STOCHASTICKÉ MODELOVÁNÍ: ODHADY ROZDĚLENÍ PRAVDĚPODOBNOSTI

Zdeněk Karpíšek

Centrum pro jakost a spolehlivost výroby (CQR)
Odbor statistiky a optimalizace, Ústav matematiky
Fakulta strojního inženýrství, Vysoké učení technické v Brně
e-mail: karpisek@fme.vutbr.cz
http://www.mat.fme.vutbr.cz/home/karpisek/

Abstrakt: Referát má přehledový charakter a je obsahově zaměřen na některé metody odhadu rozdělení pravděpodobnosti pozorovaných náhodných veličin. Konkrétně jde o empirické přístupy k odhadům a zejména o vybrané inferenční metody: Pearsonovy křivky, Gramovy-Charlierovy řady, Johnsonovy křivky, odhady pomocí kvazinorem vycházejících ze vzdáleností rozdělení a jádrové odhady hustot. Metody jsou podány popisným charakterem s naznačením teoretických základů a získané odhady jsou prezentovány na příkladech s potřebným grafickým doprovodem.

1 Náhodný výběr a odhady rozdělení

Metodami matematické statistiky řešíme v aplikacích dvě základní úlohy [1, 9]:

- ***odhady parametrů a rozdělení***,
- ***testování statistických hypotéz o parametrech a rozděleních***.

Matematická* (*inferenční, indukční*) *statistika vychází z následujícího pojetí:

- opakujeme n-krát nezávisle pokus, jehož výsledkem je pozorovaná hodnota náhodné veličiny X s distribuční funkcí $F(x, \vartheta)$, kde ϑ je reálný parametr (případně vektor parametrů anebo jejich funkce) daného rozdělení pravděpodobnosti,
- tím získáme ***statistický soubor*** s ***rozsahem*** n, tj. číselný vektor
$$\mathbf{x} = (x_1,...,x_n),$$
kde x_i je pozorovaná hodnota složky X_i, $i = 1,...,n$,
- statistický soubor je pozorovaná hodnota ***náhodného výběru*** (z náhodné veličiny X nebo jejího rozdělení pravděpodobnosti) s ***rozsahem*** n, tj. náhodného vektoru
$$\mathbf{X} = (X_1,...,X_n),$$
jehož složky jsou nezávislé náhodné veličiny X_i se stejnou distribuční funkcí (pravděpodobnostní funkcí anebo hustotou pravděpodobnosti) jako má pozorovaná náhodná veličina X.

Analogicky pak definujeme náhodný výběr z náhodného vektoru.

Náhodný výběr má simultánní distribuční funkci

$$F(\mathbf{x};\vartheta) = F(x_1,...,x_n;\vartheta) = \prod_{i=1}^{n} F(x_i;\vartheta)$$

a simultánní pravděpodobnostní funkci

$$p(\mathbf{x};\vartheta) = p(x_1,...,x_n;\vartheta) = \prod_{i=1}^{n} p(x_i;\vartheta),$$

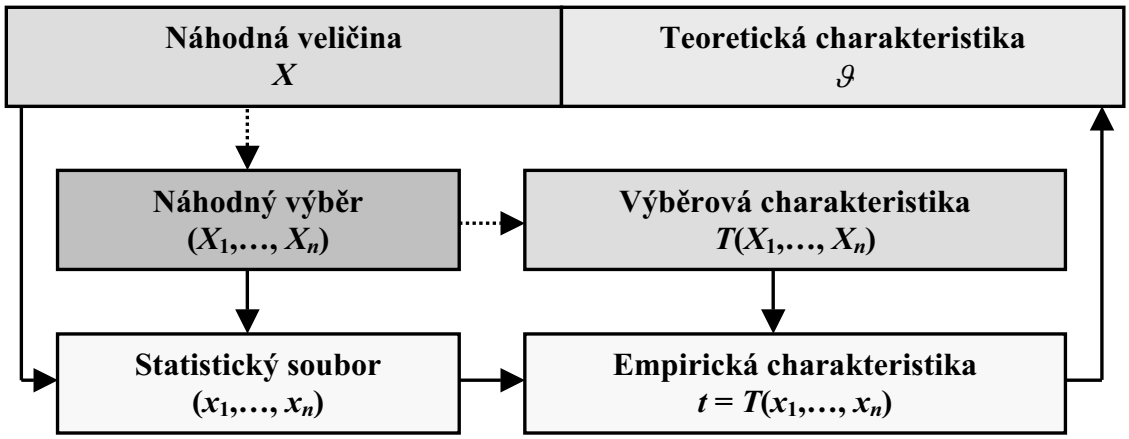
resp. simultánní hustotu pravděpodobnosti

$$f(\mathbf{x};\vartheta) = f(x_1,...,x_n;\vartheta) = \prod_{i=1}^{n} f(x_i;\vartheta).$$

Množina všech možných hodnot náhodného výběru, tj. množina všech statistických souborů, tvoří tzv. ***výběrový prostor.***

Funkce náhodného výběru $T(X_1,...,X_n)$ je ***výběrová charakteristika*** nebo ***statistika***. Její hodnota na statistickém souboru $t = T(x_1,...,x_n)$ je ***empirická charakteristika*** nebo ***pozorovaná hodnota statistiky*** T.

Základní princip statistické indukce vyjadřuje schéma:



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***.

Empirické odhady rozdělení pravděpodobnosti jsou založeny na:

- teoretických a zkušenostních předpokladech a informacích o tvaru rozdělení,
- grafických vyjádření statistických souborů (histogramy, polygony aj.).

Mezi inferenční (indukční) metody odhadů rozdělení pravděpodobnosti patří zejména:

- Pearsonovy křivky,
- Gramovy – Charlierovy řady,
- Johnsonovy křivky,
- kvazinormy,
- jádrové odhady.

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:

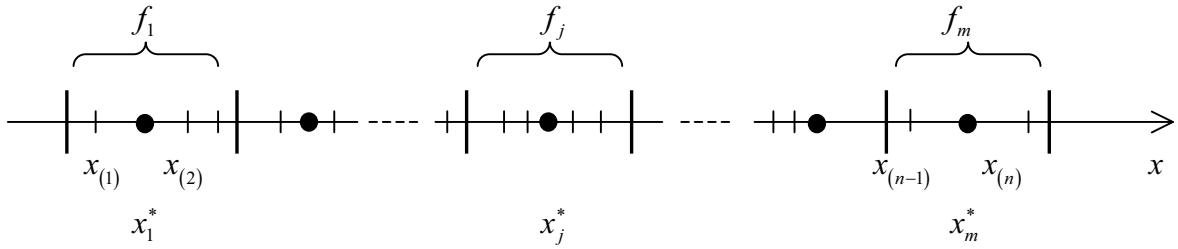
- rozhodnout, zda se jedná spojitou anebo diskrétní náhodnou veličinu s ohledem na přesnost pozorování (měření) a možný obor jejích hodnot,
- posoudit šikmost a špičatost rozdělení pozorované náhodné veličiny,
- posoudit vícemodalitu rozdělení pozorované náhodné veličiny,
- odfiltrovat extrémně odchýlené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

2 Empirické odhady rozdělení

Neroztříděný statistický soubor $(x_1,...,x_n)$ nebo uspořádaný statistický soubor $(x_{(1)},...,x_{(n)})$, $x_{(i)} \le x_{(i+1)}$, $i = 1,...,n$, s rozsahem n převedeme na roztříděný statistický soubor (variační řadu):

x_j^*	x_1^* . . . x_m^*
f_j	f_1 . . . f_m

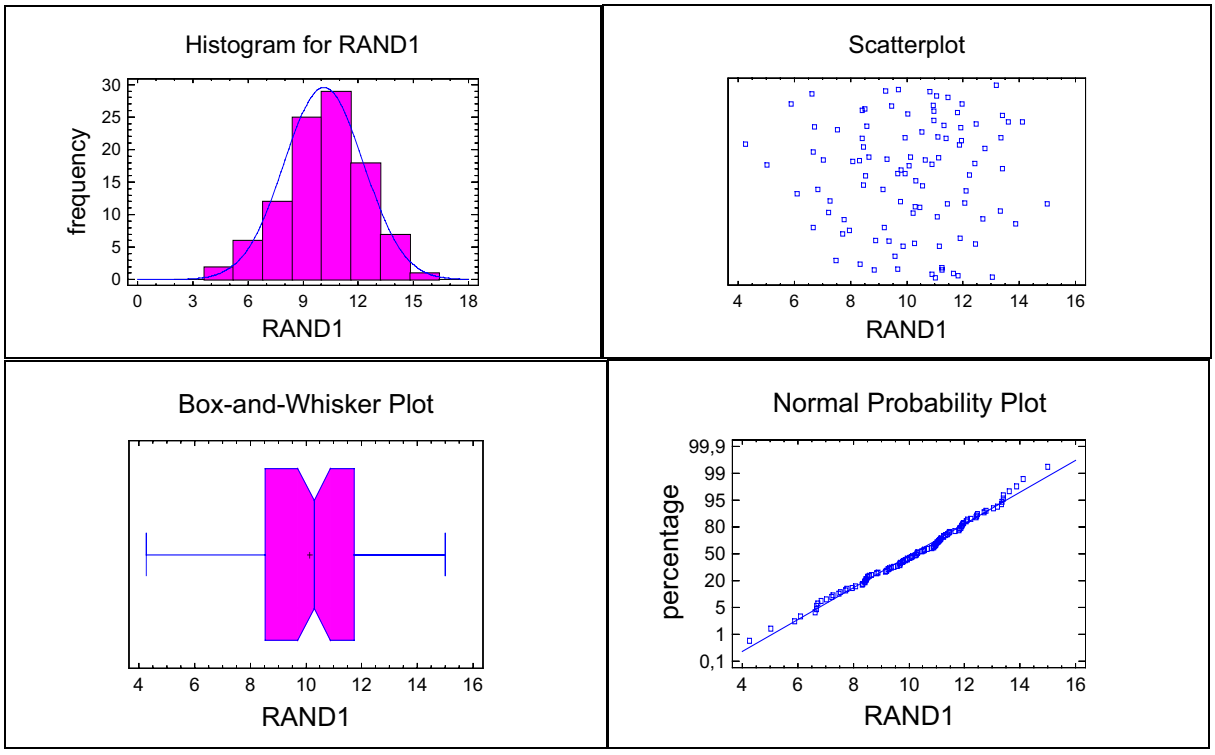
Přitom x_j^* je střed třídy, f_j je četnost hodnot $x_{(i)}$ v j-té třídě, $j = 1,...,m$; třídy jsou zleva otevřené a zprava uzavřené intervaly tvaru. Grafické vyjádření roztřídění:



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 [1, 9].

