ODHADY A TESTY ROZDĚLENÍ PRAVDĚPODOBNOSTI

Zdeněk Karpíšek

Centrum pro jakost a spolehlivost výroby MŠMT ČR

Ústav matematiky FSI VUT v Brně

Katedra aplikovaných disciplín Akademie Sting v Brně

1 Základní aspekty a problematika odhadů a testů

Odhady (fitování) rozdělení pravděpodobnosti pozorovaných náhodných veličin a náhodných vektorů mají zásadní význam jak pro odhady parametrů, tak i pro testování statistických hypotéz o parametrech. Podle způsobu jejich realizace je můžeme rozdělit na ***empirické*** a ***inferenční***.

Empirické odhady rozdělení pravděpodobnosti jsou založeny na:

- teoretických a zkušenostních předpokladech a dalších informacích o tvaru rozdělení,
- grafických vyjádřeních statistických souborů pomocí histogramů, polygonů, krabicových grafů, rozptylových grafů aj.

Inferenční (indukční) odhady rozdělení pravděpodobnosti se intenzivně rozvíjí a patří mezi ně:

- **Pearsonovy křivky**,
- **Gramovy – Charlierovy řady**,
- **Johnsonovy křivky**,
- **jádrové odhady**,
- **odhady pomocí kvazinorem**.

Postup empirického i inferenčního odhadování a následná verifikace hypotetického 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 i inferenčních odhadech musíme řešit řadu úkolů:

- **rozhodnout, zda použijeme spojitý anebo diskrétní model s ohledem na přesnost pozorování (měření) a možný obor jejich hodnot,**
- **posoudit šikmost, špičatost a vícemodalitu rozdělení pozorované náhodné veličiny,**
- **posoudit autokorelaci dat v závislosti na pořadí pozorování,**
- **odfiltrovat extrémně odchýlené pozorované hodnoty,**
- **zvážit nutnost respektování dimenze vícerozměrného statistického souboru.**

Odhady rozdělení a testy shody se provádí pomocí statistického software na PC. Jejich kvalita a věrohodnost se vyvíjí a je dosti různá (zejména s ohledem na případné odhady parametrů).

Nejčastěji používané ***testy dobré shody*** (***goodness-of-fit tests***) rozdělení pravděpodobnosti:

- **Pearsonův test (chí kvadrát test) pro jeden výběr a pro více výběrů (kategoriální analýza),**
- **Kolmogorovův – Smirnovův test pro jeden a dva výběry,**
- **modifikovaný Kolmogorovův – Smirnovův test pro jeden výběr (D test),**
- **Shapirův-Wilkův test (W test),**
- **Kuiperův test (V test),**
- **Cramerův – Misesův test (W^2 test),**
- **Watsonův test (U^2 test),**
- **Andersonův – Darlingův test (A^2 test).**

Uvedené testy se týkají až na Pearsonův test pouze odhadů rozdělení náhodných veličin – nikoli náhodných vektorů. Pearsonovu testu je blízký tzv. **Pitmanův-Hellingerův test**.

Některé testy byly původně vyvinuty pro odhady spojitých rozdělení, ale později byly upraveny i pro odhady diskrétních rozdělení.

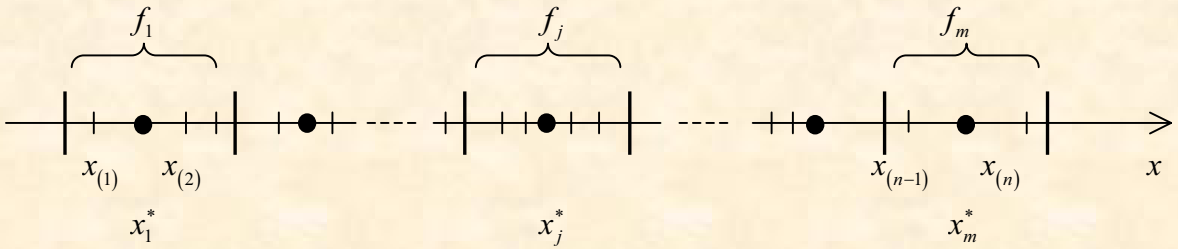
Při testování normality rozdělení se také využívá možnost testovat jeho šikmost (asymetrie) a špičatost (exces).

2 Empirické odhady rozdělení

Pro ***grafické vyjádření*** pozorovaných hodnot jednorozměrné náhodné veličiny X obvykle převádíme získaný ***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 na ***roztříděný statistický soubor*** (variační řadu):

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

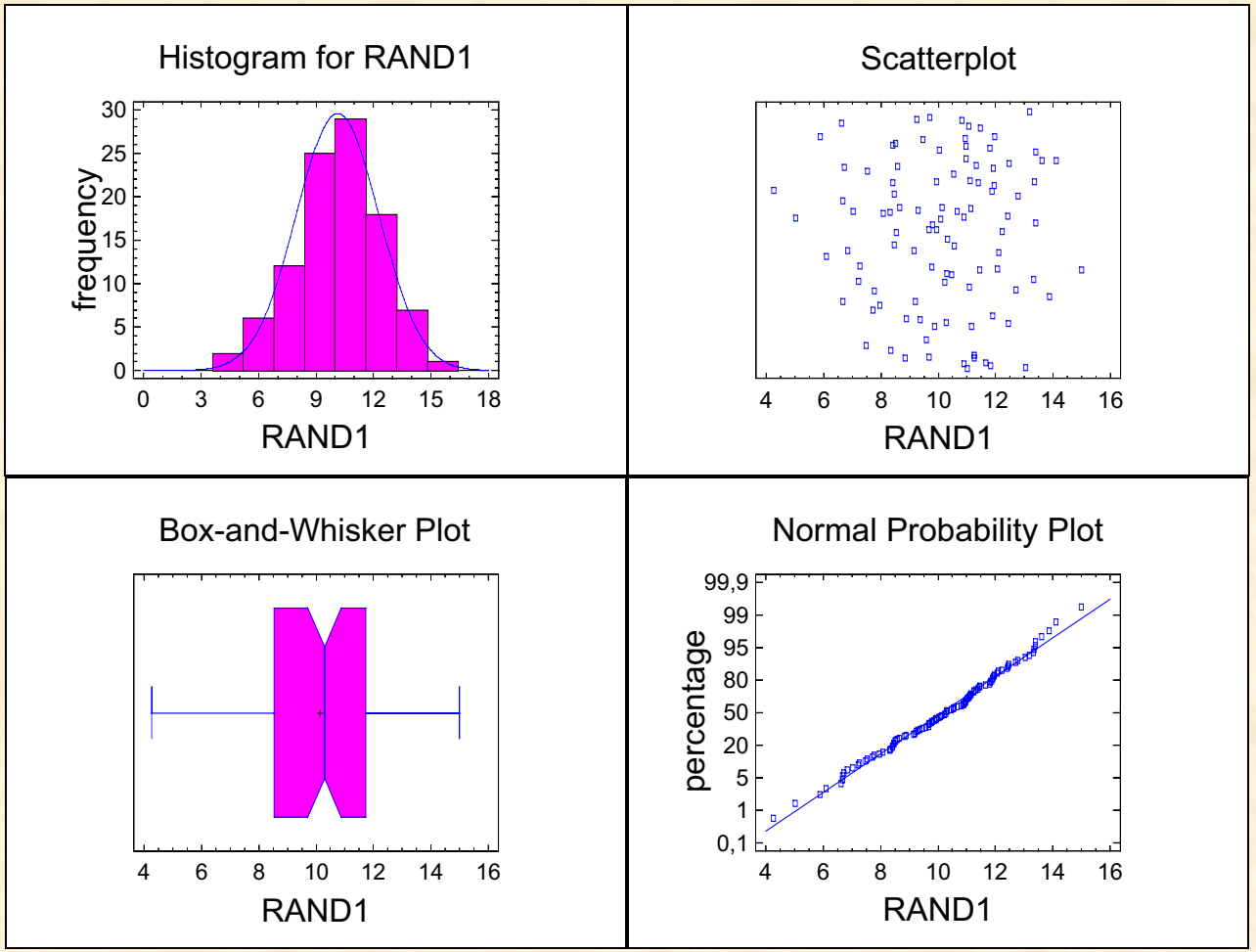
Přitom x_j^* je ***střed třídy***, f_j je ***četnost*** hodnot $x_{(i)}$ v j-té třídě, $j = 1,...,m$; ***třídy*** jsou zleva otevřené a zprava uzavřené intervaly:



