DIAGNOSTIKA ROZDĚLENÍ DAT VE STOPOVÉ ANALÝZE

Jiří Militký Katedra textilních materiálů, Textilní fakulta, Technická universita v Liberci, Liberec, e- mail jiri.militky@vslib.cz
Milan Meloun, Katedra analytické chemie, Universita Pardubice, Pardubice

Motto:

K*aždý řetěz je tak silný, jak*
silný je jeho nejslabší článek

Abstract: The main aim of this contribution is to show the application of g-h system of frequency curves for identification of data distribution type and subsequent estimation of skew ness and kurtosis. This information is used for creation of adaptive and robust rule of outliers identification and creation of mean value confidence interval. The connections of g-h system with power transformation are discussed

Abstrakt: Cílem příspěvku je ukázat použití g-h systému hustot pro diagnostiku typu rozdělení výběru a následně pro určení šikmosti resp. špičatosti výběru. Tyto informace jsou použity pro konstrukci adaptivního a robustního postupu identifikace vybočujících měření a tvorbu intervalu spolehlivosti střední hodnoty. Je diskutována také souvislost s mocninnou transformací dat.

1. Úvod

Standardně se v analytické chemii používá pouze nejjednodušších statistických metod založených na předpokladu normality dat. V řadě případů, kde se opakovaně měří za stejných podmínek konstantní parametr se s tímto přístupem vystačí , pokud se zajistí dostatečný počet opakování. Pro menší výběry a nepřesná měření lze použít jednoduché robustní techniky, které fungují dobře , pokud je rozdělení dat symetrické.

Při analýze speciálních typů dat, kde chyby měření jsou zanedbatelné ve srovnání s variabilitou měřeného materiálu resp. jednotlivé analyzované vzorky jsou silně odlišné co do koncentrace analyzované látky je rozdělení výsledků výrazně asymetrické (**zešikmené** obyčejně k vyšším hodnotám). Pak vede jak standardní tak i robustní analýza často k nesprávným závěrům resp. vylučování dat, která sice neodpovídají předpokladu symetrie, ale jsou "přijatelná". V takových případech pak bez ohledu na kvalitu analytické metody rozhoduje o výsledku kvalita zpracování dat.

Stopová analýza má zvláštní postavení mezi analytickými postupy s ohledem na koncentrační rozmezí analyzovaných látek. V řadě případů se rozmezí analyzovaných látek pohybuje v několika řádech, což omezuje použití standardních statistických metod založených na předpokladu konstantního rozptylu resp. **aditivního modelu měření.** [1]. V práci [2] bylo diskutováno o možnostech použití transformace stabilizující rozptyl nebo **multiplikativní model měření**. To vede k logaritmické transformaci dat [1]. Nevýhodou této transformace je fakt, že při nízkých koncentracích je absolutní chyba měření velmi malá (blízká 0), což odporuje realitě. Byl navržen postup kombinující oba modely měření a odstraňující jejich nevýhody [2].

Multiplikativní model měření sice vede k použití asymetrického logaritmicko normálního rozdělení ale není zdaleka universální. Jedním z obecnějších postupů eliminace asymetrie je vhodná transformace dat (obyčejně mocninná) [1]. I zde však vznikají problémy zejména se

zpětnou transformací a použitelností jen pro některé úlohy. Lze odvodit , že pro malé rozptyly ^{2} je odhad parametru mocninné transformace špatně identifikovatelný (viz [3]).

V tomto příspěvku je pozornost zaměřena na techniky odhadu tvaru rozdělení výběru a využití těchto informací pro vybrané úlohy statistického zpracování dat. Nejdříve jsou popsány základní vlastnosti výběrových kvantilů a kvantilových charakteristik, které jsou využity pro konstrukci systému g-h rozdělení. Informace o šikmosti a špičatosti rozdělení jsou použity pro konstrukci adaptivního robustního postupu identifikace vybočujících bodů a tvorbu intervalů spolehlivosti střední hodnoty. Je ukázána souvislost s mocninou transformací dat.

2. Standardní zpracování dat

Omezme se na nejfrekventovanější a *zdánlivě nejjednodušší* úlohu stanovení koncentrace analytu z výběru (x_1, x_2,..x_N) velikosti N. Jednotlivé prvky výběru přitom nejsou opakování měření ale měření na různých vzorcích. Účelem je odhad parametru polohy a stanovení jeho neurčitosti.

Standardní model měření je **aditivní**, t.j.

$$x = \mu + \varepsilon \qquad (1)$$

kde μ je skutečná hodnota měřené veličiny (koncentrace analytu) a ε je náhodná chyba měření. Tento model celkem dobře vyhovuje pro případ opakování měření, ale pokud jde o různé vzorky často selhává. Standardní statistická analýza vychází z těchto předpokladů:

- střední hodnota chyb měření je nulová, t.j. $E(\varepsilon) = 0$,
- rozptyl chyb měření je konstantní, t.j. $D(\varepsilon) = \sigma^2$
- chyby jsou vzájemně nezávislé .t.j. $E(\varepsilon_i * \varepsilon_j) = 0$
- chyby mají normální rozdělení t.j. $\varepsilon \approx N(0,\sigma^2)$

Diskuse o identifikaci a postupu při porušení prvních tří předpokladů je uvedena v práci [2].

Nejvíce restriktivní, je předpoklad, že chyby mají normální rozdělení. Tento předpoklad je potřebný pro konstrukci intervalů spolehlivosti (neurčitosti výsledků měření) resp. testování hypotéz. Pokud je k dispozici dostatek dat, lze odhadnout rozdělení chyb ε z rozdělení měření x , protože pro model (1) je tvar hustoty pravděpodobnosti totožný.

Normální rozdělení lze chápat jako jednoho z členů třídy eliptických symetrických rozdělení, pro které platí že se liší pouze délkou konců. Ve stopové analýze a tam, kde jde o měření na různých vzorcích je však častějším jevem **asymetrické rozdělení dat zešikmené k vyšším hodnotám**. Toto rozdělení je běžné u dat, kde se ve vzorcích vyskytují řádové rozdíly koncentrací (např. u dat z oblasti životního prostředí). Pro odstranění asymetrie rozdělení dat se často používá vhodná transformace $h(x)$. Ta však v případě platnosti modelu (1) vede ke vzniku nekonstantního rozptylu

$$D(h(x)) = \left[\frac{dh(x)}{dx}\right]^2 * \sigma^2 \qquad (2)$$

Např. pro běžně doporučovanou logaritmickou transformaci $h(x)=ln(x)$ vyjde

$$D(h(x)) = \left(\frac{\sigma}{x}\right)^2 = \delta^2 \qquad (3)$$

To znamená, že místo konstantní absolutní chyby je v této transformaci konstantní relativní chyba (variační koeficient), což odporuje přijatému modelu měření. Korektní analýza zde vyžaduje přímé použití zešikmeného rozdělení a konstrukci nesymetrických intervalů spolehlivosti.

Multiplikativní model měření je založen na předpokladech konstantní relativní chyby a nezápornosti měření (jde o fyzikální veličiny související s hmotou). Výsledek měření je modelován vztahem

$$x = \mu * exp(\varepsilon) \qquad (4)$$

Zde ε má stejné vlastnosti jako u modelu aditivního (rov.(1)). Po korektní logaritmické transformaci přechází tento model na aditivní model v logaritmech, tedy

$$ln(x) = ln(\mu) + \varepsilon \qquad (5)$$

Nevýhodou multiplikativního modelu je především to, že pro velmi nízké koncentrace resp.malé μ vychází absolutní chyba měření příliš nízká [4].

Pokud se použije nesprávný předpoklad o rozdělení chyb dochází ke zkreslení parametrů a následně celé statistické analýzy.