Příklad normálního rozdělení N(2;10)

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.



Uncensored Data - RAND1

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

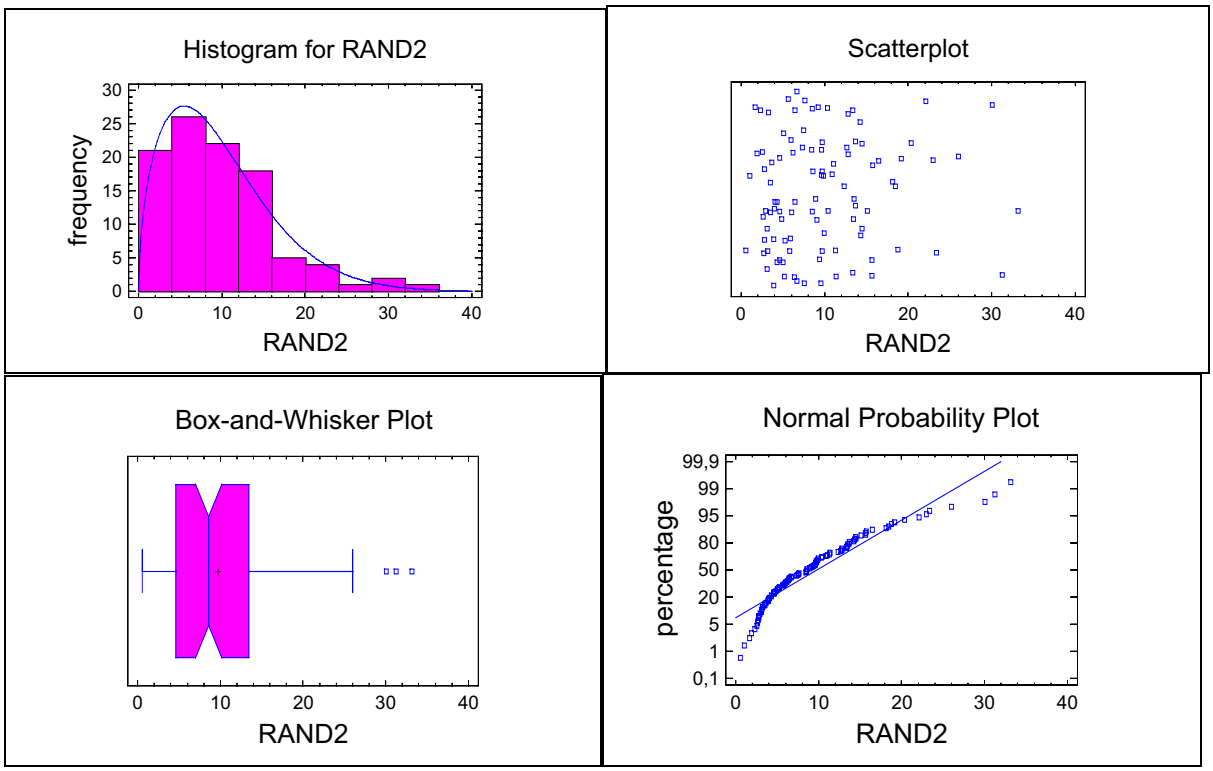
Estimated Kolmogorov statistic DPLUS = 0,0348125
Estimated Kolmogorov statistic DMINUS = 0,0698938
Estimated overall statistic DN = 0,0698938
Approximate P-Value = 0,713002

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.

Příklad Weibullova rozdělení W(1,5;10)

Příklad ilustruje základní grafické zpracování statistického souboru získaného generátorem náhodných čísel pro Weibullovo rozdělení s parametrem tvaru 1,5 a parametrem měřítka 10, odhady parametrů a následnou verifikaci pomocí testů shody. Grafy a výpočty byly realizovány pomocí statistického softwaru Statgraphics.



Uncensored Data – RAND2

```
Analysis Summary
Data variable: RAND2
100 values ranging from 0,523124 to 33,136

Fitted Weibull distribution:
     shape = 1,53209
     scale = 10,8573

Tests for Normality for RAND2

Computed Chi-Square goodness-of-fit statistic = 44,0
P-Value = 0,00233817

Shapiro-Wilks W statistic = 0,888196
P-Value = 2,73469E-10
Z score for skewness = 3,1973
P-Value = 0,00138734
Z score for kurtosis = 2,62291
P-Value = 0,00871837

Goodness-of-Fit Tests for RAND2
```

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	1,56234	2	5,00	1,80
1,56234	2,49933	3	5,00	0,80

	2,49933	3,31649	10	5,00	5,00
	3,31649	4,0789	7	5,00	0,80
	4,0789	4,81454	5	5,00	0,00
	4,81454	5,53976	5	5,00	0,00
	5,53976	6,26618	6	5,00	0,20
	6,26618	7,00346	5	5,00	0,00
	7,00346	7,76066	4	5,00	0,20
	7,76066	8,54732	3	5,00	0,80
	8,54732	9,37434	4	5,00	0,20
	9,37434	10,2552	9	5,00	3,20
	10,2552	11,2074	4	5,00	0,20
	11,2074	12,2558	2	5,00	1,80
	12,2558	13,4373	6	5,00	0,20
	13,4373	14,8122	8	5,00	1,80
	14,8122	16,4906	5	5,00	0,00
	16,4906	18,713	2	5,00	1,80
	18,713	22,2197	4	5,00	0,20
above	22,2197		6	5,00	0,20

Chi-Square = 19,2 with 17 d.f. P-Value = 0,317174

Estimated Kolmogorov statistic DPLUS = 0,0508244
Estimated Kolmogorov statistic DMINUS = 0,05203
Estimated overall statistic DN = 0,05203
Approximate P-Value = 0,949458

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0,05203	0,5203	>=0.10*
Kuiper V	0,102854	1,02854	>=0.10*
Cramer-Von Mises W^2	0,0500082	0,0510084	>=0.10*
Watson U^2	0,0441107	0,0449929	>=0.10*
Anderson-Darling A^2	0,398667	0,406641	>=0.10*

*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.

3 Inferenční metody odhadu

3.1 Pearsonovy křivky

K. Pearson [2, 3] vyšel z diferenciální rovnice pro hustotu pravděpodobnosti ve tvaru

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

Podle toho, jakých hodnot nabývají parametry rovnice, dostávají hustoty pravděpodobnosti $f(x)$ různé výrazy a v grafickém zobrazení mají různé geometrické tvary. 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

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

kde

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

a $\upsilon(x)$ značí vhodnou primitivní funkci k danému integrandu. Typově závisí integrál na hodnotách koeficientů ve jmenovateli, takže se vychází z řešení rovnice

$$c_0 + c_1 x + c_2 x^2 = 0 .$$

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

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

kde

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

přičemž $r_1 = 0, r_2 = 1$ a σ^2 je rozptyl rozdělení s hustotou $f(x)$.

Kořeny R_1 a R_2 rovnice $c_0 + c_1 x + c_2 x^2 = 0$ závisí na diskriminantu

$$D = c_1^{\,2} - 4c_0c_2 = c_1^{\,2}\left(1 - \frac{4c_0c_2}{c_1^{\,2}}\right) = c_1^{\,2}\left(1 - \frac{1}{\kappa}\right),$$

kde $\kappa = \dfrac{c_1^{\,2}}{4c_0c_2}$. Jestliže označíme

$$t = \sqrt{r_3^2\left(s+2\right)^2 + 16\left(s+1\right)},$$

pak kořeny R_1 a R_2 jsou

$$R_1 = \frac{\sigma}{4}\left(1-t\right),\ R_2 = \frac{\sigma}{4}\left(1+t\right).$$

Každé hodnotě veličiny κ pak odpovídají určité hodnoty kořenů R_1, R_2. Pro určení konkrétního typu křivky použijeme za kritérium veličinu κ, jejíž alternativní zápis pomocí momentů r_3, r_4 je

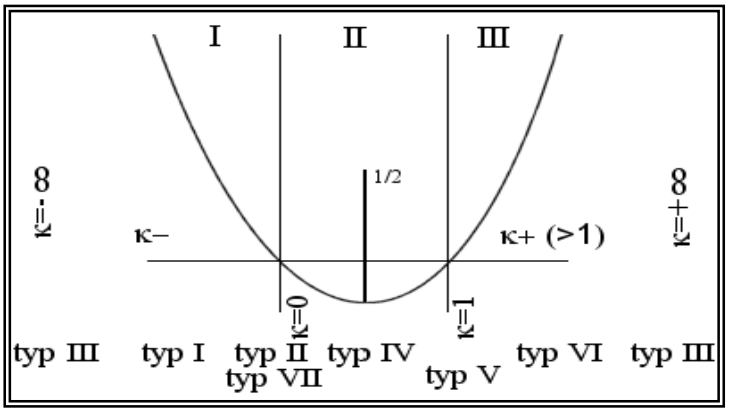
$$\kappa = \frac{r_3^2\left(r_4+3\right)^2}{4\left(4r_4 - 3r_3^2\right)\left(2r_4 - 3r_3^2 - 6\right)}$$

Význam kritéria κ vyjadřuje rozepsaný kvadratický trojčlen

$$c_0 + c_1 x + c_2 x^2 = c_2\left[\left(x + \frac{c_1}{2c_2}\right)^2 - \frac{4c_0^{\,2}}{c_1^{\,2}}\kappa\left(\kappa - 1\right)\right].$$

Následující obrázek popisuje rozdělení Pearsonových křivek na grafu paraboly $y = \kappa\left(\kappa - 1\right)$:

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.

