Volume 1. PC-Aided Statistical Data Analysis, Ellis Horwood, Chichester 1992.
[23] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*, Elsevier 1992,
[24] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxford Science Publications 1988,
[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).
[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 1994.
[27] Martens H., Naes T., *Multivariate calibration*, Wiley (1989) Chichester.
[28] Thomas E. V., *Anal. Chem.*, **66** (1994) 795A-804A.
[29] Malinowski F., Howery D., *Factor Analysis in Chemistry*, Wiley (1980) New York.
[30] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních dat*, Univerzita Pardubice, 1996.

Zlepšení rozdělení dat v chemometrii

Jiří Militký[1], Milan Meloun[2]

[1] *Technická univerzita v Liberci, Textilní fakulta, Katedra textilních materiálů, Hálkova 6, 461 17 Liberec, e- mail jiri.militky@vslib.cz*

[2] *Univezita Pardubice, Katedra analytické chemie, Čs. Legií 565, 532 10 Pardubice, e-mail: milan.meloun@upce.cz*

Motto: Všechno je jinak. Ale jak ?

Abstract: The main aim of this contribution is to show the application of data transformation for enhancing of their distribution according to the subsequent statistical analysis. The power transformation and Box Cox transformation are discussed in details. These transformations are used for construction of mean value adaptive estimator and for creation of corresponding confidence intervals

Abstrakt: Cílem příspěvku je ukázat možnosti aplikace transformace dat pro zlepšení jejich rozdělení s ohledem na následnou statistickou analýzu. Je podrobněji pojednáno o mocninné transformaci dat a Box Coxově transformaci. Tyto transformace jsou použity pro konstrukci adaptivního postupu výběru odhadu střední hodnoty a tvorbu intervalu spolehlivosti střední hodnoty.

1. Úvod

Pokud nemá být statistická analýza v chemometrii pouhým numerickým počítáním bez hlubšího smyslu, je pochopitelně třeba, aby byly ověřeny všechny předpoklady, které vedly k návrhu daného postupu analýzy.

Při zpracování výsledků rutinních měření se běžně předpokládá aditivní model měření. O datech x_i} $i = 1, \ldots N$ se apriorně soudí, že jde o nezávislé stejně rozdělené veličiny, pocházející z normálního rozdělení $N(\mu, \sigma^2)$. Tyto předpoklady jsou základem prakticky všech klasických metod analýzy experimentálních 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.

Je třeba mít na paměti, že malé porušení předpokladu normality nemusí být katastrofické s ohledem na výsledek statistické analýzy. Na druhé straně je však špatné, když odhady i testy závisejí na spíše jiných faktorech než je

chování většiny dat (na velikosti výběru, uspořádání výsledků nesledovaných proměnných atd.).

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.

Pokud data nesplňují předpoklad normality, je v řadě případů možné zlepšit jejich rozdělení vhodnou *transformací*.

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 **multiplikativního modelu 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, obyčejně mocninná, transformace dat [1]. I zde však vznikají problémy zejména se zpětnou transformací a použitelností jen pro některé úlohy. Lze odvodit, že pro malé rozptyly σ^2 je odhad parametru mocninné transformace špatně identifikovatelný (viz [3]).

V tomto příspěvku je pozornost zaměřena na techniky zlepšení tvaru rozdělení výběru a jejich využití pro základní úlohy statistického zpracování 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.

$$x = \mu + \varepsilon \tag{1}$$

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

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. V chemometrické analýze, kde jde běžně o měření na různých vzorcích, je častým 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

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

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

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

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

$$x = \mu * \exp(\varepsilon) \tag{4}$$

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

$$\ln(x) = \ln(\mu) + \varepsilon \tag{5}$$

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.

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

$$\ln(x) = \ln(\mu + \varepsilon) = \ln\mu + \ln(1 + \varepsilon/\mu) \tag{6}$$

S využitím Taylorova rozvoje lze psát

$$\ln(x) \approx \ln(\mu + \varepsilon) = \ln\mu + \varepsilon/\mu - 0.5*(\varepsilon/\mu)^2 + 0.33*(\varepsilon/\mu)^3 - \ldots \tag{7}$$

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

$$E(\ln x) = \ln\mu - 0.5*\delta^2 - 0.75*\delta^4 \tag{8}$$

a

$$D(\ln x) = \delta^2 + 2.5*\delta^4 + 4.66*\delta^6 \tag{9}$$

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**.

