^*	215	255	295	335	375	415	455
$\tilde{n}_j$	3,63	26,01	100,36	245,49	402,56	457,57	372,69
x_j^*	495	535	575	615	655	Σ	
$\tilde{n}_j$	226,81	107,89	41,16	12,45	3,36	2000	



Z tabulky i grafu je zřejmá velmi dobrá aproximace původního neznámého rozdělení. Tomu také odpovídá P-hodnota chí-kvadrát testu dobré shody pro 12 – 3 = 9 stupňů volnosti.

Pravděpodobnostní funkce pro rozdělení **typu B** má tvar

$$p_B(x) = \frac{\lambda}{x!}\mathrm{e}^{-\lambda}\left\{1 + \frac{\mu_2 - \lambda}{\lambda^2}\left[\frac{x^{[2]}}{2} - \lambda x^{[1]} + \frac{\lambda^2}{2}\right] + \frac{\mu_3 - 3\mu_2 + 2\lambda}{\lambda^3}\left[\frac{x^{[3]}}{6} - \frac{\lambda}{2}x^{[2]} - \frac{\lambda^2}{2}x^{[1]} - \frac{\lambda^3}{6}\right] + \right.$$
$$\left. + \frac{\mu_4 - 6\mu_3 + (11 - 6\lambda)\mu_2 - 3\lambda(2 - \lambda)}{\lambda^4}\left[\frac{x^{[4]}}{24} - \frac{\lambda}{6}x^{[3]} + \frac{\lambda^2}{4}x^{[2]} - \frac{\lambda^3}{6}x^{[1]} + \frac{\lambda^4}{24}\right] + \cdots\right\},$$

kde $x = 0, 1, 2, \ldots$ a $x^{[k]}$ značí variaci k-té třídy z x prvků bez opakování. Četnosti pak jsou

$$\tilde{n}_j = np_B(x).$$

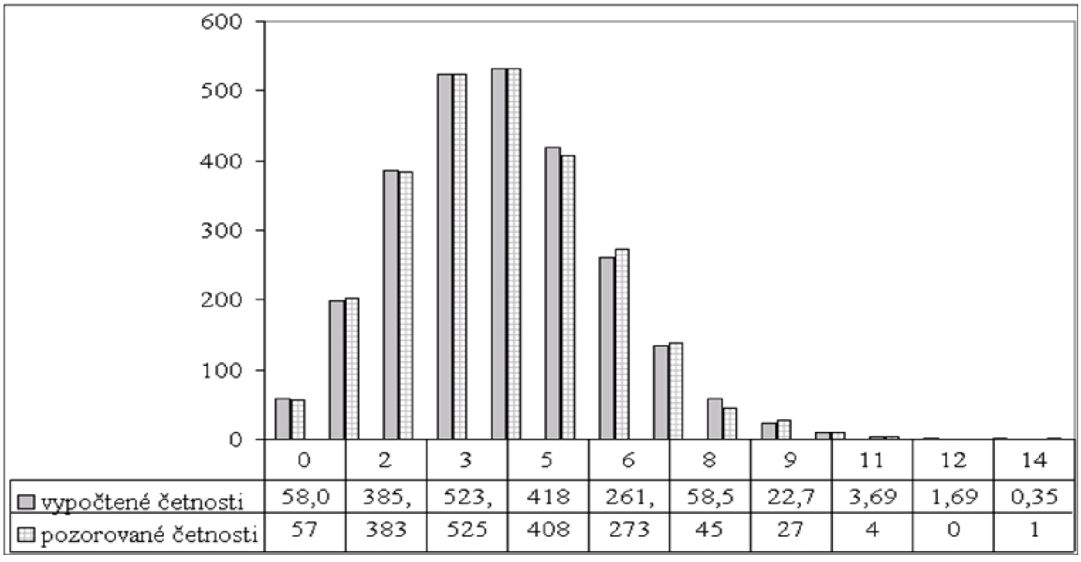
Příklad: Odhadněte diskrétní rozdělení z následujících dat, kde četnosti udávají počty α částic emitovaných poloniem v konstantních časových intervalech (1/8 minuty)

Počet α-částic x_j	0	1	2	3	4	5	6	7
Četnost n_j	57	203	383	525	532	408	273	139
Počet α-částic x_j	8	9	10	11	12	13	14	Σ
Četnost n_j	45	27	10	4	0	1	1	2608

Z tabulky vypočteme momenty $\lambda = \bar{x} = 3,872\,; \mu_2 = 3,695\,; \mu_3 = 3,398\,; \mu_4 = 47,869$ a pak po úpravě je

$$\tilde{n}_j = 2608\frac{3,872^x}{x_j!}\mathrm{e}^{-3,872}\left\{1,0695 - 0,12x_j^{[1]} + 0,0593x_j^{[2]} - 0,0114x_j^{[3]} + 0,00075x_j^{[4]}\right\}.$$

Vykreslením odhadnutých četností a původních četností obdržíme níže uvedený sloupcový graf, kde jsou také zapsány odhadnuté četnosti.



	0	2	3	5	6	8	9	11	12	14
vypočtené četnosti	58,0	385,	523,	418	261,	58,5	22,7	3,69	1,69	0,35
pozorované četnosti	57	383	525	408	273	45	27	4	0	1

2.4. Johnsonovy křivky

Jinou soustavu odhadů spojitých rozdělení pravděpodobnosti pomocí čtyřparametrických distribučních funkcí navrhl N. L. Johnson. Tyto odhady vychází z nelineární transformace normálního rozdělení $N(0;1)$. V aplikacích často vystačíme se třemi jednoduchými typy nelineární transformace. Tvarová rozmanitost rozdělení se ve všech třech případech dociluje dvěma parametry tvaru k a m, k nimž přistupují ještě parametr polohy a a parametr měřítka b.

1. Transformací
$$z_L = b \exp\left(\frac{x-k}{m}\right) + a$$
získáme Johnsonovo rozdělení typu S_L(L = lognormal) s hustotou pravděpodobnosti
$$f_L(z) = \frac{m}{\sqrt{2\pi}\,(z-a)} \exp\left\{-\frac{1}{2}\left[k + m \ln\left(\frac{z-a}{b}\right)\right]^2\right\},$$
kde $z > a,\ m > 0,\ b > 0,\ k \in \mathbb{R},\ a \in \mathbb{R}$.

2. Transformací
$$z_B = b\frac{\exp\left(\frac{x-k}{m}\right)}{1+\exp\left(\frac{x-k}{m}\right)} + a$$
získáme Johnsonovo rozdělení typu S_B (B = bounded) s hustotou pravděpodobnosti
$$f_B(z) = \frac{m}{\sqrt{2\pi}} \frac{b}{(z-a)\,(b-z+a)} \exp\left\{-\frac{1}{2}\left[k + m \ln\left(\frac{z-a}{b-z+a}\right)\right]^2\right\},$$
kde $z \in \langle a,\, a+b\rangle,\ m > 0,\ b > 0,\ k \in \mathbb{R},\ a \in \mathbb{R}$.

3. Transformací
$$z_U = b \sinh\left(\frac{x-k}{m}\right) + a$$
dostáváme Johnsonovo rozdělení typu S_U (U = unbounded) s hustotou pravděpodobnosti
$$f_U(z) = \frac{m}{\sqrt{2\pi}} \frac{1}{\sqrt{(z-a)^2 + b^2}} \exp\left\{-\frac{1}{2}\left[k + m \ln\left(\left(\frac{z-a}{b}\right) + \sqrt{\left(\frac{z-a}{b}\right)^2 + 1}\right)\right]^2\right\},$$
kde $z \in \mathbb{R},\ m > 0,\ b > 0,\ k \in \mathbb{R},\ a \in \mathbb{R}$.

Uvedená rozdělení jsou čtyřparametrická a jejich parametry obvykle určujeme metodou maximální věrohodnosti.

2.5. Jádrové odhady

Jádrové odhady hustoty spojitého rozdělení pravděpodobnosti vycházejí z tzv. **jádrové funkce K**, což je nezáporná funkce

$$\int_{-\infty}^{\infty} K(x)\ dx = 1.$$

Jádrový odhad s jádrem K hustoty pravděpodobnosti $f(x)$ pozorované spojité náhodné veličiny X je pak funkce

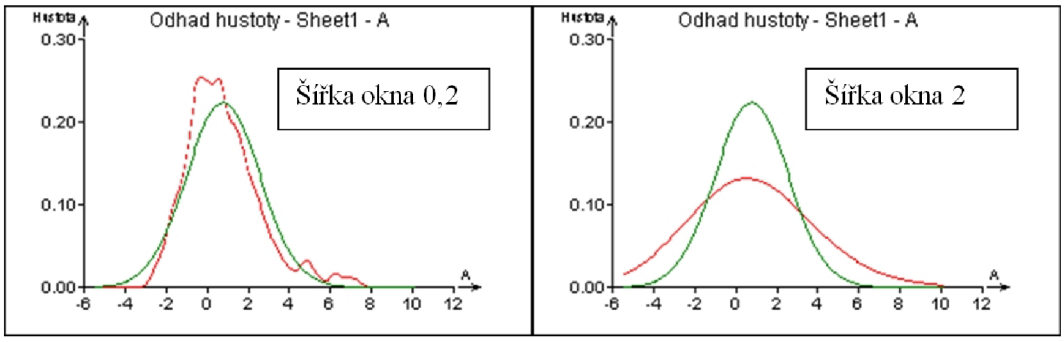
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

kde parametr h je **šířka vyhlazovacího okna** (vyhlazovací parametr) a x_i je pozorovaná hodnota X , tj. prvek statistického souboru $(x_1, \ldots, x_n)$, který netřídíme.

Nejčastěji se používají jádra

- Epanechnikovo jádro $K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}x^2\right) & \text{pro } |x| < \sqrt{5} \\ 0 & \text{jinde} \end{cases}$
- Trojúhelníkové jádro $K(x) = \begin{cases} 1 - |x| & \text{pro } |x| < 1 \\ 0 & \text{jinde} \end{cases}$
- Gaussovo jádro $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ pro $x \in (-\infty, \infty)$
- Jádro s dvojnásobnou váhou $K(x) = \begin{cases} \frac{15}{16}(1 - x^2)^2 & \text{pro } |x| < 1 \\ 0 & \text{jinde} \end{cases}$
- Obdélníkové jádro $K(x) = \begin{cases} \frac{1}{2} & \text{pro } |x| < 1 \\ 0 & \text{jinak} \end{cases}$

K základním problémům aplikace jádrových odhadů patří volba jádra a šířka vyhlazovacího okna. Dále pak jde o respektování požadavků spojitosti, příp. hladkosti získané hustoty pravděpodobnosti a důležité je také vyjádření odpovídající distribuční funkce a snadnost výpočtu kvantilů, příp. pokrytí celého rozsahu hodnot náhodné veličiny X, neboť jádrové odhady mají někdy charakter šumu na koncích rozdělení. V následujícím grafu vidíme jádrové odhady se špatně zvolenou šířkou h.



3. Test dobré shody

V této kapitole se budeme věnovat tzv. testům dobré shody a to speciálně χ^2 testu. Pomocí testů dobré shody se testuje hypotéza, zda-li výběr který máme pochází z určitého rozdělení nebo ne. Výsledná hodnota testu se pak porovná s kvantilem χ^2 rozdělení, zda-li hypotézu zamítáme, či nikoliv.
Před odvozením χ^2 testu popíšeme χ^2 rozdělení a také multinomické rozdělení, z kterého χ^2 test vychází.
V této kapitole je čerpáno především z [1].

3.1. χ^2 rozdělení

Nechť $X \sim N(0,1)$. Položme $Y = X^2$. Distribuční funkci náhodné veličiny Y označme G.
Pokud $y \leq 0$ tak platí, že $G(y) = 0$. Pokud $y > 0$ dostáváme následující:

$$G(y) = P(Y > y) = P(X^2 < y) = P(-\sqrt{y} < X < \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}),$$

kde Φ je distribuční funkce standardního normálního rozdělení. Proto platí $\Phi(-x) = 1 - \Phi(x)$. Takže dostáváme

$$G(y) = 2\Phi(\sqrt{y}) - 1.$$

Jelikož hustota je derivací distribuční funkce, dostáváme, že hustota g náh. veličiny Y rovná se 0 na $y \leq 0$ a na $y > 0$ je

$$g(y) = \frac{1}{\sqrt{y}}\varphi(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}}\mathrm{e}^{-y/2}$$

Náhodnou veličinu Y s hustotou g nazveme **chi-kvadrát rozdělení** s jedním stupněm volnosti. Střední hodnota rozdělení je pak $\mu = EY = EX^2 = 1$ a rozptyl je $\sigma^2 = varY = EY^2 - (EY)^2 = EX^4 - (EX^2)^2 = 3 - 1 = 2$.