Příklad A:

Nechť platí aditivní model (1) a na data se použije nesprávně logaritmická transformace. Pak vyjde

$$ln(x) = ln(\mu + \varepsilon) = ln\,\mu + ln(1 + \varepsilon/\mu) \qquad (6)$$

S využitím Taylorova rozvoje lze psát

$$ln(x) \approx ln(\mu + \varepsilon) = ln\,\mu + \varepsilon/\mu - 0.5*(\varepsilon/\mu)^2 + 0.33*(\varepsilon/\mu)^3 - ... \qquad (7)$$

Pro malé relativní chyby měření $\delta = \sigma/\mu$ lze pak s využitím tohoto vztahu nalézt výrazy pro střední hodnotu a rozptyl *ln(x)* ve tvaru

$$E(ln\,x) = ln\,\mu - 0.5*\delta^2 - 0.75*\delta^4 \qquad (8)$$

a

$$D(ln\,x) = \delta^2 + 2.5*\delta^4 + 4.66*\delta^6 \qquad (9)$$

Je tedy patrné, že použití nesprávného předpokladu ovlivní jak střední hodnotu tak i rozptyl. Pro μ větší než jedna vyjde střední hodnota podhodnocená a rozptyl nadhodnocený. V tab. 1 jsou uvedeny relativní odchylky

R( ) = 100(E[ln(x)]- ln( ))/ ln( )*

pro různá. a

Tabulka 1. Relativní chyba střední hodnoty při nesprávné logaritmické transformaci

		R( ) [%]
1	0.05	0.0545
10	0.1	0.2204
100	0.05	-0.0545
1	0.1	-0.2204
10	0.05	-0.0272
100	0.1	-0.1102

Je patrné, že i při 10Ý ní relativní chybě měření jsou odchylky způsobené nesprávnou volbou modelu poměrně malé.

Pro případ, že se analyzují data z různých vzorků se běžně předpokládá, že chyby měření jsou zanedbatelné vzhledem k variabilitě vzorků (měřeného materiálu) Jako model se pak používá se používá představa, že (x_i) i =..N, jsou realizace náhodné veličiny s rozdělením charakterizovaným hustotou pravděpodobnosti *f(x)* resp. distribuční funkcí *F(x*). Formálně je tedy

$$x_i = F^{-1}(p_i) \qquad (10)$$

kde p_i je hodnota distribuční funkce v místě x_i . Pokud je *f(x)* hustota pravděpodobnosti normálního rozdělení odpovídá tento model modelu (1) s tím, že je střední hodnota.

Odhadem **střední hodnoty** je pak aritmetický průměr $\bar{x}$ a odhadem **rozptýlení** je výběrový rozptyl s^2 .

Přesnosti libovolných odhadů *o* se charakterizují pomocí jejich rozptylů *D(o)*. Pro případ normálního rozdělení dat $x_i \sim N(\mu,\sigma^2)$ jsou tyto rozptyly

$$D(\bar{x}) = \frac{\sigma^2}{N} \qquad \text{a} \qquad D(s^2) = \frac{2.\sigma^4}{N-1}$$

Klasická statistická analýza je založena na odhadech $\bar{x}$, s^2 a předpokladu normality rozdělení chyb v modelu (1) resp. normality F(x) v modelu (10). Základní roli při posuzování výsledků měření hraje *100. (1 - α) %* ní interval spolehlivosti střední hodnoty, pro který obecně platí,

$$P(x_D \le \mu \le x_H) = 1-\alpha$$

kde je hladina významnosti a x_D x_H jsou náhodné meze určené z dat. (Standardně se konstruuje 95 % ní interval spolehlivosti). Pro případ normálního rozdělení chyb resp. měření je tento interval ve tvaru

$$\bar{x} - t_{1-\alpha/2}(N-1)*\frac{s}{\sqrt{N}} \le \mu \le \bar{x} + t_{1-\alpha/2}(N-1)*\frac{s}{\sqrt{N}} \qquad (11)$$

kde $t_{1-\alpha/2}(N-1)$. je kvantil Studentova rozdělení s *N-1* stupni volnosti. Pro větší výběry se tento kvantil nahrazuje kvantilem normovaného normálního rozdělení $u_{1-\alpha/2}$.

Pro jiná než normální rozdělení již nemají odhady $\bar{x}$, s^2 optimální statistické vlastnosti a interval spolehlivosti definovaný rov. (11) není rozumně použitelný. Pro asymetrická rozdělení dat je interval (11) nevhodný již proto, že je symetrický. Navíc již nebude platit , že je *100. (1 -*

α) %. Účelem je nalézt postup jak zkonstruovat lepší interval spolehlivosti pro nesymetrická rozdělení, které mohou obsahovat i vybočující hodnoty.

3. Kvantilové charakteristiky dat

Je výhodné založit analýzu malých a středních výběrů na koncepci kvantilů. Z rov. (10) je patrné , že výsledek měření x_i odpovídá hodnotě funkce inversní k distribuční v místě daném pravděpodobností p_i s jakou se daná náhodná veličina vyskytuje pod x_i. Jedná se tedy o *100 pi* procentní kvantil výběrového rozdělení dat (viz. dále). Pro precizaci toho o jaký kvantil se jedná a"neparametrické" vyjádření odhadu pravděpodobnosti p_i se využívá vzestupně setříděných hodnot prvků výběru, které se nazývají pořádkovými statistikami

$$x_{(1)} < x_{(2)} < \ldots < x_{(N)}$$

Nechť $F_e(x)$ distribuční funkce rozdělení výběru, ze kterého pocházejí x_i. Lze dokázat, že transformovaná náhodná veličina

$$z_{(i)} = F_e(x_{(i)}) \tag{12}$$

má nezávisle na typu distribuční funkce F_e beta rozdělení *Be [i, N-i+1]*. Střední hodnota této transformované veličiny je pak

$$E(z_{(i)}) = \frac{i}{N+1} \tag{13}$$

kde *E(.)* je operátor matematického očekávání. Jednotlivé $z_{(i)}$ jsou závislé a prvky V_{ij} kovarianční matice *V* jsou pro libovolné dvojice $z_{(i)}$, $z_{(j)}$ *i, j=1,...N* jednoduché funkce pouze *i,j* a *N*. Použitím zpětné transformace $E[z_{(i)}]$ resultuje vztah

$$E(x_{(i)}) = F_e^{-1}(z_{(i)}) = Q_e(P_i) \tag{14}$$

Zde $Q_e(P_i)$ označuje kvantilovou funkci a

$$P_i = \frac{i}{N+1}$$

je tzv, pořadová pravděpodobnost.

Popis kvantilové funkce a její výhody pro analýzu dat lze najít v pracích Parzena [6,7]. Pokud je $F(x) = P$ spojitá distribuční funkce a $f(x)$ je odpovídající hustota pravděpodobnosti je $Q(P) = x$ kvantilová funkce. Lze snadno ověřit, že, $F(Q(P)) = P$ pro všechna $0 \le P \le 1$ a $f(Q(P))*q(P) = 1$. Funkce $q(P) = dQ(P)/dP$ se nazývá kvantilově hustotní funkce a $f(Q(P) = 1/q(P)$ je hustotně kvantilová funkce.

Z rov. (14) je patrné, že pořádkové statistiky $x_{(i)}$ jsou hrubé odhady kvantilové funkce $Q_e(P_i)$ v místě P_i. Pro odhad kvantilu $x_P = Q_e(P)$ v místě $i/(n+1) < P < (i+1)/(n+1)$ se používá řada metod. Jednoduchá je lokálně lineární interpolace

$$x_{(P)} = (N+1)\left(\frac{PN + P - i}{N+1}\right)(x_{(i+1)} - x_{(i)}) + x_{(i)} \tag{15}$$

Rozptyl $D(x_P)$ kvantilu x_P je možno vyjádřit ve tvaru

$$D(x_P) = \frac{P(1-P)}{N f^2(x_P)} \qquad (16)$$

Symbol $f_e(x_P)$ označuje hustotu pravděpodobnosti odpovídající distribuční funkci F_e. Asymptotické rozdělení kvantilu x_P je normální se střední hodnotou $Q(P)$ a rozptylem definovaným rov. (16).