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

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

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 μ 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

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

K výraznému zkreslení rozptylu výběrového průměru může dojít v případě, že data nejsou nezávislá. To může být situace, kdy se vzorky k analýze odebírají z různých míst, které spolu nějak souvisejí (prostorová nebo časová autokorelace). Pro případ nejjednodušší autokorelace prvního řádu vyjádřené autokorelačním koeficientem ρ dojde ke zvětšení rozptylu střední hodnoty

$$D(\bar{x}) = \frac{\sigma^2}{N(1-\rho^2)}$$

Pro komplikovanější situace (prostorová závislost dlouhého dosahu) může být zkreslení způsobené závislostí v polohách odběru neúměrně vysoké.

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 α 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

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

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) %*.

Při nemožnosti použití výše uvedeného standardního postupu pro reálná data existují v zásadě dvě cesty:

I. Nalezení vhodného rozdělení pro původní data a konstrukce speciálních intervalů spolehlivosti

II. Zlepšení rozdělení původních dat tak aby bylo možno použít pro transformovaná data standardní analýzu

Obě cesty mají své výhody a nevýhody. Možné zlepšení rozdělení dat vhodnou *transformací* je logické použít zejména v případech, kdy je cílem pouze stanovení intervalu spolehlivosti parametru polohy a nikoliv konstrukce pravděpodobnostních modelů. Navíc se velmi snadno určí, zda je toto zlepšení statisticky významné či nikoliv. Na druhé straně však jedna transformace nemusí vyhovovat pro všechna data a vznikají problémy pokud je nutno realizovat zpětnou transformaci.

3. Transformace zlepšující rozdělení dat

S transformací dat se při zpracování experimentů setkáváme velmi často. Podle příčin můžeme transformaci dělit do dvou základních skupin:

A. Transformace zlepšující rozdělení dat. Zde je transformace *žádána* a přispívá ke zlepšení rozdělení dat (zjednodušuje jejich zpracování)

B. Transformace jako důsledek matematických operací (obyčejně realizace funkcí) s měřenými veličinami. To je případ, kdy známe u komplikovaných systémů vstupní náhodné veličiny a zajímá nás výstupní náhodná veličina. Patří sem tedy všechny transformace, kdy na základě experimentálních výsledků počítáme jiné veličiny (např. z hodnot poloměru plochu kruhových elementů). Zde je vlastně transformace *nežádaná*, protože deformuje původní rozdělení dat.

V případě ad A) se *hledá* vhodná transformace. V případě ad B) se hledají vhodné postupy zpracování dat, které *omezují* vliv transformace. Tato dualita způsobuje, že oblasti transformace se v literatuře nevěnuje patřičná pozornost. Nadto vede ke stavu, kdy formálně shodné (matematicky *správné*) metody poskytují značně odlišné výsledky.

Z uvedeného je zřejmé, že transformace může být buď "užitečným nástrojem", nebo "základní překážkou" při statistické analýze dat.

Jak bylo uvedeno v kap. 2, je pro statistickou analýzu dat *ideální*, pokud jsou prvky výběru *náhodné* vzájemně *nezávislé* veličiny se *stejným normálním* rozdělením. Reálné výběry se od tohoto stavu více či méně odlišují.

V jednodušším případě mají *delší konce* (vyšší špičatost), než odpovídá normálnímu rozdělení. To je často důsledek přítomnosti vybočujících měření. Zde je při statistické analýze stále střed symetrie v místě módu, který je totožný s mediánem a střední hodnotou. Efektivní odhad polohy je medián (průměr x má přibližně dvojnásobný rozptyl). Běžné statistické testy jsou vůči vyšší špičatosti dat poměrně robustní (to se týká zejména t-testu významnosti). Také valná většina robustních metod odhadu parametrů polohy a rozptýlení vychází z představy symetrického rozdělení dat, *kontaminovaného* jistým podílem vybočujících dat.

Komplikovanější je případ, kdy je rozdělení výběru *sešikmené* (obyčejně k vyšším hodnotám). Pak již není módus totožný s mediánem ani střední hodnotou a vlastní *interpretace* parametru polohy je ztížena. Efektivní odhad parametru polohy je možný jen při znalosti zákona rozdělení pravděpodobnosti (který však při analýze dat není přesně apriorně znám). Běžné statistické testy jsou vůči sešikmenému rozdělení dat obecně nerobustní. Také základní robustní metody odhadu parametrů polohy a rozptýlení zde nefungují dobře.

Je tedy zřejmé, že již *symetrizační* transformace bude pro analýzu dat velmi užitečná.