Nyní rozšíříme chí-kvadrát rozdělení z jednoho stupně volnosti na n stupňů. K tomu potřebujeme následující větu.
Věta 3.1. *Nechť X_1 a X_2 jsou nezávislé náh. veličiny. Nechť X_1 má hustotu f_1 a X_2 má hustotu f_2. Pak náh. veličina $Y = X_1 + X_2$ má hustotu*

$$h(y) = \int f_1(z) f_2(y - z) dz.$$

Mějme náhodnou veličinu $Y = X_1 + X_2 + \cdots + X_n$, kde $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny s rozdělením $N \sim N(0,1)$. Hustotu Y budeme značit $f_n(y)$.
Pro $n = 1$ už hustotu známe.

$$f_1(y) = \frac{1}{\sqrt{2\pi y}}\mathrm{e}^{-y/2} \qquad \text{pro} \quad y > 0$$

Pro $n > 1$ předpokládejme, že vzorec pro $f_1(y)$ platí pro všechna $k \leq n$. Z věty 3.1 dostáváme

$$f_{k+1}(y) = \int f_k(z) f_1(y-z) dz.$$

Pomocí indukce lze pro $n = k+1$ dostat následující vzorec

$$f_n(y) = \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n}{2}-1} \mathrm{e}^{-y/2} \qquad \text{pro} \quad y > 0. \tag{3.1}$$

Hustota $f_n(y)$ je hustotou **chi-kvadrát rozdělení s n stupni volnosti**, jež značíme χ^2_n. Střední hodnota $\mu = EY = n$ a rozptyl $\sigma^2 = varY = 2n$ jelikož pro každé i se rovnají 1 resp. 2. Z toho vidíme, že střední hodnota je rovna počtu stupňů volnosti a rozptyl je její dvojnásobek.

3.2. Multinomické rozdělení

Mějme pokus, kde mohou nastat jevy $A_1, \ldots, A_k$ takové, že jeden z nich nastat musí a nemohou nastat dva najednou. Jinak řečeno jsou disjunktní a součet všech jejich pravděpodobností je roven 1. Pravděpodobnosti těchto jevů označíme $p_i = P(A_i), i = 1, \ldots, k$. Nyní opakujme pokus n-krát nezávisle po sobě. Četnost výskytu jevu A_i označíme X_i. Pak

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \tag{3.2}$$

pro

$$x_i = 0, 1, \ldots, n \quad (i = 1, \ldots, k), \qquad x_1 + \cdots + x_k = n$$

Rozdělení dané vzorcem (3.2) se nazývá **multinomické** a značíme $\mathbf{X} \sim M(n; p_1, \ldots, p_k)$ resp. $M(n; \mathbf{p})$

Speciálně, pokud máme $k = 2$ (aneb případ úspěchu a neúspěchu pokusu) přechází multinomické rozdělení na binomické. Totéž platí i pro k obecné, máme li $A = A_1$ a $B = A_2 \cup \cdots \cup A_k$. Tudíž všechna jednorozměrná marginální rozdělení jsou binomická. Konkrétně $X_i \sim Bi(n; p_i)$.

Věta 3.2. *V multinomickém rozdělení platí*

$$EX_i = np_i, \quad varX_i = np_i(1 - P_i), \quad \text{pro} \quad i = 1, \ldots, k$$

a

$$cov(X_i, X_j) = -np_i p_j, \quad \text{pro} \quad 1 \leq i \neq k \leq k$$

Důkaz lze najít například v [1].

Věta 3.3. *Nechť* $\mathbf{X} \sim N(0, \mathbf{V})$ *je náhodný vektor, kde* $\mathbf{V}$ *je idempotentní matice (tj. platí* $\mathbf{V^2} = \mathbf{V}$*) hodnosti* $r \geq 1$*. Pak* $\mathbf{X'X} \sim \chi^2_r$.

Věta 3.4. *Označme*

$$\mathbf{p} = (\sqrt{p_1}, \ldots, \sqrt{p_k})', \qquad \mathbf{Q} = \mathbf{I} - \mathbf{pp}', \qquad \mathbf{D} = diag\{\sqrt{np_1}, \ldots, \sqrt{np_k}\}.$$

Pak matice $\mathbf{Q}$ *je idempotentní a má hodnost* $k-1$. *Variační matice* V *náhodného vektoru* $\mathbf{X} = (X_1, \ldots, X_k)$ *mající multinomické rozdělení je rovna*

$$\mathbf{V} = \mathbf{DQD}$$

Matice $\mathbf{V}$ *má také hodnost* $k-1$.

Věta 3.5 (Centrální limitní věta). *Nechť* $\mathbf{X_1}, \mathbf{X_2}, \ldots$ *jsou nezávislé náhodné vektory, které mají stejné rozdělení se střední hodnotou* $\boldsymbol{\mu}$ *a variační matici* $\mathbf{V}$ *s konečnými prvky. Pak pro* $n \to \infty$ *platí*

$$\frac{1}{\sqrt{n}}(\mathbf{X}_1 + \cdots + \mathbf{X}_k - n\boldsymbol{\mu}) \to N(0, \mathbf{V})$$

Jelikož platí 3.5, má $\mathbf{X}$ asymptoticky normální rozdělení. Položme $\mathbf{Y} = \mathbf{D}^{-1}\mathbf{X}$. Pak má také $\mathbf{Y}$ asymptoticky normální rozdělení (jelikož vzniká lineární transformací z $\mathbf{X}$). Navíc

$$E\mathbf{Y} = \mathbf{D}^{-1}E\mathbf{x} = (\sqrt{np_1}, \ldots, \sqrt{np_k})', \qquad var\mathbf{Y} = \mathbf{Q}$$

Nyní si vezmeme veličinu $(\mathbf{Y} - E\mathbf{Y})'(\mathbf{Y} - E\mathbf{Y})$ tradičně značenou χ^2. Jelikož $\mathbf{Y}$ má asymptoticky normální rozdělení a platí věta 3.3. Má i tato veličina asymptoticky rozdělení χ^2_{k-1}. $k-1$ jelikož matice $\mathbf{V}$ má hodnost právě $k-1$. Po dosazení a úpravě dostaneme následující

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}. \tag{3.3}$$

Někdy taktéž používané ve tvaru

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{k} \frac{X_i^2}{p_i} - n. \tag{3.4}$$

Veličinám X_i se nazývají empirické četnosti a np_i jsou teoretické četnosti. Tvar (3.4) se používá hlavně pokud nás zajímá pouze velikost hodnoty χ^2. Ze vzorce (3.3) lze ale lépe vyčíst, jaký vliv mají jednotlivé sčítance na celý součet. To se často hodí pro interpretaci.

3.3. χ^2 test

Pomocí veličiny χ^2 definované v (3.3), resp. upravené na tvar (3.4), můžeme testovat hypotézuH_0, že skutečné hodnoty pravděpodobnosti multinomického rozdělení jsou právě rovny číslům $p_1, \ldots, p_k$. Jakmile dostaneme $\chi^2 \geq \chi^2_{k-1}(\alpha)$, zamítneme hypotézu H_0. Tento tzv. Pearsonův χ^2test se může používat při ověřování pravidelnosti hrací kostky (tam by měla být pravděpodobnost rovna 1/6 pro každou ze šesti možností), při kontrole generátorů náhodných čísel a v řadě dalších případů. Některé i netradiční ukážeme později.

Je třeba mít na zřeteli, že test χ^2 je asymptotický, a proto ho lze doporučit jen při dostatečně velkém rozsahu výběru n. V literatuře se obvykle uvádí, že musí platit $np_i \geq 5$ pro každé $i = 1, \ldots, k$. Používá se však také Yarnoldovo kritérium. Podle toho stačí pouze když platí

$$np_i \geq 5q \quad \text{pro všechna} \quad i = 1, \ldots, k \quad \text{při} \quad k \geq 3,$$

kde q je podíl tříd, pro něž platí $np_i < 5$.

4. f-divergence a f-kvazinormy

Další možností odhadovaní rozdělení, případně testování podobností, jsou tzv. f-kvazinormy. Nyní si tedy ukážeme, co konkrétně jsou a jak vznikají. Ve zdrojích k této práci jsou pojmenovány pouze jako kvazinormy. Tento termín už je ale v matematice používán pro normu kde neplatí trojúhelníková nerovnost. Z tohoto důvodu zde budou přejmenovány na f-kvazinormy. Důvodem jejich uvedení je mimo jiné i způsob jejich zavedení, který se podobá vzniku Freeman - Tuckeyho uvedené v pozdějších kapitolách.

4.1. f-divergence

Chceme-li odhadovat rozdělení, je potřeba určitým způsobem číselně vyjádřit míru podobnosti (vzdálenosti) dvou hustot $\mathbf{p}, \mathbf{q}$. To lze například pomocí určité konvexní funkce $f : I(a,b) \to \mathbb{R}$. Pro takovouto míru se používá v literatuře název f-divergence. Více vlastností, včetně důkazů lze nalézt v [3] [6] [4].

Mějme funkci $f : I(a,b) \to \mathbb{R}^*$, která je spojitá na intervalu $(a,b) \subset I(a,b)$, spojitá zprava v bodě a a spojitá zleva v bodě b, pokud tyto body patří do intervalu $I(a,b)$. Pokud také existuje bod u_0 a $\lambda(u_0) \in \mathbb{R}$ tak, že $f(u) \geq f(u_0) + \lambda(u_0)(u - u_0)$ pro všechna $u \in I(a,b)$, $u \neq u_0$. Řekneme, že funkce f je **konvexní v bodě** $u_0 \in (a,b)$. Pokud výše uvedenou neostrou nerovnost nahradíme ostrou nerovností, řekneme, že funkce f je **striktně konvexní v bodě** u_0. Pokud je funkce f konvexní, resp. striktně konvexní, v každém bodě $u \in I(a,b)$, řekneme, že je **konvexní**, resp. **striktně konvexní**.

Nyní mějme funkci $f(u)$, která je konvexní na intervalu $(0,\infty)$, striktně konvexní v $u = 1$ a $f(1) = 0$. Pak funkcionál

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

kde klademe

$$0f\left(\frac{0}{0}\right) = 0 \quad a \quad 0f\left(\frac{p}{0}\right) = pf(*), \qquad \text{přičemž } f(*) = \lim_{u\to\infty} \frac{f(u)}{u} \in \mathbb{R}^*,$$

nazveme **f-divergencí** pravděpodobnostních modelů $(\Omega,\mathbf{p}), (\Omega,\mathbf{q})$, resp. hustot $\mathbf{p}$ a $\mathbf{q}$ na $(\Omega,\sum,P)$. Říkáme také, že funkce f generuje danou divergenci.

Věta 4.1. *Pro libovolnou f-divergenci platí nerovnost*

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

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

$f(u)$	název	$D_f(\mathbf{p},\mathbf{q})$
$u \ln u$	Shannonova divergence	$\sum_x p(x) \ln \frac{p(x)}{q(x)}$
$(u^{1/2}-1)^2$	Hellingerova divergence	$\left(1-\sum_x (p(x)q(x))\right)^{1/2}$
$(u-1)^2$	χ^2-divergence	$\sum_x \frac{(p(x)-q(x))^2}{q(x)}$

Tabulka 4.1: Nejčastější typy f-divergencí

4.2. f-kvazinormy

Před zavedením samotných f-kvazinorem je potřeba uvést následující větu.

Věta 4.2. *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 $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{q} = (q_1, \ldots, q_k)$, $k > 1$, z tohoto prostoru. Označme*

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

kde $S = \left\{ \mathbf{p} \in \mathbb{R}^k : \forall p_j \geq 0, \sum_{j=1}^{k} p_j = 1 \right\}$, 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 $\langle 0, 1\rangle$ a funkce f má spojitou druhou derivaci v $(0,\infty)$, pak $V(\mathbf{p},\mathbf{q})$ nabývá absolutního minima na S v rozdělení pravděpodobnosti

$$\mathbf{p}_0 = \left(\frac{1}{k}, \ldots, \frac{1}{k}\right).$$

Tuto větu si nyní dokážeme.
Nejprve si označíme

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

Lagrangeovu funkci pro vázaný extrém funkce $V(\mathbf{q})$ za podmínky $\sum_{j=1}^{m} q_j = 1$. Funkce f je konvexní (je to f- divergence - z definice), 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

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

f a odtud funkce $G(q_j)$ je rostoucí na $\langle 0,1\rangle$. Protože

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

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

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

$j = 1,\ldots,k$ a $i = 1,\ldots,k$. Z toho plyne, že Hessova matice funkce $V(\mathbf{q})$ je diagonální a pozitivně definitní. Jacobiho matice pro jedinou podmínku $\sum_{j=1}^{k} q_j = 1$ má hodnost 1, takže $V(\mathbf{q})$ má za dané podmínky v $\mathbf{q} = \mathbf{p}_0 = \left(\frac{1}{k},\ldots,\frac{1}{k}\right)$ absolutní minimum.