Interpolace definovaná rov. (4) se dá snadno použít pro odhad speciálních výběrových kvantilů x_{Pi} a x_{1-Pi} pro pořadové pravděpodobnosti $P_i=2^{-i}$ $i=1,...N$. Tyto kvantily se označují jako písmenové hodnoty [8]. Všechny písmenové hodnoty až na $i=1$ (medián) tvoří dvojice. Např. lze nalézt dolní kvartil $x_{0.25}$ $(P_i = 0.25)$ resp. horní kvartil $x_{0.75}$ $(P_i = 0.75)$ atd.

Pro účely průzkumové analýzy dat se často používá modifikovaného vyjádření pořadové pravděpodobnosti ve tvaru

$$P_i = \frac{i-0.375}{N+0.25} \qquad (17)$$

doporučené Blomem. Kvantilovou mírou polohy je medián $x_{0.5}$, který je určen jako střed pořádkových statistik. Rozptýlení se vyjadřuje jako rozdíl mezi horním $x_{0.75}$ $(P_i = 0.75)$ a dolním $x_{0.25}$ $(P_i = 0.25)$ Parzen [6]doporučuje tzv. kvantilovou odchylku

$$DQ = 2*(x_{0.75} - x_{0.25}) \qquad (18)$$

DQ je vybrána tak, že odpovídá derivaci kvantilové funkce $Q(P)$ pro $P=0.5$, která je přibilžně rovna hustotně kvantilové odchylce

$$DfQ = 1/f(Q(0.5)) = 1/f(x_{0.5}) = q(0.5) \qquad (19)$$

S využitím těchto odhadů rozptýlení a polohy lze provést standardizaci rozdělení dat aby byla standardizovaná kvantilová odchylka jednotková $DQ = 1$ a standardizovaný medián nulový $x_{0.5} = 0$. Standardizovaná kvantilová funkce se označuje jako tvar indikující kvantilová funkce $QI(P)$. Výběrová tvar indikující kvantilová funkce $QI(P)$ má tvar

$$QI_e(P) = \frac{(x_{(i)} - x_{0.5})}{2*(x_{0.75} - x_{0.25})} \qquad (20)$$

Hodnoty, pro které je $|QI_e(P)| \geq 1$ jsou považovány za vybočující (pro normální rozdělení) nebo za indikátory dlouhých konců. Hodnoty $QI_e(P)$ se mohou snadno použít pro určení *šikmosti* a *délky* konců výběrových rozdělení. Jako míra šikmosti se volí parametr

$$SQ = QI_e(0.25) + QI_e(0.75)$$

(pro symetrická rozdělení je roven nule). Mírami délek konců jsou $QI_e(0.05)$ a $QI_e(0.95)$.
Platí, že rozdělení dat má :

krátké konce pro $QI_e(0.95) < 0.5$,

dlouhé konce pro $QI_e(0.95) > 1$
středně dlouhé konce pro $0.5 < QI_e(0.95) < 1$.

Tyto diagnostiky jsou velmi jednoduché a umožňují snadnou diagnostiku tvaru výběrového rozdělení

Příklad B:
Byl stanoven obsah antimonu v ppm u N=17 vzorků měděné rudy
4,5,7,7,7,8,8.3,8.4,9.4,9.5,10,10.5,12,12.8,13,22,23

Standardní statistická analýza vede k odhadům. Průměr aritmetický = 10.406, průměr geometrický =9.421,rozptyl = 26.83 ,šikmost = 1.399, špičatost = 4.272.
Kvantilové míry jsou medián = 9.5, dolní kvartil = 7, horní kvartil = 12, DQ = 10, SQ = 0.02, $QI_e(0.95) = 1.28$. Kvantilová analýza vede k závěrům, že rozdělení je mírně zešikmené a má velmi dlouhé konce

4. G-h systém rozdělení pravděpodobnosti

Jak je ukázáno v dalších odstavcích lze při znalosti šikmosti a špičatosti dat nalézt způsoby statistického zpracování výběrových rozdělení silně odlišných od normálního. Je proto vhodné vybrat vhodný systém hustot s parametry tvaru souvisejícími s kvantilovými charakteristikami který umožňuje snadnou diagnostiku a kvantifikaci tvaru rozdělení. Jeden z takových systémů se označuje jako g-h.

Systém g-h empirických rozdělení je založen na monotónní transformaci standardizované veličiny s normálním rozdělením a využívá vhodných kvantilových funkcí pro definici tvaru. Jde o jeden z translačních systémů rozdělení, kdy se nejprve provádí standardizace původních náhodných proměnných (kvantilů) *x* na normalizované kvantily

$$Y = (x - x_{0.5}) / R \tag{21}$$

kde $x_{0.5}$ je parametr polohy (medián) a *R* je vhodný parametr měřítka (např. *DQ* nebo interkvartilový odhad směrodatné odchylky z rov (41)). Proměnná *Y* má stejný tvar rozdělení jako proměnná *x a je monotónní transformací* standardizované náhodné veličiny *z* mající normální rozdělení. Formálně je tedy *Y = Y(z)*.

Pro třídu g-rozdělení (pouze zešikmených) se volí obecná transformace typu

$$Y = Q_{g,0}(z) = G(z) \cdot z \tag{22}$$

kde *G(z)* je lichá funkce, pro kterou musí platit, že

$$\begin{aligned} &G_{g,0}(0) = 0 \\ &\lim_{z \to 0} Q_{g,0}(z) \approx z \end{aligned} \tag{23}$$

G(z) je vlastně operátor šikmosti , který závisí na parametru šikmosti *g*. Je vhodné, aby pro $g>0$ vycházela rozdělení zešikmená vpravo, pro $g<0$ rozdělení zešikmená vlevo a pro $g=0$ normální rozdělení. Všem těmto požadavkům vyhovuje jednoduchá funkce

$$G(z) = [exp(g.z) - 1]/(g.z) \qquad (24)$$

Lze jednoduše nalézt, že hustota pravděpodobnosti g-rozdělení $f_g(x)$ je dána vztahem

$$f_g(x) = 1/\left[\sqrt{2\pi} \cdot \left|(x - \tilde{x}_{0,5}) \cdot g + R\right|\right] \cdot exp\left(-\{ln[(x - \tilde{x}_{0,5}) \cdot g / R + 1]\}^2 / (2g^2)\right)$$

Jde tedy o lognormální rozdělení (rozdělení typu S_L v Johnsonově systému hustot).

Pro třídu h-rozdělení (symetrických s různou špičatostí) se volí transformace typu

$$Y = Q_{0,h}(z) = H(z).z \qquad (25)$$

Operátor špičatosti $H(z)$ musí být rostoucí kladná sudá funkce závislá na parametru špičatosti h. Pro $h=0$ jde o normální rozdělení a čím je $h>0$ větší, tím má odpovídající rozdělení delší konce. Tomu vyhovuje volba

$$H(z) = exp(hz^2/2) \qquad (26)$$

Z této rovnice plyne, že parametr h může být i záporný. Podmínka monotónnosti funkce $Q_{g,0}$ je však narušena, pokud je $z^2 > -1/h$. Třída h-rozdělení definovaná rov. (26) se v oblasti konců chová jako Paretovo rozdělení. Volba faktoru1/2 v rov. (26) také zajišťuje, že pro h≈1 je $Q_{0,h}$ rozdělení blízké Cauchyho rozdělení [5].Hustotu pravděpodobnosti h rozdělení lze však získat pouze numericky, protože nelze nalézt analyticky transformaci inverzní k rov. (26).

