ZPRACOVÁNÍ VÝBĚRŮ Z ASYMETRICKÝCH ROZDĚLENÍ

JIŘÍ MILITKÝ , Katedra textilních materiálů, Technická universita v Liberci, 461 17 Liberec
MILAN MELOUN , Katedra analytické chemie, Universita Pardubice, Pardubice

Abstrakt

Jsou popsány základní postupy pro určování parametru polohy (střední hodnoty) a odpovídajícího intervalu spolehlivosti pro typická data z oblasti monitorování úrovně škodlivin v životním prostředí. Pozornost je zaměřena jak na metody transformace dat, tak i na postupy vycházející z eliminace šikmosti rozdělení Studentovy t- statistiky. Jsou vybrány metody, které jsou jednoduchým rozšířením klasických postupů a lze je snadno realizovat i bez náročnějších výpočtů.

1. Úvod

Jednou ze základních úloh analytické chemie v oblasti životního prostředí je monitorování úrovně škodlivin v ovzduší, vodě a půdě. Cílem je zjištění, zda daná škodlivina nepřekračuje povolenou úroveň kontaminace. Standardně se postupuje tak, že se na základě měření (x_1..... x_N) stanoví vhodný odhad střední hodnoty μ a porovná se s povolenou úrovní μ_o. S ohledem na variabilitu měření je vhodné ověřit, zda μ_o padne do intervalu spolehlivosti *CI* parametru μ či nikoliv. Data z oblasti životního prostředí mají standardně některé specifické zvláštnosti:

I. Obsahují často extrémně velké hodnoty, které však nejsou důsledkem chyb měření
II. Mohou být cenzurována zdola s ohledem na limitu detekce přístrojů
III. Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám
IV. Jejich počet je omezen díky drahému vzorkování a složitému analytickému vyhodnocení
V. Jsou často prostorově nebo časově závislá, protože zdroj znečištění ovlivňuje okolí
VI. Není možní opakování stanovení (tj. vzorkování a měření) za stejných podmínek, protože se koncentrace škodlivin mění jak v čase, tak i v prostoru.

Tyto zvláštnosti pak omezují použití různých technik založených na průzkumové analýze a identifikaci vybočujících měření. Také robustní techniky zde selhávají, protože eliminují extrémy, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

Standardní statistická analýza zde vede k přehnaně optimistickým závěrům. Platí totiž , že i když zde může být aritmetický průměr x_A asymptoticky nevychýleným odhadem je s velkou pravděpodobností menší než skutečná hodnota parametru polohy μ. Dochází tedy k podcenění odhadu střední hodnoty a tím ke „zdánlivému“ menšímu zjištěnému obsahu škodlivin. To může mít až katastrofické následky v případech, kdy se jedná o životu nebezpečné látky. Standardně se tento problém řeší tak, že se místo aritmetického průměru použije horní mez odpovídajícího intervalu spolehlivosti (viz. např. doporučení US Environmental Protection agency –EPA). Pokud je velikost výběru malá a data jsou silně zešikmená nezajišťuje standardní interval spolehlivosti požadované pokrytí a navíc je systematicky vychýlený.

V tomto příspěvku jsou navrženy metody pro dílčí problémy spojené s velikostí výběru a zešikmením rozdělení dat. Problémy s cenzurovanými výběry a závislostí dat lze řešit pomocí některých postupů popsaných např. v [1, 2].

2. **Základní pojmy**

Standardní způsob zpracování jednorozměrných výběrů spočívá ve výpočtu aritmetického průměru x_A a výběrového rozptylu s^2. Je známo, že pokud zpracovávaný výběr velikosti N prochází z ne - normálního rozdělení se střední hodnotou μ a rozptylem σ^2 *(<∞)* má náhodná veličina

$$Z = \sqrt{N} * (x_A - \mu) / \sigma \qquad (1)$$

asymptoticky normální rozdělení. Pokud není σ^2 známo, nahrazuje se výběrovou směrodatnou odchylkou *s*. Pak má tzv. Studentova náhodná veličina