Nyní si vezměme dvě diskrétní rozdělení pravděpodobnosti na pravděpodobnostním prostoru (Ω,Σ,P) a to $\mathbf{p} = (p_1,\ldots,p_k)$ a $\mathbf{p}_0 = (\frac{1}{k},\ldots,\frac{1}{k}), k > 1$. Dále f-divergenci na stejném pravděpodobnostním prostoru. **f-kvazinormou** pak rozumíme f-divergenci ve tvaru

$$D_f(\mathbf{p},\mathbf{p}_0)$$

Diskrétní rovnoměrné rozdělení pravděpodobnosti se používá proto, že je nejvíce neurčité a minimalizuje $V(\mathbf{q})$ aneb integrál všech f-divergencí rozdělení $\mathbf{p}$ od pevně zvoleného rozdělení $\mathbf{q}$. Navíc má také maximální entropii. [15]

Věta 4.3. *f-kvazinormy mají následující vlastnosti*

a) $D_f(\mathbf{p},\mathbf{p}_0) = \frac{1}{k}\sum_{j=1}^{k} f(kp_j)$,

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*,* $j = 1,\ldots,m$,

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

Důkaz lze nalézt v [4].

4.3. Odhad diskrétního rozdělení pravděpodobnosti

Předpokládáme, že pozorovaná diskrétní náhodná veličina X na pravděpodobnostním prostoru (Ω,Σ,P),jejíž rozdělení pravděpodobnosti chceme odhadnout (fitovat), nabývá

nejvýše konečně mnoha různých hodnot $x_j^* \in \mathbb{R}$, (tj. $\Omega = \{x_1^*, \ldots, x_k^*\} \subset \mathbb{R}$) s neznámými pravděpodobnostmi

$$p_j = P(X = x_j^*), \quad j = 1, \ldots, k, \quad k > 1.$$

Pozorováním náhodné veličiny X získáme statistický soubor $(x_1, \ldots, x_n)$ a jeho roztříděním dostaneme statistický soubor

$$\left((x_1^*, \frac{f_1}{n}), \ldots, (x_k^*, \frac{f_k}{n}) \right),$$

kde f_j je absolutní četnost pozorované hodnoty x_j^*. Dále předpokládáme $n > k > 1$ a $f_j > 0$ pro všechna $j = 1, \ldots, k$ (jestliže $f_j = 0$, pak j-tou třídu vynecháme). Pro odhad rozdělení $\mathbf{p}$ požadujeme, aby toto rozdělení navíc splňovalo nějaké zadané podmínky, jejichž počet je $K \geq 1$. Mezi tyto podmínky nezařazujeme zřejmou podmínku $\sum_{j=1}^{k} p_j = 1$, ale další např. momentové podmínky. Hledáme pak takové rozdělení $\mathbf{p}$, které má minimální f-kvazinormu $D_f(\mathbf{p}, \mathbf{p}_0)$.

Rozdělení pravděpodobnosti $\mathbf{p} = (p_1, \ldots, p_k$ pozorované diskrétní náhodné veličiny X má na pravděpodobnostním prostoru (Ω, Σ, P), kde $\Omega = \{x_1^*, \ldots, x_k^*\}, k > 1$ a Σ je množina všech podmnožin Ω, minimální f-kvazinormu $D_f(\mathbf{p}, \mathbf{p}_0)$ za K počátečních momentových podmínek

$$\sum_{j=1}^{k} p_j (x_j^*)^l = M_l, \quad l = 0, \ldots, K,$$

jestliže jeho f-kvazinorma $D_f(\mathbf{p}, \mathbf{p}_0)$ nabývá za těchto podmínek minimální hodnotu.

Pro $K < k-1$ obdržíme odhady pravděpodobností $p_j(\boldsymbol{\lambda})$ pomocí Lagrangeovy funkce

$$\Lambda(\mathbf{p}, \boldsymbol{\lambda}) = D_f(\mathbf{p}, \mathbf{p}_0) + \sum_{l=1}^{K} \lambda_l \left(\sum_{j=1}^{k} p_j (x_j^*)^l - M_l \right).$$

Lagrangeovy multiplikátory λ_l je možno určit pomocí nelineární soustavy rovnic odpovídající nulovému gradientu Lagrangeovy funkce, anebo přímo aplikovat některou metodu nelineární optimalizace pro určení jejího minima.

Má- li pozorovaná náhodná veličina X empirické rozdělení $\mathbf{f} = (\frac{f_1}{n}, \ldots, \frac{f_k}{n})$, pak statistika

$$\chi^2(\mathbf{f}, \mathbf{p}(\boldsymbol{\lambda})) = \sum_{j=1}^{k} \frac{(f_j - np_j(\boldsymbol{\lambda}))^2}{np_j(\boldsymbol{\lambda})} = \frac{1}{n} \sum_{j=1}^{k} \frac{f_j^2}{p_j(\boldsymbol{\lambda})} - n$$

má pro $n \to \infty$ asymptoticky rozdělení chí-kvadrát s $k - K - 1$ stupni volnosti. Asymptotickou vlastnost můžeme použít k testování vhodnosti nalezeného rozdělení pravděpodobnosti $\mathbf{p}(\boldsymbol{\lambda}) = (p_1(\boldsymbol{\lambda}), \ldots, p_k(\boldsymbol{\lambda}))$. Pro praktické využití požadujeme, aby bylo $np_j(\boldsymbol{\lambda}) > 5$ pro všechna $j = 1, \ldots, k$.

Postupným přidáváním momentových podmínek a opakovaným odhadem rozdělení pravděpodobnosti pomocí minimální pravděpodobnosti pomocí minimální f-kvazinormy lze určit minimální potřebný počet K těchto podmínek tak, aby platilo $\chi^2(\mathbf{f}, \mathbf{p}(\boldsymbol{\lambda})) \leq \chi^2_{n-1}$, kde χ^2_{n-1} je $(1-\alpha)$ kvantil rozdělení chí-kvadrát s daným počtem stupňů volnosti pro hladinu významnosti α.

4.4. Cressie-Readovy kvazinormy

V této kapitole si ukážeme a odvodíme celou třídu f-kvazinorem (a s nimi související f-divergence.) za pomocí tzv. Cressie-Readovy statistiky.

Věta 4.4. *Jestliže $\lambda \in \mathbb{R}$, pak Cressie-Readova statistika*

$$2nI^{\lambda} = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{k} f_j \left\{ \left(\frac{f_j}{np_j} \right)^{\lambda} - 1 \right\}, \tag{4.1}$$

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

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

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

má asymptotický chí-kvadrát rozdělení s $k-1$ stupni volnosti pro všechna λ.

Nyní uvedeme důkaz, ale pouze pro $\lambda \neq 0, \lambda \neq -1$. Pro tyto dva speciální případy lze nalézt důkaz např. v [4].

Výraz (4.1) přepíšeme do tvaru

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

$\frac{f_j - np_j}{np_j}$ označíme jako V_j a $(1+V_j)^{\lambda+1}$ rozvineme v Taylorovu řadu. Dostaneme

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

$$= \frac{2n}{\lambda} \sum_{j=1}^{k} p_j q \left(V_j + \frac{\lambda V_j^2}{2} \right) + o_p(1) = \sum_{j=1}^{k} p_j (n^{1/2} V_j)^2 + o_p(1),$$

přičemž $o_p(1)$ konverguje k nule pro $n \to \infty$. Pak

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

Z třídy Cressie-Readových statistik lze odvodit vzhledem k λ třídu f- divergencí a f-kvazinorem vedoucí na asymptotický chí-kvadrát test následovně.

Výraz (4.1) lze upravit a přeznačit na následující výraz

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

Pro $\lambda = 0$ a $\lambda = -1$ analogicky.

Jedná se o f-divergenci s následující generující funkcí

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

Nyní lze vytvořit **Creesie-Readovu** f-kvazinormu

$$CR_f(\mathbf{p}, \mathbf{p}_0) = \frac{1}{\lambda(\lambda+1)k} \sum_{j=i}^{k} \left\{ \left(\frac{1}{kp_j}\right)^{\lambda} - 1 \right\} \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\}$$

Speciálně pro $\lambda = 0$

$$CR_f(\mathbf{p}, \mathbf{p}_0) = \frac{1}{k} \sum_{j=1}^{k} \ln \frac{1}{k} - \ln p_j$$

a pro $\lambda = -1$

$$CR_f(\mathbf{p}, \mathbf{p}_0) = \sum_{j=1}^{k} p_j \ln p_j + \ln k.$$

Z Cressie-Readových f-kvazinorem (f-divergencí) dostaneme pro $\lambda = 1$ Pearsonovu, pro $\lambda = -\frac{1}{2}$ Hellingerovu a pro $\lambda = -1$ Shannonovu f-kvazinormu (f-divergenci).

Pro duální f-divergenci platí $D_F(\mathbf{p}, \mathbf{q}) = D_f(\mathbf{q}, \mathbf{p})$

Věta 4.5. *Cressie-Readova f-kvazinorma tvoří uzavřenou třídu kvazinorem vzhledem k dualitě.*

Důkaz: Duální Cressie-Readovu f-kvazinormu

$$CR_F(\mathbf{p}, \mathbf{p}_0) = \frac{1}{\lambda(\lambda+1)k} \sum_{j=1}^{k} \left\{ (kp_j)^{\lambda+1} - 1 \right\} \text{ pro } \lambda \in \mathbb{R} - \{-1, 0\}$$

dostaneme z

$$CR_f(\mathbf{p}, \mathbf{p}_0) = \frac{1}{\lambda(\lambda+1)k} \sum j = 1^{k} \left\{ \left(\frac{1}{kp_j}\right)^{\lambda} - 1 \right\}$$

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 f-kvazinorma pro $\lambda = -1$.

5. Hledání minima f-kvazinorem

Při odhadování rozdělení, ať už pomocí f-kvazinorem či jiného funkcionálu posuzujícím podobnost, vždy minimalizujeme danou funkci. To se provádí téměř vždy numericky. Z tohoto důvodu si zde uvedeme základní podmínky pro řešení nelineární úlohy s podmínkami ať už rovnosti, či nerovnosti. Dále si také uvedeme numerickou metodu, která bude použita pro hledání minima v příkladech v dalších kapitolách. Zde je čerpáno hlavně z [5] a [8], kde lze nalézt i podrobnější informace včetně některých důkazů.

5.1. Vázané extrémy a podmínky optimality

Dle [8] dostatečná i nutná podmínka, že bod $\mathbf{x}_0$ je globálním minimem diferencovatelné konvexní funkce f na konvexní množině C, je nerovnice

$$\nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x_0}) \geq \mathbf{0}, \forall \mathbf{x} \in \mathbf{C}.$$

Nyní si ukážeme další nutné podmínky existence extrémů, jelikož tato nezahrnuje možné vazby.

Symbolem $\nabla \mathbf{g}(\mathbf{x})^T$ budeme značit transponovanou Jacobiho matici mající n řádků a m sloupců s prvky $\frac{\partial g_i(\mathbf{x})}{\partial x_j}$ v j-tém řádku a i-tém sloupci. Tedy matice $\nabla \mathbf{g}(\mathbf{x})^T$ má ve svých funkcích gradienty funkcí $g_i(\mathbf{x})$. Podobně budeme značit $\nabla \mathbf{h}(\mathbf{x})^T$.

Věta 5.1 (**Lagrange**). *Nechť $\mathbf{x}_0$ je přípustné řešení minimalizační úlohy s omezení ve tvaru rovnic $\min\{f(\mathbf{x})|\mathbf{h}(\mathbf{x}) = \mathbf{0}\}$. Nechť funkce f a $\mathbf{h}$ mají spojité parciální derivace 1. řádu v okolí bodu $\mathbf{x}_0$. Nechť sloupce matice $\nabla \mathbf{h}(\mathbf{x})^T$ jsou lineárně nezávislé (podmínka regularity). Je-li $\mathbf{x}_0$ bodem lokálního minima řešené úlohy, potom existuje vektor $\mathbf{v}$ takový, že platí*

$$\nabla f(\mathbf{x}_0) + \nabla \mathbf{h}(\mathbf{x})^T \mathbf{v} = \mathbf{0}.$$

Věta 5.1 nám určuje nutné podmínky existence extrému. Pro samotný výpočet se používá tzv. **Lagrangeova funkce (lagrangian)**

$$\mathbf{L}(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{h}(\mathbf{x})^T \mathbf{v} = f(\mathbf{x}) + \sum_{i=1}^{m} v_i h_i(\mathbf{x}).$$

Přičemž hledáme stacionární body Lagrangianu, splňující podmínku $\nabla L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$, tj. $\nabla_x L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$ a $\nabla_v L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$.