Pro obecnou třídu g-h rozdělení se volí dvou parametrová transformace

$$Y = Q_{g,h}(z) = G(z).H(z).z \qquad (27)$$

kde $Q(z)$ je voleno podle rov.(24) a $H(z)$ je definováno rov. (26). Také v tomto případě je třeba odpovídající hustotu pravděpodobnosti počítat numericky.

Pro generaci náhodných čísel z g-h rozdělení se zvolenými parametry g a h postačuje generovat kvantily z_p standardního normálního rozdělení (zde p určuje 100p % kvantil) a dosazovat do rov. (27) za $z=z_p$.

Pro účely využití tohoto systému při zpracování asymetrických výběrových rozdělení je nutné znát výrazy pro první čtyři momenty. Ty jsou funkcemi parametrů g, h . Pro standardizované náhodné proměnné $Y=Q_{g,h}(z)$ pocházející z **g-h rozdělení** je střední hodnota definovaná vztahem

$$E(Y) = \frac{1}{g\sqrt{h-1}}\left[exp\left(\frac{g^2}{2(1-h)}\right) - 1\right] \qquad (28)$$

$0 \le h \ge 1$

a pro rozptyl platí

$$D(Y) = \frac{1}{g^2\sqrt{1-2h}}\left[exp(2\omega) - 2\,exp\left(\frac{\omega}{2}\right) + 1\right] - \frac{1}{g^2(1-h)}\left[exp\left(\frac{\omega}{2}\right) - 1\right]^2 \qquad (29)$$

$0 \le h \prec 05$

kde $\omega=g^2/(1-2h)$. Pro případ **g-rozdělení** je $h=0$ a pak z rov. (28) vyjde

$$E(Y)=\left[exp\left(\frac{g^2}{2}\right)-1\right]/g \qquad (30)$$

a z rov (29) je

$$D(Y)=exp(g^2).[exp(g^2)-1]/g^2 \qquad (31)$$

Šikmost $_1$ je u g-rozdělení vyjádřitelná ve tvaru

$$\beta_1=[exp(3g^2)-3exp(g^2)+2]/\sqrt{(exp(g^2)-1)^3} \qquad (32)$$

a pro špičatost $_2$ platí

$$\beta_2=[exp(6g^2)-4exp(3g^2)+6exp(g^2)-3]/(exp(g^2)-1)^2 \qquad (33)$$

Třetí centrální moment , který je v některých případech potřebný je dán vztahem

$$E[(Y-E(Y))]^3=exp(3g^2/2)[exp(3g^2)-3exp(g^2)+2]/g^3 \qquad (34)$$

Pro případ h-rozdělení je $g=0$, takže $E(Y)=g_1=0$. Pro rozptyl pak platí

$$g\bar{a}^1(x)=x-\frac{\beta_1(x^2/3+1/6)}{\sqrt{N}} \qquad (35)$$

a špičatost je ve tvaru

$$\beta_2=3(1-2h)^3/\sqrt{(1-4h)^5} \qquad (36)$$

Při výpočtu špičatosti vyšších momentů g-h rozdělení je možné použít Martinezova vztahu, který platí pro $g\neq0$ a $0\leq h\leq 1/n$ [5]

$$E(Y^n)=\frac{1}{g^n\sqrt{1-nh}}\sum_{i=0}^{n}(-1)^i\binom{n}{i}exp\left\{\frac{[(n-i)g]^2}{2(1-nh)}\right\}$$

S výhodou se pro odhad parametrů g, h respektive obou užívá výběrového mediánu $x_{0,5}$ a dvojic kvantilů x_p , x_{1-p} pro vhodné $\mathbf{0<p<0.5}$. (toto rozmezí p se používá v dalších vztazích) U malých výběrů je možné použít všech pořádkových statistik $x_{(i)}$. U větších výběrů se používá. písmenových hodnot, kde $p=2^{-i}$ $i=2,3$ (odpovídající kvantilům, oktilům, sedecilům atd.).
V případě g-rozdělení je možné pro zvolenou pravděpodobnost p určit parametr g (a také R) ze dvou podmínek

$$x_p = x_{0.5} + R.Y = x_{0.5} + R\left[\frac{exp(g.z_p)-1}{g}\right]$$
$$x_{1-p} = x_{0.5} + R\left[\frac{exp(-gz_p)-1}{g}\right] \qquad (37)$$

kde z_p jsou standardizované kvantily normálního rozdělení (vzhledem k symetrii je z_p=-z_{1-p}). Po jednoduchých úpravách vyjde

$$g_p = -\frac{1}{z_p}.ln\left[\frac{x_{1-p} - x_{0.5}}{x_{0.5} - x_p}\right] \qquad (38)$$

Index *p* zde ukazuje, že parametr šikmosti obecně může záviset na poloze kvantilů vzhledem k mediánu. Pokud je g_p přibližně konstantní (resp. není funkcí *p*) postupuje se tak, že se určí medián *g* ze všech g_p . Jednoduše je pak možné odhadnout parametr *R* jako směrnici v Q-Q grafu, kde se vynášejí hodnoty x_p vs. $Q_{g,o}(z_p)$. přímka v tomto grafu indikuje, že lze použít g-rozdělení s konstantní hodnotou *g*.

Pro stejný účel je možné použít také graf symetrie, kdy se vynášejí polosumy $0.5(x_p + x_{1-p})$ vs. $z_p^2/2$. S využitím rov. (37) se snadno určí teoretická závislost

$$0.5(x_p + x_{1-p}) = x_{0,5} + 0.5\frac{R}{g}\left[exp(gz_p) + exp(-gz_p) - 2\right] \approx x_{0.5} + R.g.z_p^2/2$$

Platí-li tedy předpoklad konstantnosti parametru šikmosti *g*, vyjde v grafu symetrie přibližně lineární závislost. Tento graf se však dá použít jen pro malé g. Při vyšších *g* se projevuje nelinearita.

Pokud není g_p konstantní, předpokládá se obyčejně, že je to jednoduchý polynom vzhledem k normalizovaným kvantilům z_p resp. jejich čtvercům. Jednoduché je zobecnění typu

$$g_p(z_p) = g_o + g_1 z_p^2 \qquad (39)$$

Rov. (39) lze snadno ověřit na základě grafu šikmosti, kdy se vynáší g_p vs. Z_p^2. Pokud vyjde v tomto grafu přibližně lineární závislost, je třeba uvažovat obecnější rozdělení s nekonstantním g_p vyjádřeným rov. (39). Parametry g_0 a g_1 v rov. (39) lze snadno určit z úseku a směrnice v grafu šikmosti (doporučuje se použití robustní regrese).

Pro případ **h-rozdělení** lze při znalosti parametru polohy $x_{0.5}$ a parametru měřítka *R* určit pro každé x_p odpovídající hodnotu h_p přímo z definičního vztahu (26). Vyjde

$$h_p = \left[2ln\left(\frac{(x_p - x_{0,5})/R}{z_p}\right)\right] / z_p^2 \qquad (40)$$

V případě, že je výběrové rozdělení symetrické, musí pochopitelně být h_p=h_{1-p}. Jako vhodný parametr měřítka se doporučuje interkvartilový odhad směrodatné odchylky mediánu

$$R = 0.926 \cdot (x_{0,75} - x_{0,25})/\sqrt{n} \qquad (41)$$

kde *n* je rozsah výběru. Za předpokladu symetrického rozdělení lze přímo z definičních vztahů

$$x_p = x_{0,5} + R \cdot z_p \cdot exp\left(h \cdot z_p^2 / 2\right)$$
$$x_{1-p} = x_{0,5} - R \cdot z_p \cdot exp\left(h \cdot z_p^2 / 2\right)$$

(zde $z_p<0$ a $z_{1-p}= -z_p$) dospět k lineární funkci vzhledem k *h*

$$ln\left[\frac{x_{1-p} - x_p}{-2z_p}\right] = \ln R + hz_p^2 / 2 \qquad (42)$$

Z rov (42) je patrné, že vynesením $ln[(x_{1-p} -x_p)/(-2z_p)]$ vs. $z_p^2/2$ vyjde v případě platnosti h-rozdělení s konstantním parametrem špičatosti *h* přibližně lineární závislost. Tato závislost se označuje jako graf pseudosigma a umožňuje odhad *h* i *R* ze směrnice resp. úseku regresní přímky.