$$t = \sqrt{N} * (x_A - \mu) / s \qquad (2)$$

Studentovo rozdělení s *(N - 1)* stupni volnosti. Asymptotická normalita veličiny Z resp. Studentovo rozdělení veličiny *t* umožňuje konstrukci intervalu spolehlivosti střední hodnoty μ. Při tzv. frekventistickém přístupu je 100 (1 - α) % na interval spolehlivosti *CI* definován vztahem

$$P(CID \le \mu \le CIH) = 1 - \alpha \qquad (3)$$

Symbol P (.) označuje pravděpodobnost a α je tzv. hladina významnosti. Obyčejně se volí α = 0.05 nebo α = 0.01 s tím, že čím je α menší, tím je interval (CID, CIH) širší. Při znalosti rozptylu σ^2 je možno interval spolehlivosti *CI* vyjádřit ve tvaru

$$x_A - z_{1-\alpha/2} * \frac{s}{\sqrt{N}} \le \mu \le x_A + z_{1-\alpha/2} * \frac{s}{\sqrt{N}} \qquad (4)$$

kde $z_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení. Pokud není σ^2 známo lze použít vztah

$$x_A - t_{1-\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \le \mu \le x_A - t_{\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \qquad (5)$$

kde $t_{1-\alpha/2}(N-1)$. a $t_{\alpha/2}(N-1)$ jsou kvantily Studentova rozdělení s *N-1* stupni volnosti. Pro případ normálního rozdělení mají intervaly (4) resp. (5) přesně 100*(1-*α*) % ní pokrytí střední hodnoty*.To znamená, že jen v 100α*/2 %* případů je *střední hodnota menší než CI* (nejistota *NP* zprava) a v 100α*/2 %* případů je větší než *CI* (nejistota *NL* zprava).Pro případ ne normálního rozdělení platí tyto intervaly pouze asymptoticky tedy pro dostatečně vysoká *N*.Pro jak vysoká *N* lze tyto intervaly použít závisí silně na šikmosti $g_1(x)$ rozdělení z kterého data pocházejí [3].

Pro kvantifikaci vlivu šikmosti na rozdělení náhodné veličiny *Z* definované rov. (1) je možno použít prvního členu Edgeworthova rozvoje pro, který platí

$$P(Z \le x) = F_n(x) - \frac{g_1(x) * (x^2 - 1)}{6\sqrt{N}} f_n(x) \qquad (6)$$

Zde $F_n(x)$ je distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je odpovídající hustota pravděpodobnosti. Šikmost náhodné veličiny Z je dána vztahem

$$g_1(Z) = g_1(x)/\sqrt{N} \qquad (7)$$

Čím je $g_1(Z)$ blíže k nule, tím je rozdělení veličiny Z bližší normálnímu. Z rov. (6) je patrné, že pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ kladné), je také rozdělení náhodné veličiny Z zešikmené k vyšším hodnotám (tj. $g_1(Z)$ kladné). Interval spolehlivosti (4) pak má vyšší horní mez *CIH* a menší dolní mez *CID* než odpovídá reálnému rozdělení statistiky Z. Např. pro výběr rozsahu $N=10$ ze standardizovaného exponenciálního rozdělení, kdy je $g_1(x)=2$, je 97.5 % ní kvantil rozdělení veličiny Z určený z rov. (6) roven 2.24 a odpovídající kvantil normovaného normálního rozdělení je pouze 1.96. Podobně lze určit, že 2.5 % ní kvantil Z je pouze –1.65 oproti odpovídajícímu kvantilu normovaného normálního rozdělení -1.96. Interval spolehlivosti definovaný rov. (4) je tedy celý <u>posunut doprava</u> oproti skutečnému [3].