Věta 5.2 (**Karush-Kuhn-Tucker**). *Nechť funkce f a $\mathbf{g}$ jsou diferencovatelné v bodě $\mathbf{x}_0$, který je bodem lokálního minima úlohy $\min\{f(\mathbf{x})|\mathbf{g}(\mathbf{x}) \leq 0\}$. Nechť dále platí, že sloupce matice $\nabla \mathbf{g}(\mathbf{x})^T$, odpovídající gradientům aktivních omezení jsou lineárně nezávislé. Potom existují koeficienty $\mathbf{u}$ tak, že platí:*

$$\nabla f(\mathbf{x}) + \nabla \mathbf{g}(\mathbf{x})^T \mathbf{u} = \mathbf{0}, \qquad \mathbf{u}^T \mathbf{g}(\mathbf{x}_0) = 0, \qquad \mathbf{u} \geq \mathbf{0}.$$

Věty 5.1 i 5.2 jsou obě použitelné pouze na $\mathbf{x} \in \mathbb{R}^n$, resp na konvexní množině. Pokud bychom ovšem uvažovali obecnější případ, kde by bylo řešení $\mathbf{x}$ na otevřené neprázdé množině, daly by se obě zobecnit na následující větu.

Věta 5.3 (**Nutné podmínky 1. řádu**). *Uvažujme následující základní předpoklady.*

1. *Je dána $X \subset \mathbb{R}^n$ neprázdná otevřená množina. Nechť jsou dány funkce $f : \mathbb{R}^n \to \mathbb{R}, \mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ a $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^l$. Složky $\mathbf{g}$ značíme g_i a složky $\mathbf{h}$ značíme h_i.*
2. *Řešíme úlohu $\min\{f(\mathbf{x}) | \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{h}(\mathbf{x}) = \mathbf{0}, \mathbf{x} \in X\}$.*
3. *Mějme $\mathbf{x}_0$ přípustné řešení úlohy, označme $I = \{i | g_i(\mathbf{x}_0 = 0\}$ množinu indexů aktivních omezení tvaru nerovnic.*
4. *Nechť f a g_i pro $i \in I$ jsou diferencovatelné v $\mathbf{x}_0$, dále g_i jsou spojité pro $i \notin I$ v $\mathbf{x}_0$ a také h_i jsou spojité diferencovatelné pro $i = 1, \ldots, l$ v $\mathbf{x}_0$.*
5. *Dále předpokládáme, že gradienty $\nabla g_i(\mathbf{x}_0)$ pro $i \in I$ a $\nabla h_i(\mathbf{x}_0)$ pro $i = 1, \ldots, l$ jsou lineárně nezávislé (podmínka regularity).*

Jestliže $\mathbf{x}_0$ je lokální minimum, potom existují koeficienty u_i pro $i \in I$ a v_i pro $i = 1, \ldots, l$ takové, že platí KKT podmínky:

$$\nabla f(\mathbf{x}_0) + \sum_{i \in I} u_i \nabla g_i(\mathbf{x}_0) + \sum_{i=1}^{l} v_i \nabla h_i(\mathbf{x}_0) = \mathbf{0}, \quad u_i \geq 0 \quad \text{pro } i \in I.$$

5.2. Kvazinewtonovské metody

Kvazinewtonovské metody, někdy též nazývané metody proměnné metriky, využívají tzv. sdružených směrů. jsou založeny na myšlence vyjádřit spádový směr ve tvaru $\mathbf{d}_j = -\mathbf{D}_j \nabla f(\mathbf{x})$, kde D_j je symetrická pozitivně definitní matice, aproximující inverzní matici $\mathbf{H}^{-1}(\mathbf{x})$ ($\mathbf{H}(\mathbf{x})$ je tzv. Hessova matice druhých parciálních derivací). Metody konvergují za značně obecných předpokladů a řád konvergence je superlineární.

Nyní si pro představu uvedeme jednu z možností jak volit spádové směry i když to pravděpodobně není konkrétně ta, kterou používá Microsoft EXCEL.

Davidon-Fletcher-Powellův algoritmus: Zvolíme $\varepsilon > 0$, počáteční bod $\mathbf{x}_1$, symetrickou, pozitivně definitní $\mathbf{D}_1$. Dále $\mathbf{y}_1 := \mathbf{x}_1, j := 1, k := 1$.

1. Je-li $||\nabla f(\mathbf{y}_j)|| < \varepsilon$ potom STOP, jinak $\mathbf{d}_j := -\mathbf{D}_j \nabla f(\mathbf{y}_j)$ a řešíme $\min\{f(\mathbf{y}_j + \lambda \mathbf{d}_j) | \lambda \geq 0\}$. Řešení λ_j použijeme pro $\mathbf{y}_{j+1} := \mathbf{y}_j + \lambda_j \mathbf{d}_j$. Je-li $j = n$, tak $\mathbf{x}_{k+1} := \mathbf{y}_{n+1}, \mathbf{y_1} := \mathbf{x}_{k+1}, k := k + 1$, $j := 1$ a GOTO 1. V opačném případě když $j < n$, pokračujeme dále.
2. Přiřadíme $\mathbf{D}_{j+1} := \mathbf{D}_j + \mathbf{C}_j^{DFP}$, kde

$$\mathbf{C}_j^{DFP} = \frac{\mathbf{p}_j \mathbf{p}_j^T}{\mathbf{p}_j^T \mathbf{q}_j} - \frac{\mathbf{D}_j \mathbf{q}_j \mathbf{q}_j^T \mathbf{D}_j}{\mathbf{q}_j^T \mathbf{D}_j \mathbf{q}_j}$$

$$\mathbf{p}_j = \lambda_j \mathbf{d}_j = \mathbf{y}_{j+1} - \mathbf{y}_j, \quad \mathbf{q}_j = \nabla f(\mathbf{y}_{j+1}) - \nabla f(\mathbf{y}_j)$$

dále $j := j + 1$ a GOTO 1.

6. F-T pseudokvazinorma

6.1. Testovací kritérium

V knize [16] je uveden málo známý test shody diskrétního rozdělení pravděpodobnosti, jehož autory jsou Freeman a Tukey. Testovací kritérium má tvar

$$T = \sum_{i=1}^{k} \left(\sqrt{O_i} + \sqrt{O_i + 1} - \sqrt{4E_i + 1} \right)^2, \tag{6.1}$$

kde O_i značí pozorovanou četnost náhodného jevu A_i s pravděpodobností p_i pro $i = 1, \ldots, k$. E_i pak značí očekávanou četnost.

Jestliže jde o shodu pravděpodobnostního diskrétního rozdělení, pak T má asymptotické rozdělení chí-kvadrát s $k - s - 1$ stupni volnosti, kde s je počet odhadovaných parametrů daného rozdělení. V uvedené knize je odkaz na článek [17], ve krém však důkaz o vlastnostech kritéria T není, pouze je naznačeno, jaké vlastnosti má transformace $\sqrt{X} + \sqrt{X + 1}$ náhodné veličiny X s Poissonovým rozdělením z hlediska stabilizace rozptylu. Autoři se v článku zabývají stabilizací rozptylu Poissonova rozdělení různými transformacemi.

$$Y = \sqrt{X}$$
$$Y = \sqrt{X + 1}$$
$$Y = \sqrt{X + \frac{3}{8}}$$
$$Y = \sqrt{X + \frac{1}{2}}$$

a po výpočtech v závislosti na velikosti náhodného výběru a chytré aproximace, dospívají k názoru, že nejlépe stabilizuje rozptyl transformace

$$Y = \sqrt{X} + \sqrt{X + 1}.$$

Dále však informace o odvození kritéria T nejsou. Tudíž se naskýtá otázka členu kritéria ve tvaru $\sqrt{4E_i + 1}$. Vzhledem k vlastnostem absolutních četností [1] platí, že $O_i \to \infty$, $E_i \to \infty$ pro rozsah výběru $n \to \infty$, jinak (s pravděpodobností 0) $\frac{O_i}{n} \to 0$, $\frac{E_i}{n} \to 0$.

Jestliže hypotéza o tvaru pozorovaného multinomického rozdělení pravděpodobnosti $M(n, p_1, \ldots, p_k)$, $p_i \geq 0$, $p_1 + \cdots p_k = 1$ platí, pak $p_i = \frac{E_i}{n}$ a s pravděpodobností 1 je

$$\lim \frac{O_i}{n} = p_i$$

navíc $\frac{O_i}{n}$ je nestranný odhad pravděpodobnosti p_i.
Protože

$$\lim_{x \to \infty} \frac{\sqrt{x} + \sqrt{x + 1}}{\sqrt{4x + 1}} = \lim_{x \to \infty} \frac{\sqrt{x}(1 + \sqrt{1 + \frac{1}{x}})}{\sqrt{x}\sqrt{4 + \frac{1}{x}}} = 1.$$

Odtud pak pro velká n platí

$$\sqrt{O_i} + \sqrt{O_i + 1} \sim \sqrt{4O_i + 1}. \tag{6.2}$$

Věta 6.1. *Freemanova - Tukeyho statistika* (6.1) *má asymptotické chí-kvadrát rozdělení s* $k - s - 1$ *stupni volnosti.*

Z (6.2) pro veliká n platí

$$\left(\sqrt{O_i} + \sqrt{O_i + 1} - \sqrt{4E_i + 1}\right)^2 \sim \left(\sqrt{4O_i + 1} - \sqrt{4_i + 1}\right)^2 =$$
$$= \left(2\sqrt{O_i}\sqrt{1 + \frac{1}{4O_i}} - 2\sqrt{E_i}\sqrt{1 + \frac{1}{4E_i}}\right)^2 \sim 4\left(\sqrt{O_i} - \sqrt{E_i}\right)^2$$

Statistika

$$\sum_{i=1}^{k} 4\left(\sqrt{O_i} - \sqrt{E_i}\right)^2$$

je pak Hellingerova. Patří mezi Cressie-Readovy statistiky a dle věty 4.4 má asymptotické chí-kvadrát rozdělení. V [10]lze také nalézt přímý důkaz, že Hellingerova statistika má asymptotické rozdělení chí-kvadrát s $k - s - 1$ stupni volnosti.

6.2. Pseudokvazinorma

Nyní si opět vezměme původní tvar kritéria, který ale upravíme na tvar obsahující jednotlivé pravděpodobnosti p_i a q_i. Dostaneme divergence

$$D_1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \left(\sqrt{np_i} + \sqrt{np_i + 1} - \sqrt{4nq_i + 1}\right)^2 \tag{6.3}$$

$$D_2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \left(\sqrt{nq_i} + \sqrt{nq_i + 1} - \sqrt{4np_i + 1}\right)^2 \tag{6.4}$$

Divergence D_1 a D_2 jsou jedna k druhé duální, jelikož platí $D_2(\mathbf{p}, \mathbf{q}) = D_1(\mathbf{q}, \mathbf{p})$.

Všimněme si také podobnosti mezi výše uvedenými výrazy a f-divergencemi. Oboje popisuje „vzdálenost“ (neboli divergenci) mezi dvěma hustotami rozdělení, ale zde není generována funkcí f.

Nyní si uvedeme větu analogickou k větě 4.2.