Pro normální rozdělení je graf pseudosigma prakticky horizontální přímka s nulovou směrnicí a úsekem *lnR* .

Pro nesymetrická rozdělení vycházejí grafy pseudosigma nelineární. Pokud není h_p konstantní (ale $h_p \approx h_{1-p}$) předpokládá se, že jde opět o jednoduchou kvadratickou závislost typu

$$h_p = h_0 + h_1 z_p^2 \qquad (43)$$

Rov. (43) lze jednoduše ověřit na základě grafu špičatosti, kdy se vynáší h_p v závislosti na z_p^2. Pokud vyjde v tomto grafu přibližně lineární závislost, znamená to, že platí obecnější rozdělení s nekonstantním parametrem špičatosti vyjádřeným rov, (43). Další možností je přímé využití h rozdělení pro konstrukci asymetrických rozdělení, kdy se uvažují dvě větve s různým h. Toto HH rozdělení je použitelné, pokud tvoří graf $ln[(x_p - x_{0.5})/(z_p)]$ vs. $z_p^2/2$ dvě přímky protínající se v bodě $z_p=0$. Této podmínce vyhovují také další jednoduchá zobecnění h rozdělení (HQ, HHH a HR *viz.* [11]).

Nejčastější je případ, kdy je rozdělení dat nesymetrické a ani po symetrizační (zde logaritmické) transformaci neodpovídá délkou konců normálnímu rozdělení. Pro tento případ je nutné použít **g-h rozdělení** Vzhledem k volbě $Q_{g,h}(z)$ ve tvaru (27) a funkcím *G(z)*, *H(z)*, lze snadno určit, že rov. (38) platí nezávisle na velikosti *h*. To znamená, že lze nejprve nalézt odhad parametru špičatosti *g* (stejně jako u „čistého“ g-rozdělení) pak provést opravu na šikmost před odhadem *h*. Při znalosti *g* můžeme snadno určit s využitím z rov. (27) polorozpětí

$$x_{1-p} - x_{0,5} = \frac{R}{g}\left(exp\left(-gz_p\right) - 1\right) \cdot exp\left(hz_p^2 / 2\right) \qquad (44)$$

Po úpravě vyjde lineární závislost vzhledem k *h*

$$y^* = ln\left[\frac{g\left(x_{1-p} - x_{0,5}\right)}{exp\left(-gz_p\right) - 1}\right] = \ln R + hz_p^2 / 2 \qquad (45)$$

Vynesením y^* vs. $z_p^2/2$ (modifikovaný graf pseudosigma) vyjde v případě platnosti g-h rozdělení s konstantními *g, h* přibližně lineární závislost.

Jednoduše lze také postupovat v případě, že g_p je vyjádřeno rov. (39) a h_p je vyjádřeno rov. (43). Protože je odhad g_p nezávislý na h_p , lze stejně jako u „čistého“ g-rozdělení nalézt parametry g_0, g_1. Pro odhad parametrů h_0, h_1 však již nelze použít rov. (43) ale je třeba provést opravu na šikmost (stejně jako u rov. (44)). Po úpravách vyjde vztah

$$y_p^* = \ln R + \frac{h_0}{2} z_p^2 + \frac{h_1}{2} z_p^4 \qquad (46)$$

Nejdříve se opět provede korekce na šikmost a určí se

$$y_p^* = \ln\left[\frac{\left(x_{1-p} - x_{0,5}\right) g_p(z_p)}{\exp\left(-z_p * g_p(z_p)\right) - 1}\right] \qquad (47)$$

při znalosti g_0,g_1 pro každé p. Závislost y_p^* vs. $z_p^2/2$ se označuje jako zobecněný graf špičatosti. Vyjde-li parabolický, znamená to, že je nutné použít zobecněné g-h rozdělení s nekonstantními parametry g, h.

Pro ověřování platnosti různých typů g-h rozdělení je vhodné pochopitelně použít formální aparát lineární regrese a testovat významnost směrnic, respektive úseků ve výše uvedených grafech. Pro účely analýzy dat však běžně postačuje posouzení vlastních grafů a pouze v případě konstantních parametrů g, h určit odpovídající šikmosti a špičatosti pro další analýzu.

Pro posouzení tvaru rozdělení s ohledem na šikmost je možné konstruovat grafy:

- Symetrie tj. závislosti $0.5(x_p+x_{1-p})$ na $z^2_p/2$
- *Šikmosti tj. závislosti g_p na z^2_p*

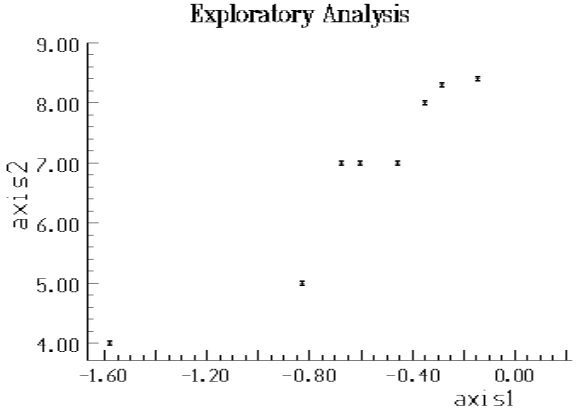
Pro posouzení tvaru s ohledem na špičatost jsou to grafy:

- Pseudosigma tj. závislosti $ln[(x_{1-p} - x_p)/(-2z_p)]$ na. $z_p^2/2$
- *Špičatosti tj. závislosti h_p na z^2_p*

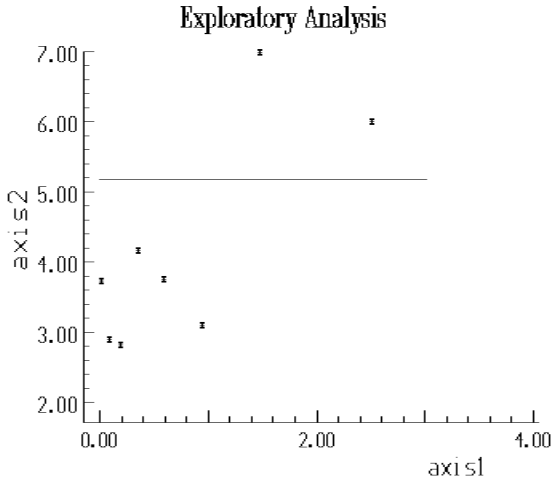
S ohledem na asymetrii rozdělení dat umožňuje tento systém rozdělení zpracovat pouze vybrané typy úloh. Na druhou stranu lze dobře komplexně posoudit vhodnost přijatých parametrů šikmosti a špičatosti. Zejména možnost nalezení pouze g nebo pouze h rozdělení je poměrně zajímavá. Pro tvorbu intervalů spolehlivosti a mocninnou analýzu často postačuje nalezení parametru šikmosti g jako mediánu ze všech g_p z rov (38).

Příklad B (pokračování):

Pro zadaná data jsou grafy šikmosti a špičatosti na obr 1 a 2.



Obr 1 Graf Šikmosti



Obr 2 Graf Špičatosti

Je patrný lineární trend zejména u grafu šikmosti, což ukazuje, že nepostačuje jeden parametr šikmosti g.