Také pro kvantifikaci vlivu šikmosti na rozdělení náhodné veličiny t definované rov. (2) je možno použít prvního členu Edgeworthova rozvoje

$$P(t \le x) = F_n(x) + \frac{g_1(x)*(2x^2+1)}{6\sqrt{N}} f_n(x) \qquad (8)$$

Zde je opět $F_n(x)$ distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je odpovídající hustota pravděpodobnosti. Při porovnání s rov. (6), je patrné opačné znaménko korekčního členu, což znamená, že pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ kladné), je rozdělení náhodné veličiny t zešikmené k nižším hodnotám (tj. $g_1(t)$ záporné). Interval spolehlivosti (5) pak má nižší horní mez *CIH* a větší dolní mez *CID* než odpovídá reálnému rozdělení statistiky t. Interval spolehlivosti definovaný rov. (5) je tedy celý <u>posunut doleva</u> oproti skutečnému [3]. To je zvláště nepříjemné u dat silně zešikmených vpravo a vede to k k přehnaně optimistickým závěrům o úrovni kontaminace. ***Postup doporučený EPA pak vlastně nevyčísluje horní mez 95 % ního intervalu spolehlivosti, ale jinou mez závislou na šikmosti dat a velikosti výběru.***

Důvodem toho rozdílu mezi chováním náhodné veličiny Z a t je v korelaci mezi odhady x_A a s . Asymptotický korelační koeficient je roven [3].

$$\rho(x_A, s) = \frac{g_1(x)}{\sqrt{(g_2(x)-1}} \qquad (9)$$

kde $g_2(x)$ je šikmost rozdělení dat.

Je patrné, že problémy s výpočtem intervalů spolehlivosti střední hodnoty nastávají pokud je rozdělení dat <u>ne-normální</u> (zešikmené vpravo) a velikost výběru je malá. Přitom, co je malá velikost výběru závisí na šikmosti rozdělení dat.

Problémem je nejen posun intervalu spolehlivosti definovaného rov (5) směrem k nižším hodnotám, ale také to, že pro pozitivně zešikmená rozdělení je odhad x_A s velkou pravděpodobností menší, než μ. Na druhé straně bylo určeno, že interval spolehlivosti definovaný rov.(5) je poměrně robustní.

V dalším se omezíme na základní techniky omezení vlivu zešikmení dat

A. Snížení asymetrie rozdělení náhodné veličiny *t*
B. Výpočet korigovaného průměru
C. Symetrizační transformace dat

V případech (A) a (C) jde o použití vhodné transformace vedoucí ke zlepšení statistických vlastností testovacích statistik (A), resp. původních dat (B). Ani jeden z těchto postupů není prost jistých omezení a vždy je využito rozvoje do řady a použití několik prvních členů. V případě (B) se používá klasický interval spolehlivosti pro korigovaný průměr, který je blíže střední hodnotě (vyšší než x_A).

3. **Omezení asymetrie rozdělení Studentovy statistiky**

Asymetrie rozdělení t statistiky je zřejmá z Edgeworthova rozvoje definovaného rov. (8). Johnson navrhl nahradit čitatel rov (2) několika členy inverzního Cornish Fisherova rozvoje.

$$t_J = \sqrt{N} * [(x_A - \mu) + \frac{g_1(x) * s}{6N} + \frac{g_1(x)}{3s}(x_A - \mu)^2] / s \qquad (10)$$

Pro tuto transformaci již přibližně platí, že

$$P(t_J \le x) \approx F_n(x) \qquad (11)$$

Johnsonova transformace *t* statistiky však není obecně ani monotónní ani v neupravené formě invertovatelná. Tyto problémy eliminují transformace navržené Hallem [4]

$$t_H = K + \frac{g_1(x) * K^2}{3} + \frac{g_1(x)^2 * K^3}{27} + \frac{g_1(x)}{6N} \qquad (12)$$

resp.

$$t_{H1} = \frac{g_1(x)}{6N} + \frac{3 * \sqrt{N} * exp(\frac{2 * K * g_1(x)}{3\sqrt{N}} - 1)}{2 * g_1(x)} \qquad (13)$$

Zde

$$K = \frac{x_A - \mu}{s}$$

Obě tyto transformace násobené faktorem $N^{0.5}$ splňují rov (11) tj. vedou k přibližné normalitě (redukci šikmosti) a jsou invertovatelné . Inverzní forma statistiky t_H se zahrnutou násobivou konstantou má tvar

$$t_H^{-1}(y) = \frac{3 * \sqrt{N}}{g_1(x)} [(1 + g_1(x) * (\frac{y}{\sqrt{N}} - \frac{g_1(x)}{6N}))^{1/3} - 1] \qquad (14)$$