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

Příklad normálního rozdělení N(10;2)

Příklad ilustruje základní grafické zpracování statistického souboru o rozsahu 100 získaného generátorem náhodných čísel pro normální rozdělení se střední hodnotou $\mu = 10$ a rozptylem $\sigma^2 = 4$, odhady parametrů a následnou verifikaci pomocí testů shody. Grafy a výpočty byly realizovány pomocí statistického softwaru Statgraphics.



Uncensored Data - RAND1

Analysis Summary

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	7,89310	4	5,00	0,20
7,89310	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	10,1310	6	5,00	0,20
10,1310	10,4024	4	5,00	0,20
10,4024	10,6781	4	5,00	0,20
10,6781	10,9630	4	5,00	0,20
10,9630	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

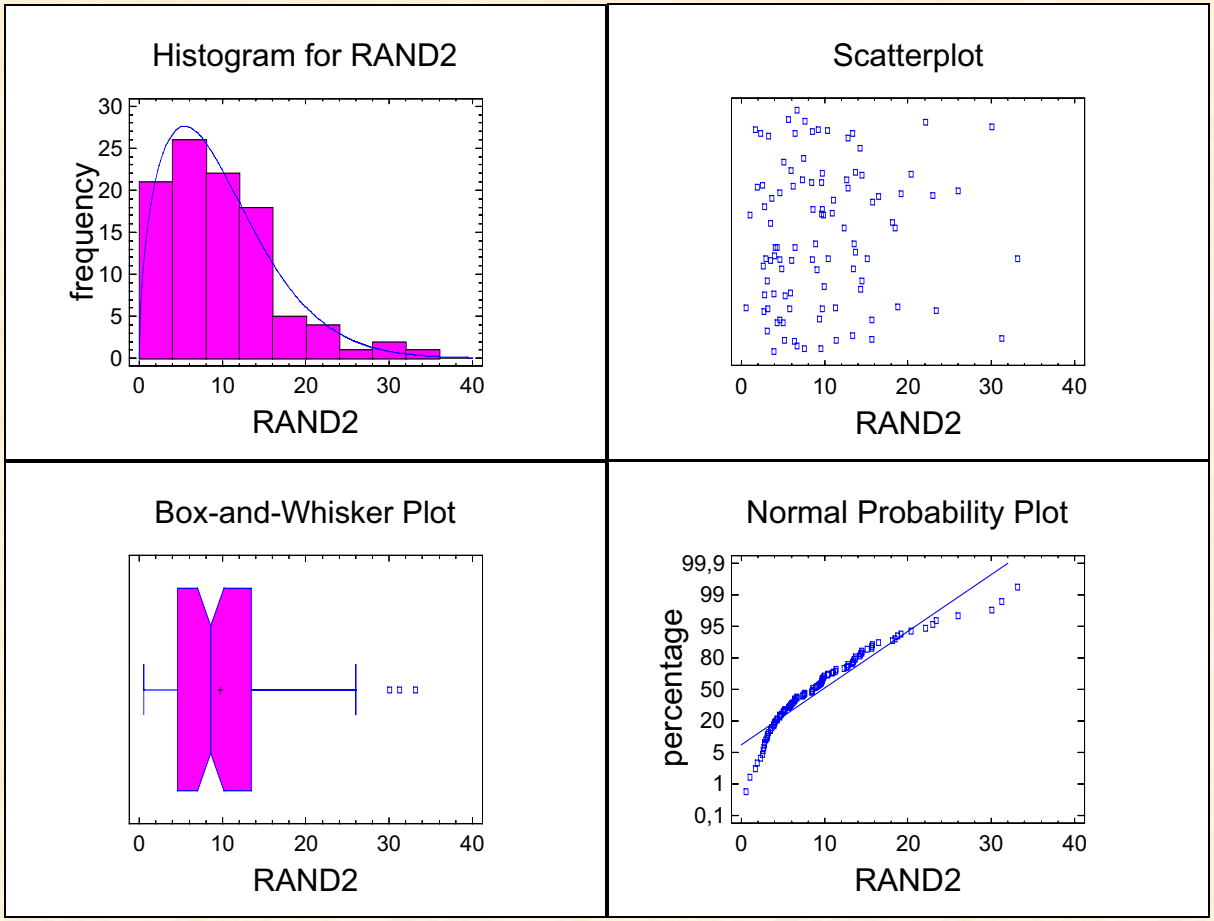
Estimated Kolmogorov statistic DPLUS = 0,0348125
Estimated Kolmogorov statistic DMINUS = 0,0698938
Estimated overall statistic DN = 0,0698938
Approximate P-Value = 0,713002

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 o rozsahu 100 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	4,07890	7	5,00	0,80
4,07890	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	18,7130	2	5,00	1,80
18,7130	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 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 $f(x)$.