Průvodním zjevem u řady "nenormálně" rozdělených výběrů je *nekonstantnost rozptylu* (pouze pro normální rozdělení platí, že střední hodnota je nezávislá na rozptylu).

Transformace stabilizující rozptyl je tedy zároveň transformací vedoucí k *normalitě*.

Otázky spojené s existencí transformace vedoucí k normalitě jsou teoreticky řešeny v práci [5].

4. Transformace stabilizující rozptyl

Nekonstantnost rozptylu je průvodním jevem u řady měření. Indikuje buď neplatnost aditivního modelu měření typu rov. (1) nebo nenormalitu rozdělení náhodné veličiny, ze které byl realizován výběr.

Zde se omezíme na případ, kdy je rozptyl $D(x)$ jistou funkcí velikosti x, což můžeme formálně vyjádřit vztahem

$$D(x) = g(x) \tag{12}$$

Při známém (předpokládaném) tvaru g(x) se pak hledá stabilizující transformace $h(x)$, pro kterou již bude rozptyl konstantní. Elementární vztah pro rozptyl funkce náhodné veličiny je definován rov. (2). Protože je požadavkem výběr takové funkce $h(x)$, aby $D(h(x)) = konst.$ a $D(x) = g(x)$, lze z rovnice (12) snadno nalézt, že

$$h(x) \approx const. \int \frac{dx}{\sqrt{g(x)}} \tag{13}$$

Řešením tohoto integrálu (konstanta const. není důležitá pro tvar transformace) můžeme pak snadno určit transformaci stabilizující rozptyl.

V řadě případů je měření realizováno za podmínky konstantnosti relativní chyby, tj. konstantnosti variačního koeficientu $CV = [\sigma_x/x].\ 10^2$. Rozptyl σ_x^2

v místě x je pak zřejmě $\sigma_x^2 = [CV/10^2]x^2$, funkce (12) je tedy $g(x) = x^2$. Po dosažení do rovnice (13) a analytické integraci pak dostáváme $h(x) = ln(x)$. Použitím logaritmické transformace zde tedy eliminujeme nekonstantnost rozptylu (obecně platí, že tato transformace je vždy výhodná, pokud se jednotlivé *prvky výběru* mění v rozmezí několika řádů).

Častým případem je, že $g(x) = x^P$ (obecná mocninná závislost rozptylu). Pak lze při znalosti konkrétního P z rovnice (13) nalézt stabilizující transformaci $Z(x) = x^{1-P/2}$. Pro silně sešikmená data (jako χ^2 rozdělení) se doporučuje odmocninová transformace. Pro gamma rozdělení je zase stabilizující transformace třetí odmocniny $Z(x) = \sqrt[3]{x}$.

5. Mocninná transformace

Mocninná transformace je poměrně široce využitelná pro řešení celé řady problémů. Platí, že 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

$$h(x) = h(\mu) + \varepsilon \tag{14}$$

U pravděpodobnostního modelu (10) lze vhodnou transformací dat **stabilizovat rozptyl, přiblížit šikmost rozdělení k nule** a tvar rozdělení k **normálnímu rozdělení**. Cílem je na základě znalostí o výběru x_i, $i = 1, \ldots N$ nalézt vhodnou mocninu, resp. vhodný člen (pokud se použije celá rodina transformací).

Nejjednodušší je *prostá mocninná transformace*

$$hp(x) = sign(x)*abs(x)^{\lambda} \text{ pro } \lambda \neq 0 \tag{15}$$

$$hp(x) = ln(x) \text{ pro } \lambda = 0$$

$$hp(x) = exp(c*x) \text{ pro } \lambda \to \infty$$

kde $abs(x)$ je absolutní hodnota a $sign(x)$ je znaménková funkce

$$sign(x) = 1 \text{ pro } x > 0, \quad sign(x) = -1 \text{ pro } x < 0, \quad sign(x) = 0 \text{ pro } x = 0$$

Tato transformace nezachovává měřítko a ani není vzhledem k všude spojitá. Zachovává však pořadí dat ve výběru (jako všechny mocninné transformace).

Používá se jako jednoduchá *symetrizující* transformace a proto se hledá optimální mocnina λ tak, aby byly minimalizovány vhodné míry symetrie výběru. Je možno použít přímo výběrovou šikmost $g_1(y)$, nebo její robustní verzi $g_{R1}(y)$ viz. [1]. Stejně jednoduché je sledovat rozdíl mezi průměrem a medianem v transformaci.