Při sledování úrovně škodlivin je prakticky zajímavý pouze 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 Pro horní mez pravostranného intervalu spolehlivosti pak platí,že

$$\mu \le x_A + t_H^{-1}(z_{1-\alpha}) * \frac{s}{\sqrt{N}} \qquad (15)$$

Inverzní forma pro t_{H1} je uvedena c práci [4].
Místo normovaného normálního kvantilu z se doporučuje použít odpovídajícího kvantilu určeného z Bootstrap výběrů (viz. [4]). Místo transformace definované rov. (14) lze požít zjednodušenou versi

$$t_a^{-1}(y) = y - \frac{g_1(x) * (y^2/3 + 1/6)}{\sqrt{N}} \qquad (16)$$

Tato transformace se pak dosadí do rov (15). Opět je možno použít Bootstrap kvantilů. Jak je patrné znalost šikmosti výběrového rozdělení je zde nezbytnou podmínkou pro použití korekcí.

V práci [3] byl na rozsáhlém simulačním experimentu určen vztah mezi nejistotou pokrytí zleva , zprava a z obou stran. Nejistota pokrytí zprava *NP* vyjadřuje pravděpodobnost, že skutečná střední hodnota je nižší než meze intervalu spolehlivosti. Pro nejistotu pokrytí zleva *NL* se určuje pravděpodobnost, že skutečná střední hodnota je vyšší než meze intervalu spolehlivosti. Nejistota pokrytí z obou stran *NC* je pak sjednocení obou chyb pokrytí, tj. *NC=NP+NL*.

Pro širokou třídu rozdělení bylo nalezeno, že

$$PR = \alpha/2 + [-0.73 + 0.71 * exp(-\alpha/2)] * g_1/\sqrt{N} \qquad (17)$$

a

$$PL = \alpha/2 + [0.19 + 0.026 * ln(\alpha/2)] * g_1/\sqrt{N} \qquad (18)$$

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

Další možností použití výše uvedených vztahů je fixovat nejistotu pokrytí na zvolené hodnotě a pro známé *N* i $g_1(x)$ nalézt pravděpodobnost * pro výpočet kvantilu Studentova rozdělení. Takto opravené kvantily se pak dosadí do rov (5). Klasický pravostranný interval spolehlivosti má tvar

$$\mu \le x_A + t_{1-\alpha}(N-1) * \frac{s}{\sqrt{N}} \qquad (19)$$

Po dosazení do rov (18) za $PL = 0.05$ rezultuje výraz

$$0 = \alpha^* + [0.19 + 0.026 * ln(\alpha^*)] g_1(x)/\sqrt{N} - 0.05 = f(\alpha^*)$$

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

4. Výpočet korigovaného průměru

Jednoduchá možnost jak počítat korigovaný průměr pro stanovení intervalu spolehlivosti u asymetrických rozdělení je založena na Johnsonově transformaci. Opravený průměr x_O má tvar

$$x_O = (x_A + \frac{s * g_1}{6N}) \tag{20}$$

Je patrné, že velikost korekce opět souvisí se šikmostí a počtem měření. Na rozdíl od předchozího postupu se však mění poloha centra.

Další možností je použití odhadů minimalizujících penále za přecenění resp. nedocenění odhadu střední hodnoty. Chenová zavedla tzv. MCE odhad x_{MCE} ve tvaru

$$x_{MCE} = x_A + d * s \tag{21}$$

kde *d* se počítá podle vztahu

$$d = 0.5 * \left[b - \frac{2\sqrt{N}}{g_1(x)} + \sqrt{4 - \frac{b^2}{3} + \frac{4 * N}{g_1(x)} + \frac{8 * log(a) * \sqrt{N}}{b * g_1(x)}} \right] \tag{22}$$

