

ZÁKLADNÍ ANALÝZA DAT S OHLEDEM NA LIMITU DETEKCE

JIŘÍ MILITKÝ,

Katedra textilních materiálů, Technická universita v Liberci,
461 17 Liberec

MILAN MELOUN,

Katedra analytické chemie, Universita Pardubice, Pardubice

Motto:

I pořadí nese informaci

Abstrakt:

Jsou popsány základní postupy analýzy dat pro určování parametru polohy (střední hodnoty koncentrace analytu), rozptýlení a odpovídajícího intervalu spolehlivosti pro normální a pro lognormální rozdělení (typická data z oblasti stopové analýzy). Jsou popsány metody založené na metodě maximální věrohodnosti a metodách využívajících pořádkové statistiky které umožňují zpracovat výběry, kde některá měření jsou pod prahem detekce. Na typovém příkladu je provedeno porovnání jednotlivých postupů

1.Úvod

Jednou ze základních úloh analytické chemie v oblasti stopové analýzy je monitorování úrovně analytů v oblasti nízkých koncentrací. Obvykle se požaduje určení odhadu střední hodnoty a odpovídajícího intervalu spolehlivosti. S ohledem na omezenou přesnost a citlivost měřicích přístrojů dochází často k případům, kdy jsou některé měřené hodnoty pod limitou detekce a je tedy k dispozici pouze kvalitativní informace o koncentraci analytu. Tato data mají běžně některé další specifické zvláštnosti:

- I.** Obsahují často extrémně velké hodnoty, které však nemusí být důsledkem chyb měření
- II.** Mohou být cenzurována také zhora s ohledem na omezený rozsah přístrojů
- III.** Jsou vždy kladná a výrazně zešikmená k vyšším hodnotám
- IV.** Jejich počet je často omezen díky malému počtu vzorků a složitému způsobu jejich získávání.
- V.** Jsou často prostorově nebo časově závislá.
- VI.** Je obtížné realizovat opakování stanovení (tj. vzorkování a měření) za stejných podmínek, protože se koncentrace stopových látek mění jak v čase, tak i v prostoru.

Uvedené zvláštnosti pak omezují použití běžných technik určování parametrů polohy resp. rozptýlení a identifikace vybočujících měření. Také robustní techniky obvykle selhávají, protože eliminují extrémní hodnoty, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

V tomto příspěvku jsou uvedeny problémy spojené s velikostí výběru a zešikmením rozdělení dat.

Jsou navrženy postupy pro případy symetrických i asymetrických rozdělení, kdy jsou některá data pod limitou detekce (cenzurované výběry). Jsou ukázány možnosti použití metody maximální věrohodnosti a metod založených na pořádkových statistikách. Jsou uvedeny i grafické postupy založené na principech konstrukce Q-Q grafů [1].

2. Základní pojmy

Vyjděme z předpokladu, že máme k dispozici náhodný výběr reprezentovaný N -tící dat (x_1, x_2, \dots, x_N) . Účelem je odhadnout střední hodnotu a rozptyl resp. konstruovat interval spolehlivosti pro střední hodnotu. Tyto úlohy je možné efektivně řešit jen na základě vhodných předpokladů o pravděpodobnostním modelu měření a specifikaci typu rozdělení. Existují dvě mezní situace:

1. Rozmezí analyzovaných dat se pohybuje v rámci jednoho řádu. To umožňuje použití standardních statistických metod založených na předpokladu konstantního rozptylu resp. **aditivního modelu měření**.
2. Rozmezí analyzovaných dat se pohybuje v rozmezí několika řádů. Pak se může použít transformace stabilizující rozptyl nebo **multiplikativní model měření**. To vede k logaritmické transformaci dat

Nevýhodou logaritmické transformace je fakt, že při nízkých koncentracích je absolutní chyba měření velmi malá (blízká 0) aby byla relativní chyba konstantní. To odporuje realitě.