Pro určení konkrétního typu křivky použijeme za kritérium veličinu κ, jejíž 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 $y = \kappa(\kappa - 1)$:

I. pro $\kappa < 0$

II. pro $\kappa = 0$ a $r_4 < 3$

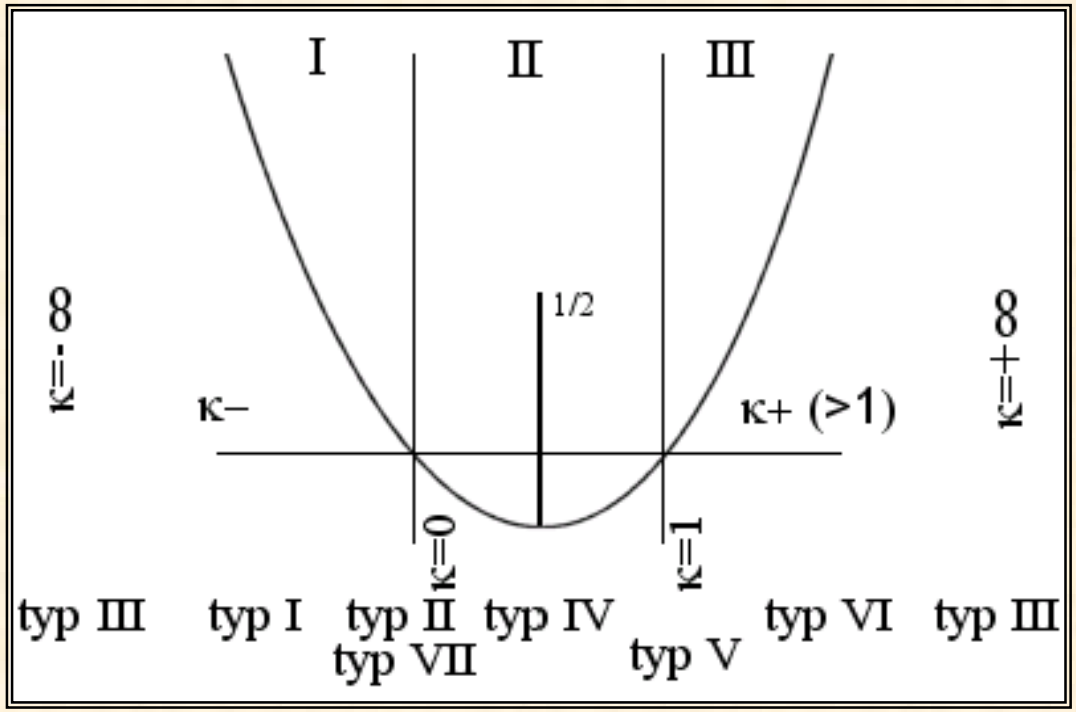
III. pro $\kappa = \pm\infty$

IV. pro $0 < \kappa < 1$

V. pro $\kappa = 1$

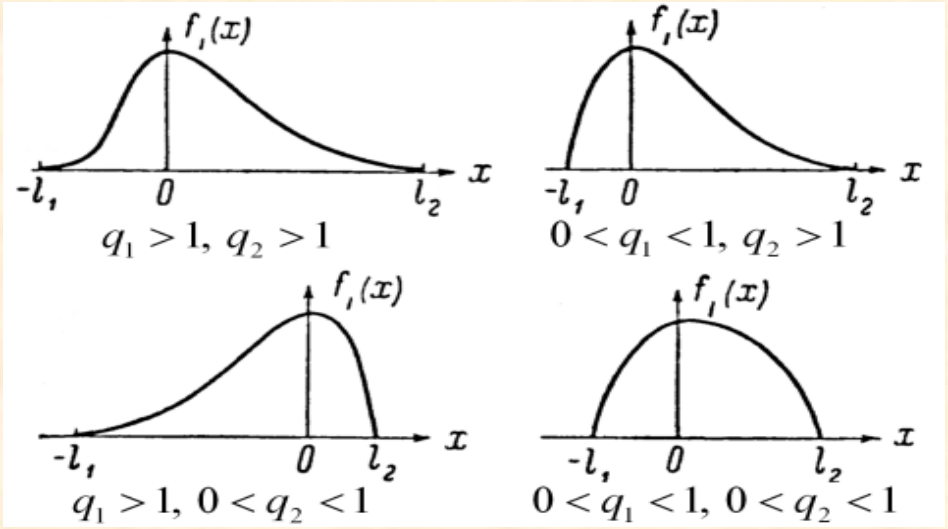
VI. pro $\kappa > 1$

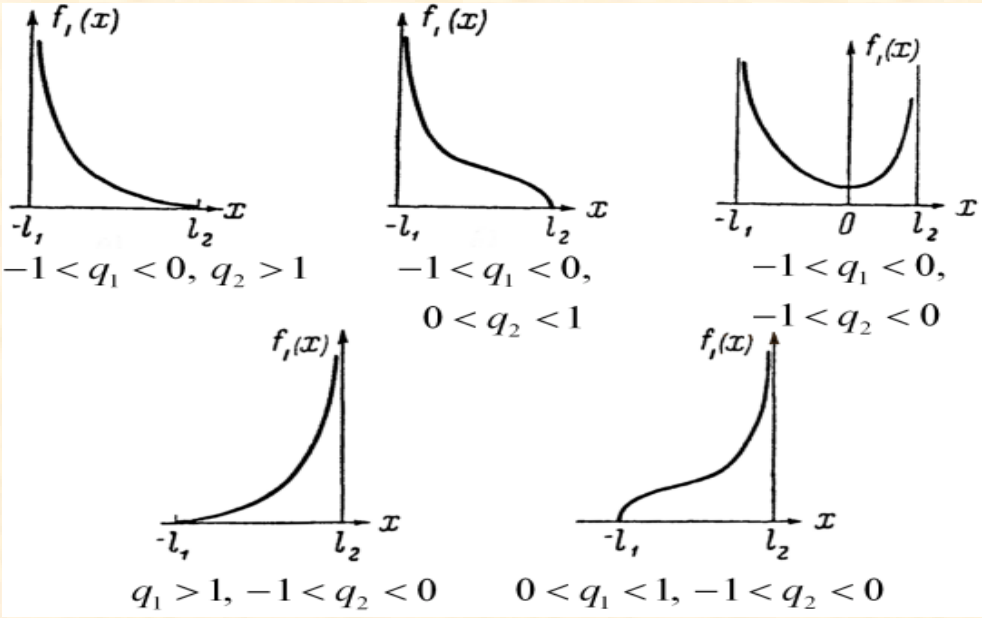
VII. pro $\kappa = 0$ a $r_4 > 3$



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

Ukázka typu I





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

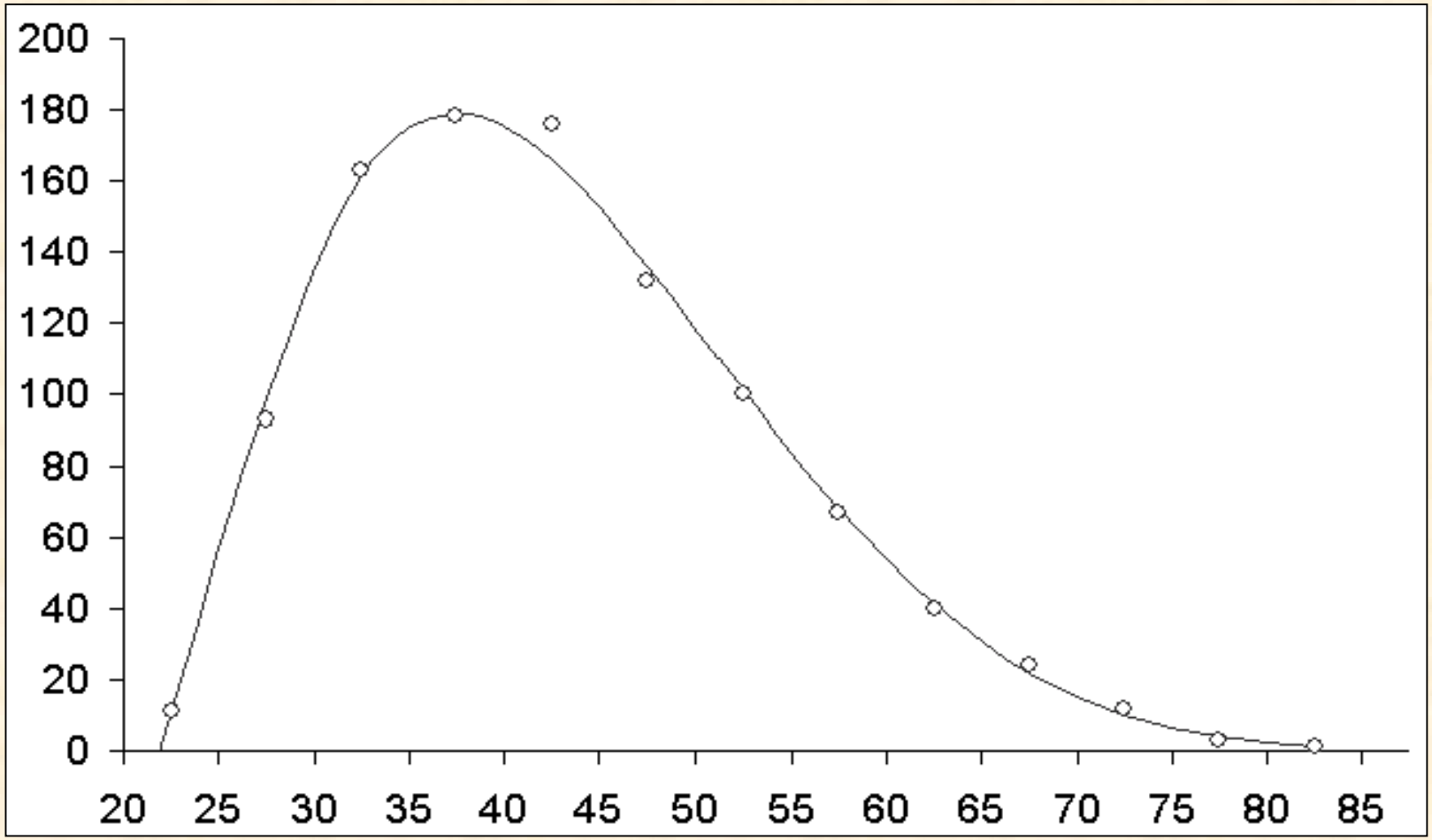
Z tabulky se získají výpočtem popisné charakteristiky daného statistického souboru a kritérium $\kappa = -0,26$. Protože je $\kappa < 0$, použijeme Pearsonovu křivku typu I. Dalším výpočtem obdržíme odhady četnosti

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

Vypočtené četnosti jsou:

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

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



Z 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:

- 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**).

Rozdělení typu A

Odhadovaná hustota pravděpodobnosti má tvar