Volba *a* a *b* souvisí se zvoleným penále. Doporučuje se *a = 1* a *b = 2* i když na základě simulací vychází spíše *a = 10* a *b = 3.* Zajímavé je použití koncepce vycházející z kompromisu mezi vychýlením odhadu a pravděpodobností, že bude ležet nad střední hodnotou. Na tomto základě byl navržen penalizovaný průměr x_P , pro který platí, že

$$x_P = x_A + \frac{4.5 * s^2}{\sqrt{N}} f(x_A)[1 - F(x_A)] \tag{23}$$

Zde $f(x_A)$ resp $F(x_A)$ jsou hodnoty hustoty pravděpodobnosti a distribuční funkce, které se nahrazují neparametrickými odhady. Pro určení $f(x_A)$ se doporučuje vztah

$$f(x_A) = \frac{int(\sqrt{N})}{2 * N * A(x_A)} \tag{24}$$

Zde $A(x_A)$ se bere jako *k*- tá nejmenší hodnota rozdílů $w_i = abs(x_i - x_A)$, *kde* $k = int(N^{0.5})$. Jde vlastně o k tou pořádkovou statistiku. Hodnota distribuční funkce se počítá jako počet hodnot prvků výběru ležících pod x_A dělený *N*. Je možné použít i dalších neparametrických odhadů založených např. na pořádkových statistikách. Dalším zlepšením je použití upraveného výběru uvažujícího extrémy. V upraveném výběru se nejvyšší pořádková statistika $x_{(N)}$ nahrazuje hodnotou $x_A + 4.5\ s$, pokud je větší.Tato modifikace se doporučuje pro silně zešikmená rozdělení, kde se vyskytují hodnoty, sice extrémně vysoké, ale patřící do výběru.

7. Transformace dat

Je známo, že vhodnou transformací dat *h(x)* lze stabilizovat rozptyl, přiblížit šikmost nule a tvar rozdělení normálnímu rozdělení [1]. S výhodou se jako funkce *h(.)* používá Box Coxova třída polynomických transformací ve tvaru

$$h(x) = \frac{(x-1)^\lambda}{\lambda} \quad \lambda \neq 0 \tag{25}$$

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

Pro rozdělení zešikmená k vyšším hodnotám postačuje uvažovat interval $0 \le \lambda \le 1$. Lze ukázat, že vhodným odhadem parametru μ (neznámá koncentrace) je výběrový medián, který je invariantní vůči monotónní´transformaci.

Transformace *h(x)* vyjádřená rov. (25) 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

$$h^{-1}[E(h(x))] = hp^{-1}[E(hp(x))] \qquad (26)$$

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

$$M = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^{\lambda} \right)^{1/\lambda} \quad \text{pro } \lambda \neq 0 \qquad (27)$$

resp.

$$M = \left(\prod_{i=1}^{N} x_i \right)^{1/N} \quad \text{pro } \lambda = 0 \qquad (28)$$

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 v tansformaci docíleno normality a nezávislosti má logaritmus věrohodnostní funkce tvar

$$ln\, L(\lambda) = \sum (\lambda - 1) * ln(x_i) - \frac{1}{2\sigma^2} \sum [h(x_i) - h(\mu)]^2 \qquad (29)$$

Pro pevné λ lze určit maximálně věrohodný odhad rozptylu ve tvaru

$$\sigma_c^2 = \frac{1}{N} \sum [h(x_i) - h(\mu)]^2 \qquad (30)$$

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

$$h(\mu) \approx \frac{1}{N} \sum h(x_i) \qquad (31)$$

Po dosazení do věrohodnostní funkce resultuje vztah

$$ln\, L^*(\lambda) = \sum (\lambda - 1) * ln(x_i) - \frac{N * ln\, \sigma_c^2}{2} \qquad (32)$$

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*. Z druhé derivace věrohodnostní funkce lze určit rozptyl maximálně věrohodného odhadu mocninné transformace[7] Po úpravách vyjde :

$D(\hat{\lambda}) = 2(1-0.333*g_1^2 +0.388\ g_2)/(3Nw)$, kde $w =$ $/(1+$).