V některých situacích nelze získat všechny výsledky měření, protože se některé vyskytují pod mezí detekce $DL = x_L$, kde $DL = x_L$ odpovídá koncentraci na mezí detekce. Z celkového počtu měření N nechť $N - n_1$ je číselných (leží nad limitou detekce) o zbytku n_1 měření je známo pouze to, že leží někde pod limitou detekce. Tento problém se nazývá tzv. prosté cenzorování zdola. Pokud dochází k omezení shora vlivem omezení přístroje (maximálně měřitelná hodnota je $UL = x_U$), takže je celkem n_2 měření nad limitou pracovního intervalu stanovení resultuje výběr s prostým cenzorováním shora. Konečně lze očekávat i případ oboustranného cenzorování, kdy jsou známy numerické hodnoty pouze pro $N - n_1 - n_2$ prvků výběru. Omezme se pro zjednodušení na prosté cenzorování zdola. Cohen definuje pro tento případ tři základní úlohy [2]:

1. Data pod limitou detekce nejsou zaznamenána. Není tedy známo n_1 a ani N . Pouze je k dispozici $N - n_1$ je číselných hodnot. Této situaci odpovídá tzv. **jednoduché uřezání**.
2. Je zaznamenám počet hodnot ležících pod mezí detekce n_1 o kterých je známo to, že leží pod zadanou hodnotou x_L . Jsou také k dispozici numerické hodnoty $N - n_1$ prvků výběru celkové velikosti N . To je případ **cenzorování typu I**.
3. Nejmenších n_1 prvků výběru nemá numerickou hodnotu, protože ležely pod mezí detekce, která není přesně známa. Pouze se ví, že x_L je menší než než nejmenší hodnota prvků výběru nad limitou detekce $x_{(n_1+1)}$, kde $x_{(n_1+1)}$ označuje pořádkovou statistiku. To je případ **cenzorování typu II**.

Standardně se řeší pouze cenzorování typu I s předpokladem, že všechna data (i ta nezaznamenaná) pocházejí ze stejného rozdělení pravděpodobnosti.

3. Vliv zešikmení dat na intervaly spolehlivosti

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

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

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

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

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 $100(1 - \alpha) \%$ na interval spolehlivosti CI definován vztahem

$$P(CID \leq \mu \leq CIH) = 1 - \alpha \quad (3)$$

Symbol $P(\cdot)$ označuje pravděpodobnost a α je tzv. hladina významnosti. Obvykle se volí $\alpha = 0.05$ nebo $\alpha = 0.01$ 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

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

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

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

kde $t_{1-\alpha/2}(N-1) = -t_{\alpha/2}(N-1)$ jsou kvantily Studentova rozdělení s $N-1$ stupni volnosti. Pro případ normálního rozdělení mají intervaly (4) resp. (5) přesně $100(1-\alpha) \%$ ní pokrytí střední hodnoty. To znamená, že jen v $100\alpha/2 \%$ případů je střední hodnota menší než CI (nejistota NP zprava) a v $100\alpha/2 \%$ případů je větší než CI (nejistota NL zleva). Pro případ ne-normálního rozdělení platí tyto intervaly pouze asymptoticky tedy pro dostatečně vysoká N . Dostatečná velikost N 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

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

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

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

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 vyšší 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 posunut doprava oproti skutečnému [3].

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

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

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 C_{IH} a nižší dolní mez C_{ID} než odpovídá reálnému rozdělení statistiky t . Interval spolehlivosti definovaný rov. (5) je tedy celý posunut doleva oproti skutečnému [3].

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

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

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

Je patrné, že problémy s výpočtem intervalů spolehlivosti střední hodnoty nastávají pokud je rozdělení dat ne-normální (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 dvě základní techniky omezení vlivu zešikmení dat pro:

A. Snížení asymetrie rozdělení náhodné veličiny t

B. Výpočet korigovaného průměru

Často používaná symetrizační transformace dat je podrobně popsána v knize [1].

V případě (A) jde o použití vhodné transformace vedoucí ke zlepšení statistických vlastností testovacích statistik (A). Tento postup 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).

4. Omezení asymetrie rozdělení Studentovy statistiky

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

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

Pro tuto transformaci již přibližně platí, že

$$P(t_j \leq x) \approx F_n(x) \quad (11)$$

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

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

resp.