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

kde

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

je hustota normovaného normálního rozdělení $N(0;1)$ a $f^{(n)}(x)$ je n-tá derivace $f(x)$ podle x.

Po řadě úprav dostaneme pomocí normovaných momentů r_k aproximaci neznámé hustoty

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

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.

Odhady četností pro neznámou náhodnou veličinu v případě roztříděného souboru určíme pro středy tříd x_j^* ze vztahu

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

Zde značíme četnosti n_j, resp. jejich odhady $\tilde{n}_j$, místo f_j, resp. $\tilde{f}_j$. Při rozsahu souboru $n > 400$ se doporučuje aplikovat Sheppardovy korekce odhadů centrálních momentů na roztříděný soubor.

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:

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	

Sheppardovy korekce jsou:

$$\breve{\mu}_2 = \mu_2 - \frac{1}{12}h^2,$$

$$\breve{\mu}_3 = \mu_3,$$

$$\breve{\mu}_4 = \mu_4 - \frac{1}{2}\mu_2 h^2 + \frac{7}{240}h^4.$$

Do vztahu s prvními třemi členy

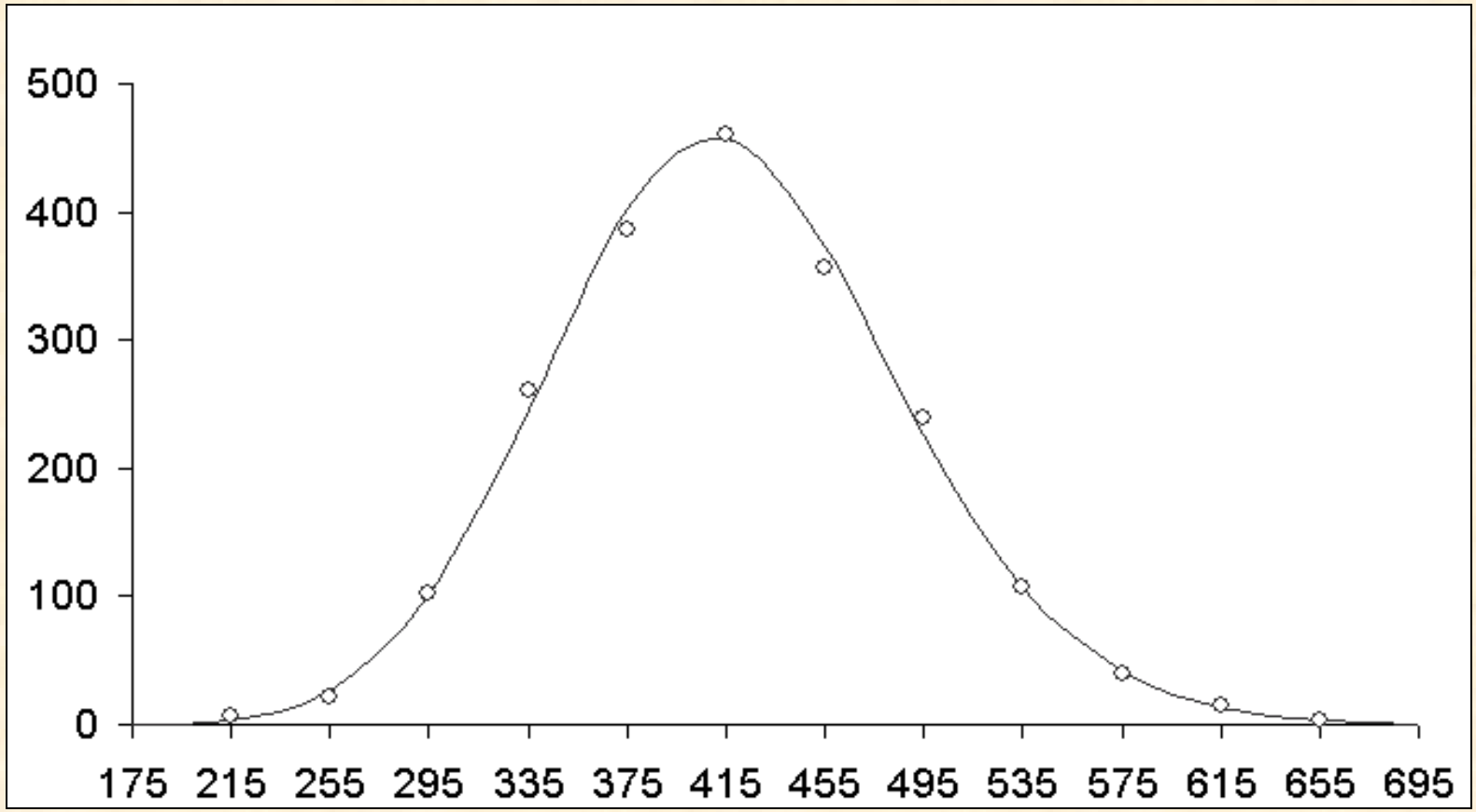
$$f_A(x) = f(x) - \frac{r_3}{6}f^{(3)}(x) + \frac{r_4 - 3}{24}f^{(4)}(x)$$