Zde ^{2} , g_1 a g_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

$$ln L(\lambda) \geq ln L(\hat{\lambda}) - 0.5 * \chi^2_{1-\alpha}(1) \qquad (33)$$

Všechna λ splňující tuto nerovnost leží v intervalu spolehlivosti a jsou tedy přijatelná. V rovnici (22) označuje $\chi^2_{1-\alpha}(1)$ kvantil chí kvadrát rozdělení s 1 stupněm volnosti.

Parametr mocninné transformace zřejmě souvisí s šikmostí rozdělení dat. Pro kvantifikaci tohoto vztahu lze dosadit do podmínky (32) 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

$$\lambda \approx 1 - \frac{E(x) * \sigma * g_1(x)}{6} \qquad (34)$$

Je patrné, že pro data zešikmená k vyšším hodnotám vyjde parametr transformace podstatně menší než jedna.
Při znalosti parametru transformace lze vyčíslit střední hodnotu *E(x)* původních dat jako nelineární funkci střední hodnoty μ_T a rozptylu σ_T^2 v transformaci.

$$E(X) = \int_{-\lambda/2}^{\infty} \frac{1}{\sigma} \sqrt{1+\lambda y} * f_n\left(\frac{y-\mu_T}{\sigma_T}\right) dy \qquad (35)$$

Zde f_n je hustota pravděpodobnosti normovaného normálního rozdělení. Pro $\lambda = 0$ vyjde po dosazení do rov. (35) a integraci, že

$$E(x) = exp(\mu_T + 0.5\sigma_T^2) \qquad (36)$$

a pro $\lambda = 0.5$ je

$$E(x) = [0.5\mu_T + 1]^2 + 0.5\sigma_T^2 \qquad (37)$$

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

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

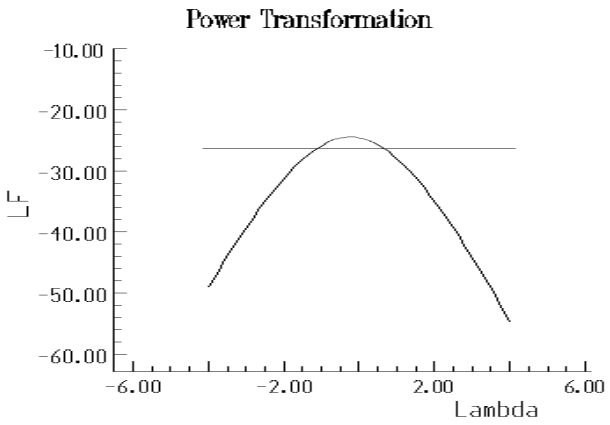
Pro určení intervalu spolehlivosti lze využít asymptotické normality střední hodnoty v transformaci. Výsledný interval má tvar

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

Tento interval však již nemusí obsahovat uprostřed parametr polohy. Modifikovaný postup je popsán v práci [9].

Příklad:
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.

Pro zadaná data je znázorněn průběh věrohodnostní funkce na obr. 1.



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 Ỳ 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***“ 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] Shuway, R.M., Atazi, A.S., Johnson, P.: Technometrics 31, 347 (1989)
[3] Boos D.D. , Hughes-Oliver J. M.: Amer. Statist. **54**, 121 (2000)
[4] Hall, P.: J.R. Stat. Sor. 54, 221 (1992)
[5] Chen L. : Environmetrica **6,** 181 (1995)
[6] Chen L.: J. Appl. Statist. **25**, 739 (1998)
[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] Berger G., Cassela. R.: Amer. Statist ***46***, 279 (1992)

Název souboru:	Asym1
Adresář:	E:\Konference\Konfer-prednasky\2000\Kom-Org-Militky
Šablona:	D:\Program Files\Microsoft Office\Sablony\Normal.dot
Název:	Asymetrické rozdělení
Předmět:	
Autor:	katedra textilních materiálů
Klíčová slova:	
Komentáře:	
Datum vytvoření:	11.10.00 13:50
Číslo revize:	2
Poslední uložení:	11.10.00 13:50
Uložil:	Milan Meloun
Celková doba úprav:	5 min.
Poslední tisk:	11.10.00 14:36

Jako poslední úplný tisk

Počet stránek:	10
Počet slov:	2 783 (přibližně)
Počet znaků:	15 865 (přibližně)