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

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

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

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

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

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 požit zjednodušenou versi

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

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

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

a

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

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 α^* pro výpočet kvantilu Studentova rozdělení. Takto opravené kvantily se pak dosadí do rov (5). Klasický pravostranný interval spolehlivosti má tvar

$$\mu \leq x_A + t_{1-\alpha}(N-1) * \frac{S}{\sqrt{N}} \quad (19)$$

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

$$0 = \alpha^* + [0.19 + 0.026 * \ln(\alpha^*)] g_1(x) / \sqrt{N} - 0.05 = f(\alpha^*) \quad (19a)$$

Kořenem funkce $f(\alpha^*)$ je pak α^* , pro které se spočítá opravený kvantil Studentova rozdělení, tj. hodnota $t_{1-\alpha^*}(N-1)$, která se dosadí do rov (18). Pro hledání kořene lze s výhodou použít derivační metodu sečen protože je první derivace $f'(\alpha^*)$ rovna

$$f'(\alpha^*) = 1 + \frac{0.026 * g_1(x)}{\alpha^* \sqrt{N}}$$

5. Výpočet korigovaného průměru

Jednoduchá možnost jak počítat korigovaný průměr pro stanovení intervalu spolehlivosti u asymetrických rozdělení je založena na Johnsonově transformaci. Opravený průměr x_0 má tvar

$$x_0 = (x_A + \frac{s^* g_1}{6N}) \quad (20)$$

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. nedocenení odhadu střední hodnoty. Chenová [5,6] zavedla tzv. MCE odhad x_{MCE} ve tvaru

$$x_{MCE} = x_A + d^* s \quad (21)$$

kde d se počítá podle vztahu

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

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

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

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

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

Zde $A(x_A)$ se bere jako k -tá nejmenší hodnota rozdílů $w_i = \text{abs}(x_i - x_A)$, kde $k = \text{int}(N^{0.5})$. 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 zešíkmená rozdělení, kde se vyskytují hodnoty, sice extrémně vysoké, ale patřící do výběru.

6. Modely měření

Jak bylo demonstrováno výše je zdánlivě jednoduchý úkol stanovení intervalu spolehlivosti střední koncentrace analytu silně komplikován pro případy, kdy rozdělení dat není normální. V této kapitole jsou ukázány základní modely měření, ze kterých rozdělení dat vlastně vychází.

Aditivní model měření

Pro tento nejčastěji používaný model je výsledek měření x roven

$$x = \mu + \varepsilon \quad (25)$$

kde μ je skutečná hodnota měřené veličiny (koncentrace analytu) a ε je náhodná chyba s rozdělením charakterizovaným hustotou pravděpodobnosti $f(\varepsilon)$. Běžně se dále předpokládá, že

- 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. $\varepsilon \approx N(0, \sigma^2)$

Pokud lze všechny tyto předpoklady přijmout má výsledek měření x normální rozdělení $x \approx N(\mu, \sigma^2)$.

V případě, kdy $E(\varepsilon) = k$, se tato konstanta vlastně přičte k μ a resultují systematicky vychýlené odhady střední hodnoty, protože vyjde, že $E(x) = \mu + k$.

Předpoklad o konstantnosti rozptylu, je ekvivalentní požadavku konstantní aditivní chyby měřicího přístroje $\Delta \approx \sigma$. Pro aditivní model je tedy relativní chyba

$$\delta = \Delta / x \quad (26)$$

Předpoklad o nezávislosti měření je v řadě případů splněn. Potíže mohou nastat např. u dynamických experimentů, kde může dojít vlivem vzorkování na jedné soustavě ke vzniku autokorelace I.řádu

$$\varepsilon_i = \rho * \varepsilon_{i-1} + u_i \quad (27)$$

Zde u_i je náhodná veličina s konstantním rozptylem Platí, že $u_0 = 0$. Autokorelační koeficient ρ je korelační koeficient mezi dvojicemi x_j a x_{j+1} , $i = 1, \dots, M - 1$. Lze určit, že

$$D(u) = D(x) = \frac{\sigma^2}{1 - \rho^2} \quad (28)$$

Je patrné, že v případě výrazné autokorelace dojde ke zvýšení rozptylu. Autokorelační koeficient ρ se obvykle odhaduje pomocí vztahu

$$\hat{\rho} = \frac{\sum_{j=1}^{N-1} (x_j - x_A) * (x_{j+1} - x_A)}{[s^2 (N - 1)]} \quad (29)$$

kde s^2 je výběrový rozptyl a x_A je aritmetický průměr. Orientačně platí, že pokud leží $\hat{\rho}$ v intervalu

$$- 2 / \sqrt{M} \leq \hat{\rho} \leq 2 / \sqrt{M}$$

Lze považovat ρ za nevýznamný. Předpoklad, že chyby mají normální rozdělení 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 x měření, protože pro model (25) je tvar hustoty pravděpodobnosti totožný.