Ukázka typu I

Odhady četností výskytu v jednotlivých třídách je možno vypočítat ze vztahů (v tomto oddílu značíme četnosti n_j, resp. jejich odhady $\tilde{n}_j$, místo f_j, resp. $\tilde{f}_j$):

$$\tilde{n}_j = \tilde{n}_0 \left(1+\frac{x}{l_1}\right)^{q_1} \left(1-\frac{x}{l_2}\right)^{q_2} \qquad \text{pro } q_1 > 0,\ q_2 > 0,$$

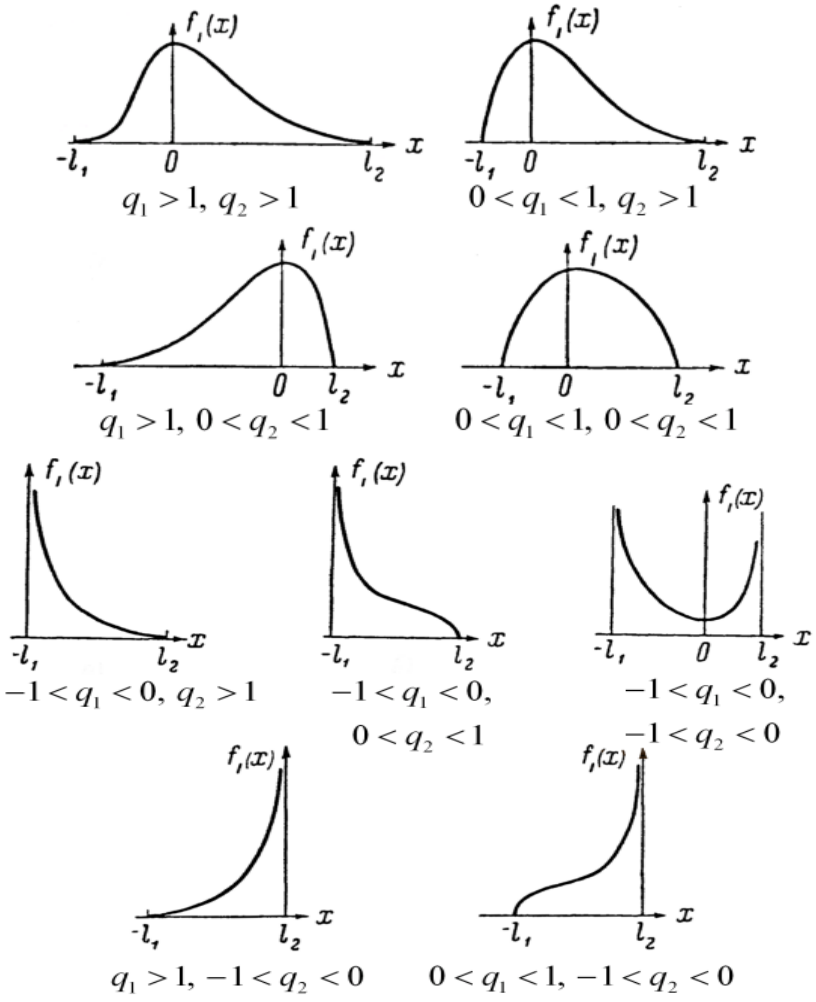
$$\tilde{n}_j = \tilde{n}_0 \left(1+\frac{x}{l_1}\right)^{-q_1} \left(1-\frac{x}{l_2}\right)^{q_2} \qquad \text{pro } q_1 < 0,\ q_2 > 0,$$

$$\tilde{n}_j = \tilde{n}_0 \left(1+\frac{x}{l_1}\right)^{-q_1} \left(1-\frac{x}{l_2}\right)^{-q_2} \qquad \text{pro } q_1 < 0,\ q_2 < 0,$$

kde pro odchylku od střední hodnoty $x' - m_1$ je $x = x' - m_1 + \frac{\sigma r_3}{2}\frac{s+2}{s-2}$ a konstanta

$$\tilde{n}_0 = \frac{n}{l}\frac{q_1^{q_1} q_2^{q_2}}{(s-2)^{s-2}}\frac{\Gamma(s)}{\Gamma(q_1+1)\Gamma(q_2+1)}.$$

Pro rozpětí l křivek rozdělení je $l = 2\sigma\sqrt{(s+1)(1-\kappa)}$, $l = \frac{\sigma t}{2}$, $l = l_1 + l_2$, $l_1 = \frac{q_1 l}{s-2}$, $l_2 = \frac{q_2 l}{s-2}$, kde $q_{1,2} = \frac{1}{2}\left[(s-2) \mp s(s+2)\frac{r_3}{t}\right]$. Grafy možných Pearsonových křivek (hustot rozdělení typy I jsou následující:



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 [3, 5] .

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 se získají výpočtem následující statistiky:

$$m_1 = 0,087, \quad \bar{x} = 42,935, \quad \sigma = 2,203,$$
$$m_2 = 4,861, \quad \mu_2 = 4,853, \quad r_3 = 0,605,$$
$$m_3 = 7,737, \quad \mu_3 = 6,469, \quad r_3^2 = 0,366,$$
$$m_4 = 72,385, \quad \mu_4 = 62,912, \quad r_4 = 2,968.$$

Na základě těchto statistik vypočteme veličiny s, t a kritérium κ:

$$s = 8,272, \quad t = 13,674, \quad \kappa = -0,26.$$

V tomto případě je $\kappa < 0$, takže použijeme Pearsonovu křivku typu I. Dalším výpočtem obdržíme modus $\hat{x}$, rozpětí rozdělení l s krajními hodnotami l_1 a l_2, exponenty q_1, q_2 pro výpočet odhadů četností $\tilde{n}_j$ a $\tilde{n}_0$:

$$\hat{x} = 37,480, \quad l = 15,06, \quad q_1 = 1,256, \quad q_2 = 5,016, \quad l_1 = 3,017, \quad l_2 = 12,043.$$

Konstanta $\tilde{n}_0 = 179,43$, takže odhadnuté četnosti jsou

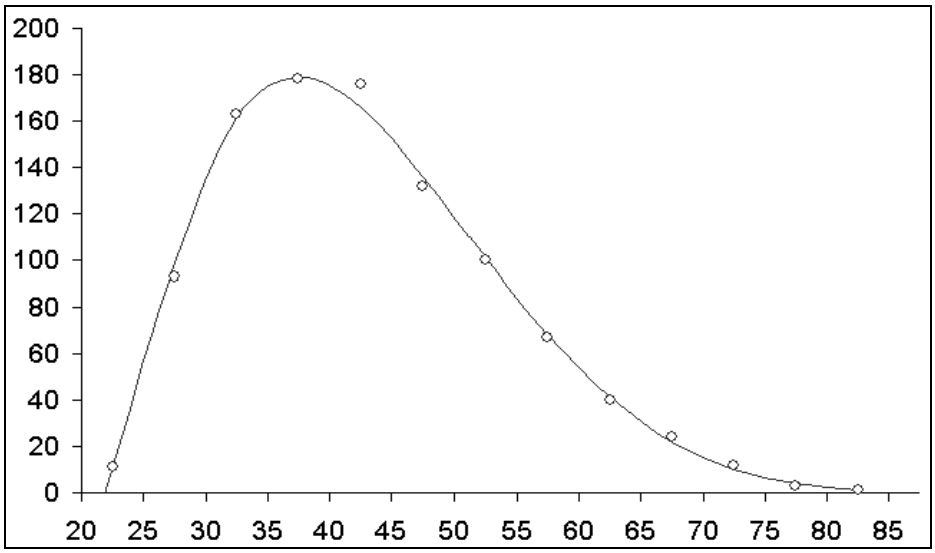
$$\tilde{n}_j = 179,43\left(1 + \frac{x_j^*}{3,017}\right)^{1,256}\left(1 - \frac{x_j^*}{12,043}\right)^{5,016}.$$

Vypočtené četnosti jsou:

j	Třídy	x_j^*	n_j	$\tilde{n}_j$
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

Vykreslením původních četností n_j a křivky prokládající vypočtené četnosti $\tilde{n}_j$ do grafu dostáváme výsledný tvar rozdělení:



Z tabulky i grafu je zřejmá dobrá aproximace původního neznámého rozdělení.