Věta 6.2. *Nechť* (Ω, Σ, P) *je konečný pravděpodobnostní prostor, kde* P *je libovolná pravděpodobnostní míra na* Ω *a* D_1 *je divergence uvedená v* (6.3). *Přičemž* $\mathbf{p}$ *a* $\mathbf{q}$ *jsou z tohoto pravděpodobnostního prostoru a* $k > 1$. *Označme*

$$V(\mathbf{q}) = \int_S D_1(\mathbf{p}, \mathbf{q}) dS,$$

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*

$$G(q_j) = \frac{\partial V(\mathbf{q})}{\partial q_j} = \int\limits_S \frac{\partial D_1(\mathbf{p}, \mathbf{q})}{\partial q_j} dS,$$

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

$$\mathbf{p}_0 = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right).$$

Důkaz: Opět využijeme Lagrangeovu funkci pro vázaný extrém funkce $V(\mathbf{q})$ za podmínky $\sum_{j=1}^{k} q_j = 1$

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

Tu zderivujeme

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

$$\begin{aligned}
G(q_j) &= \int\limits_S \frac{\partial}{\partial q_j} \left(\sum_{j=1}^{k} (\sqrt{np_j} + \sqrt{np_j + 1} - \sqrt{4nq_j + 1})^2\right) dS \\
&= \int\limits_S \frac{\partial}{\partial q_j} \left(\sqrt{np_j} + \sqrt{np_j + 1} - \sqrt{4nq_j + 1}\right)^2 dS \\
&= 2n \int\limits_S \left(\sqrt{p_j} + \sqrt{p_j + \frac{1}{n}} - \sqrt{4q_j + \frac{1}{n}}\right) \left(\frac{4}{2\sqrt{4q_j + \frac{1}{n}}}\right) \\
&= 4n \int\limits_S \left(-\frac{\sqrt{p_j} + \sqrt{p_j + \frac{1}{n}}}{\sqrt{4q_j + \frac{1}{n}}} + 1\right) dS \\
&= 4n \left(-\frac{a_1}{\sqrt{4q_j + \frac{1}{n}}} + b_1\right),
\end{aligned}$$

kde

$$\begin{aligned}
a_1 &= \int\limits_S \left(\sqrt{p_j} + \sqrt{p_j + \frac{1}{n}}\right) dS, \quad a_1 \text{ existuje a } a_1 > 0, \\
b_1 &= \int\limits_S dS, \quad b_1 \text{ existuje a } b_1 = \frac{1}{(k-1)!}.
\end{aligned}$$

Takže

$$\frac{\partial \Lambda}{\partial q_j} = 4n \left(-\frac{a_1}{\sqrt{4q_j + \frac{1}{n}}} + b_1 \right) + \lambda.$$

Nyní položíme $\frac{\partial \Lambda}{\partial q_j}$ rovno 0 pro nalezení bodu podezřelého z extrému, tedy

$$\begin{aligned}
4n \left(-\frac{a_1}{\sqrt{4q_j + \frac{1}{n}}} + b_1 \right) + \lambda =& 0, \\
-\frac{4na_1}{\sqrt{4q_j + \frac{1}{n}}} =& -4nb_1 - \lambda, \\
\sqrt{4q_j + \frac{1}{n}} =& \frac{4na_1}{4nb_1 + \lambda}, \\
4q_j =& \left(\frac{4na_1}{4nb_1 + \lambda} \right)^2 - \frac{1}{n}, \\
q_j =& \left(\frac{4na_1}{4nb_1 + \lambda} \right)^2 - \frac{1}{4n}.
\end{aligned}$$

Jelikož $\sum_{j=1}^{k} = 1$ platí, že $q_j = \frac{1}{k}$. Funkce $G(q_j)$ je rostoucí na [0,1], takže $q_j = \frac{1}{k}$ je jediný kořen rovnice $G(q_j) + \lambda = 0$.

Dále je

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

přičemž $j, k = 1, \ldots, k$. Dále je

$$\begin{aligned}
G'(q_j) &= \frac{\partial}{\partial q_j} \left[4n \left(-\frac{a_1}{\sqrt{4q_j + \frac{1}{n}}} + b_1 \right) \right] \\
&= 4n \frac{2a_1}{\sqrt{4q_j + \frac{1}{n}}} \\
&> 0.
\end{aligned}$$

Z toho plyne že Hessova matice Lagrangeovy funkce je diagonální a pozitivně definitní. Jacobiho matice pro podmínku $\sum_{j=1}^{k} = 1$ má hodnost 1. Odtud plyne, že $V(\mathbf{q})$ má za dané podmínky v $\mathbf{q} = \mathbf{p}_0 = \left(\frac{1}{k}, \ldots, \frac{1}{k} \right)$ absolutní minimum.

Stejný postup nyní aplikujme na duální formuli.

$$D_2(\mathbf{p},\mathbf{q}) = D_1(\mathbf{q},\mathbf{p}) = \sum_{j=1}^{k} \left(\sqrt{nq_j} + \sqrt{nq_j+1} - \sqrt{4np_j+1}\right)^2,$$

$$V(\mathbf{q}) = \int_S D_2(\mathbf{p},\mathbf{q})dS,$$

$$G(q_j) = \frac{\partial V(\mathbf{q}}{\partial q_j} = \int_S \frac{\partial D_2(\mathbf{p},\mathbf{q})}{\partial q_j} dS,$$

kde S je stejná oblast jako u D_1.

Nyní opět zaveďme Lagrangian

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

Zderivujeme

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

$$\begin{aligned}
G(q_j) &= \int_S \frac{\partial}{\partial q_j}\left(\sum_{j=1}^{k}(\sqrt{nq_j} + \sqrt{nq_j+1} - \sqrt{4np_j+1})^2\right) dS \\
&= \int_S \left(\sqrt{nq_j} + \sqrt{nq_j+1} - \sqrt{4np_j+1}\right)^2 dS \\
&= 2n\int_S \left(\sqrt{q_j} + \sqrt{q_j+\frac{1}{n}} - \sqrt{4p_j+\frac{1}{n}}\right)\left(\frac{1}{2\sqrt{q_j}} + \frac{1}{2\sqrt{q_j+\frac{1}{n}}}\right) dS \\
&= n\left(\frac{1}{\sqrt{q_j}} + \frac{1}{\sqrt{q_j+\frac{1}{n}}}\right)\left[\left(\sqrt{q_j} + \sqrt{q_j+\frac{1}{n}}\right) b_2 - a_2\right],
\end{aligned}$$

kde

$$a_2 = \int_S \sqrt{4p_j+\frac{1}{n}} dS, \quad a_2 > 0,$$

$$b_2 = \int_S dS, \quad b_2 = \frac{1}{(k-1)!}\ .$$

Nyní opět položíme $\frac{\partial \Lambda}{\partial q_j}$ rovno 0, takže

$$n\left(\frac{1}{\sqrt{q_j}}+\frac{1}{\sqrt{q_j+\frac{1}{n}}}\right)\left[\left(\sqrt{q_j}+\sqrt{q_j+\frac{1}{n}}\right)b_2-a_2\right]+\lambda=0,$$

$$n\left(\frac{1}{\sqrt{q_j}}+\frac{1}{\sqrt{q_j+\frac{1}{n}}}\right)\left[\left(\sqrt{q_j}+\sqrt{q_j+\frac{1}{n}}\right)b_2-a_2+\frac{\lambda}{n}\left(\frac{1}{\sqrt{q_j}}+\frac{1}{\sqrt{q_j+\frac{1}{n}}}\right)^{-1}\right]=0,$$

$$\left(\sqrt{q_j}+\sqrt{q_j+\frac{1}{n}}\right)b_2-a_2+\frac{\lambda}{n}\left(\frac{1}{\sqrt{q_j}}+\frac{1}{\sqrt{q_j+\frac{1}{n}}}\right)^{-1}=0.$$

Levá strana rovnice je rostoucí, tudíž existuje nejvíce jeden kořen. Navíc také platí, že $\sqrt{4p_j+1/n}>1$. Takže záporná hodnota existuje. Zda-li existuje i kladná ovšem závisí na velikosti λ, které nevíme jak je veliké. Pokud by i kladná hodnota existovala, existoval by i kořen. Minimem by pak pravděpodobně bylo opět $q_j=\frac{1}{k}$, což potvrzují i experimentální výsledky.

Nyní obdobně jako u tvorby f-kvazinorem vytvoříme pseudokvazinormu a to položením $\mathbf{q}$ rovno rovnoměrnému rozdělení pravděpodobnosti $\mathbf{p}_0=(\frac{1}{k},\ldots,\frac{1}{k})$, které minimalizuje $V(\mathbf{q})$ aneb integrál divergence rozdělení $\mathbf{p}$ od pevně zvoleného $\mathbf{q}$. Dostáváme tedy pseudokvazinormy

$$T_1=D_1(\mathbf{p},\mathbf{p}_0)=\sum_{i=1}^{k}\left(\sqrt{np_i}+\sqrt{np_i+1}-\sqrt{4\frac{n}{k}+1}\right)^2, \tag{6.5}$$

$$T_2=D_2(\mathbf{p},\mathbf{p}_0)=\sum_{i=1}^{k}\left(\sqrt{\frac{n}{k}}+\sqrt{\frac{n}{k}+1}-\sqrt{4nq_i+1}\right)^2. \tag{6.6}$$

Nyní je vhodné položit si otázku, zda - li je pseudokvazinorma nulová pro $\mathbf{p}=\mathbf{p}_0$ stejně jako f-kvazinormy. Po dosazení dostaneme vzdálenost

$$T_1=T_2=\sum_{i=1}^{k}\left(\sqrt{\frac{n}{k}}+\sqrt{\frac{n}{k}+1}-\sqrt{4\frac{n}{k}+1}\right)^2.$$

Z tabulky vidíme, že vzdálenost nulová není. Hodnota závisí na n a k. Zvětšením n hodnota klesá, naopak zvětšením k se hodnota pseudokvazinormy zvyšuje.

Obrázky 6.1 a 6.2 ukazují grafy izočar (křivky spojující body v prostoru se stejnou hodnotou) pseudokvazinormy T_1 a duálního tvaru T_2 pro $n=10$ a $k=3$ to jest $p_1\in\langle 0,1\rangle$, $p_2\in\langle 0,1-p_1\rangle$ a $p_3=1-p_2-p_1$. Grafy pak odpovídají rovnicím

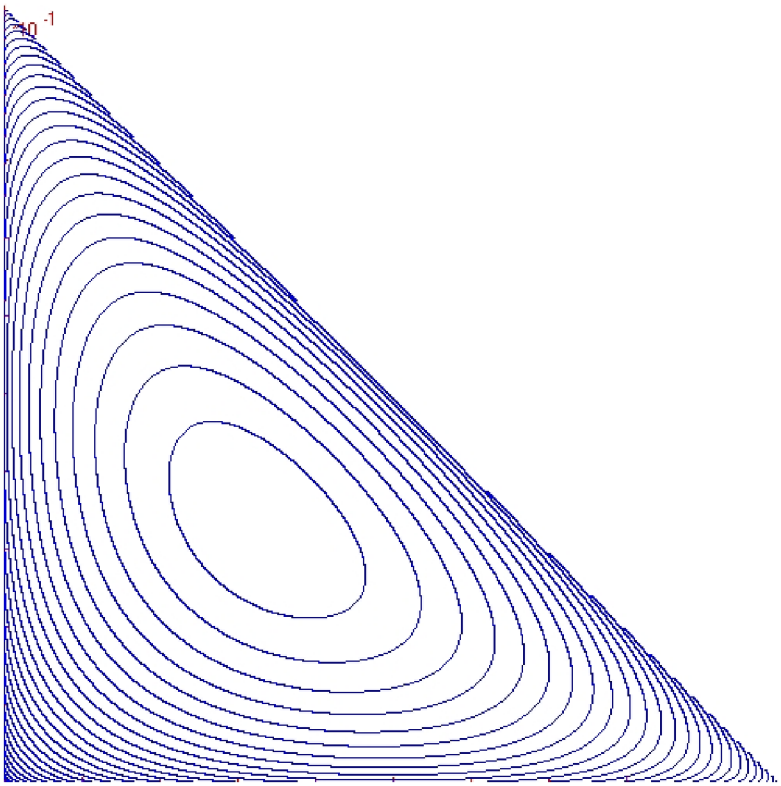
$$\sum_{j=1}^{3}\left(\sqrt{10p_j}+\sqrt{10p_j+1}-\sqrt{\frac{40}{3}+1}\right)^2=c,\ \text{pro obrázek 6.1},$$

$$\sum_{j=1}^{3}\left(\sqrt{\frac{10}{3}}+\sqrt{\frac{10}{3}+1}-\sqrt{40p_j}\right)^2=c,\ \text{pro obrázek 6.2},$$

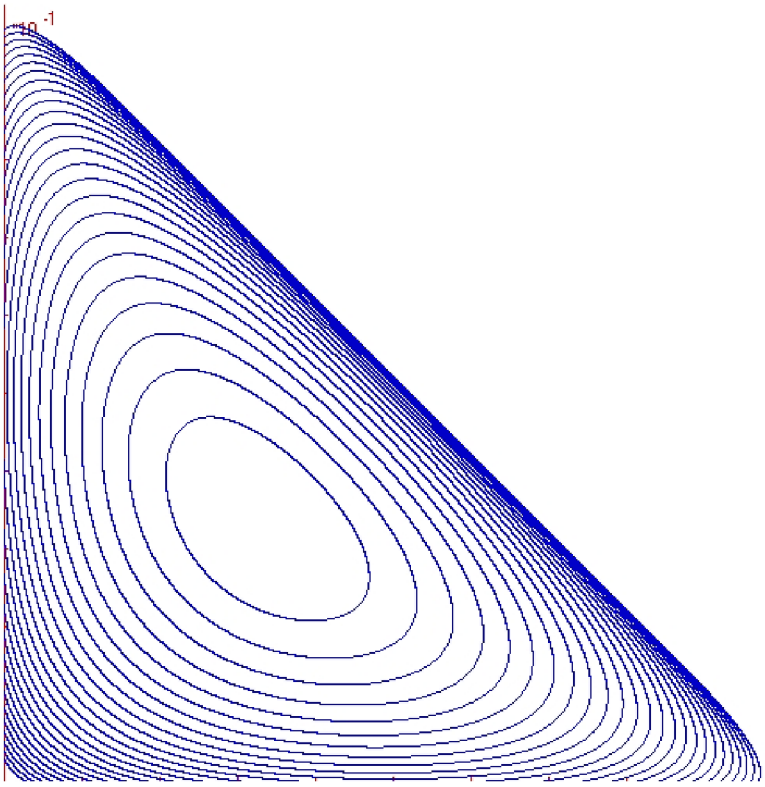
$n\backslash k$	2	3	4	5	10
5	0,0366353	0,0721801	0,1136287	0,1586792	0,3992038
10	0,0212106	0,0442641	0,0732706	0,1069663	0,3173585
50	0,0048302	0,0106855	0,0186815	0,0287117	0,1060530
100	0,0024569	0,0054807	0,0096605	0,0149668	0,0574234
1000	0,0002496	0,0005610	0,0009965	0,0015557	0,0061957

kde c nabývá pak hodnot $0, \ldots, 20$ pro 50 izočar.