V chemické analýze je častějším jevem asymetrické rozdělení dat zešikmené k vyšším hodnotám. Pro odstranění této asymetrie se často používá vhodná transformace $h(x)$. Ta však v případě platnosti modelu (1) vede ke vzniku nekonstantního rozptylu

$$D(h(x)) = \left[\frac{dh(x)}{dx} \right]^2 * \sigma^2 \quad (30)$$

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

$$D(h(x)) = \left(\frac{\sigma}{x} \right)^2 = \delta^2 \quad (31)$$

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).

$$x = \mu * \exp(\varepsilon) \quad (32)$$

Zde ε má stejné vlastnosti jako u modelu aditivního (rov.(25)).

Po korektní logaritmické transformaci přechází tento model na aditivní model v logaritmech

$$\ln(x) = \ln(\mu) + \varepsilon \quad (33)$$

Nevýhodou multiplikativního modelu je především to, že pro velmi nízké koncentrace resp.malé μ vychází absolutní chyba měření příliš nízká [1]. Pokud platí, že $\ln(x) \approx N(\nu, \tau^2)$ má výsledek měření x *lognormální rozdělení* s parametry

$$\mu = \exp(\nu + \tau^2 / 2) \quad \sigma = \mu^2 (\exp(\tau^2) - 1)$$

Kromě střední hodnoty se používá pro charakterizaci polohy také geometrický průměr

$$\mu_G = \exp(E(\ln(x))) = \exp(\nu)$$

Lze snadno dokázat, že platí

$$P(x \leq \mu_G) = P(x \leq \exp(\nu)) = P(\ln(x) \leq \nu) = 0.5$$

To znamená, že geometrický průměr je roven mediánu

$$med(x) = \exp(\nu) = \mu_G$$

Pro odhady parametrů se používá dat v logaritmické transformaci

$$\hat{\nu} = \frac{1}{N} \sum_{i=1}^N \ln x_i \quad \hat{\tau} = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\ln x_i - \hat{\nu})^2} \quad \hat{\mu} = \exp(\hat{\nu} + \hat{\tau}^2 / 2) \quad (34)$$

Pro rozptyly odhadů platí

$$D(\hat{\mu}) = \frac{\mu^2}{N[\exp(\tau^2) - 1]} \quad D(\hat{\mu}_G) = E(\hat{\mu}_G)^2 [\exp(\tau^2 / N) - 1] \quad (35)$$

Střední hodnota geometrického průměru je $E(\mu_G) = \exp(\nu + \tau^2 / 2N)$. Oba průměry jsou vychýlené odhady. Geometrický průměr je pro tento model měření výhodnější než aritmetický průměr [7].

V případě, že se použije „nevhodná“ logaritmická transformace na model (25) vyjde

$$\ln(x) = \ln(\mu + \varepsilon) = \ln \mu + \ln(1 + \varepsilon / \mu) \quad (36)$$

S využitím Taylorova rozvoje lze psát $\ln(1 + x) = x - x^2 / 2 + x^3 / 3 - x^4 / 4$ a pak

$$\ln(x) \approx \ln(\mu + \varepsilon) = \ln \mu + \varepsilon / \mu - 0.5 * (\varepsilon / \mu)^2 + (\varepsilon / \mu)^3 / 3 - (\varepsilon / \mu)^4 / 4 \dots$$

Pokud má náhodná veličina S rozdělení $N(\mu, \sigma^2)$ platí pro její centrální momenty

$$\mu_r = E[(S - \mu)^r]$$

tyto vztahy

$$\mu_r = 0 \quad \text{pro } r \text{ liché}$$

$$\mu_r = \frac{r!}{(r/2)!} \frac{\sigma^r}{2^{r/2}}$$

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

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

a

$$D(\ln x) = \delta^2 + 2.5 * \delta^4 + 4.66 * \delta^6 + 6 * \delta^8$$

Pokud vychází šikmost větší než 0.35 doporučuje se volit tři parametrické lognormální rozdělení s prahovou hodnotou A [8]. Při znalosti A lze místo x použít x-A ve všech výše uvedených vztazích. Pro rychlý odhad A lze použít pořádkových statistik $x_{(i)}$ tj. vzestupně seřazených hodnot výběru.

$$A = \frac{x_{(1)} * x_{(n)} - \tilde{x}_{0.5}^2}{x_{(1)} + x_{(n)} - 2 * \tilde{x}_{0.5}}$$

Přesnější odhad A je uveden v knize [1].