dosadíme normované momenty r_3 a r_4 a obdržíme odhad hustoty

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

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

Odhadovaná pravděpodobnostní funkce má tvar

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

kde $\psi(x)$ je pravděpodobnostní funkce Poissonova rozdělení $\psi(x) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x = 0, 1, 2, \ldots$, a polynom

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

kde $x^{[k]}$ je variace *k*-té třídy z *x* prvků bez opakování. Pak odhad pravděpodobnostní funkce je

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

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

Odhady četnosti pro jednotlivé třídy roztříděného souboru jsou $\tilde{n}_j = n p_B(x)$.

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

Po výpočtu dostaneme

$$\tilde{n}_j = 2608\frac{3,872^{x_j}}{x_j!}e^{-3,872}\left\{1-0,0118\left[\frac{x_j^{[2]}}{2}-3,872x_j^{[1]}+7,4962\right]+\right.$$

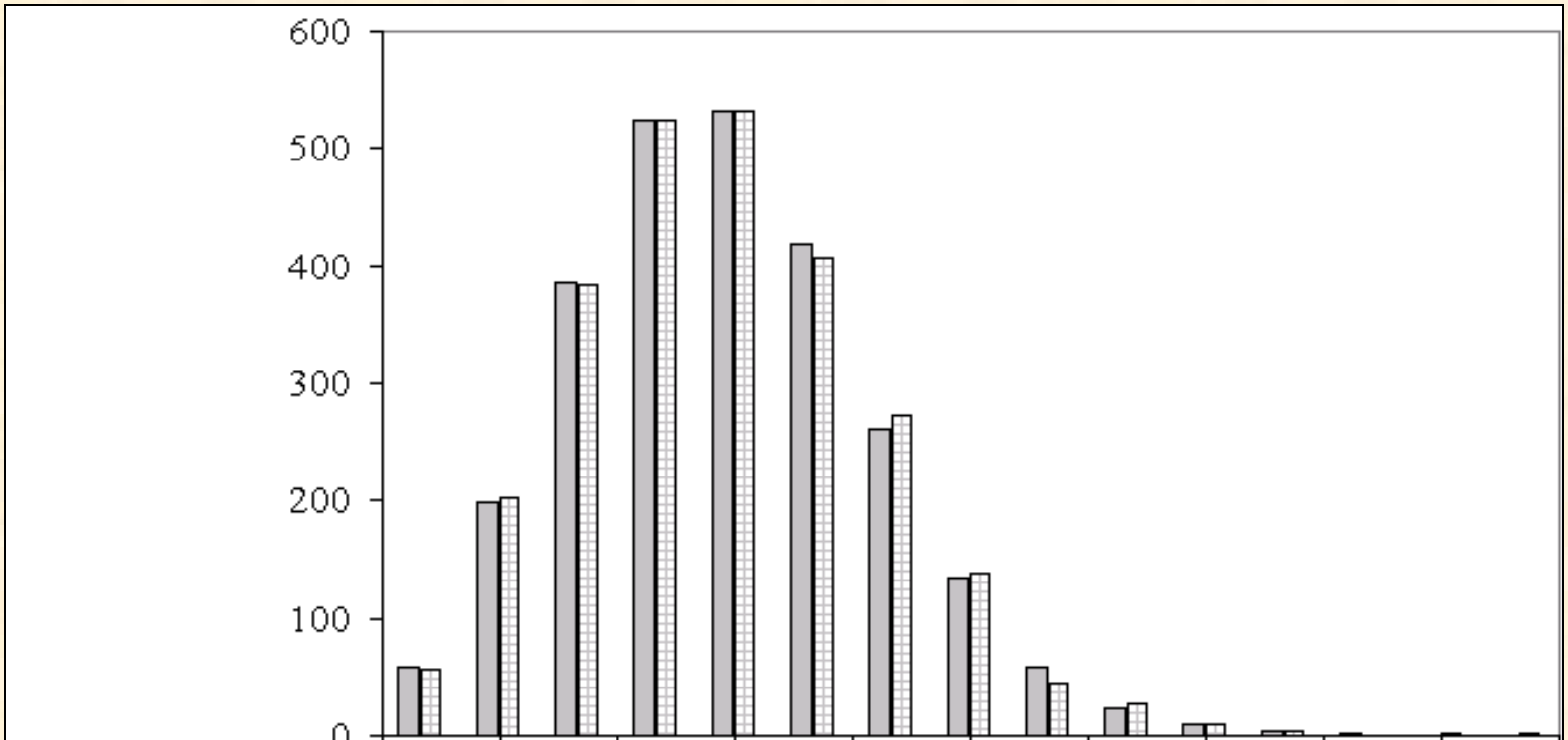
$$+0,001\left[\frac{x_j^{[3]}}{6}-1,936x_j^{[2]}+7,4962x_j^{[1]}-9,6751\right]+$$

$$\left.+0,0179\left[\frac{x_j^{[4]}}{24}-0,6353x_j^{[3]}+3,7481x_j^{[2]}-9,6751x_j^{[1]}+9,3655\right]\right\}=$$

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

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	$\frac{n_j}{n}$	$1,0695 - 0,12x_j^{[1]} + 0,0593x_j^{[2]} - -0,0114x_j^{[3]} + 0,00075x_j^{[4]}$	$p_B(x_j)$	$\tilde{n}_j$
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

Johnsonovy čtyřparametrické odhady spojitých rozdělení pravděpodobnosti vychází z nelineární transformace normovaného normálního rozdělení s hustotou

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

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. Transformací $z_L=b\exp\left(\frac{x-k}{m}\right)+a$ dostaneme 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 $z_B = b\dfrac{\exp\left(\dfrac{x-k}{m}\right)}{1+\exp\left(\dfrac{x-k}{m}\right)} + a$ 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 $z_U = b\sinh\left(\dfrac{x-k}{m}\right) + a$ 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 Jádrové odhady

Jádrové odhady hustoty spojitého rozdělení pravděpodobnosti 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*** $K(x)=\begin{cases}\frac{1}{2} & \text{pro } |x|<1\\ 0 & \text{jinde}\end{cases}$

Základní problémy aplikace jádrových odhadů:

- volba jádra a šířka vyhlazovacího okna,
- respektování požadavků spojitosti, příp. hladkosti získané hustoty pravděpodobnosti,
- vyjádření odpovídající distribuční funkce,
- snadnost výpočtu kvantilů,
- pokrytí celého rozsahu hodnot náhodné veličiny X (jádrové odhady mají někdy charakter šumu na koncích rozdělení).