Je vidět, že poblíž rovnoměrného rozdělení $\mathbf{p}_0$ jsou si obě pseudokvazinormy velmi podobné. Čím blíže ale jsme u hranice množiny S, tím více T_2 roste oproti T_1.



Obrázek 6.1: T_1 pseudokvazinorma



Obrázek 6.2: T_2 pseudokvazinorma

6.3. Odhady diskrétního rozdělení pomocí T_1 a T_2

Opět předpokládejme že pozorovaná diskrétní náhodná veličina X na pravděpodobnostním prostoru (Ω, Σ, P), jejíž rozdělení pravděpodobnosti chceme odhadnout (fitovat), nabývá nejvýše konečně mnoha různých hodnot $x_j^* \in \mathbb{R}$, (tj. $\Omega = \{x_1^*, \ldots, x_k^*\} \subset \mathbb{R}$) s neznámými pravděpodobnostmi

$$p_j = P(X = x_j^*), \quad j = 1, \ldots, k, \quad k > 1.$$

Pozorováním náhodné veličiny X získáme statistický soubor $(x_1, \ldots, x_n)$ a jeho roztříděním dostaneme statistický soubor

$$\left((x_1^*, \frac{f_1}{n}), \ldots, (x_k^*, \frac{f_k}{n}) \right),$$

kde f_j je absolutní četnost pozorované hodnoty x_j^*. Dále předpokládáme $n > k > 1$ a $f_j > 0$ pro všechna $j = 1, \ldots, k$ (jestliže $f_j = 0$, pak j-tou třídu vynecháme). Pro odhad rozdělení $\mathbf{p}$ požadujeme, aby toto rozdělení navíc splňovalo nějaké zadané podmínky, jejichž počet je $K \geq 1$. Mezi tyto podmínky nezařazujeme zřejmou podmínku $\sum_{j=1}^{k} p_j = 1$, ale další např. momentové podmínky.

Hledáme pak takové rozdělení $\mathbf{p}$, které má minimální pseudokvazinormu T_1 resp. T_2.

Rozdělení pravděpodobnosti $\mathbf{p} = (p_1, \ldots, p_k$ pozorované diskrétní náhodné veličiny X má na pravděpodobnostním prostoru (Ω, Σ, P), kde $\Omega = \{x_1^*, \ldots, x_k^*\}, k > 1$ a Σ je

množina všech podmnožin Ω, minimální pseudokvazinormu T_1 resp. T_2 za K počátečních momentových podmínek

$$\sum_{j=1}^{k} p_j (x_j^*)^l = M_l, \quad l = 0, \ldots, K,$$

jestliže jeho pseudokvazinorma T_1 resp. T_2 nabývá za těchto podmínek minimální hodnoty.

Pro $K < k-1$ obdržíme odhady pravděpodobností $p_j(\boldsymbol{\lambda})$ pomocí Lagrangeovy funkce

$$\Lambda_1(\mathbf{p}, \boldsymbol{\lambda}) = T_1 + \sum_{l=1}^{K} \lambda_l \left(\sum_{j=1}^{k} p_j (x_j^*)^l - M_l \right),$$

resp.

$$\Lambda_2(\mathbf{p}, \boldsymbol{\lambda}) = T_2 + \sum_{l=1}^{K} \lambda_l \left(\sum_{j=1}^{k} p_j (x_j^*)^l - M_l \right).$$

Lagrangeovy multiplikátory λ_l je možno určit pomocí nelineární soustavy rovnic odpovídající nulovému gradientu Lagrangeovy funkce, anebo přímo aplikovat některou metodu nelineární optimalizace pro určení jejího minima.

Má- li pozorovaná náhodná veličina X empirické rozdělení $\mathbf{f} = (\frac{f_1}{n}, \ldots, \frac{f_k}{n})$, pak statistika

$$\chi^2(\mathbf{f}, \mathbf{p}(\boldsymbol{\lambda})) = \sum_{j=1}^{k} \frac{(f_j - np_j(\boldsymbol{\lambda}))^2}{np_j(\boldsymbol{\lambda})} = \frac{1}{n} \sum_{j=1}^{k} \frac{f_j^2}{p_j(\boldsymbol{\lambda})} - n$$

má pro $n \to \infty$ asymptoticky rozdělení chí-kvadrát s $k-K-1$ stupni volnosti. Asymptotickou vlastnost můžeme použít k testování vhodnosti nalezeného rozdělení pravděpodobnosti $\mathbf{p}(\boldsymbol{\lambda}) = (p_1(\boldsymbol{\lambda}), \ldots, p_k(\boldsymbol{\lambda}))$. Pro praktické využití požadujeme aby bylo $np_j(\boldsymbol{\lambda}) > 5$ pro všechna $j = 1, \ldots, k$.

Postupným přidáváním momentových podmínek a opakovaným odhadem rozdělení pravděpodobnosti pomocí minimální pravděpodobnosti pomocí minimální f-kvazinormy lze určit minimální potřebný počet K těchto podmínek tak, aby platilo $\chi^2(\mathbf{f}, \mathbf{p}(\boldsymbol{\lambda})) \leq \chi^2_{n-1}$, kde χ^2_{n-1} je $(1-\alpha)$ kvantil rozdělení chí-kvadrát s daným počtem stupňů volnosti pro hladinu významnosti α.

7. Aplikace pseudokvazinorem

V této kapitole si ukážeme fungování obou pseudokvazinorem na příkladech. Odhadneme rozdělení a ověříme zda-li je odhad dobrý pomocí chí-kvadrát statistiky i pomocí Freeman-Tuckyho statiky. Výpočty jsou provedeny v programu Microsoft Office EXCEL modulem řešitel, který využívá kvazinewtonovské metody.

Příklad 1

Máme statistický soubor, kde hodnoty udávají počty α-částic emitovaných poloniem v konstantních časových intervalech (1/8 minuty):

Počet α-částic x_j	0	1	2	3	4
Četnost f_j	57	203	383	525	532
Počet α-částic x_j	5	6	7	8	Σ
Četnost f_j	408	273	139	88	2608

Hledáme minimum pseudokvazinorem za vedlejších podmínek daných prvními pěti momenty.

$$M_0 = \frac{1}{n}\sum_{j=1}^{k} f_j = 1, \qquad M_1 = \frac{1}{n}\sum_{j=1}^{k} f_j x_j = 3,844709,$$

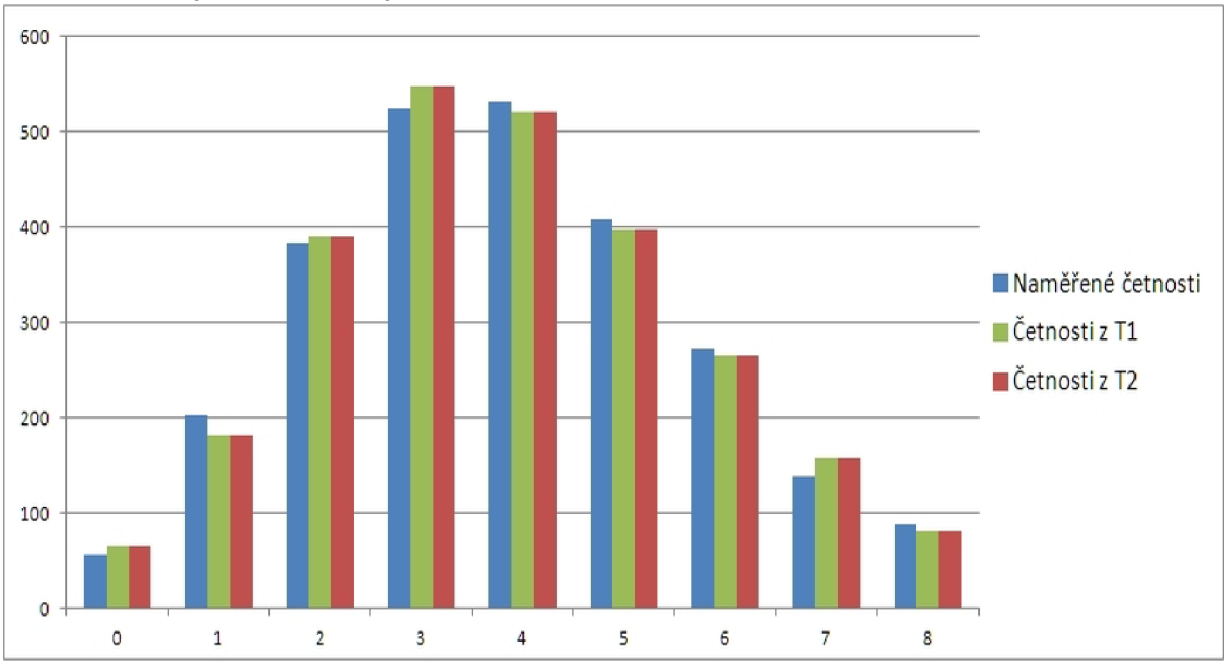
$$M_2 = \frac{1}{n}\sum_{j=1}^{k} f_j x_j^2 = 18,19133, \qquad M_3 = \frac{1}{n}\sum_{j=1}^{k} f_j x_j^3 = 97,46587,$$

$$M_4 = \frac{1}{n}\sum_{j=1}^{k} f_j x_j^4 = 570,5686.$$

Odhadnuté četnosti pak vycházejí pro T_1 a T_2:

Počet α-částic x_j	0	1	2	3	4
Odhadnuté $\tilde{f}_j$ z T_1	64,935	180,900	390,469	548,328	521,335
Odhadnuté $\tilde{f}_j$ z T_2	64,957	180,832	390,510	548,390	521,271
Počet α-částic x_j	5	6	7	8	Σ
Odhadnuté $\tilde{f}_j$ z T_1	397,107	266,258	157,585	81,082	2608
Odhadnuté $\tilde{f}_j$ z T_2	397,085	266,290	157,588	81,076	2608

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 8,274582276 pro T_1 a 8,307665297 pro T_2.

F-T kritérium má hodnotu 8,273830826 pro T_1 a 8,305966765 pro T_2.

Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 4 stupně volnosti má hodnotu 9,487729037. Obě kritéria jsou menší, než chí-kvadrát kvantil, tudíž je odhad vhodný. Navíc jsou pro obě pseudokvazinormy téměř stejné.

Z tabulky a grafu lze vidět, že odhady pomocí T_1 a T_2 jsou téměř totožné. Toto je způsobeno hlavně vysokou hodnotou n a asymptotickým charakterem pseudokvazinorem.

Nyní si ukažme chování pseudokvazinorem s podstatně menším n.