Pro posouzení kvality transformace, resp. nalezení optimálního λ je také možno použít grafu rozptýlení s kvantily (GRK), resp. kvantilových grafů (Q-Q grafů), jejichž konstrukce je popsána v [1].

Nevýhody prosté mocninné transformace (zejména nespojitost v okolí nuly a nesrovnatelnost měřítek v transformaci) odstraňuje *rodina Box-Coxových transformací h(x)*, která je lineární transformací prosté mocninné transformace *hp(x)*. Box Coxova třída polynomických transformací má tvar

$$h(x) = \frac{x^{\lambda} - 1}{\lambda} \quad \lambda \neq 0 \tag{16}$$

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

kde λ je parametr transformace. Pro $\lambda = 1$ resultuje aditivní model měření a pro $\lambda = 0$ model multiplikativní. S využitím Taylorova rozvoje lze odvodit, že v tomto případě je

$$x \approx \mu + \varepsilon / \mu^{1-\lambda} \tag{17}$$

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.

Pokud $h(x)$ je lineární transformací $hp(x)$ platí pro re-transformované střední hodnoty

$$h^{-1}[E(h(x))] = hp^{-1}[E(hp(x))] \tag{18}$$

Pro obě transformace je pak odhadem re-transformované střední hodnoty **zobecněný průměr**

$$M = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^{\lambda} \right)^{1/\lambda} \quad \text{pro } \lambda \neq 0 \tag{19}$$

resp.

$$M = \left(\prod_{i=1}^{N} x_i \right)^{1/N} \quad \text{pro } \lambda = 0 \tag{20}$$

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

$$B = E(h(x)) - h(\mu) \approx \frac{\sigma^2}{2!}\frac{d^2h(x)}{dx^2} \downarrow (x=\mu) \qquad (21)$$

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

Prostá mocninná transformace je invariantní vůči změně měřítka a Box Coxova transformace není invariantní vůči změně měřítka. Detaily lze nalézt v práci [6]. Pro eliminaci této nevýhody lze použít modifikované transformace

$$h(x,p) = \frac{x^\lambda - p^\lambda}{\lambda} \qquad \lambda \neq 0$$

$$h(x,p) = \ln(x/p) \qquad \lambda = 0$$

kde parametr p se volí jako aritmetický průměr, geometrický průměr resp. medián původních dat. Z uvedeného také přímo plyne, že obě transformace jsou závislé na posunu. Tedy mocninná transformace (x+a) poskytne jiné výsledky než mocninná transformace x.

Lze se snadno přesvědčit, že:

- rodina transformací definovaných rovnicí (16) je vzhledem k mocnině λ spojitá. V okolí nuly platí $lim\ (x-1)^\lambda/\lambda = lim\ x.ln(x) = ln(x)$ všechny transformační závislosti $h(x)$ procházejí jedním bodem o souřadnicích y = 0, x = 1 a mají v tomto bodě společnou směrnici (jsou zde, co do průběhu, shodné)
- Mocninné transformace s exponenty -2, -3/2, -1, -0,5, 0, 0,5, 1, 3/2, 2 jsou co do křivosti rovnoměrně rozmístěné.
- Vlivem transformace (16) se však obecně mění charakteristiky polohy a rozptýlení, což komplikuje porovnání různě transformovaných výběrů (nevadí pochopitelně pro přiblížení k normalitě, resp., zesymetričtění výběru).

Pro zajištění toho, aby měla transformovaná data přibližně stejnou polohu a rozptýlení jako data netransformovaná, *je možné použít* dostatečné lineární transformace (viz /2/).

Z hlediska analýzy dat je transformace vždy žádoucí, pokud $x_{(N)}/x_{(1)} > 20$ (kde se předpokládají kladná data).

Rovnice (16) je použitelná pouze pro kladná data. Pokud je znám jiný počátek x_0, pod kterým se data nemohou vyskytovat, volí se zobecněná *mocninná transformace*

$$h(x) = \frac{(x+c)^\lambda - 1}{\lambda} \qquad \lambda \neq 0 \qquad (22)$$

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

Zde c>= x_0. Obecně se hledají u této transformace dva parametry. S ohledem na to, že dosavadní transformace platí pro zdola omezené rozdělení dat, není zřejmě možné, aby jejich rozdělení bylo striktně normální. Pro odstranění této (prakticky nepříliš důležité) nevýhody doporučují Bickel a Doksum *rozšířenou* Box-Coxovu transformaci (pro parametr $\lambda > 0$), která pokrývá celou reálnou osu

$$h(x) = \frac{sign(x) * abs(x)^\lambda - 1}{\lambda} \qquad \lambda \neq 0 \qquad (23)$$

Nevýhodou je, že tato transformace neobsahuje logaritmickou transformaci. Tato transformace je již nezávislá na měřítku.