5. Problém vybočujících hodnot

Indikace vybočujících měření je velmi důležitý a zároveň ošidný problém. Pokud data obsahují vybočující hodnoty je jejich identifikace a odstranění jednou z nejdůležitějších operací , protože zlepší statistické vlastnosti odhadů, umožní jejich korektní interpretaci a zabrání nesprávné interpretaci výsledků. Bohužel je v praktických úlohách velmi obtížné rozlišit mezi vybočujícími a „obvyklými“ (očekávanými) daty bez znalostí o tvaru rozdělení dat. Celá řada jednoduchých i komplikovanějších pravidel funguje dobře jen pro normálně rozdělená nebo symetrická data Komplikovanější postupy vyžadují nejen a priorní znalost počtu vybočujících hodnot ale také specifikaci jejich rozdělení. Poměrně jednoduché a robustní je pravidlo, kdy se jako vybočující berou měření příliš vzdálená od mediánu. Definuje se rozhodná hodnota C^u jejíž překročení indikuje vybočující prvky výběru. Pro případ, že se identifikují vybočující hodnoty jako extrémně vysoké platí, že

$$C^u = x_{0.5} + k_2 * (x_{0.75} - x_{0.25}) \qquad (48)$$

Pro extrémně nízké hodnoty se v rov (48) provede pouze změna znaménka u druhého sčítance. Parametr k_2 závisí obecně na typu rozdělení dat charakterizovaného šikmostí $_1$ a špičatostí $_2$ a na přípustném podílu hodnot *pp* [%]. které mohou být nesprávně zařazeny jako vybočující, i když výběr žádné neobsahuje Na základě simulací byl nalezen empirický vztah platný pro pp=0.2 %. [8]

$$k_2 = \frac{17.63N - 23.64}{N\left\{7.74 - 3.71/N - 0.83\beta_1 - 0.48\beta_1^2 - 0.48(\beta_2 - 3) + 0.04(\beta_2 - 3)^2\right\}} \qquad (49)$$

Speciálně pro Gaussovo rozdělení vyjde jednoduchý vztah

$$k_2 = \frac{17.63N - 23.64}{7.74N - 3.71} \qquad (50)$$

Při znalosti parametrů šikmostí $_1$ a špičatostí $_2$ je pak možné přímo určit veličinu k_2 a stanovit mez C^u nad kterou se prvky výběru považují za vybočující. Odhady šikmosti a špičatosti je možné určit při znalosti typu g-h rozdělení dle vztahů z kap. 5.

Příklad B (pokračování):
Pro zadaná data vyjde po dosazení do rov (49) , že $k_2 = 3.331$. Z rov. (48) je pak $C^u = 26.05$. Výběr tedy neobsahuje vybočující hodnoty.

6. Intervaly spolehlivosti pro asymetrická rozdělení

Je známo, že pro asymetrická rozdělení je interval spolehlivosti definovaný rov. (11) nepokrývá s pravděpodobností (1-) skutečnou střední hodnotu. Kvalita pokrytí závisí na šikmosti rozdělení a na počtu dat N. V práci [9] byl na rozsáhlém simulačním experimentu určen vztah mezi chybou pokrytí zleva , zprava a z obou stran. Chyba pokrytí zprava *PR* vyjadřuje pravděpodobnost, že skutečná střední hodnota je nižší než meze intervalu spolehlivosti. Pro chybu pokrytí zleva *PL* se určuje pravděpodobnost, že skutečná střední hodnota je vyšší než meze intervalu spolehlivosti. Chyba pokrytí z obou stran *PC* je pak sjednocení obou chyb pokrytí, tj. $PC=PR+PL$.

Pro širokou třídu rozdělení bylo nalezeno, že

$$PR = \alpha/2 + [-0.73 + 0.71 * exp(-\alpha/2)] * \beta_1 / \sqrt{N} \qquad (51)$$

a

$$PL = \alpha/2 + [0.19 + 0.026 * ln(\alpha/2)] * \beta_1 / \sqrt{N} \qquad (52)$$

Z těchto rovnic se dá např. určit potřebná velikost výběru, aby byla zachována chyba pokrytí jako rozdíl mezi požadovanou pravděpodobností pokrytí (např. *0.95*) a dosaženou pravděpodobností pokrytí (např. *0.94*).

Příklad C:
Nechť se konstruuje pravostranný interval spolehlivosti a postačuje místo zvolené pravděpodobnosti 0.95 pouze 0.94. Hodnota $PL=0.06$. Účelem je určit minimální velikost výběru Nm. Po dosazení do rov. (52) a úpravách rezultuje vztah $Nm = [11.2 * \beta_1]^2$. Tedy pokud je šikmost rozdělení dat $_1 = 1$ je třeba provést minimálně 126 měření.

Další možností použití výše uvedených vztahů je fixovat chyby pokrytí na zvolené hodnotě a pro známé N i $_1$ nalézt pravděpodobnost * pro výpočet kvantilu Studentova rozdělení. Takto opravené kvantily se pak dosadí do rov (11).

Příklad D

Omezme se na prakticky často zajímavý pravostranný interval spolehlivosti (jednostranný interval spolehlivosti zprava tj. horní hranici střední hodnoty).Tento interval se často používá u rozdělení zešikmených vpravo k určení povolené horní hranice např. znečištění. Klasický interval spolehlivosti má tvar

$$\mu \le \bar{x} + t_{1-\alpha}(N-1) * \frac{s}{\sqrt{N}} \qquad (53)$$

Po dosazení do rov (51) za *PL = 0.05* rezultuje výraz

$$0 = \alpha^* + [0.19 + 0.026 * \ln(\alpha^*)]\beta_1 / \sqrt{N} - 0.05 = f(\alpha^*)$$

Kořenem funkce $f(\alpha^*)$ je pak * , pro které se spočítá opravený kvantil Studentova rozdělení, tj. hodnota $t_{1-*}(N-1)$.

Další jednoduchou možností jak počítat interval spolehlivosti pro asymetrická rozdělení je korekce založena na Johnsonově transformaci kdy se pro pravostranný interval spolehlivosti definovaný rov (53) použije opravený průměr

$$\mu \le (\bar{x} + \frac{s * \beta_1}{6N}) + t_{1-\alpha}(N-1) * \frac{s}{\sqrt{N}} \qquad (54)$$

Je patrné, že velikost korekce opět souvisí se šikmostí a počtem měření. Na rozdíl od předchozího postupu se však mění poloha centra. Johnsonova transformace založená na Edgeworthově rozvoji t statistiky však není obecně ani monotónní ani v neupravené formě invertovatelná. Tyto problémy eliminuje Hallova transformace (viz. [10]), jejíž inverzní forma má tvar

$$g^{-1}(x) = \frac{\sqrt{N}}{0.33\beta_1}[(1 + \beta_1 * (\frac{x}{\sqrt{N}} - \frac{\beta_1}{6N}))^{1/3} - 1] \qquad (55)$$

Pro horní mez pravostranného intervalu spolehlivosti pak platí,že

$$\mu \le \bar{x} + g^{-1}(z_{1-\alpha}) * \frac{s}{\sqrt{N}} \qquad (56)$$

Místo normovaného normálního kvantilu z se doporučuje použít odpovídajícího kvantilu určeného z Bootstrap výběrů (viz. [10]). Místo transformace definované rov. (55) lze požít zjednodušenou versi

$$ga^{-1}(x) = x - \frac{\beta_1(x^2/3 + 1/6)}{\sqrt{N}} \qquad (57)$$

Tato transformace se pak dosadí do rov (56). Opět je možno použít Bootstrap kvantilů.
Jak je patrné znalost šikmosti výběrového rozdělení je zde nezbytnou podmínkou pro použití korekcí jak kvantilů tak i průměrů.