7. Odhady parametrů

Z výše uvedeného je patrné, že rozdělení výsledků měření pro oba základní modely obsahuje střední hodnotu a rozptyl, res. Ještě další parametry. Lze tedy předpokládat, že i- té měření pochází z rozdělení $f(x_i, \mu, \sigma^2, a)$ s neznámými parametry μ, σ^2, a . Pro získání odhadů parametrů rozdělení výsledků měření na základě náhodného výběru velikosti N lze použít celé řady metod. Tradiční **momentová metoda** je založená na porovnání výběrových

momentů a teoretických momentů rozdělení $f(x_i, \mu, \sigma^2, a)$. Výsledkem je soustava m rovnic o m neznámých μ, σ^2, a . Pro případ normálního rozdělení a aditivního modelu měření vyjdou jako odhad střední hodnoty aritmetický průměr x_A a jako odhad rozptylu výběrový rozptyl s^2 . V některých případech lze místo momentů použít jiných charakteristik jako jsou např. vybrané kvantily. Nevýhodou těchto modelů je složitější určování jejich rozdělení a přesnosti.

Tyto problémy odstraňuje **metoda maximální věrohodnosti**, kdy se získávají odhady maximalizující věrohodnostní funkci L resp. její logaritmus. Věrohodnostní funkce je obecně sdružená hustota pravděpodobností všech měření ve výběru. Za předpokladu nezávislých chyb měření ε_i je věrohodnostní funkce resp. její logaritmus ve tvaru

$$\ln L(\mu, \sigma^2, a) = \sum \ln f(x_i, \mu, \sigma^2, a) \quad (37)$$

Maximálně věrohodné odhady parametrů jsou pak určeny z podmínek maxima věrohodnostní funkce, tedy soustavy rovnic

$$\frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \mu} = 0 \quad \frac{\partial \ln L(\mu, \sigma^2, a)}{\partial \sigma^2} = 0 \quad \frac{\partial \ln L(\mu, \sigma^2, a)}{\partial a} = 0$$

Jde opět o obecně o soustavu m nelineárních rovnic pro m proměnných.

Pro případ normálního rozdělení a nezávislých měření má logaritmus věrohodnostní funkce tvar

$$\ln(L) = (-N/2) * [\ln(2\pi) + \ln(\sigma^2)] - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$$

Pro první derivace logaritmu věrohodnostní funkce pak platí, že

$$\frac{\partial \ln(L)}{\partial \mu} = 2 \sum_i x_i - 2 * N * \mu \quad \frac{\partial \ln(L)}{\partial \sigma^2} = \frac{\sum_i (x_i - \mu)^2}{[\sigma^2 \sqrt{2}]^2} - \frac{N}{2\sigma^2}$$

Maximálně věrohodné odhady střední hodnoty a rozptylu pro normální rozdělení jsou tedy opět (pro větší rozsahy měření N) totožné s výběrovým průměrem a výběrovým rozptylem. Výhodou použití metody maximální věrohodnosti je možnost stanovení rozptylů odhadů z matice druhých derivací věrohodnostní funkce, asymptotická nevychýlenost a asymptotická normalita pro libovolné rozdělení dat [1].

Pro některá rozdělení související s normálním lze získat odhady parametrů polohy a rozptýlení s využitím **pořádkových statistik**.

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

kteří jsou vlastně vzestupně seřazenými hodnotami prvků výběru. Je zřejmé, že $x_{(1)}$ je nejmenší prvek výběru a $x_{(N)}$ je největší prvek výběru. Dá se ukázat, že pokud je $F_e(x)$ neznámá výběrová distribuční funkce má náhodná veličina $z_{(i)} = F_e(x_{(i)})$ nezávisle na této distribuční funkci rozdělení Beta $Be [i, N-i+1]$. Platí tedy, že střední hodnota je

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

kde $E(\cdot)$ je operátor střední hodnoty. Pořádkové statistiky $z_{(i)}, z_{(j)}$, $i, j=1, \dots, N$ jsou korelované s kovariancí závislou na i, j a N . Pro původní pořádkové statistiky pak platí, že

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

$Q_e(P_i)$ je kvantilová funkce a $P_i = \frac{i}{N+1}$. Pořádková statistika $x_{(i)}$ je tedy hrubým odhadem kvantilu $Q_e(P_i)$ pro pravděpodobnost P_i . $x_p = Q_e(P)$. Pokud je třeba získat odhad kvantilové funkce mimo polohy pořádkových statistik, tedy pro zadanou pravděpodobnost p , kde $i/(n+1) < P < (i+1)/(n+1)$ volí se běžně lineární interpolace..