Pro odhady parametrů v těchto rodinách transformací lze opět použít různých charakteristik šikmosti a špičatosti.

V případě **jednoparametrických rodin transformací** se lze zaměřit pouze na jednu charakteristiku tvaru (obyčejně šikmost). Výhodnější je použití testů normality dat po mocninné transformaci. Známý Shapiro-Wilkův test je úměrný testu významnosti směrnice v Q-Q grafu, takže lze také posuzovat linearitu v Q-Q grafech.

S ohledem na požadavek, aby se rozdělení výběru v transformaci co nejvíce blížilo normálnímu rozdělení, lze pro odhad optimálního použít metodu maximální věrohodnosti.

Pokud platí předpoklady aditivního modelu měření (normalita a nezávislost) má logaritmus věrohodnostní funkce tvar

$$\ln L(\lambda) = \sum(\lambda-1) * \ln(x_i) - \frac{1}{2\sigma^2}\sum\left[h(x_i) - h(\mu)\right]^2 \qquad (24)$$

Pro pevné λ lze určit maximálně věrohodný odhad rozptylu ve tvaru

$$\sigma_c^2 = \frac{1}{N}\sum\left[h(x_i) - h(\mu)\right]^2 \qquad (25)$$

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

$$h(\mu) \approx \frac{1}{N}\sum h(x_i) \qquad (26)$$

Po dosazení do věrohodnostní funkce resultuje vztah

$$\ln L^*(\lambda) = \sum(\lambda-1) * \ln(x_i) - \frac{N * \ln\sigma_c^2}{2} \qquad (27)$$

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 λ vyjde přibližný výraz

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

kde δ je variační koeficient. Je zřejmé, že pro pevné λ 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 $D(\hat{\lambda})\to\infty$ 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 (27) vyjádřit ve tvaru

$$\frac{d\ln(L)}{d\lambda}=\sum_i \ln(x_i)-\frac{1}{\sigma^2}\sum_i\left(h(x_i)-\frac{1}{N}\sum_i h(x_i)\right)*\frac{dh(x_i)}{d\lambda}=0 \qquad (28)$$

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

Z druhé derivace věrohodnostní funkce lze určit rozptyl maximálně věrohodného odhadu mocninné transformace[7] Po úpravách vyjde: $D(\hat{\lambda}) = 2(1-0.333*\beta_1^2+0.388\,\beta_2)/(3Nw)$, kde $w = \lambda\sigma/(1+\lambda)$. Zde σ^2, β_1 a β_2 jsou rozptyl, šikmost a špičatost původních dat. Je patrné, že pro $\sigma^2\to 0$ roste rozptyl odhadu mocninné transformace nade všechny meze.

Na základě asymptotického $(1-\alpha)$% ního intervalu spolehlivosti parametru mocninné transformace lze sestavit nerovnost

$$\ln L(\lambda)\geq \ln L(\hat{\lambda})-0.5*\chi^2_{1-\alpha}(1) \qquad (29)$$

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 (29) 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].

S výhodou lze využít grafického záznamu ln L(λ) na se zakresleným (obyčejně 95 %ním) konfidenčním intervalem. Z takového grafu lze již snadno odhadnout jak kvalitu transformace, tak i posoudit, v jakých mezích se může hodnota λ pohybovat. (Platí, že čím jsou tyto meze užší, je kvalita transformace vyšší, pokud v nich neleží $\lambda = 1$).

Parametr mocninné transformace zřejmě souvisí s šikmostí rozdělení dat. Pro kvantifikaci tohoto vztahu lze dosadit do podmínky (28) místo h(x) jeho rozvoj do Taylorovy řady a určit maximálně věrohodný odhad analyticky. V práci [8] je toto odvození provedeno. Výsledek lze zapsat ve tvaru

$$\lambda\approx 1-\frac{E(x)*\sigma*\beta_1}{6} \qquad (30)$$

kde β_1 je šikmost původních dat. Je patrné, že pro data zešikmená k vyšším hodnotám vyjde parametr transformace podstatně menší než jedna.

Pomocí vztahu (30) 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.