Příklad 2

Hodnoty v následující tabulce udávají šířku řezu elektroerozivní drátovou řezačkou. Hodnoty jsou z původního souboru roztříděny do pěti tříd.

střed třídy x_j^*	0,345	0,355	0,365
četnost f_j	6	14	9
střed třídy x_j^*	0,375	0,385	Σ
četnost f_j	3	1	33

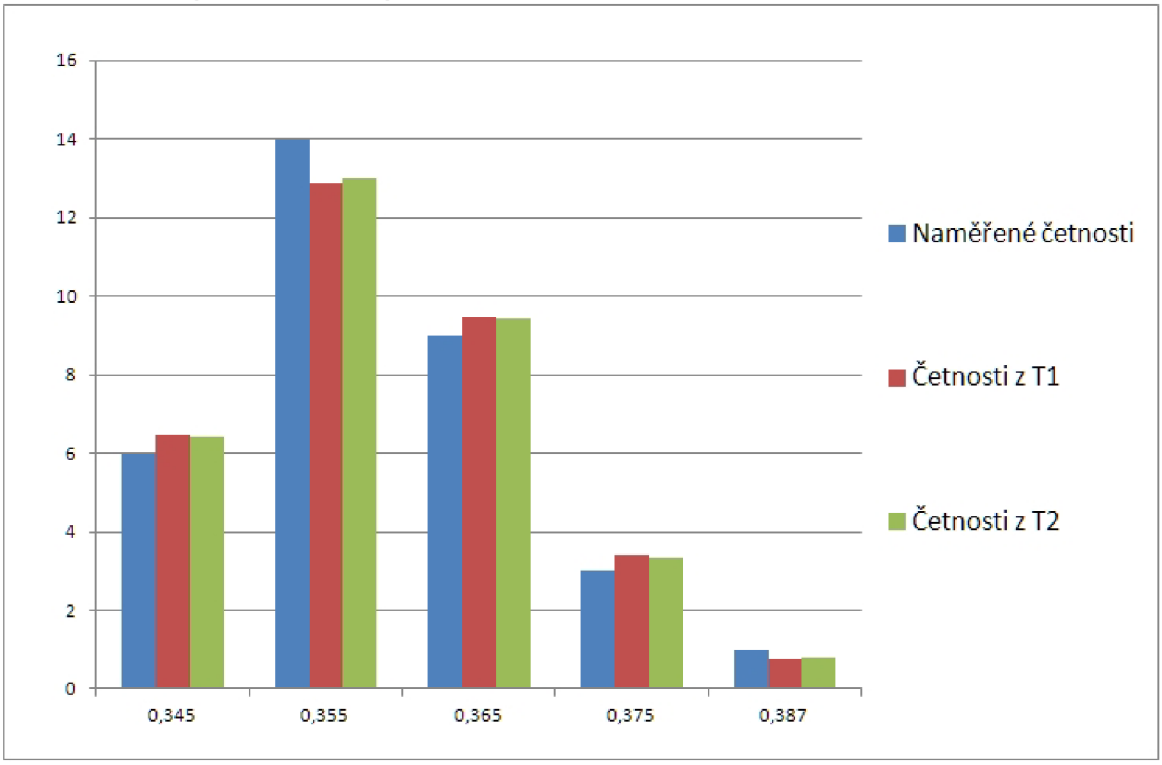
Budeme hledat minimum při splnění prvních tří momentů.

$$M_0 = 1, \quad M_1 = 0,35869697 \quad \text{a} \quad M_2 = 0,128762697.$$

Odhadnuté četnosti pak vycházejí následovně:

střed třídy x_j^*	0,345	0,355	0,365
četnost $\tilde{f}_j$ z T_1	6,476400634	12,89512541	9,485802155
četnost $\tilde{f}_j$ z T_2	6,425167733	13,01357887	9,434624628
střed třídy x_j^*	0,375	0,385	Σ
četnost $\tilde{f}_j$ z T_1	3,388294188	0,754377618	33
četnost $\tilde{f}_j$ z T_2	3,345558756	0,781070016	33

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 0,279062551 pro T_1 a 0,219983567 pro T_2.

F-T kritérium má hodnotu 0,321428036 pro T_1 a 0,269052639 pro T_2.

Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 2 stupně volnosti má hodnotu 5,991464547. Obě kritéria jsou menší, než chí-kvadrát kvantil, tudíž je odhad vhodný. Vzhledem k nižším hodnotám kritérií by se mohla T_2 považovat za vhodnější.

Z grafu a tabulky lze vidět, že tentokrát už jsou rozdíly viditelné, i když pořád velmi malé.

Příklad 3

Zde máme četnosti známek resp. bodů z předmětu „Pokročilé metody rozhodování“ na fakultě podnikatelské VUT v Brně. Známka A odpovídá 19 a 20 bodům, B odpovídá 17 a 18, C odpovídá 15 a 16, D odpovídá a 14 bodům, E odpovídá 10 až 12. Známku F pro méně jak 10 bodů vyřadíme, jelikož má četnost 0. Pro výpočet nahradíme známku středem jejího rozsahu bodů.

střed třídy x_j^*	11	13,5	15,5
četnost f_j	13	26	43
střed třídy x_j^*	17,5	19,5	Σ
četnost f_j	17	3	102

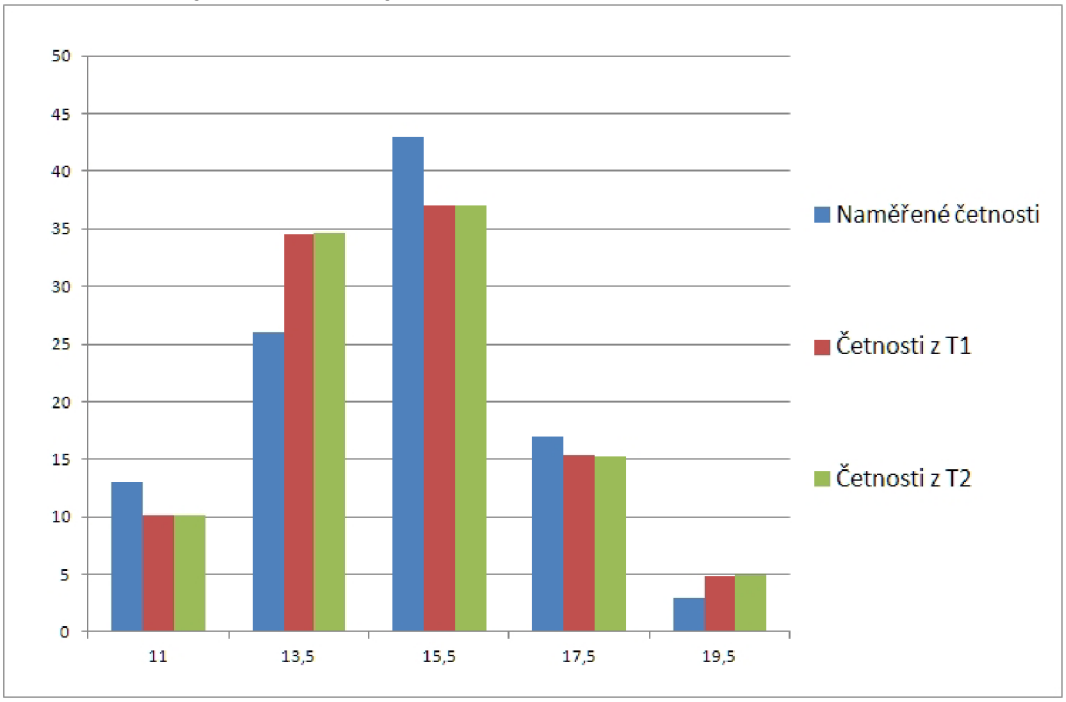
Budeme Odhadovat četnosti opět při splnění prvních tří momentů.

$$M_0 = 1, \quad M_1 = 14,86764706 \quad \text{a} \quad M_2 = 225,3848039.$$

Odhadnuté četnosti pak pro T_1 a T_2 vycházejí

střed třídy x_j^*	11	13,5	15,5
četnost $\tilde{f}_j$ z T_1	10,13445522	34,59328356	37,010318
četnost $\tilde{f}_j$ z T_2	10,11045472	34,63571876	37,04876577
střed třídy x_j^*	17,5	19,5	Σ
četnost $\tilde{f}_j$ z T_1	15,37807867	4,883864562	102
četnost $\tilde{f}_j$ z T_2	15,27587959	4,92918115	102

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 4,811979917 pro T_1 a 4,884565449 pro T_2.

F-T kritérium má hodnotu 4,910936431 pro T_1 a 4,99063443 pro T_2.

Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 2 stupně volnosti má hodnotu 5,991464547. Obě kritéria jsou menší, než chí-kvadrát kvantil, tudíž je odhad vhodný.

Příklad 4

Následující tabulka uvádí počet usmrcených osob na tisíc nehod na silnicích v roce 2013 způsobených danou věkovou kategorií.

Věk řidiče	16	19	22,5	29,5
Počet nehod	19,7	10,2	9,7	8,3
Věk řidiče	39,5	49,5	59,5	70
Počet nehod	8,6	7,3	9,9	12,9

Budeme hledat minimum při splnění prvních tří momentů.

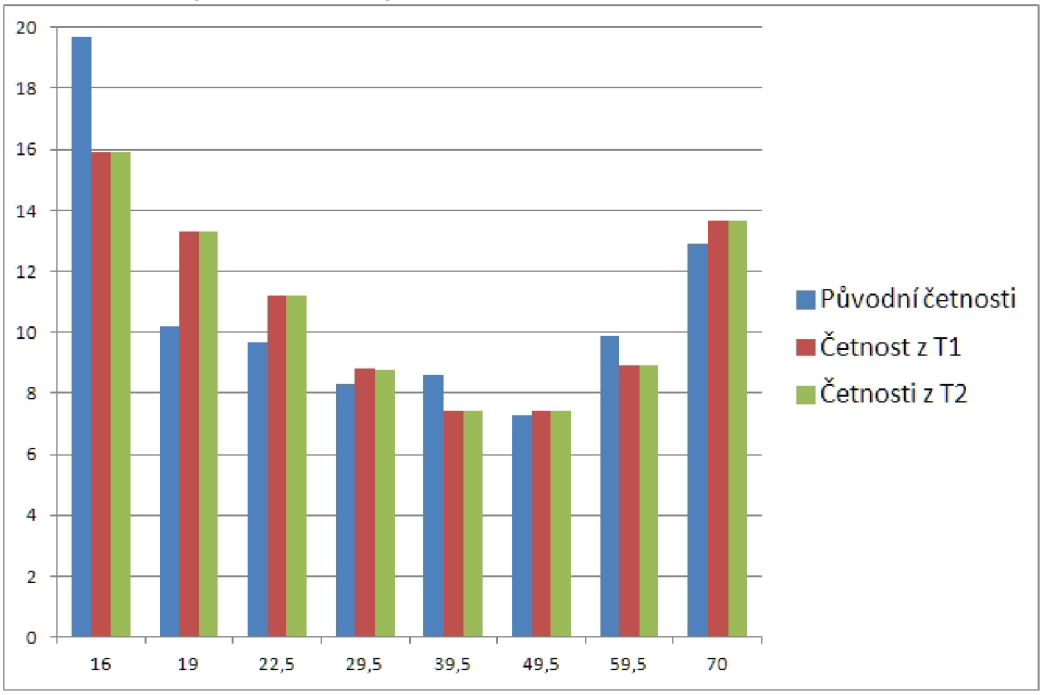
$$M_0 = 1, \quad M_1 = 36,54965358 \quad \text{a} \quad M_2 = 1736,980947.$$

Odhadnuté četnosti pak pro T_1 a T_2 vycházejí

Věk řidiče	16	19	22,5	29,5
četnost $\tilde{f}_j$ z T_1	15,90993174	13,28415871	11,19745805	8,787189976
četnost $\tilde{f}_j$ z T_2	15,92460374	13,27733568	11,18643783	8,78270262

Věk řidiče	39,5	49,5	59,5	70
četnost $\tilde{f}_j$ z T_1	7,417212847	7,447913855	8,905134705	13,65100011
četnost $\tilde{f}_j$ z T_2	7,422325622	7,454400755	8,902464575	13,64972917

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 2,190195087 pro T_1 a 2,175370151 pro T_2.

F-T kritérium má hodnotu 2,139559337 pro T_1 a 2,125545304 pro T_2.

Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 5 stupňů volnosti má hodnotu 11,07049775.