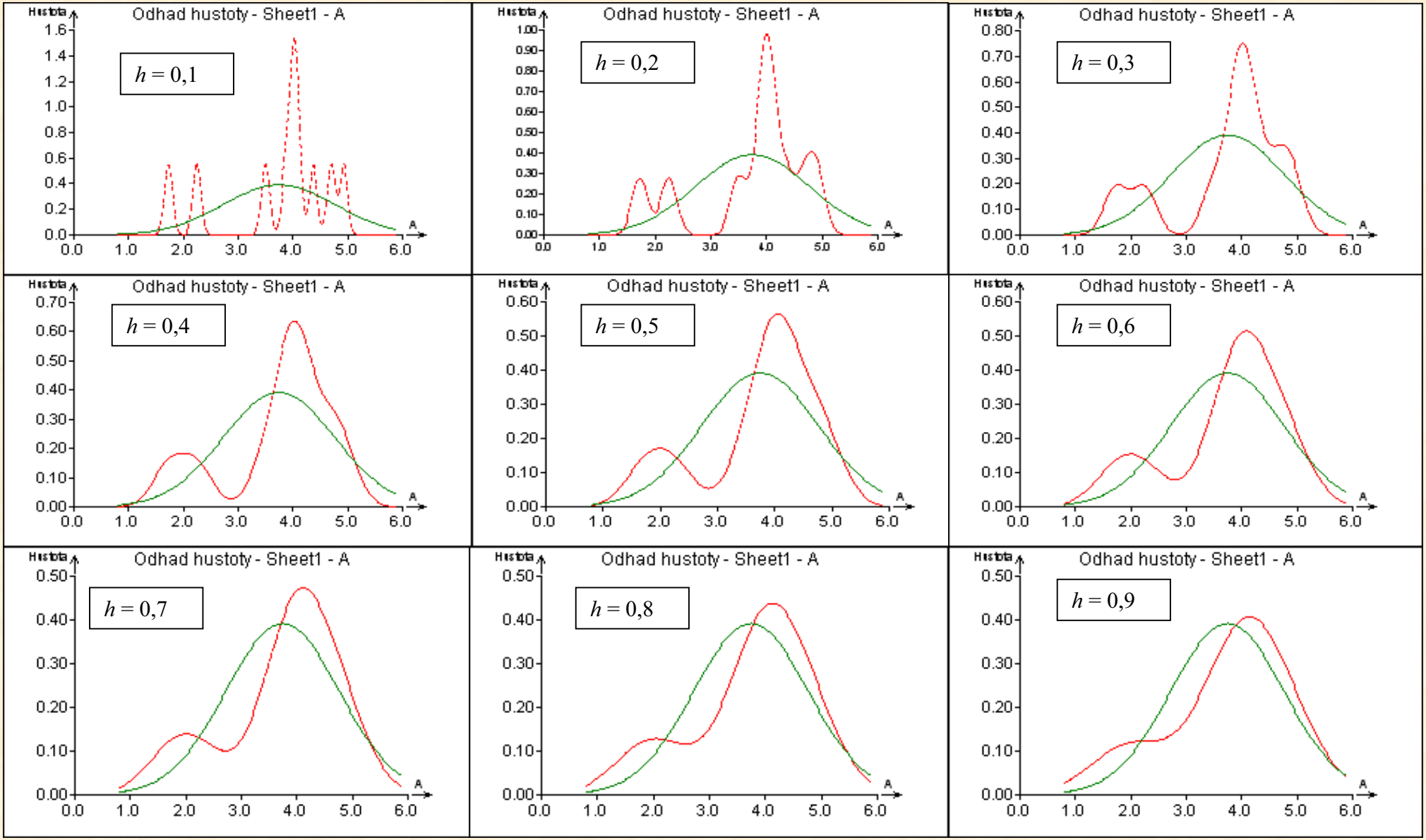
Příklad

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					

Pro náhodnou veličinu X s neznámým rozdělením pravděpodobnosti je typická bimodalita. Pro modelování její hustoty pravděpodobnosti by bylo možno uvažovat směs dvou rozdělení, avšak bylo by nutno expertně odhadnout jejich tvar.

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
Hustota
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

3.5 Odhady diskrétních rozdělení pomocí kvazinorem

Nechť funkce $f(u)$ je konvexní na $(0,\infty)$, striktně konvexní v $u=1$ a $f(1)=0$. ***f–divergencí*** rozdělení pravděpodobností $\mathbf{p}$, $\mathbf{q}$ na diskrétním pravděpodobnostním prostoru (Ω,Σ,P) rozumíme funkcionál

$$D_f(\mathbf{p},\mathbf{q}) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right).$$

Nechť $\mathbf{p}=(p_1,\ldots,p_m)$ a $\mathbf{p}_0=\left(\frac{1}{m},\ldots,\frac{1}{m}\right)$ pro $m>1$ jsou diskrétní rozdělení z pravděpodobnostního prostoru (Ω,Σ,P), a D_f je f-divergence definovaná na daném prostoru. ***Kvazinormou*** rozdělení pravděpodobnosti $\mathbf{p}=(p_1,\ldots,p_m)$ na (Ω,Σ,P) rozumíme f–divergenci $D_f(\mathbf{p},\mathbf{p_0})$. Platí:

- $D_f(\mathbf{p},\mathbf{p}_0) = \frac{1}{m}\sum_{j=1}^{m} f(mp_j)$,
- $D_f(\mathbf{p},\mathbf{p}_0)$ je nezáporná symetrická funkce proměnných p_j, $j=1,\ldots,m$.