Speciálně pro účely průzkumové analýzy dat (viz. [1]) byl navržen postup, který umožňuje grafické posouzení vhodnosti mocninné transformace. Je použito jednoduché třídy transformací typu

$$h(x)=a*x^{\lambda}+b \quad \lambda\neq 0 \qquad (31)$$

$$h(x)=c*\ln(x)+d \qquad \lambda=0$$

Parametry a, b, c, d volí Emerson a Stotto [9] tak, aby byla zachována přibližná linearita transformace v okolí mediánu, tj.

$$med(x^{\lambda})\approx med(x) \qquad \frac{d}{dx}(med(x^{\lambda})\approx 1$$

Pro určení vhodné transformace se vychází z výběrových kvantilů x_p a mediánu $x_{0,5}$.

Vynesením $y^* = (x_p + x_{1-p}) / 2$ na $x^* = [(x_{1-p} - x_{0,5})^2 + (x_{0,5} - x_p)^2]/(4\, x_{0,5})$ rezultuje v případě možnosti symetrizační transformace lineární závislost, procházející počátkem typu $y^* = (1 - \lambda)\, x^*$. Ze směrnice této závislosti tedy můžeme přímo nalézt odhad parametru transformace λ. Při praktické aplikaci tohoto postupu se volí jednotlivé písmenové hodnoty (viz [1]), pro které je $P_i = 2^{-(i+1)}$, i = 1, ... Pro robustní odhad směrnice $(1 - \lambda)$ se doporučuje počítat pro všechny body směrnice $k_i = y_i^*/x_i^*$ a jako optimální vzít pak medián ze všech k_i.

Uvedený postup je vhodný pro málo a středně sešikmená rozdělení. Cameron [10] ukázal, že pro silně sešikmená rozdělení a kvantily x_p vzdálené od mediánu vzniká na grafu y^* vs. x^* systematická křivost. Pak je vhodné provádět iterativní hledání optimálního, kdy se výsledek z prvého určení směrnice (y^* vs. x^*) dosadí do transformace (31) a v dalších vyneseních se místo kvantilů proměnné x používají transformované kvantily $h(x)$ (určené z předchozího grafu). Také je výhodné v prvních fázích brát spíše směrnice určené z kvantilů ($P_i = 0{,}25$). Z toho plyne, že Emerson-Stottův postup není zcela automatický a vyžaduje často iterativní hledání vhodného λ, kde v každé iteraci se konstruuje graf typu y^* na x^*. Na druhé straně je tento postup velmi jednoduchý a umožňuje posouzení vlivu případných vlivných bodů na výsledek transformace.

Lze se snadno přesvědčit, že všechny uváděné typy transformace jsou členy obecné *Johnsonovy rodiny* transformací J(x). Lze ukázat, že pouze pro tři funkce h(.)pokrývá transformace J(x) celé rozmezí šikmosti a špičatosti.

6. Zpracování transformovaných dat

Pokud vedla transformace dat k přibližné normalitě, lze pro veličiny $y=h(x)$ určit průměr y_p, rozptyl s_y^2, konfidenční interval střední hodnoty $y_p \pm t_{1-\alpha/2} * s_y / \sqrt{N}$ a případně provádět testy významnosti. V řadě případů je dosaženo tímto postupem adekvátních výsledků (i když je lépe použít t-testů vycházejících z d-uřezaného průměru [12]). I přes některé teoretické problémy (viz práce [13] a diskuse k ní), lze tedy v korektní transformaci provádět základní statistickou analýzu dat velmi snadno.

Problém však je, že je často požadováno určit jak statistické charakteristiky, tak i konfidenční intervaly v původních proměnných. Při znalosti parametru transformace λ lze vyčíslit střední hodnotu $E(x)$ původních dat jako nelineární funkci střední hodnoty μ_T a rozptylu σ_T^2 v transformaci.

$$E(X) = \int_{-\lambda/2}^{\infty} \frac{1}{\sigma}\sqrt{1+\lambda y} * f_n\left(\frac{y-\mu_T}{\sigma_T}\right) dy \qquad (32)$$

Zde f_n je hustota pravděpodobnosti normovaného normálního rozdělení. Pro $\lambda = 0$ vyjde po dosazení do rov. (32) a integraci, že

$$E(x) = \exp(\mu_T + 0.5\sigma_T^2) \qquad (33)$$

a pro $\lambda = 0.5$ je

$$E(x) = [0.5\mu_T + 1]^2 + 0.5\sigma_T^2 \qquad (34)$$

Přesnější aproximace $E(x)$ pro logaritmickou transformaci má tvar

$$E(x) = \exp(\mu_T + 0.5\sigma_T^2) * \left[1 - \frac{\sigma_T^2\left(\sigma_T^2 + 2\right)}{4N} + \frac{\sigma_T^4(3\sigma_T^4 + 44\sigma_T^2 + 84}{96N^2}\right]$$