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

Rozptyl $D(x_p)$ lze vyčíslit ze vztahu

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

Pořádkové statistiky se dají použít pro současné řešení dvou úloh, tj. stanovení typu rozdělení výběru a odhad jeho vybraných parametrů při využití tzv. **Q-Q grafů**. Základní myšlenka vychází z faktu, že pokud je rozdělení výběru (F_e) totožné s předpokládaným teoretickým rozdělením (F_t) musí platit shoda kvantilů, tedy

$$x_{(i)} = F_t^{-1}(P_i)$$

Tato závislost (Q-Q graf) je tedy přímka. Pro výpočet $F_t^{-1}(P_i)$ je třeba znát obecně všechny parametry teoretického rozdělení. V řadě případů je však možná standardizace

$$s = \frac{x - Q}{R}$$

kde R je parametr rozptýlení resp. měřítka (pro normální rozdělení je to σ^2) a Q je parametr polohy resp. prahová hodnota (pro normální rozdělení je to μ). Standardizované kvantilové funkce $Q_s(P_i) = F_{s_t}^{-1}(P_i)$ pak již obsahují jen tvarové faktory. V případě shody obou rozdělení pak resultuje přímková závislost

$$x_{(i)} = Q + R * Q_s(P_i) = a + b * Q_s(P_i)$$

Pro normální rozdělení jsou $Q_s(P_i) = \Phi^{-1}(P_i)$ rovny kvantilům normovaného normálního rozdělení. Odhad střední hodnoty pak odpovídá absolutnímu členu a a odhad směrodatné odchylky směrnici b regresní přímky. Pro odhad parametrů z Q-Q grafů je možno použít buď neváženou metodu nejmenších čtverců nebo váženou metodu kdy váhy odpovídají korelaci mezi pořádkovými statistikami a jejich rozptylům. Pro dvou-parametrové lognormální rozdělení se využívá logaritmu pořádkových statistik $\ln(x_{(i)})$. Pro případ tří-parametrového normálního rozdělení je výhodné použít zápisu hustoty pravděpodobnosti ve tvaru [9]

$$f(x) = \frac{1}{\sqrt{2} * \pi * (1 + \sigma * Y) * \tau} \exp\left\{-\frac{1}{2\sigma^2} [\ln(1 + \sigma * Y)]^2\right\}$$

kde $Y = x - \mu / \tau$ a platí, že $x \geq (\mu - \tau) / \sigma$. Při použití transformace

$$z = \ln[(1 + \sigma * Y)^{1/\sigma}]$$

má náhodná veličina z normované normální rozdělení $N(0,1)$. Pořádková statistika $x_{(i)}$ pak souvisí s pořádkovou statistikou normovaného normálního rozdělení $z_{(i)}$ podle vztahu

$$x_{(i)} = \mu + \tau * \left[\frac{\exp(\sigma * z_{(i)}) - 1}{\sigma} \right] \approx \mu + \tau * g_i(\sigma)$$

Blom doporučuje určení odhadu $g(\sigma)$ s využitím pořadové pravděpodobnosti P_i , kdy

$$g_i(\sigma) = \left[\frac{\exp(\sigma * \Phi^{-1}(P_i)) - 1}{\sigma} \right]$$

kde je vhodné volit pořadovou pravděpodobnost s korekcí [9]

$$P_i = \frac{i + 0.3133\sigma - 0.3927}{N + 0.2146} \quad (38)$$

V případě známého σ je tedy možné odhadnout parametry μ, τ jako úsek a směrnici regresní přímky $x_{(i)} \approx \mu + \tau * g_i(\sigma)$. Pro korekci (38) lze určit prvky váhové matice V pro metodu vážených nejmenších čtverců ve tvaru

$$V_{ii} = 2(N+1)(N+2) * h_i^2 \quad V_{i+1,i} = V_{i,i+1} = -(N+1)(N+2) * h_i * h_{i+1} \quad V_{i,j} = 0 \text{ jinde}$$

$$\text{kde } h_i = \frac{\Phi[\Phi^{-1}(i/(N+1))]}{\exp[\sigma * \Phi^{-1}(i/(N+1))]}$$

Pokud není σ známo, lze použít iterativní postup jeho postupného zpřesňování z prvních odhadů [9]. Váhová matice V ukazuje na vychýlení vznikající použitím nevážených nejmenších čtverců.