3.2 Gramovy – Charlierovy řady

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

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

je charakteristická funkce binomického rozdělení Bi(*n*,*p*), pomocí ortogonálního systému funkcí tvořícího bázi založenou [3]:

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

Jde o Fourierovu transformaci, takže zpětná transformace je

$$p(x;n,p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} \varphi(t)\, dt\,.$$

Rozdělení typu A

Hustota pravděpodobnosti má tvar [3]

$$f_A(x) = f(x) + a_1 f^{(1)}(x) + a_2 f^{(2)}(x) + \ \ldots\ + a_n f^{(n)}(x) + \ \ldots\ ,$$

kde

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

je hustota normovaného normálního rozdělení a $f^{(n)}(x)$ je *n*-tá derivace $f(x)$ podle *x*.
Pro Čebyševův-Hermitův polynom

$$H_n(x) = x^n - \frac{n^{[2]}}{2 \cdot 1!} x^{n-2} + \frac{n^{[4]}}{2^2 \cdot 2!} x^{n-4} - \frac{n^{[6]}}{2^3 \cdot 3!} x^{n-6} + \dots,$$

kde $n^{[k]}$ je variace *k*-té třídy z *n* prvků bez opakování, platí

$$f^{(n)}(x) = (-1)^n H_n(x) f(x),$$

takže

$$H_n(x) = (-1)^n \sqrt{2\pi}\, e^{x^2/2} f^{(n)}(x).$$

Obecně pak pro *n*-tou derivaci platí

$$f^{(n)}(x) = (-1)^n \left(x^n - \frac{n^{[2]}}{2 \cdot 1!} x^{n-2} + \frac{n^{[4]}}{2^2 \cdot 2!} x^{n-4} - \frac{n^{[6]}}{2^3 \cdot 3!} x^{n-6} + \dots \right) f(x).$$

Derivace $f^{(n)}(x)$ a mnohočleny $H_n(x)$ jsou ortogonální, neboť

$$\int_{-\infty}^{\infty} f^{(n)}(x) H_m(x)\, dx = \begin{cases} 0 & \text{pro } m \neq n, \\ (-1)^n n! & \text{pro } m = n. \end{cases}$$

Pak je

$$\int_{-\infty}^{\infty} f_A(x) H_n(x)\, dx = \sum_{m=0}^{\infty} a_m \int_{-\infty}^{\infty} f^{(m)}(x) H_n(x)\, dx = a_n \int_{-\infty}^{\infty} f^{(n)}(x) H_n(x)\, dx = (-1)^n a_n n!,$$

odkud

$$a_n = \frac{(-1)^n}{n!} \int_{-\infty}^{\infty} f_A(x) H_n(x)\, dx, \quad n = 0, 1, 2, \dots .$$

Za pomocí obecného normovaného momentu

$$r_n = \int_{-\infty}^{\infty} x^n f_A(x)\, dx,$$

je

$$a_n = \frac{(-1)^n}{n!} \int_{-\infty}^{\infty} f_A(x) \left[x^n - \frac{n^{[2]}}{2 \cdot 1!} x^{n-2} + \frac{n^{[4]}}{2^2 \cdot 2!} x^{n-4} - \frac{n^{[6]}}{2^3 \cdot 3!} x^{n-6} + \dots \right] dx,$$

takže

$$a_n = \frac{(-1)^n}{n!} \sum_{h=0}^{n} \frac{n^{[2h]}}{2^h h!} r_{n-2h}.$$

Protože $r_0 = r_2 = 1$, $r_1 = 0$, pak prvních šest koeficientů je

$$a_0 = 1, \qquad a_1 = a_2 = 0,$$
$$a_3 = -\frac{1}{3!} r_3, \qquad a_4 = -\frac{1}{4!}(r_4 - 3),$$
$$a_5 = -\frac{1}{5!}(r_5 - 10 r_3), \quad a_6 = -\frac{1}{6!}(r_6 - 15 r_4 + 30).$$

Odtud

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

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 (v tomto oddílu značíme četnosti n_j, resp. jejich odhady $\tilde{n}_j$, místo f_j, resp. $\tilde{f}_j$):

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

Příklad

Měřením mezi pevnosti $X\left(kg/cm^2\right)$ při stlačení vzorku z borového dřeva podél vláken byly po roztřídění získány hodnoty v následující tabulce [3, 5]:

Mez pevnosti x_j^* (kg/cm^2)	215	255	295	335	375	415	455
Počet zkoušek n_j	7	22	102	260	386	461	356
Mez pevnosti x_j^* (kg/cm^2)	495	535	575	615	655	Σ	
Počet zkoušek n_j	239	108	40	15	4	2000	

Pro výpočet potřebujeme následující statistiky:

$$\begin{aligned}
&m_1 = 0,041, \quad \mu_2 = 3,155, \quad \sigma = 1,752,\\
&m_2 = 3,157, \quad \breve{\mu}_2 = 3,072, \quad r_3 = 0,205,\\
&m_3 = 1,490, \quad \mu_3 = 1,102, \quad r_4 = 3,021,\\
&m_4 = 30,271, \quad \mu_4 = 30,059,\\
&\overline{x} = 416,64, \quad \breve{\mu}_4 = 28,510.
\end{aligned}$$

Statistiky $\breve{\mu}_2$, $\breve{\mu}_4$ jsou odhady centrálních momentů pro původní netříděná data použitím Sheppardových korekcí pro třídění [1]:

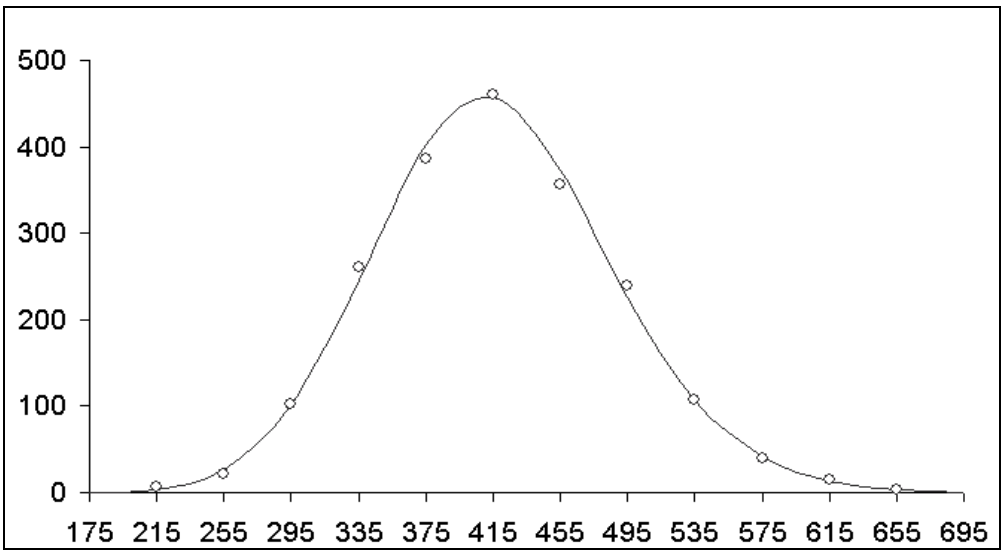
$$\begin{aligned}
\breve{\mu}_2 &\doteq \mu_2 - \frac{1}{12} h^2,\\
\breve{\mu}_3 &\doteq \mu_3,\\
\breve{\mu}_4 &\doteq \mu_4 - \frac{1}{2}\mu_2 h^2 + \frac{7}{240} h^4.
\end{aligned}$$

Do vztahu s prvními třemi členy dosadíme normované momenty r_3 a r_4 a obdržíme

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

Postupný výpočet četností z tohoto vzorce je v dále uvedené tabulce. Vykreslením původních četností n_j a křivky prokládající vypočtené četnosti $\tilde{n}_j$ do grafu dostáváme výsledný tvar rozdělení.

x_j^*	n_j	$x' - m_1$	$x = \frac{x' - m_1}{\sigma}$	$f(x)$	$f^{(3)}(x)$	$f^{(4)}(x)$	$f_A(x)$	$\tilde{n}_j = \frac{n}{\sigma} f_A(x)$
175	0	-6,041	-3,4480	0,0010	0,0320	0,0763	1,7E-05	0,01
215	7	-5,041	-2,8772	0,0063	0,0965	0,1389	0,0031	3,62
255	22	-4,041	-2,3065	0,0279	0,1493	-0,0172	0,0227	26,01
295	102	-3,041	-1,7357	0,0884	0,0019	-0,5306	0,0879	100,36
335	260	-2,041	-1,1649	0,2024	-0,3873	-0,6681	0,2150	245,49
375	386	-1,041	-0,5941	0,3343	-0,5259	0,3365	0,3526	402,56
415	461	-0,041	-0,0234	0,3988	-0,0280	1,1951	0,4008	457,57
455	356	0,959	0,5473	0,3434	0,5076	0,4437	0,3264	372,69
495	239	1,959	1,1181	0,2135	0,4177	-0,6273	0,1986	226,81
535	108	2,959	1,6889	0,0958	0,0238	-0,5729	0,0945	107,89
575	40	3,959	2,2597	0,0310	-0,1478	-0,0485	0,0360	41,16
615	15	4,959	2,8304	0,0072	-0,1030	0,1388	0,0109	12,45
655	4	5,959	3,4012	0,0012	-0,0357	0,0827	0,0025	2,87
695	0	6,959	3,9720	0,0001	-0,0075	0,0235	0,0004	0,49
Σ	2000	---	---	1,7518	---	---	1,7520	2000,04



Z tabulky i grafu je zřejmá velmi dobrá aproximace původního neznámého rozdělení.

Rozdělení typu B

Pravděpodobnostní funkce má tvar [3]

$$p_B(x) = \psi(x) \sum_{m=0}^{\infty} b_m G_m(x),$$

kde $\psi(x)$ je pravděpodobnostní funkce Poissonova rozdělení

$$\psi(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots ,$$

a polynom $G_m(x)$ je analogie Čebyševova – Hermitova polynomu $H_n(x)$. V polynomech $G_n(x)$ se místo normální funkce rozdělení bere funkce $\psi(x)$ a místo *n*-té derivace *n*-tá diference

$$\Delta^n \psi(x-n) = \psi(x) - \frac{n^{[1]}}{1!}\psi(x-1) + \frac{n^{[2]}}{2!}\psi(x-2) - \ldots + (-1)^n \psi(x-n),$$

kde $n^{[k]}$ je variace *k*-té třídy z *n* prvků bez opakování.