Příklad 5

Počítačovou simulací diskrétní náhodné veličiny X s Poissonovým rozdělením pravděpodobnosti s parametrem $\lambda = 1,5$ jsme získali statistický soubor pozorovaných hodnot $x_i, i, \ldots, 100$. Po jeho roztřídění a sloučení původních tříd s malými četnostmi pro $x_j^* = 4, 5, 6$ dostaneme roztříděný statistický soubor, který je uveden v tabulce:

střed třídy x_j^*	0	1	2	3	5
četnost f_j	21	36	27	9	7

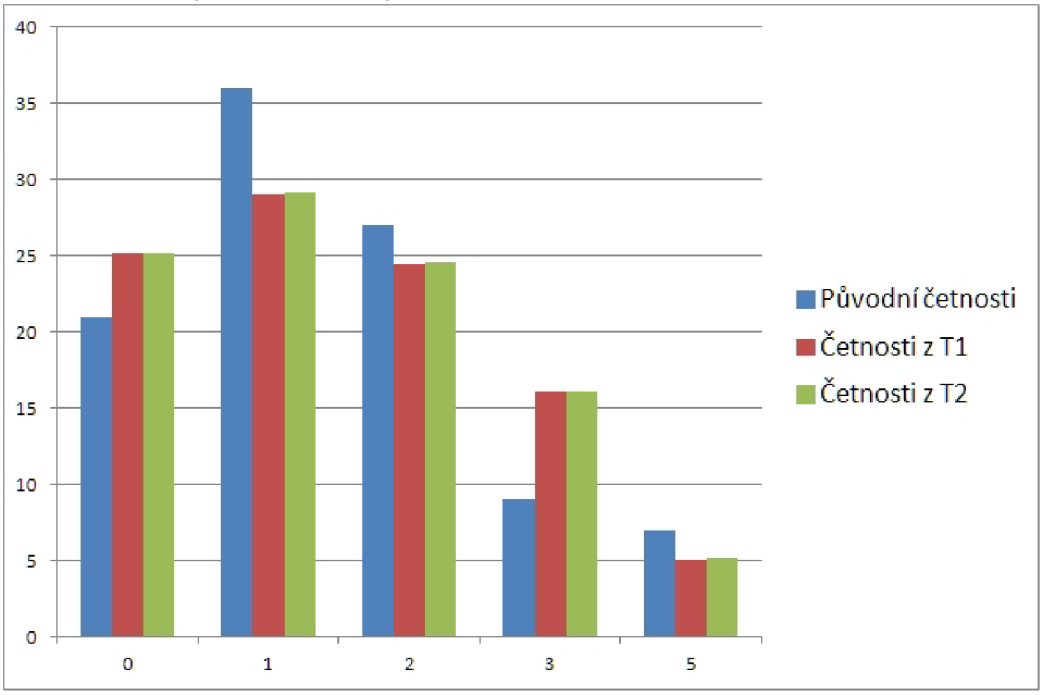
Budeme hledat minimum při splnění prvních tří momentů.

$$M_0 = 1, \quad M_1 = 1,52 \quad \text{a} \quad M_2 = 4.$$

Odhadnuté četnosti pak pro T_1 a T_2 vycházejí

střed třídy x_j^*	0	1	2	3	5
četnost $\tilde{f}_j$ z T_1	25,21385026	29,03547803	24,50280729	16,14020736	5,107657064
četnost $\tilde{f}_j$ z T_2	25,16899438	29,11372927	24,51841672	16,07243043	5,126429199

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 6,489092149 pro T_1 a 6,367386414 pro T_2.

F-T kritérium má hodnotu 7,003201099 pro T_1 a 6,870962999 pro T_2.

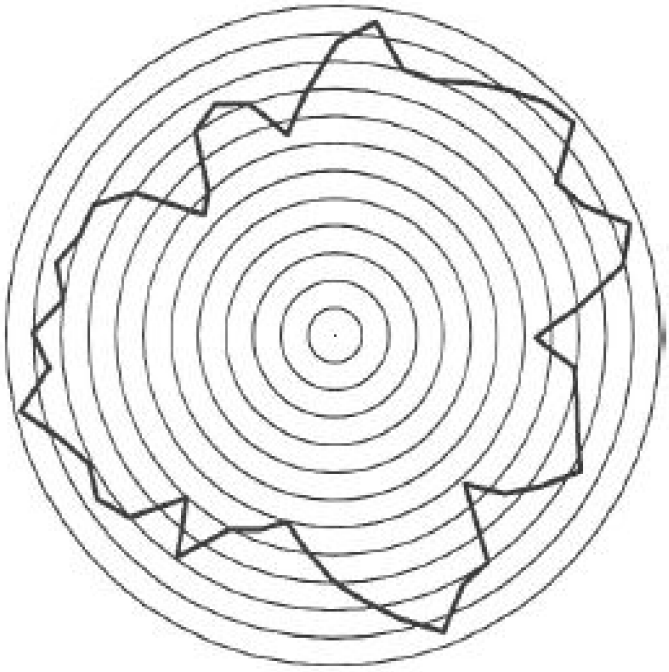
Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 3 stupně volnosti má hodnotu 7,814727764.

Příklad 6

Sledovali jsme rozložení obyvatelstva ve vybrané městské části (Obrázek 7.1). Čtvrť byla pokryta soustřednými ekvidistantními mezikružími a byl zjištěn počet obyvatel v jednotlivých mezikružích. V následující tabulce jsou zjištěné četnosti f_j a x_j^* jsou poloměry jednotlivých mezikruží [14].

Odhadujeme rozdělení při splnění prvních pěti momentů

$$M_0 = 1, \quad M_1 = 5,277591973, \quad M_2 = 39,05016722,$$
$$M_3 = 353,09699 \text{ a } M_4 = 3616,053512.$$



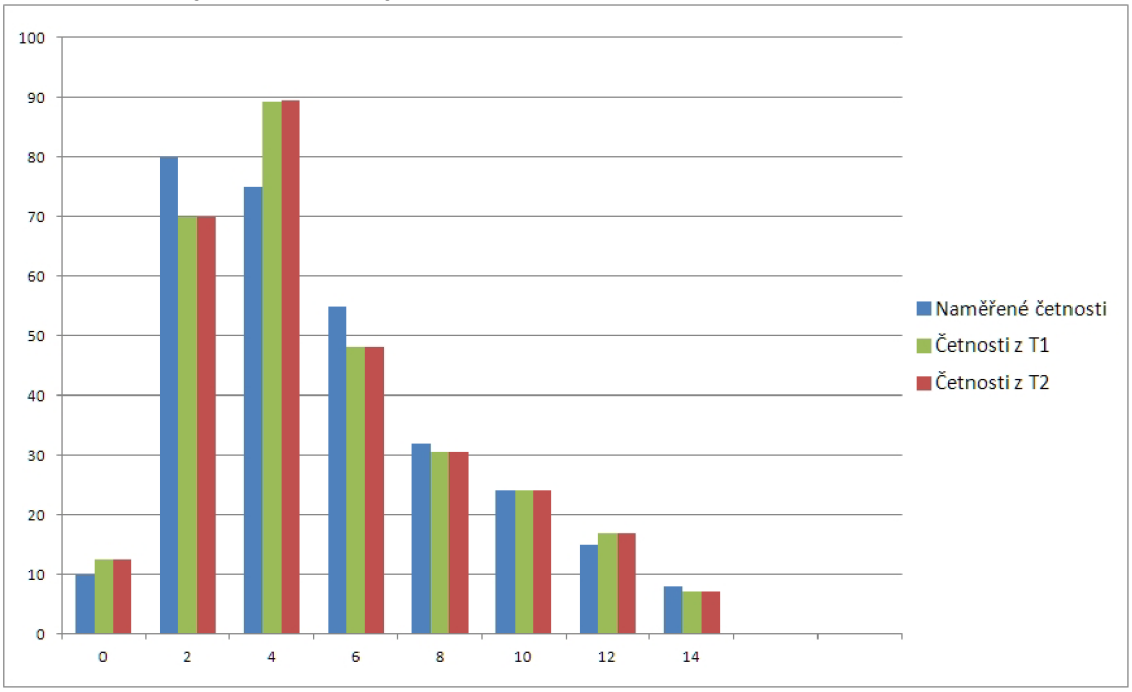
Obrázek 7.1: Městská čtvrť

x_j^*	0	2	4	6
f_j	10	80	75	55
x_j^*	8	10	12	14
f_j	32	24	15	8

x_j^*	0	2	4	6
$\tilde{f}_j$ z T_1	112,45001934	69,95017724	89,34002998	48,29651448
$\tilde{f}_j$ z T_2	12,45943007	69,90187863	89,43487012	48,2174175
x_j^*	8	10	12	14
$\tilde{f}_j$ z T_1	30,65942049	24,13928707	17,00575774	7,158793658
$\tilde{f}_j$ z T_2	30,66400371	24,17857172	16,97954255	7,164285699

Odhady pomocí T_1 a T_2 vypadají následovně:

Na následujícím obrázku jsou naměřené a odhadnuté četnosti.



χ^2 kritérium má hodnotu 5,553005615 pro T_1 a 5,615948043 pro T_2.

F-T kritérium má hodnotu 5,24949566 pro T_1 a 5,649785266 pro T_2.

Kvantil chí-kvadrát rozdělení pro $\alpha = 0.05$ a 3 stupňů volnosti má hodnotu 7,814727764.

8. Závěr

Diplomová práce se zabývá problematikou fitování diskrétního rozdělení pravděpodobnosti pozorovaných náhodných veličin z jejich hodnot. Hlavním cílem práce bylo rozšířit doposud získané výsledky v dané oblasti, kterých dosáhli doktoranti ústavu matematiky FSI VUT v Brně. Šlo o nalezení nových kvazinorem diskrétního rozdělení pravděpodobnosti [3], [4], [6], při zachování principu maximálního neurčitého rozdělení, které ještě splňuje testovací kritérium a současně vyhovuje zadaným momentovým podmínkám, konkrétně počátečním momentům co nejmenšího stupně.

Vychází se z málo známého a v podstatě nepoužívaného testovacího kritéria, jehož autory jsou Freeman a Tukey [17]. Ani po rozsáhlé rešerši se nepodařilo najít publikaci, v níž by bylo dokázáno, že toto testové kritérium má asymptotické chí-kvadrát rozdělení. Proto bylo toto tvrzeni v práci dokázáno. Dané kritérium lze chápat také jako míru vzdálenosti dvou diskrétních rozdělení pravděpodobnosti a to umožňuje vzít jej jako zobecnění f-divergence dvou rozdělení. Tak byla definována nová míra neurčitosti diskrétního rozdělení pravděpodobnosti a nazvána jako pseudokvazinorma. Byly nalezeny její základní vlastnosti a pro fitování rozdělení byla tato pseudokvazinorma aplikována na konkrétních příkladech. V práci je prokázána její použitelnost pro řešení konkrétních úloh, např. použitím řešiče z Excelu v MS Windows.Bude vhodné se v budoucnu zabývat dalšími vlastnostmi pseudokvazinormy a případně využít zpětně uvedený postup pro zobecnění pojmu kvazinormy.

Literatura

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

[2] Anděl, J.: Statistické metody, Praha:MATHFYZPRESS, 2007. ISBN 978-80-7378-003-6.

[3] Šácha, J.: Kvazinormy diskrétních rozdělení pravděpodobnosti a jejich aplikace. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2012. Vedoucí dizertační práce doc. RNDr. Zdeněk Karpíšek,CSc.

[4] Lacinová, V.:Odhady diskrétního rozložení pravděpodobnosti a bootstrap. Brno: Vysoké učení technické v Brně, Fakulta strojního inženýrství, 2014. Vedoucí dizertační práce doc. RNDr. Zdeněk Karpíšek,CSc.

[5] Bazaraa, M., S., Sherali, H., D., Shetty, C., M. Nonlinear programming: Theory and Algorithms USA: Whiley-Interscience, 2006, ISBN 978-0-471-48600-8.

[6] Jurák, P. Odhady diskrétních rozdělení pravděpodobnosti pomocí kvazinorem. Brno: Vysoké u˘cení technické v Brně, Fakulta strojního inženýrství, 2008. 85 s. Vedoucí disertační práce Doc. RNDr. Zdeněk Karpíšek, CSc.

[7] Gray, R. M.: Entropy and information theory. Berlin: Springer-Verlag, 1990.

[8] Klapka, J., Dvořák, J., Popela, P.: Metody operačcního výzkumu. Brno: PC-DIR, 1996.

[9] Paulov, J.: Teória informacií a regionálna anlýza. PMFA, 1982.

[10] Pitman, E. J. G.: Some basic theory for statistical inference. New York: John Wiley & Sons, 1978.

[11] Karpíšek, Z., Jurák, P. a Šácha, J.: Divergences fo discrete probability distribution estimations. In Summer School DATASTAT '06, Proceedings, Masaryk University, Brno, 2007,pp. 109-120, ISBN 978-80-210-4493-7.

[12] Neradová,V.: Progesivní metody odhadu rozdělení pravděpodobnosti. Diplomová práce(vedoucí Karpíšek, Z.), FSI VUT v Brně, Brno 2007.

[13] Cressie, N. and Read, T. R. C.: J. R. Statist. Soc. B, 1984, p. 440-464.

[14] Vajda, I.: Theory of statistical inference and information. London. Kluver Academic Press, 1989, ISBN 9027727813.

[15] Buck,B. and Macaulay, V. A.: Maximum entropy in action. New York, 1992, Oxford University Press. ISBN 0-19-853941-X.

[16] Kotz, Johnson and Read :Encyyclopedia of Statistical Sciances, vol. 3, 1989, p.210

[17] Freeman, M. F. and Tukey. J. W.:Transformations related to angluar and square root. Annals of Mathematical Statistics,1950,21,607-611.