7. Mocninná transformace dat

Aditivní i multiplikativní model lze vyjádřit jako speciální případy mocninné třídy modelů měření, která je charakterizována tím, že transformací obou stran pomocí funkce *h(.)* vyjde aditivní model

$$h(x) = h(\mu) + \varepsilon \qquad (58)$$

U pravděpodobnostního modelu (10) lze vhodnou transformací dat stabilizovat rozptyl, přiblížit šikmost nule a tvar rozdělení normálnímu rozdělení. S výhodou se jako funkce *h(.)* používá Box Coxova třída polynomických transformací ve tvaru

$$h(x) = \frac{(x-1)^{\lambda}}{\lambda} \quad \lambda \neq 0 \qquad (59)$$

$$h(x) = ln(x) \qquad \lambda = 0$$

kde λ je parametr transformace. Pro $\lambda = 1$ resultuje aditivní model měření a pro $\lambda = 0$ model multiplikativní. S využitím Taylorova rozvoje lze vyjádřit rov. (16) ve tvaru

$$x \approx \mu + \varepsilon / \mu^{1-\lambda} \qquad (60)$$

Pro případ, že rozptyl $D(\varepsilon) = \sigma^2$ je malý jde o aditivní model s nekonstantními chybami, pro který lze použít jako odhad μ vážený aritmetický průměr s vahami úměrnými $\mu^{-(1-\lambda)/2}$.

Lze ukázat, že vhodným odhadem parametru μ (neznámá koncentrace) je výběrový medián, který je invariantní vůči monotónní´transformaci. Transformace *h(x)* vyjádřená rov. (59) je lineární transformací tzv. prosté mocninné transformace

hp(x) = x pro 0 resp hp(x) = ln (x).

Lze dokázat, že pokud *h(x)* je lineární transformací *hp(x)* platí pro retransformované střední hodnoty

$$h^{-1}[E(h(x))] = hp^{-1}[E(hp(x))] \qquad (61)$$

Pro obě transformace je pak odhadem retransformované střední hodnoty např. **zobecněný průměr**

$$M = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^{\lambda} \right)^{1/\lambda} \quad \text{pro } \lambda \neq 0 \qquad (62)$$

resp.

$$M = \left(\prod_{i=1}^{N} x_i \right)^{1/N} \quad \text{pro } \lambda = 0 \qquad (63)$$

Pokud se použije mocninná transformace na **aditivní model** měření vyjde $h(x) = h(\mu + \varepsilon)$. Z Taylorova rozvoje pak resultuje odhad vychýlení vlivem této nekorektnosti

$$B = E(h(x)) - h(\mu) \approx \frac{\sigma^2}{2!} \frac{d^2 h(x)}{dx^2} \downarrow (x = \mu) \qquad (64)$$

Tak např. pro logaritmickou transformaci vyjde $B = -0.5\ ^2$, kde $= /$ je variační koeficient. To odpovídá prvnímu členu v rov. (8).

Prostá mocninná transformace je invariantní vůči změně měřítka, protože , pro konstantu *b* je
$(b*x) = b\ *x$ a tedy $D((b*x)\ ) = D((x)\ )$. Pro případ Box Coxovy transformace je však

$$\frac{(x*b)^{\lambda} - 1}{\lambda} = b^{\lambda} * \frac{x^{\lambda} - 1}{\lambda} + \frac{b^{\lambda} - 1}{\lambda}$$

Tento výraz již nelze upravit stejně jako u prosté mocninné transformace a Box Coxova transformace tedy není invariantní vůči změně měřítka. Detaily lze nalézt v práci [12]. Z uvedeného také přímo plyne, že obě transformace jsou závislé na posunu. Tedy mocninná transformace (x+a) poskytne jiné výsledky než mocninná transformace *x*. (viz dále)
Pro odhad parametru λ je možno použít metodu maximální věrohodnosti. Pokud je ε rozdělené v souladu s předpoklady aditivního modelu měření (normalita a nezávislost) má logaritmus věrohodnostní funkce tvar

$$\ln L(\lambda) = \sum (\lambda - 1) * \ln(x_i) - \frac{1}{2\sigma^2} \sum [h(x_i) - h(\mu)]^2 \qquad (65)$$

Pro pevné λ lze určit maximálně věrohodný odhad rozptylu ve tvaru

$$\sigma_c^2 = \frac{1}{N} \sum [h(x_i) - h(\mu)]^2 \qquad (66)$$

kde se za $h(\mu)$ dosazuje aritmetický průměr transformovaných dat

$$h(\mu) \approx \frac{1}{N} \sum h(x_i) \qquad (67)$$

Po dosazení do věrohodnostní funkce resultuje vztah

$$\ln L^*(\lambda) = \sum (\lambda - 1) * \ln(x_i) - \frac{N * \ln \sigma_c^2}{2} \qquad (68)$$

Maximalizací $\ln L^*(\lambda)$ podle λ (viz.[1]) lze pak snadno určit maximálně věrohodný odhad $\hat{\lambda}$ parametru transformace λ. Je patrné, že je tato úloha ekvivalentní minimalizaci rozptylu v transformovaných proměnných σ_c^2. Na základě Taylorova rozvoje funkce *h(x)* pro pevné vyjde přibližný výraz

$$D(\frac{x^{\lambda} - 1}{\lambda}) = \frac{1}{\lambda^2} D(x^{\lambda}) \approx E(x)^{2\lambda - 2} D(x) = E(x)^{2\lambda} \delta^2$$

kde je variační koeficient. Je zřejmé, že pro pevné bude rozptyl v transformaci tím vyšší, čím bude větší rozptýlení dat. To umožní identifikaci extrému (minima). Pro málo rozptýlená data bude rozptyl v transformaci malý a identifikace extrému bude obtížnější. V práci [3] bylo ukázáno, že pro D(x) 0 je rozptyl D($\hat{\lambda}$) a podobně i rozptyl zobecněného průměru roste nade všechny meze.Pro snadnou identifikovatelnost transformace je tedy výhodné mít větší rozptýlení dat jak je např. běžné u výběrů s asymetrických rozdělení.
Formálně lze úlohu maximalizace rov (68) vyjádřit ve tvaru

$$\frac{d\,ln(L)}{d\lambda} = \sum_i ln(x_i) - \frac{1}{\sigma^2}\sum_i \left(h(x_i) - \frac{1}{N}\sum_i h(x_i) \right) * \frac{dh(x_i)}{d\lambda} = 0 \qquad (69)$$

kde $\frac{dh(x_i)}{d\lambda} = \frac{(1+\lambda * x_i)ln(1+\lambda * x_i) - \lambda * x_i}{\lambda^2}$

Z druhé derivace věrohodnostní funkce lze určit rozptyl maximálně věrohodného odhadu mocninné transformace[13] Po úpravách vyjde : *D($\hat{\lambda}$) = 2(1-0.333* $_1^2$ +0.388 $_2$)/(3Nw),* kde *w = /(1+).* Zde ^{2} , $_1$ a $_2$ jsou rozptyl, šikmost a špičatost původních dat. Je patrné, že pro ^{2} *0* roste rozptyl odhadu mocninné transformace nade všechny meze.
Na základě asymptotického $(1-\alpha)$ % ního intervalu spolehlivosti parametru mocninné transformace lze sestavit nerovnost

$$ln\,L(\lambda) \geq ln\,L(\hat{\lambda}) - 0.5 * \chi^2_{1-\alpha}(1) \qquad (70)$$

Všechna λ splňující tuto nerovnost leží v intervalu spolehlivosti a jsou tedy přijatelná. Toho lze snadno využít pro rozlišení mezi aditivním a multiplikativním modelem měření. V rovnici (22) označuje $\chi^2_{1-\alpha}(1)$ kvantil chí kvadrát rozdělení s 1 stupněm volnosti.
Platí, že:

- pokud obsahuje 95% ní interval spolehlivosti také jedničku, volí se aditivní model.
- pokud obsahuje 95% ní interval spolehlivosti nulu a nikoliv jedničku, volí se multiplikativní model.
- v ostatních případech je možné zvolit pravděpodobnostní model (10) a použít pro další analýzu postup navržený v .[1].