Volíme:

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

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

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

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

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

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

Pozorováním náhodné veličiny X, jejíž rozdělení pravděpodobnosti $\mathbf{p}=(p_1,\ldots,p_m)$ chceme odhadnout, získáme statistický soubor $(x_1,\ldots,x_n)$

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

f_j … absolutní četnost pozorované hodnoty x_j^*

n … počet pozorování

Rozdělení pravděpodobnosti **p** má tzv. ***minimální kvazinormu za K počátečních momentových podmínek,*** jestliže příslušná kvazinorma nabývá minimální hodnoty pro

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

Jestliže $K<m-1$, pak obdržíme pro minimální

a) Hellingerovu kvazinormu $p_j = \dfrac{1}{m\left(\sum_{k=0}^{K} \lambda_k x_j^{*k}\right)^2}$,

b) Shannonovu kvazinormu $p_j = \exp\left(-1-\sum_{k=0}^{K} \lambda_k x_j^{*k}\right)$,

c) Pearsonovu kvazinormu $p_j = \dfrac{1}{m\left(\sqrt{\sum_{k=0}^{K} \lambda_k x^{*k}_j}\right)}$,

$j=1,\ldots,m$, kde λ_k, $k=0,\ldots,K$, jsou Lagrangeovy multiplikátory pro Lagrangeovu funkci

$$\Lambda(\mathbf{p},\boldsymbol{\lambda}) = D_f(\mathbf{p},\mathbf{p}_0) + \sum_{k=0}^{K} \lambda_k \left(\sum_{j=1}^{m} p_j x_j^{*k} - M_k\right), \quad \boldsymbol{\lambda} = (\lambda_0,\ldots,\lambda_K).$$

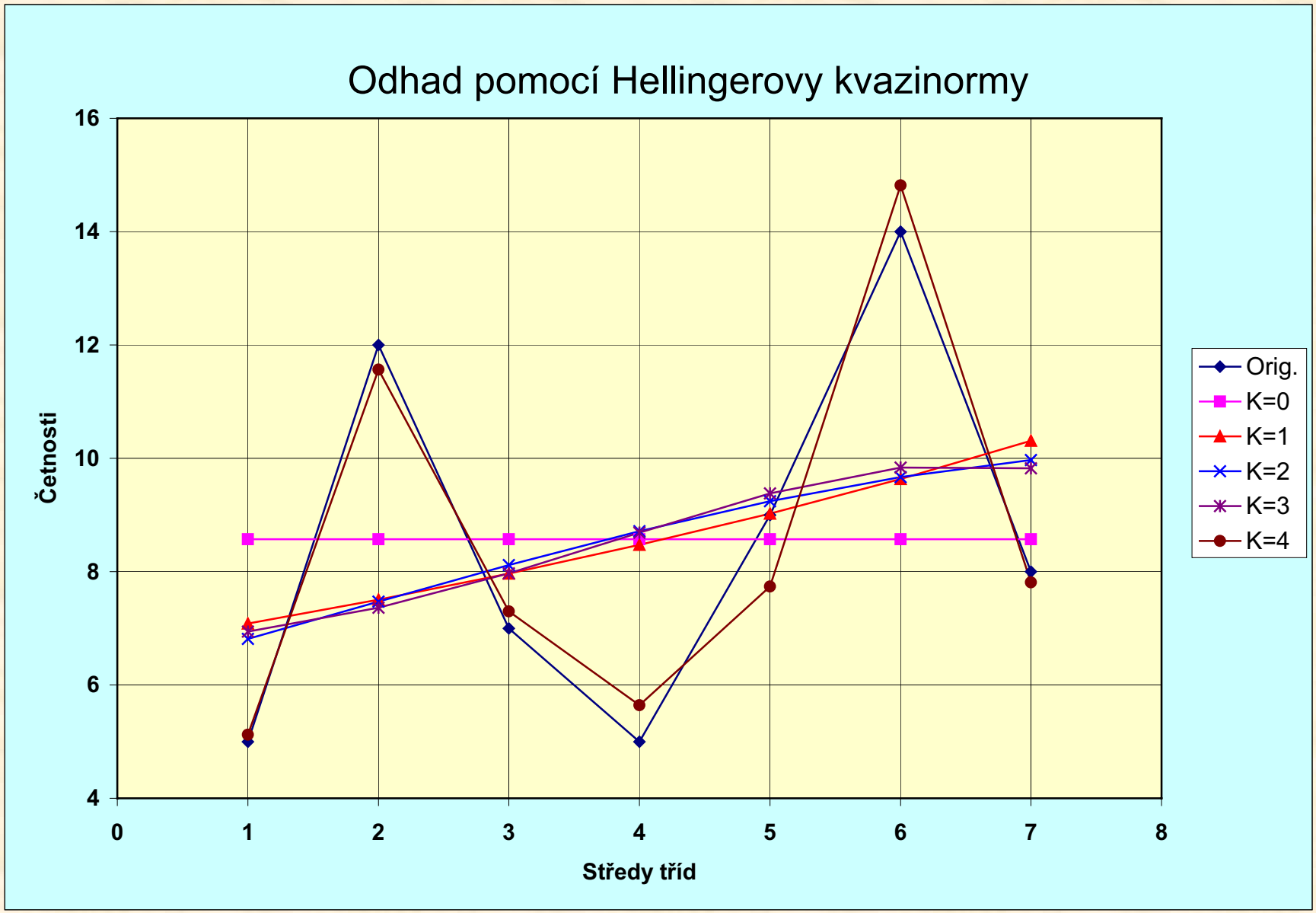
Příklad

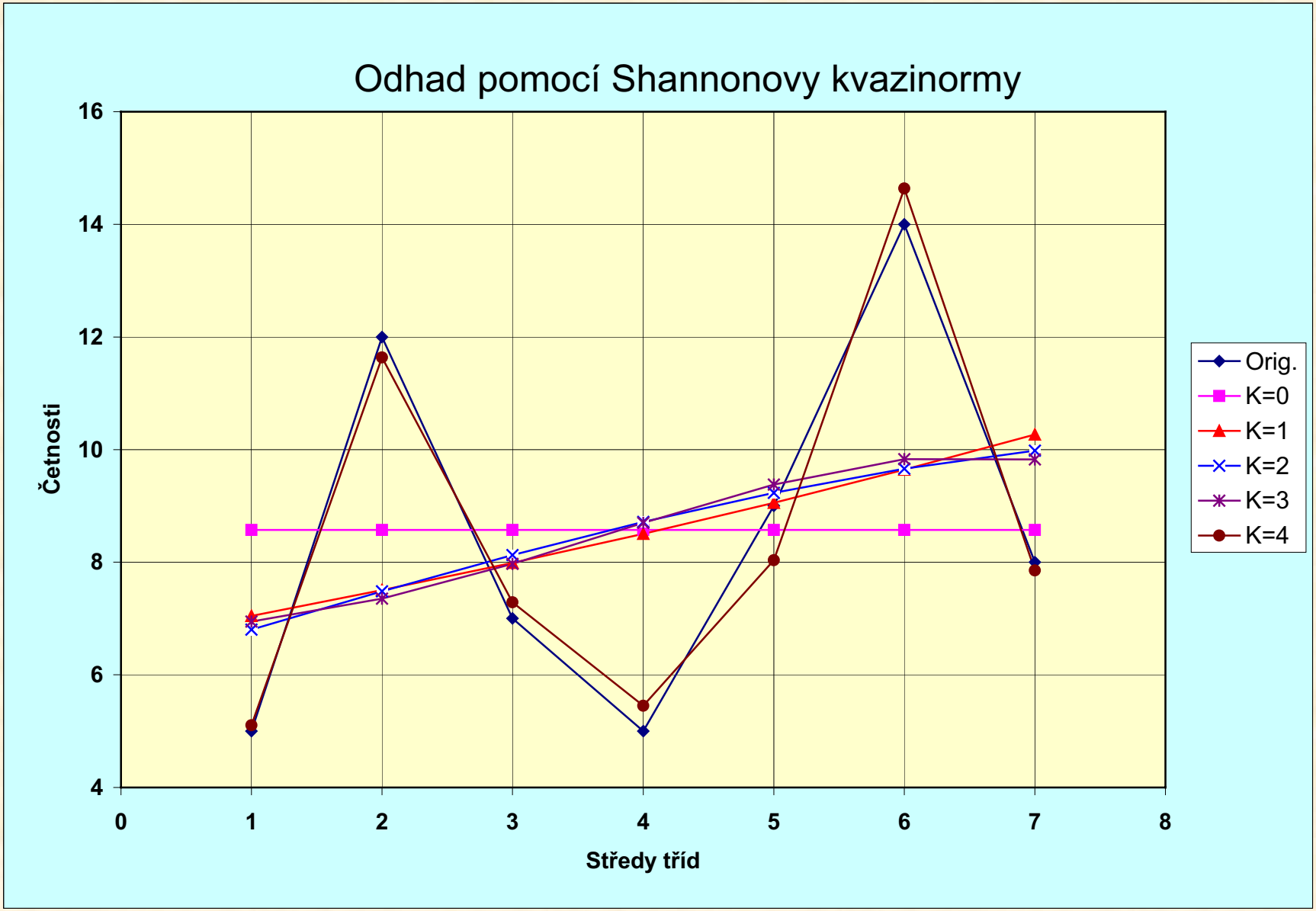
Sledováním diskrétní náhodné veličiny X jsme získali statistický soubor o rozsahu $n = 60$, který je bimodální:

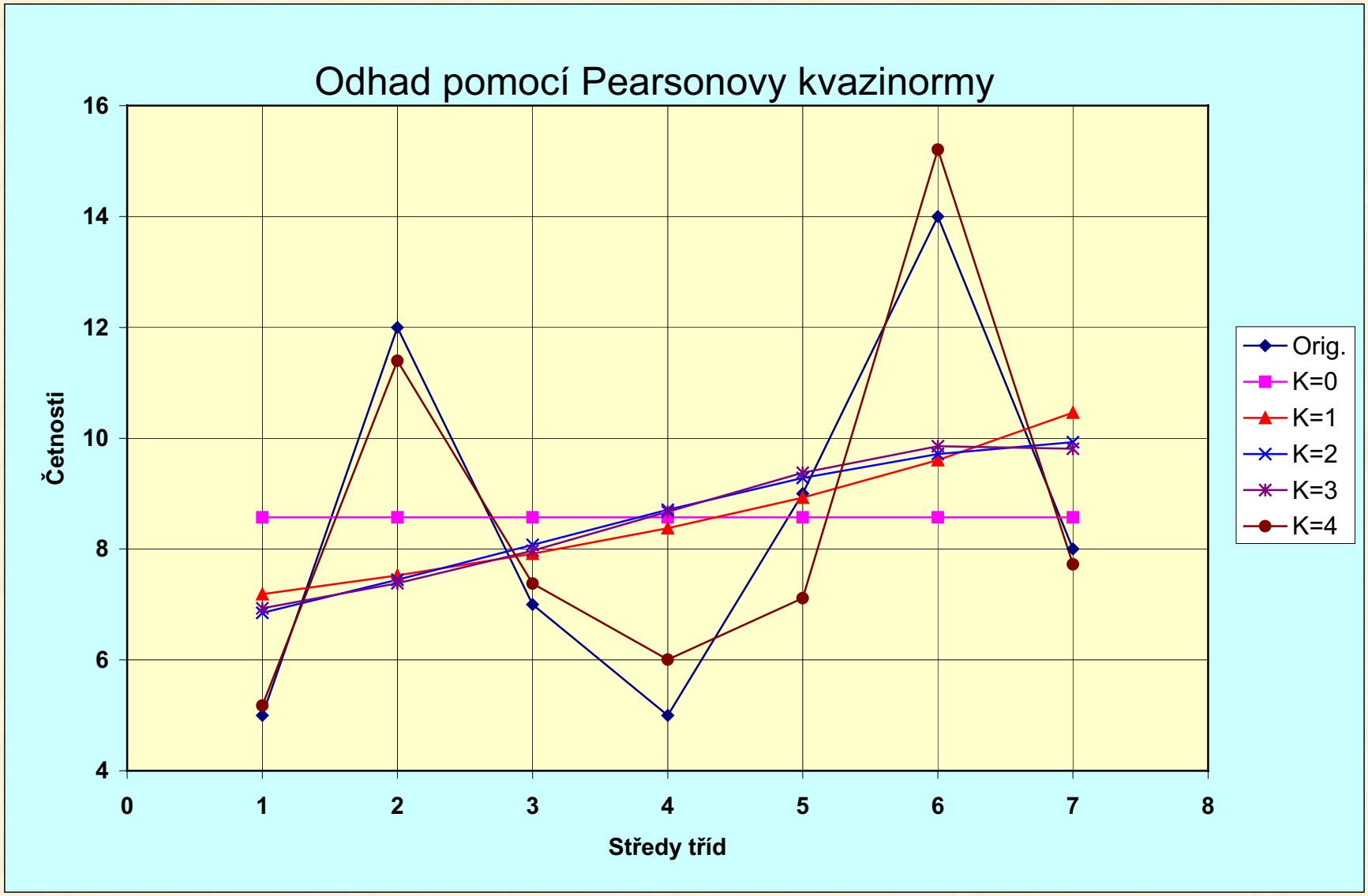
x_j^*	1	2	3	4	5	6	7
f_j	5	12	7	5	9	14	8

Výsledky za postupného přidávání momentových podmínek pro všechny kvazinormy jsou ilustrovány tabulkou a obrázky.

Vzdálenost	M_0	M_0, M_1	M_0, M_1, M_2	M_0, M_1, M_2, M_3	M_0, M_1, M_2, M_3, M_4
$\mathrm{D}(\mathbf{p}, \mathbf{p}_0)$	0	0,0039139	0,0040676	0,0041245	0,0318283
$S(\mathbf{p}, \mathbf{p}_0)$	0	0,0078253	0,0080530	0,0081838	0,0647159
$P(\mathbf{p}, \mathbf{p}_0)$	0	0,0156410	0,0168758	0,0169991	0,1250675


Odhad pomocí Hellingerovy kvazinormy
Četnosti
16
14
12
10
8
6
4
0
1
2
3
4
5
6
7
8
Středy tříd
Orig.
K=0
K=1
K=2
K=3
K=4


Odhad pomocí Shannonovy kvazinormy
16
14
12
10
8
6
4
Četnosti
0
1
2
3
4
5
6
7
8
Středy tříd
Orig.
K=0
K=1
K=2
K=3
K=4


Odhad pomocí Pearsonovy kvazinormy
Četnosti
16
14
12
10
8
6
4
0
1
2
3
4
5
6
7
8
Středy tříd
Orig.
K=0
K=1
K=2
K=3
K=4

DĚKUJI ZA POZORNOST !!!