Polynomy $G_n(x)$, závislé na proměnné *x*, jsou pro hodnoty 0, 1, 2, ... , definované rovností

$$G_n(x) = (-1)^n \frac{\Delta^n \psi(x-n)}{\psi(x)},$$

takže

$$G_n(x) = (-1)^n \frac{n!}{\lambda^n} \sum_{h=0}^{n} \frac{(-1)^h \lambda^h}{h!} \frac{x^{[n-h]}}{(n-h)!}.$$

Koeficienty b_m najdeme pomocí polynomů $G_m(x)$ podobně jako byly nalezeny koeficienty a_n pomocí polynomů $H_n(x)$. Platí

$$b_0 = 1, \quad b_1 = 0, \quad b_2 = \frac{1}{2!}(\mu_2 - \lambda),$$

$$b_3 = -\frac{1}{3!}(\mu_3 - 3\mu_2 + 2\lambda),$$

$$b_4 = \frac{1}{4!}\left[\mu_4 - 6\mu_3 + (11 - 6\lambda)\mu_2 - 3\lambda(2-\lambda)\right].$$

Pak je pravděpodobnostní funkce pro rozdělení typu *B*

$$p_B(x) = \frac{\lambda^x}{x!}e^{-\lambda}\left\{1 + \frac{\mu_2 - \lambda}{\lambda^2}\left[\frac{x^{[2]}}{2} - \lambda x^{[1]} + \frac{\lambda^2}{2}\right] + \frac{\mu_3 - 3\mu_2 + 2\lambda}{\lambda^3}\left[\frac{x^{[3]}}{6} - \frac{\lambda}{2}x^{[2]} + \frac{\lambda^2}{2}x^{[1]} - \frac{\lambda^3}{6}\right] + \right.$$

$$\left. + \frac{\mu_4 - 6\mu_3 + (11 - 6\lambda)\mu_2 - 3\lambda(2-\lambda)}{\lambda^4}\left[\frac{x^{[4]}}{24} - \frac{\lambda}{6}x^{[3]} + \frac{\lambda^2}{4}x^{[2]} - \frac{\lambda^3}{6}x^{[1]} + \frac{\lambda^4}{24}\right] + \ldots\right\}.$$

Četnost v jednotlivých třídách roztříděného souboru je

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

Příklad

Odhadněte diskrétní rozdělení z následujících dat, kde četnosti udávají počty α-částic emitovaných poloniem v konstantních časových intervalech (1/8 minuty) [3, 5]:

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

Pro výpočet potřebujeme statistiky

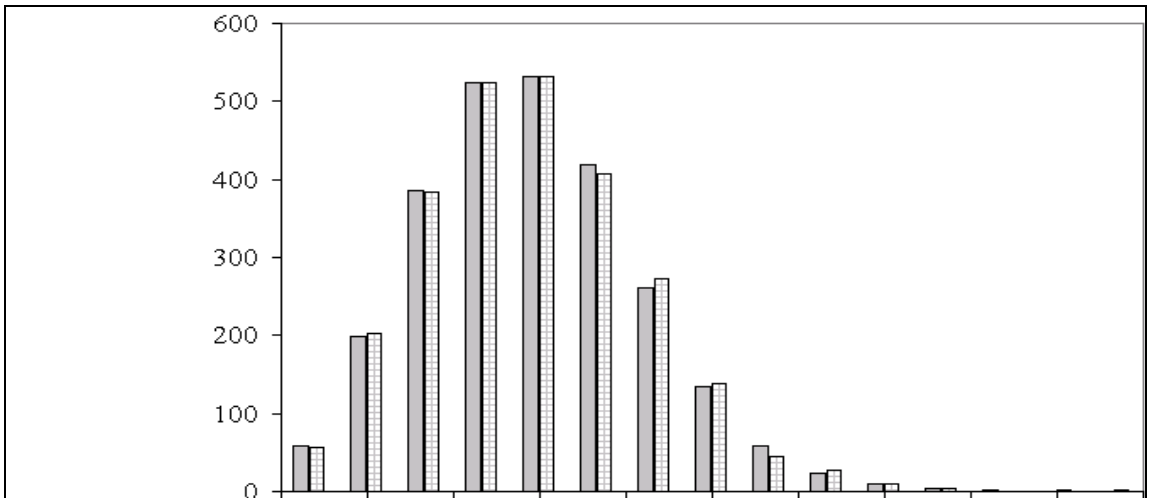
$$m_1 = -0{,}1285, \quad \lambda = 3{,}872,$$
$$m_2 = 3{,}7113, \quad \mu_2 = 3{,}695,$$
$$m_3 = 1{,}9720, \quad \mu_3 = 3{,}398,$$
$$m_4 = 46{,}4889, \quad \mu_4 = 47{,}869.$$

Pak je

$$\tilde{n}_j = 2608\frac{3{,}872^{x_j}}{x_j!}e^{-3{,}872}\left\{1 - 0{,}0118\left[\frac{x_j^{[2]}}{2} - 3{,}872x_j^{[1]} + 7{,}4962\right] + \right.$$
$$+ 0{,}001\left[\frac{x_j^{[3]}}{6} - 1{,}936x_j^{[2]} + 7{,}4962x_j^{[1]} - 9{,}6751\right] +$$
$$\left. + 0{,}0179\left[\frac{x_j^{[4]}}{24} - 0{,}6353x_j^{[3]} + 3{,}7481x_j^{[2]} - 9{,}6751x_j^{[1]} + 9{,}3655\right]\right\} =$$
$$= 2608\frac{3{,}872^{x}}{x_j!}e^{-3{,}872}\left\{1{,}0695 - 0{,}12x_j^{[1]} + 0{,}0593x_j^{[2]} - 0{,}0114x_j^{[3]} + 0{,}00075x_j^{[4]}\right\}.$$

Výpočet četností z tohoto vzorce je v následující tabulce. Vykreslením odhadnutých četností $\tilde{n}_j$ a původních četností n_j obdržíme dále uvedený sloupcový graf .

j	n_j	x_j	$\frac{n_j}{n}$	$1,0695 - 0,12x_j^{[1]} + 0,0593x_j^{[2]} - -0,0114x_j^{[3]} + 0,00075x_j^{[4]}$	$p_B\left(x_j\right)$	$\tilde{n}_j$
0	57	0	0,020817	1,0695	0,0222	58,06
1	203	1	0,080602	0,9495	0,0765	199,59
2	383	2	0,156046	0,9481	0,1479	385,84
3	525	3	0,201403	0,9969	0,2007	523,63
4	532	4	0,194958	1,0455	0,2038	531,58
5	408	5	0,150976	1,0615	0,1602	417,96
6	273	6	0,097430	1,0305	0,1004	261,84
7	139	7	0,053893	0,9561	0,0515	134,38
8	45	8	0,026084	0,8599	0,0224	58,49
9	27	9	0,011222	0,7785	0,0087	22,78
10	10	10	0,004345	0,7785	0,0033	8,82
11	4	11	0,001529	0,9265	0,0014	3,69
12	0	12	0,000494	1,3191	0,0006	1,69
13	1	13	0,000147	2,0679	0,0003	0,79
14	1	14	4,07E-05	3,3025	0,0001	0,35
Σ	2608	---	1,000000	---	1,0005	2609,54



	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

Z tabulky i grafu je zřejmá velmi dobrá aproximace původního neznámého rozdělení.

3.3 Johnsonovy křivky

Jinou soustavu odhadů spojitých rozdělení pravděpodobnosti pomocí čtyřparametrických distribučních funkcí navrhl N. L. Johnson [4, 5]. Tyto odhady vychází z nelineární transformace normálního rozdělení. V aplikacích se ukazuje, že často vystačíme se třemi jednoduchými typy nelineární transformace:

$$z_L = \exp x\,,\ z_B = \frac{\exp x}{1+\exp x}\,,\ z_U = \sinh x\,.$$

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. Transformace

$$z_L = b\exp\left(\frac{x-k}{m}\right)+a$$

převádí normální hustotu pravděpodobnosti

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

na hustotu 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 $z > a$, $m > 0$, $b > 0$, $k \in \mathbb{R}$, $a \in \mathbb{R}$. Jde vlastně o tříparametrické lognormální rozdělení s prahovou hodnotou a, které se označuje jako Johnsonovo rozdělení typu S_L (L = lognormal).

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.

3.4 Fitování rozdělení pomocí kvazinorem

Odhadujeme diskrétní rozdělení pravděpodobnosti za vedlejších momentových podmínek tak, aby vzdálenost vypočteného rozdělení $\mathbf{p} = (p_1, \ldots, p_m)$ od nějakého daného rozdělení $\mathbf{q} = (q_1, \ldots, q_m)$ byla minimální. K odhadu můžeme použít ***Hellingerovu vzdálenost***

$$\mathrm{D}(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{m} \left(\sqrt{p_j} - \sqrt{q_j} \right)^2$$

nebo ***Shannonovu pseudovzdálenost***

$$\mathrm{S}(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{m} \left(p_j \ln p_j - q_j \ln q_j \right)$$

rozdělení $\mathbf{p} = (p_1, \ldots, p_m)$ a $\mathbf{q} = (q_1, \ldots, q_m)$ [6, 7, 8].

Pro odhad předpokládáme, že pozorovaná diskrétní náhodná veličina X nabývá konečně mnoha navzájem různých reálných hodnot x_j^* s neznámými pravděpodobnostmi

$$p_j = P\left(X = x_j^* \right),\ j = 1, \ldots, m,\ m > 1.$$

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

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

kde f_j je absolutní četnost pozorované hodnoty x_j^*.

Hellingerova kvazinorma

Rozdělení pravděpodobnosti $\mathbf{p} = (p_1, \ldots, p_m)$ diskrétní náhodné veličiny X má na pravděpodobnostním prostoru $(\Omega, \mathbf{p})$, kde $\Omega = \{ x_1^*, \ldots, x_m^* \}$ a $m > 1$, tzv. ***minimální Hellingerovu kvazinormu za K počátečních momentových podmínek***

$$\sum_{j=1}^{m} p_j x_j^{*k} = M_k,\ k = 0, \ldots, K,$$

jestliže jeho Hellingerova vzdálenost

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

od rozdělení pravděpodobnosti $\mathbf{p}_0=\left(\frac{1}{m},\ldots,\frac{1}{m}\right)$ je minimální pro

$$M_k=\sum_{j=1}^{m}\frac{f_j}{n}x_j^{*k},\ k=0,\ldots,K.$$

Pak je

$$p_j=\frac{1}{m\left(1+\sum_{k=0}^{K}\lambda_k x_j^{*k}\right)^2},\ j=1,\ldots,m,$$

kde $\lambda_k,\ k=0,\ldots,K$, jsou Lagrangeovy multiplikátory pro Lagrangeovu funkci

$$\Lambda(\mathbf{p},\boldsymbol{\lambda})=\mathrm{D}(\mathbf{p},\mathbf{p}_0)+\sum_{k=0}^{K}\lambda_k\left(\sum_{j=1}^{m}p_j x_j^{*k}-M_k\right)$$

a $\boldsymbol{\lambda}=(\lambda_0,\ldots,\lambda_K)$.