Pro určení intervalu spolehlivosti lze využít asymptotické normality střední hodnoty v transformaci. Výsledný interval má tvar

$$h^{-1}(\mu_T - t_{1-\alpha/2}(N-1)\sigma_T/\sqrt{N}) \le \mu \le h^{-1}(\mu_T - t_{1-\alpha/2}(N-1)\sigma_T/\sqrt{N})$$

Tento interval však již nemusí obsahovat uprostřed parametr polohy. Modifikovaný postup je popsán v práci [14]. Postup založený na re-transformaci je také popsán v knize [1].

Z uvedeného je zřejmé, že zpětná transformace je dosti komplikovaný problém. Většina odhadů střední hodnoty je *vychýlená* a mají také *větší rozptyly*. Proto je vždy výhodné pracovat jen s transformovanými hodnotami (pokud není nezbytně nutné znát charakteristiky původních veličin). Tento přístup vyhovuje zejména při realizaci testů významnosti, kde může být celá analýza v transformaci. Také při pravděpodobnostních úvahách lze pracovat pouze v transformaci.

7. Příklad

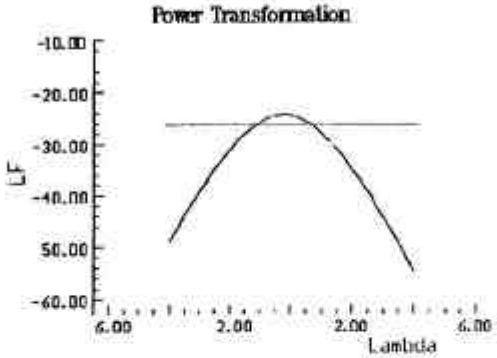
Byl stanoven obsah antimonu v ppm u N=17 vzorků měděné rudy. Setříděná data jsou:

4,5,7,7,7,8,8.3,8.4,9.4,9.5,10,10.5,12,12.8,13,22,23

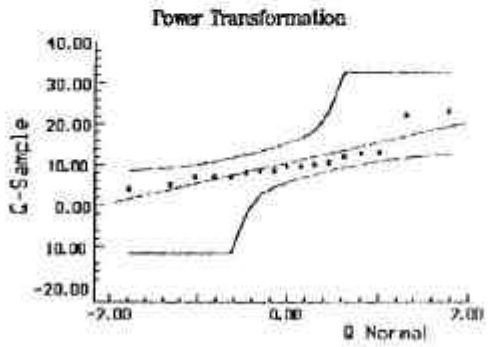
Standardní statistická analýza vede k odhadům. Průměr aritmetický = 10.406, průměr geometrický =9.421,rozptyl = 26.83 ,šikmost = 1.399, špičatost = 4.272.

Kvantilové míry jsou medián = 9.5, dolní kvartil = 7, horní kvartil = 12. Kvantilová analýza vede k závěrům, že rozdělení je mírně zešikmené a má velmi dlouhé konce.

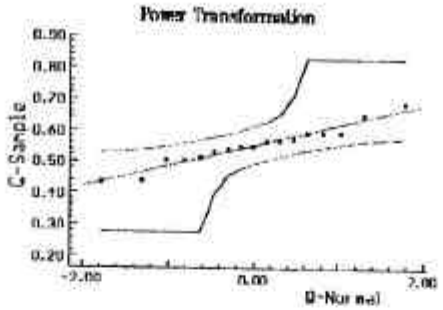
Pro zadaná data je znázorněn průběh věrohodnostní funkce na obr.1. Rankitový graf pro původní data je na obr.2 a pro transformovaná data na obr. 3.



Obr 1. Věrohodnostní funkce



Obr 2. Rankitový graf pro původní data



Obr 3. 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 % ni 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

Tyto výsledky byly získány pomocí programu ADSTAT 1.25. Pro ilustraci jsou uvedeny také výsledky programu ACOX v jazyce MATLAB.