Parametr mocninné transformace zřejmě souvisí s šikmostí rozdělení dat. Pro kvantifikaci tohoto vztahu lze dosadit do podmínky (67) místo h(x) jeho rozvoj do Taylorovy řady a určit maximálně věrohodný odhad analyticky. V práci [14] je toto odvození provedeno. Výsledek lze zapsat ve tvaru

$$\lambda \approx 1 - \frac{E(x) * \sigma * \beta_1}{6} \qquad (71)$$

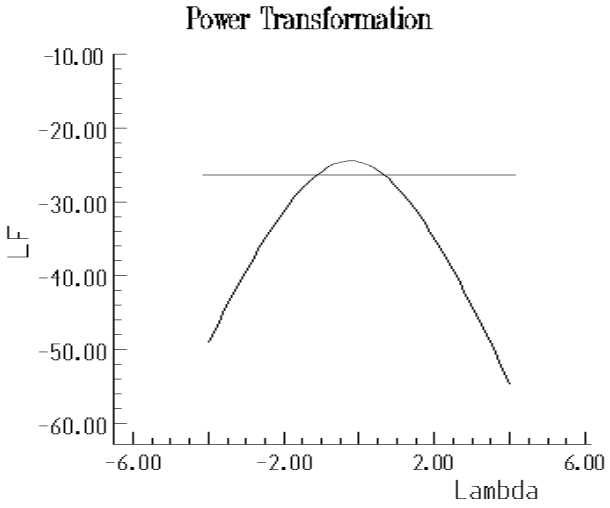
Pomocí toho vztahu můžeme např. snadno posoudit vliv posunu dat na parametr mocninné transformace. Např. pro případ, že data posuneme o konstantu a tj. *y = a*x* vyjde, že

$$\lambda_y = \lambda_x - (a * \sigma * \beta_1)/6$$

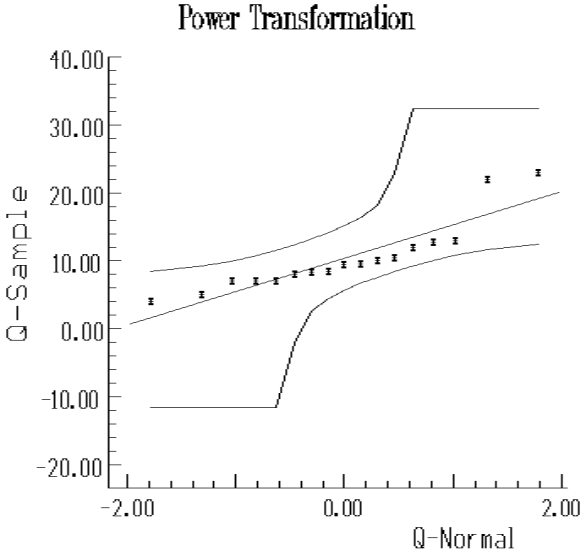
Jak je patrné, je třeba při použití postupu mocninné transformace brát v úvahu také případné lineární transformace dat a jejich rozmezí.

Příklad B (pokračování):

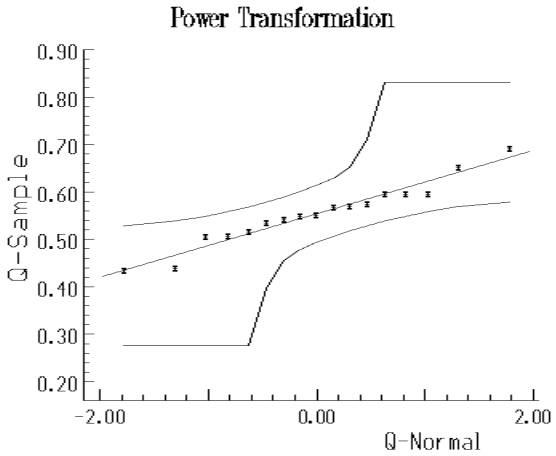
Pro zadaná data je znázorněn průběh věrohodnostní funkce na obr. 3. Rankitový graf pro původní data je na obr 4 a pro transformovaná data na obr 5.



Obr 3. Věrohodnostní funkce



Obr 4. Rankitový graf pro původní data



Obr 5. Rankitový graf pro transformovaná data

Optimální mocnina vyšla –0.23. s mezemi (–1.13, 0.67).
Protože tento interval obsahuje nulu lze provádět další analýzu v logaritmické transformaci resp. volit multiplikativní model měření.Pro 95 Ý ní intervaly spolehlivosti pak vyjde

Z předpokladu normality	Dolní mez: 7.74	Horní mez: 13.069
Z kvantilů (robustní)	Dolní mez: 6.72	Horní mez: 12.55.
Z Box Coxovy transformace	Dolní mez: 7.36	Horní mez: 11.62.

8. Závěr

Je patrné, že statistické zpracování dat v analytické chemii a speciálně ve stopové analýze má celou řadu specifických zvláštností, které je třeba brát v úvahu. Je vždy výhodné začít průzkumovou analýzou a porovnáním resp. selekcí modelů měření a až poté zvolit další cestu. Ve shodě s koncepcí „***satistical methods mining***“ [6] je často nezbytné kombinovat různé přístupy jako je transformace, robustní metody a počítačově intenzivní metody k dosažení rozumných výsledků. Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy bez hlubšího rozboru zde může vést ke katastrofickým výsledkům

Poděkování: Tato práce vznikla s podporou grantu MŠMT č. VS 97084, grantu GAČR . 106/99/1184 a výzkumného záměru MŠMT č.J11/98:244101113

9. Literatura

[1] Meloun M., Militký J.: *Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Militký J., Meloun M.: Konference Mikroelementy "99 ,Řež u Prahy , listopad 1999
[3] Bickel P.J., Doksum K.A.: *J. Amer. Stat Assoc.* 76, 296 (1981)
[4] Massart D.L. a kol. : Chemometrics a textbook, Elsevier Amsterdam 1988
[5] Hoaglin D. C., Mosteler F., Tukey J.W, Eds. : Exploring Data Tables Trends and Shapes , J. Wiley New York 1985, kap. 11
[6] Parzen E.: Proc Ninth Int. conf. on quantitative methods for environmental science, July 1988, Melbourne
[7] Parzen E.: J. Amer. Statist. Assoc. 74, 105 (1985)
[8] Carlig K.: Comput. Statist. and Data Anal. **33**, 249 (2000)
[9] Boos D.D. , Hughes-Oliver J. M.: Amer. Statist. **54**, 121 (2000)
[10] Zhou X.H., Gao S.: Amer. Statist. . **54**, 100 (2000)
[11] Morgenhalter S., Tukey J.W. : J. Comput. Graph. Statist. **9**, 180(2000)
[12] Schlesselman J.: J. Roy Stat. Soc. **B33**, 307 (1971)
[13] Draper N.R., Cox D. R.: J. Roy Stat. Soc. **B31**, 472 (1969)
[14] Box G. E. P., Cox D. R.: J. Roy Stat. Soc. **B26**, 211 (1964)

Název souboru:	ghroy1
Adresář:	E:\Pom
Šablona:	D:\Program Files\Microsoft Office\Sablony\Normal.dot
Název:	Zpracování dat ve stopové analýze
Předmět:	
Autor:	Militky
Klíčová slova:	
Komentáře:	
Datum vytvoření:	14.09.00 13:39
Číslo revize:	2
Poslední uložení:	14.09.00 13:39
Uložil:	Milan Meloun
Celková doba úprav:	0 min.
Poslední tisk:	14.09.00 13:43
Jako poslední úplný tisk	
Počet stránek:	20
Počet slov:	5 677 (přibližně)
Počet znaků:	32 360 (přibližně)