Lagrangeovy multiplikátory λ_k je možno určit pomocí nelineární soustavy rovnic anebo přímo aplikovat nelineární optimalizaci. Např. pro $K=0$ je $\lambda_{0,1}=0$, resp. $\lambda_{0,2}=-2$, a $p_j=\frac{1}{m},\ j=1,\ldots,m$, a $\mathrm{D}(\mathbf{p},\mathbf{p}_0)=0$. Pro $K=m$ jde o interpolaci $p_j=\frac{f_j}{m},\ j=1,\ldots,m$.

Hellingerova vzdálenost umožňuje zavést nepříliš známý tzv. Hellingerův – Pitmanův test shody, který spočívá ve skutečnosti, že statistika

$$4n\,\mathrm{D}(\mathbf{p},\mathbf{f})=4n\sum_{j=1}^{m}\left(\sqrt{p_j}-\sqrt{\frac{f_j}{n}}\right)^2$$

má pro $n\to\infty$ asymptoticky rozdělení chí-kvadrát s $m-k-1$ stupni volnosti. Postupným přidáváním momentových podmínek a opakovaným odhadem rozdělení pravděpodobnosti pomocí minimální Hellingerovy kvazinormy lze určit minimální potřebný počet *K* těchto podmínek tak, aby platilo $\chi^2\le\chi^2_{1-\alpha}$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$ - kvantil rozdělení chí−kvadrát s daným počtem stupňů volnosti pro hladinu významnosti α [7, 8].

Shannonova kvazinorma

Rozdělení pravděpodobnosti $\mathbf{p}=(p_1,\ldots,p_m)$ diskrétní náhodné veličiny *X*, má na pravděpodobnostním prostoru $(\Omega,\mathbf{p})$, kde $\Omega=\{x_1^*,\ldots,x_m^*\}$ a $m>1$, tzv. ***minimální Shannonovu kvazinormu za K momentových podmínek***

$$\sum_{j=1}^{m}p_j x_j^{*k}=M_k,\ k=0,\ldots,K,$$

jestliže jeho Shannonova pseudovzdálenost

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

od rozdělení pravděpodobnosti $\mathbf{p}_0=\left(\frac{1}{m},\ldots,\frac{1}{m}\right)$ je minimální pro

$$M_k = \sum_{j=1}^{m} \frac{f_j}{n} x_j^{*k}, \ k = 0,\ldots,K.$$

Pak je

$$p_j = \exp\left(-1 - \sum_{k=0}^{K} \lambda_k x_j^{*k}\right), j = 1,\ldots,m,$$

kde λ_k, $k = 0,\ldots,K$, jsou Lagrangeovy multiplikátory pro Lagrangeovu funkci

$$\Lambda(\mathbf{p},\boldsymbol{\lambda}) = \mathrm{S}(\mathbf{p},\mathbf{p}_0) + \sum_{k=0}^{K} \lambda_k \left(\sum_{j=1}^{m} p_j x_j^{*k} - M_k \right),$$

a $\boldsymbol{\lambda} = (\lambda_0,\ldots,\lambda_K)$.

Lagrangeovy multiplikátory λ_k je možno opět určit pomocí nelineární soustavy rovnic anebo přímo aplikovat nelineární optimalizaci. Navíc jsou odhady parametrů λ_k maximálně věrohodné, neboť jde současně o odhady parametrů modifikovanou metodou minimálního chí–kvadrát. Dále pak můžeme aplikovat chí-kvadrát test shody rozdělení. Postupným přidáváním momentových podmínek a opakovaným odhadem rozdělení pravděpodobnosti pomocí minimální Shannonovy kvazinormy lze určit minimální potřebný počet *K* těchto podmínek tak, aby platilo $\chi^2 \le \chi^2_{1-\alpha}$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$- kvantil rozdělení chí–kvadrát s $m-k-1$ stupňů volnosti pro hladinu významnosti α. Navíc má každé takto postupně získané rozdělení pravděpodobnosti $\mathbf{p}(\boldsymbol{\lambda})$ vždy maximální entropii pro dané momentové podmínky [7, 8].

Příklad

Sledováním určitého statistického znaku (diskrétní náhodné veličiny X) jsme získali statistický soubor o rozsahu $n = 100$. Po jeho roztřídění jsme obdrželi diskrétní empirické rozdělení náhodné veličiny X dané následující tabulkou, kde jsou x_j^* středy tříd a f_j pozorované absolutní četnosti [7]:

x_j^*	1	2	3	4	5	6	7
f_j	6	10	14	18	22	20	10

Hledáme minimum Hellingerovy, resp. Shannonovy kvazinormy, za vedlejších podmínek daných prvními třemi obecnými momenty M_0, M_1, M_2 od rozdělení $\mathbf{p}_0 = \left(\frac{1}{m},\ldots,\frac{1}{m}\right)$.

Z rovnosti mezi teoretickými a empirickými obecnými momenty plyne, že

$$M_0 = \frac{1}{n}\sum_{j=1}^{m} f_j = \frac{1}{100}\sum_{j=1}^{7} f_j = 1, \quad M_1 = \frac{1}{n}\sum_{j=1}^{m} f_j x_j^* = \frac{1}{100}\sum_{j=1}^{7} f_j x_j^* = 4{,}4,$$

$$M_2 = \frac{1}{n}\sum_{j=1}^{m} f_j x_j^{*2} = \frac{1}{100}\sum_{j=1}^{7} f_j x_j^{*2} = 22{,}2.$$

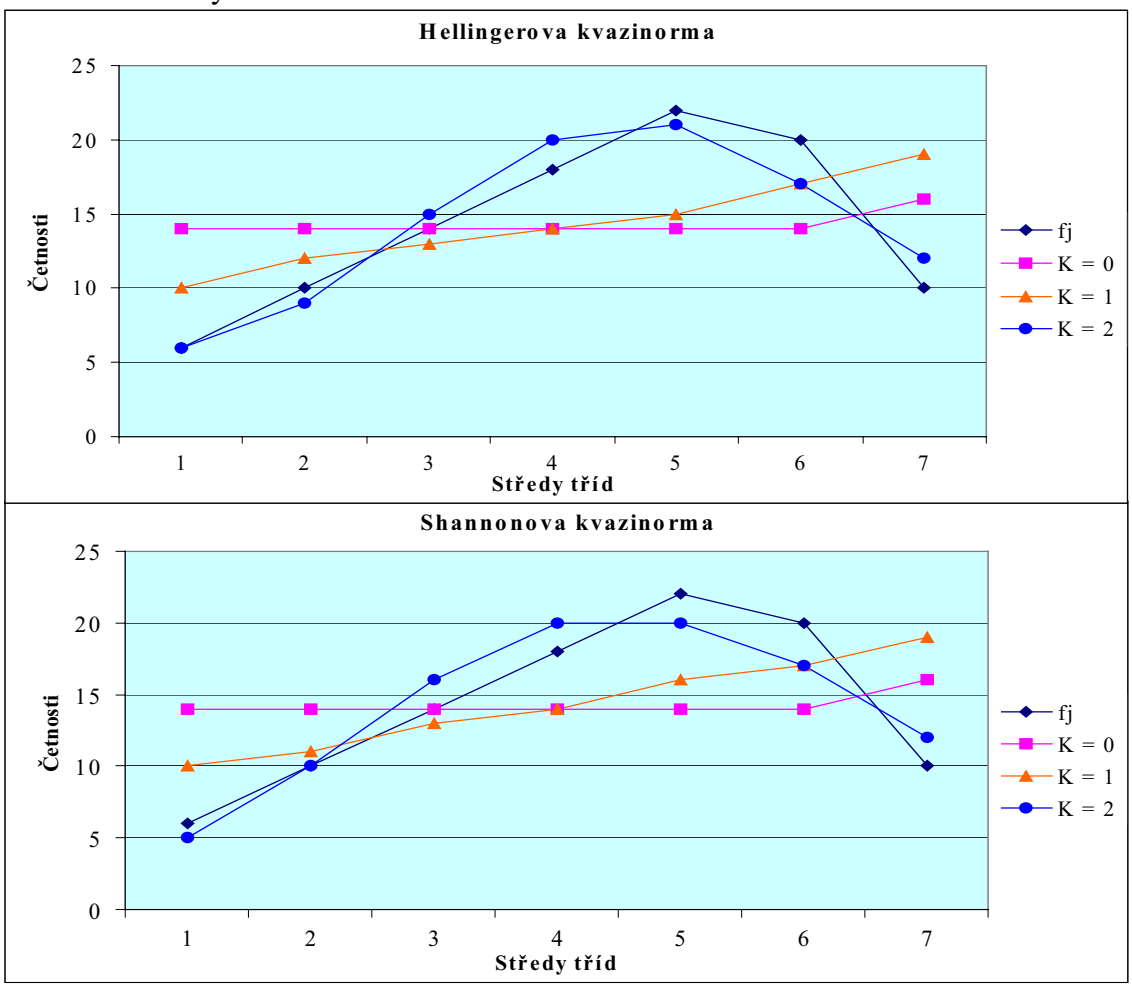
Výpočty byly provedeny pomocí speciálního softwaru s optimalizačním jádrem z programu GAMS. Výsledky za postupného přidávání momentových podmínek pro obě kvazinormy jsou ilustrovány následující tabulkou.