8. Cenzorované výběry

Pro odhady parametrů v cenzorovaných výběrech lze použít jak metodu maximální věrohodnosti, tak i metody založené a pořádkových statistikách. Omezme se na případ cenzorování typu I, kdy známe limitu detekce x_L (mez pod kterou se již zaznamenává pouze „přítomnost“ měření) a předpokládejme, známe rozdělení dat charakterizované hustotou pravděpodobnosti $f(x)$ resp. distribuční funkcí $F(x)$. Pro cenzorovaná měření lze při znalosti distribuční funkce měření určit pouze pravděpodobnost s jakou leží pod mezí detekce, která je rovna $F(x_L)$. Veličina n_l se uvažuje jako náhodná, protože nelze a priori předpovědět, v kterém vzorku dojde k překročení meze detekce a v kterém ne. Všechny možné kombinace

n_1 prvků, které ve výběru velikosti N leží pod limitou detekce jsou dány binomickým koeficientem $N!/(n_1!(N-n_1)!)$. Věrohodností funkce má pro tento případ tvar

$$\ln(L) = \frac{N!}{n_1!(N-n_1)!} F(X_L)^{n_1} * \prod_{i=n_1+1}^N f(x_{(i)}) \quad (39)$$

Pro případ, kdy se uvažuje omezení zprava, tj, n_2 prvků leží nad pracovním rozmezím přístroje x_U má věrohodnostní funkce tvar

$$\ln(L) = \frac{N!}{n_2!(N-n_2)!} [1-F(x_U)]^{n_2} * \prod_{i=1}^{N-n_2} f(x_{(i)})$$

Konečně pro případ omezení z obou stran platí, že

$$\ln(L) = \frac{N!}{n_2!n_1!} F(x_L)^{n_1} * [1-F(x_U)]^{n_2} * \prod_{i=n_1+1}^{N-n_2} f(x_{(i)})$$

Pro známé rozdělení dat lze tedy dosadit do těchto vztahů a po logaritmování resp. derivacích nalézt maximálně věrohodné odhady parametrů.

Pro případ normálního rozdělení dat nalezl Cohen [2] vztahy pro odhad střední hodnoty x_{AC} a rozptylu s_C^2 odpovídající maximalizaci rov (39) s využitím odhadů střední hodnoty a rozptylu z necenzorované části dat

$$x_{AN} = \frac{1}{N-n_1} \sum_{i=n_1+1}^N x_{(i)} \quad s_N^2 = \frac{1}{N-n_1-1} \sum_{i=n_1+1}^N (x_{(i)} - x_{AN})^2$$

Platí, že

$$x_{AC} = x_{AN} - \lambda * (x_{AN} - x_L) \quad s_C^2 = s_N^2 + \lambda * (x_{AN} - x_L)^2 \quad (40)$$

Parametr λ závisí na odhadnutém podílu cenzorovaných dat $h = n_1 / N$ a na parametru $g = s_N^2 / (x_{AN} - x_L)^2$. Jeho hodnoty jsou tabelovány a existují také empirické vztahy [11]. Algoritmus numerického určení parametru λ je popsán v práci [12].

Je tedy patrné, že zanedbání cenzorované části dat vede k nadhodnocení aritmetického průměru a podhodnocení rozptylu v závislosti na velikosti parametru λ .

Místo optimálního řešení často postačuje jedнокrokový odhad založený na předpokladu, že počet hodnot pod limitou detekce má binomické rozdělení [10,11]. Pro odhady střední hodnoty x_{ACJ} a rozptylu s_{CJ}^2 pak platí, že

$$x_{ACJ} = x_{AN} - q * s_N \quad s_{CJ}^2 = \frac{\sum_{i=n_1+1}^N x_{(i)}^2}{N-n_1} - (x_{AN})^2 - s_N^2 * (q * \Phi^{-1}(h) - q^2) \quad (41)$$