****

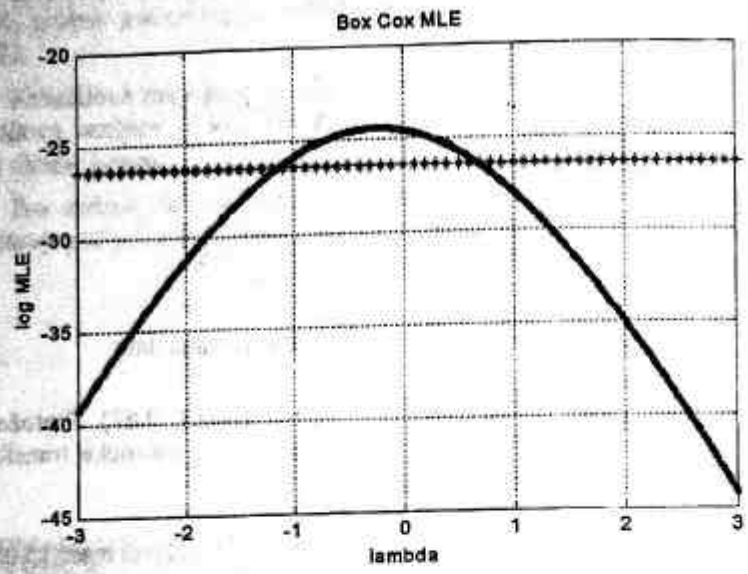
Puvodni data .
Prumer arit. = 10.4059.
Prumer geom. = 9.42055.
Rozptyl = 26.8343.
Sikmost = 1.27746.
Spicatost = 3.78427.
Odhad lambda presny = 0.571042.
Odhad lambda hruby = 0.517177.

****

Optim lambda Box Cox-MLE -0.23.
Konf. interval -1.13 < lambda< 0.67.
Transformace .
Prumer = 1.73925.
Rozptyl = 0.0711426.
Sikmost = 0.000514101.
Spicatost = 2.70961.

****

Optim lambda Box Cox-Shapir Wilk -0.25.
Smernice Q Q grafu 1.11064.



Obr 4. Věrohodnostní funkce MATLAB

Je patrné, že oba programy poskytují pro tytéž parametry shodné výsledky. Program ACOX počítá také odhady z šikmosti a ze směrnice v Q-Q grafu, které byly podle simulací v práci [15] nalezeny jako nejlepší pro širokou skupinu různých rozdělení. Program ACOX je k dispozici u prvního autora tohoto sdělení.

8. Závěr

Je patrné, že statistické zpracování dat v chemometrii 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**" [11] je často nezbytné kombinovat různé přístupy jako je transformace, robustní metody a počítačově intenzivní metody k dosažení rozumných výsledků.

Speciálně při transformaci dat je třeba mít na paměti, že jde o datově orientovaný přístup a pro dva různé výběry z téhož rozdělení lze získat různé odhady parametru transformace. Vodítkem může být kvalita transformace vyjádřená intervalem spolehlivosti parametru λ. Také vztah k šikmosti dat vyjádřený rov. (30) indikuje často vhodnost transformace s ohledem na možné vybočující hodnoty (lze počítat šikmost ze všech dat a bez vybočujících bodů a porovnat odhady λ).

Je varující, že formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy bez hlubšího rozboru zde může vést ke zkresleným závěrům.

Poděkování:

Tato práce vznikla s podporou grantu GAČR 106/99/1184 a výzkumného záměru MŠMT č.J11/98:244100003

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] Efron B.: *Annals of Statist.* **10**, 323 (1982)

[6] Schlesselman J.: *J. Roy Stat. Soc.* **B33**, 307 (1971)

[7] Draper N.R., Cox D. R.: *J. Roy Stat. Soc.* **B31**, 472 (1969)

[8] Box G. E. P., Cox D. R.: *J. Roy Stat. Soc.* **B26**, 211 (1964)

[9] Emerson J.D., Stotto M.A: *J.Amer.Stat.Assoc.* **77**, 103 (1982)

[10] Cameron M.: *J.Amer. Statist. Assoc.* **79**, 107 (1984)

[11] Parzen E.: Proc *Ninth Int. conf. on quantitative methods for environmental science*, July 1988, Melbourne

[12] Doksum K., Wong Ch.W.: *J.Amer.Statist. Assoc.* **78**, 411 (1983)

[13] Hinkley D.V., Runger G.: *J.Amer.Statist.Assoc.* **79**, 302 (1984)

[14] Berger G., Cassela. R.: *Amer. Statist* **46**, 279 (1992)

[15] Gaudard M., Karson M.: *Commun. Statist. Simula.* **29**, 559 (2000)