Vzdálenost	M_0	$M_0 + M_1$	$M_0 + M_1 + M_2$
$\mathrm{D}(\mathbf{p}, \mathbf{p}_0)$	0	0,0100505	0,0382016
$\mathrm{S}(\mathbf{p}, \mathbf{p}_0)$	0	0,0200841	0,0713272

V následující tabulce jsou výsledky:

Četnosti\Třídy		1	2	3	4	5	6	7
f_j		6	10	14	18	22	20	10
f_j^H	$K = 0$	14	14	14	14	14	14	16
	$K = 1$	10	12	13	14	15	17	19
	$K = 2$	6	9	15	20	21	17	12
f_j^S	$K = 0$	14	14	14	14	14	14	16
	$K = 1$	10	11	13	14	16	17	19
	$K = 2$	5	10	16	20	20	17	12

Grafická ilustrace výsledků:



Z tabulky i grafů jsou zřejmé dobré aproximace původního neznámého rozdělení pomocí obou kvazinorem pro $K = 2$.

3.5 Jádrové odhady

Jádrové odhady hustoty spojitého rozdělení pravděpodobnosti [10] vycházejí z tzv. ***jádrové funkce*** K, což je nezáporná funkce na $\mathbb{R}$ vyhovující podmínce

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

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

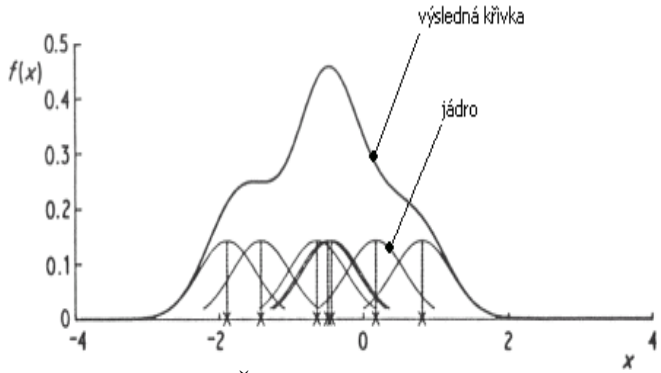
$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h} \right),$$

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.

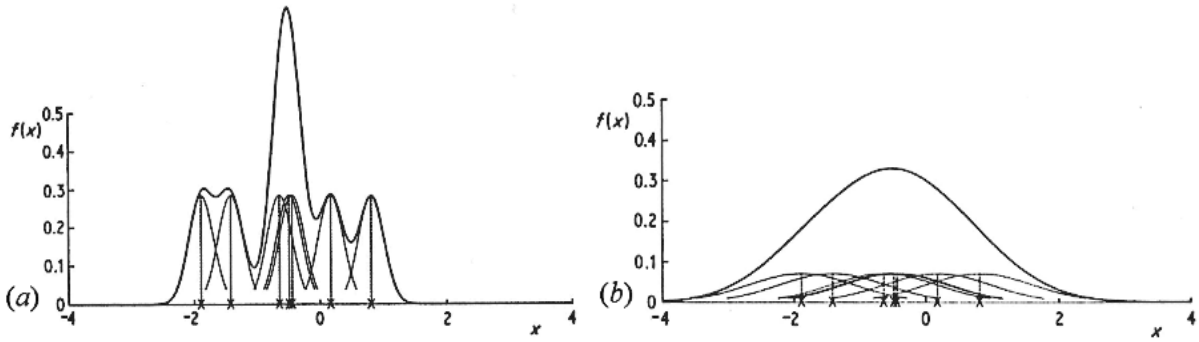
Nejčastěji se užívají jádra:

- ***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*** $K(x) = \begin{cases} \frac{15}{16}\left(1 - x^2\right)^2 & \text{pro } |x| < 1 \\ 0 & \text{jinde} \end{cases}$
- ***Obdélníkové jádro*** $K(x) = \begin{cases} \frac{1}{2} & \text{pro } |x| < 1 \\ 0 & \text{jinde} \end{cases}$

Vliv šířky vyhlazovacího okna na vyhlazení je ilustrován na následujících obrázcích:

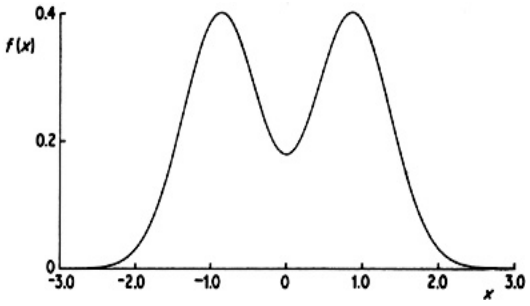


Šířka okna 0,4.

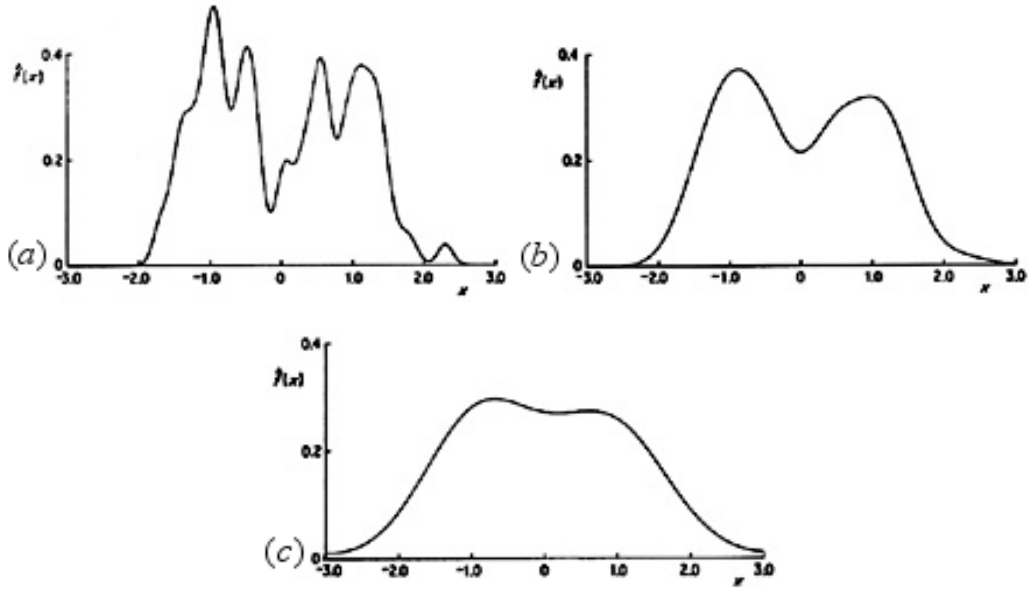


Šířka okna (a) 0,2; (b) 0,8.

Vliv šířky jádra také ilustrují simulovaná data podle bimodální hustoty:



Odpovídající jádrové odhady s Gaussovým jádrem pro šířky okna (a) 0,1, (b) 0,3, (c) 0,6 jsou:



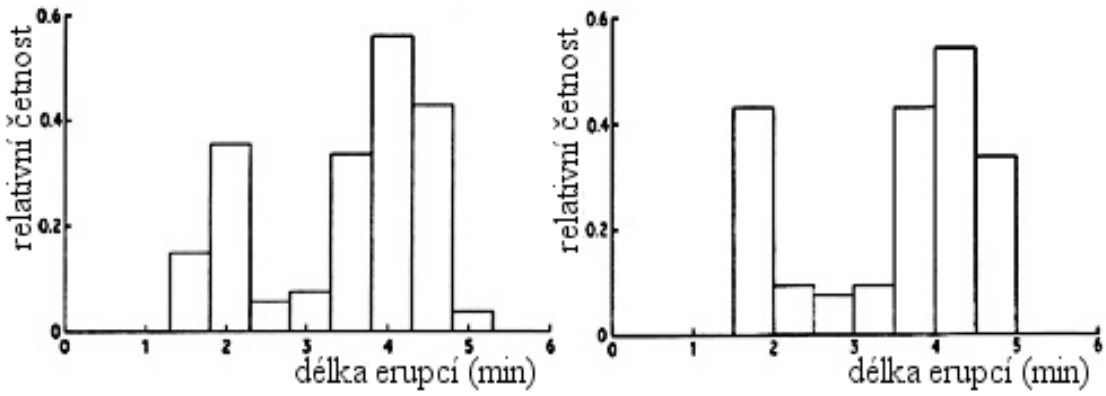
Z předcházejících obrázků je zřejmé, že k základním problémům aplikace jádrových odhadů patří volba jádra a šířka vyhlazovacího okna. Dále pak jde o respektování požadavků spojitosti, příp. hladkosti získané hustoty pravděpodobnosti a důležité je také vyjádření odpovídající distribuční funkce a snadnost výpočtu kvantilů, příp. pokrytí celého rozsahu hodnot náhodné veličiny X, neboť jádrové odhady mají někdy charakter šumu na koncích rozdělení.

Příklad

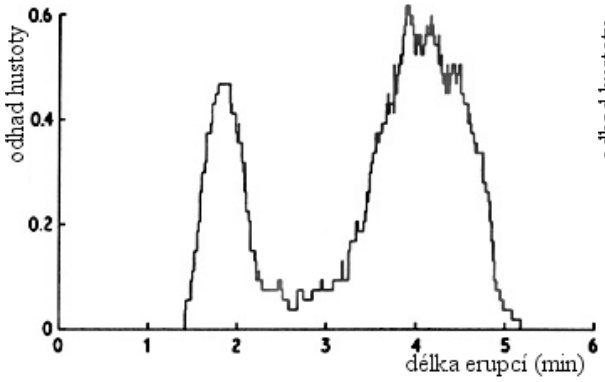
Data v tabulce vyjadřují délky erupcí x_i (v minutách), $i = 1, \ldots, 107$, gejzíru Old Faithful v Yellowstonském národním parku (USA):

4,37	3,87	4,00	4,03	3,50	4,08	2,25	4,70	1,73	4,93	1,73	4,62	3,43	4,25
1,68	3,92	3,68	3,10	4,03	1,77	4,08	1,75	3,20	1,85	4,62	1,97	4,50	3,92
4,35	2,33	3,83	1,88	4,60	1,80	4,73	1,77	4,57	1,85	3,52	4,00	3,70	3,72
4,25	3,58	3,80	3,77	3,75	2,50	4,50	4,10	3,70	3,80	3,43	4,00	2,27	4,40
4,05	4,25	3,33	2,00	4,33	2,93	4,58	1,90	3,58	3,73	3,73	1,82	4,63	3,50
4,00	3,67	1,67	4,60	1,67	4,00	1,80	4,42	1,90	4,63	2,93	3,50	1,97	4,28
1,83	4,13	1,83	4,65	4,20	3,93	4,33	1,83	4,53	2,03	4,18	4,43	4,07	4,13
3,95	4,10	2,72	4,58	1,90	4,50	1,95	4,83	4,12					

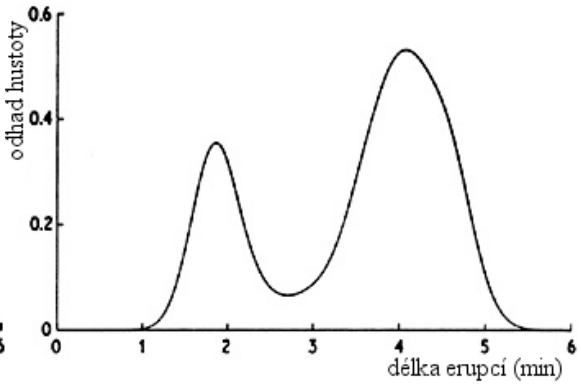
Uvedená data jsou hodnoty náhodné veličiny X s neznámým rozdělením pravděpodobnosti a typická pro tuto náhodnou veličinu je bimodalita jejího rozdělení. Pro modelování její hustoty pravděpodobnosti by bylo možno uvažovat směs dvou rozdělení, avšak v tomto případě by bylo nutno expertně odhadnout jejich tvar. Dokládají to také následující histogramy, které se liší volbou pokrytí třídami:



Výpočtem pomocí WWW apletu [11] obdržíme následující jádrové odhady:

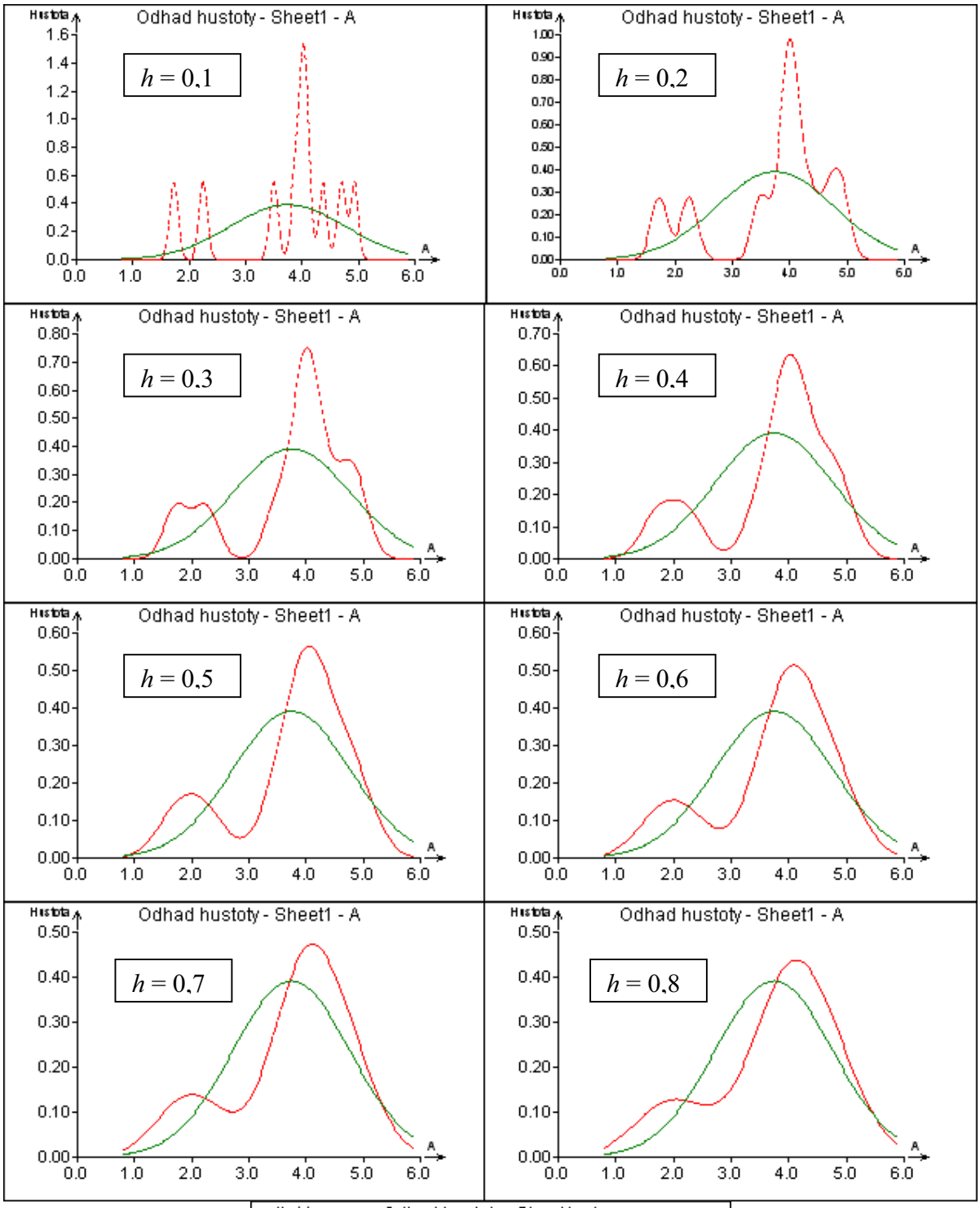


Odhad pomocí obdélníkového jádra.



Odhad pomocí Gaussova jádra se šířkou okna 0,25.

Na následujících obrázcích jsou znázorněny jádrové odhady hustoty pro uvedená data pomocí Gaussova jádra pořízené demoverzí programu QCExpert pouze z prvních deseti hodnot délky erupcí x_i gejzíru Old Faithful z výše uvedené tabulky pro různé šířky jádra h (hustota normálního rozdělení je pro porovnání různých měřítek). Za optimální lze vzít šířku okna h lze vzít hodnotu od 0,4 do 0,6.


Odhad hustoty - Sheet1 - A
h = 0,1
h = 0,2
h = 0,3
h = 0,4
h = 0,5
h = 0,6
h = 0,7
h = 0,8
h = 0,9
A

4 Závěr

Empirický přístup k odhadům rozdělení pravděpodobnosti vyžaduje při řešení konkrétních úloh dostatečnou dávku zkušeností a nelze přitom spoléhat na profesionální statistické softwarové produkty, které navíc obsahují pouze nevelké množství různých typů rozdělení. Odhady využívající inferenční přístup k pozorovaným náhodným veličinám jsou rigoróznější a bývají účinnější než empirické odhady, avšak i při jejich použití povětšinou zůstávají problémy s multimodalitou a eliminací extrémně odchýlených hodnot. V současné době se stále více využívají v aplikacích jak ve výzkumu, tak i výrobě, finančnictví a ekonomických modelech. Např. Pearsonovy i Johnsonovy křivky našly uplatnění ve statistickém řízení procesů a jsou implementovány do softwaru (Statistica aj.). Úspěšně se také rozvíjí teorie jádrových odhadů, které přinesly nové postupy pro časové řady a regresní analýzu, a běžně se již užívají pro grafický softwarový fiting hustoty (QCExpert, Statgraphics). Odhadům založeným na vzdálenostech rozdělení pravděpodobnosti, resp. jejich kvazinormách, patří z důvodu jejich flexibility a vhodnosti i pro vícerozměrné statistické soubory blízká budoucnost.

Předložený přehled inferenčních odhadů rozdělení pravděpodobnosti není zdaleka úplný a v současné době se na fitování rozdělení pravděpodobnosti v matematické statistice usilovně pracuje.

Literatura

1. ANDĚL, J. Statistické metody. Praha: MATFYZPRESS, 1993.
2. GRUSKA, G. F., MIRKHANI, K., and LAMBERSON, L. R. Non-Normal data Analysis. Garden City, MI: Multiface Publishing, 1989.
3. MITROPOLSKIJ, A. K. Těchnika statističeskich vyčislenij. Moskva: Izd. Nauka, 1971.
4. HAHN, G. J. – SHAPIRO, S. S. Statistical Models in Engineering. New York: John Wiley. 1967.
5. ADAMČÍK, J. Fitování rozdělení pravděpodobnosti materiálových charakteristik. Diplomová práce. Ústav matematiky FSI VUT v Brně, 2004.
6. PITMAN, E. J. G. Some Basic Theory for Statistical Inference. New York: John Wiley, 1978.
7. KARPÍŠEK, Z. Statistical Properties of Discrete Probability Distributions with Maximum Entropy. *Folia Fac. Sci. Nat. Univ. Masarykianae Brunensis, Mathematica 9*, Brno 2001, ISBN 80-210-2544-1.
8. KARPÍŠEK, Z. – JURÁK, P. Minimální kvazinormy Hellingerova a Shannonova typu. *Sborník celostátního semináře Analýza dat 2002/II.* Lázně Bohdaneč 26.-29.11.2002, ISBN 80-239-0204-0.
9. KARPÍŠEK, Z. Matematika IV – Statistika a pravděpodobnost. Učební text. FSI VUT v CERM Brno, Brno 2003 (druhé doplněné vydání), ISBN 80-214-2522-9.
10. SILVERMAN, B.W. Density Estimation for Statistics and Data Analysis. London: Chapman & Hall, 1986.
11. http://www.stat.sc.edu/rsrch/gasp/

Referát je součástí řešení dílčího výzkumného úkolu národního projektu MŠMT České republiky čís. 1M06047 Centrum pro jakost a spolehlivost výroby (CQR).