Korekční faktor q má tvar

$$q = \frac{N}{(N-n_1)\sqrt{2\pi}} \exp(-0.5 * [\Phi^{-1}(h)]^2)$$

Odhady x_{ACJ} a s_{CJ}^2 lze tedy určit relativně snadno bez nutnosti použití speciálních tabulek. Pokud bude předpokládané rozdělení dat dvou-parametrové logaritmicke normální stačí místo hodnot $x_{(i)}$ použít jejich logaritmy $\ln(x_{(i)})$ a logaritmovat i limitu detekce. Pro tří-parametrové

logaritmicko normální rozdělení a cenzorování z obou stran našel vhodné iterativní řešení Tiku [13]. Existuje pochopitelně také možnost hledat přímo maximum věrohodnostní funkce (39) s využitím nelineárních optimalizačních metod.

Velmi jednoduše je možné použít pro odhady parametrů cenzorovaných rozdělení Q-Q grafů, protože pořádkové statistiky jsou založeny na pořadí a pro konstrukci Q-Q grafu stačí jen vybrané pořádkové statistiky (ty, které jsou nad limitou detekce). Pro případ normálního rozdělení a prostého cenzorování zleva se tedy vynášejí hodnoty $x_{(i)}$ pro $i = n_1 + 1$ proti kvantilům normovaného normálního rozdělení $\Phi^{-1}(P_i)$ pro $i = n_1 + 1$. Pořádkové statistiky je však třeba upravit s ohledem na cenzorování, tedy

$$P_i = \frac{n_1}{N} + \frac{N - n_1}{N} \left(\frac{i - 0.375 - n_1}{N + 0.25 - n_1} \right)$$

Opět lze použít jak vážené tak nevážené metody nejmenších čtverců. Po logaritmické transformaci je možno odhadovat parametry dvou parametřového logaritmicko normálního rozdělení resp. použít různých speciálních postupů pro tři parametřové logaritmicko normální rozdělení. Nevýhodou tohoto postupu je vychýlenost odhadů a malá přesnost pro malé výběry. Pokud je podíl cenzorovaných měření h menší jak 0,25 je možno použít pro statistickou analýzu použít medián a interkvartilové rozpětí protože pro jejich určení postačuje znalost počtu hodnot pod limitou detekce. Navíc dojde ke zrobustnění odhadů. Pro porovnání odhadů získaných z Q-Q grafu metodou robustní regrese sa odhady x_{ACJ} a s_{CJ}^2 byl sestaven program LCENZ v jazyce MATLAB. Tento program by použit při řešení příkladu 1.

9. Praktické doplňky

Při zpracování experimentálních i neexperimentálních dat záleží na množství informací, které jsou před vlastní analýzou k dispozici. Existují tři základní skupiny s ohledem na úroveň informací:

A. Víme vše – tj. známe pravděpodobnostní model – pak stačí jen ověření předpokladů jeho platnosti před vlastní konfirmativní statistickou analýzou

B. Nevíme nic – buduje se datově závislý pravděpodobnostní model – pak se provádí komplexní analýza dat (průzkumová, transformace, porovnání výběrového rozdělení s teoretickými atd.)

C. Něco tušíme – konstruuje se empirický model zahrnující jak známé tak i datově závislé informace – pak se realizuje jak analýza dat tak ověřování předpokladů

Zdánlivě nejjednodušší úlohou je odhad intervalu spolehlivosti střední hodnoty na základě výběru (x_1, x_2, \dots, x_n) z (ne)známého rozdělení $f(x)$. Základní problém je nenulová šikmost ($g_1 \neq 0$) a špičatost odpovídající nenormálnímu rozdělení. ($g_2 \neq 3$). Vybrané techniky byly diskutovány s ohledem na data z oblasti životního prostředí v předchozím textu. V obecném případě lze použít také další postupy:

- robustní metody
- použití zešikmených rozdělení
- počítačově intenzivní metody
- generalizovaná lineární regrese

Jak tyto tak i další postupy konstrukce intervalu spolehlivosti střední hodnoty jsou založeny na nějakých předpokladech a nejsou univerzální pro všechny situace. Většina postupů je uvedena v knize [1]. Pro případ cenzorovaných měření je třeba volit speciálnější postupy, které zabrání podcenění rozptylu a přecenění odhadu střední hodnoty.

10. Příklad *Určení koncentrace nečistot v surovině*

V rámci monitorování kvality suroviny byla sledována koncentrace nečistot v $\mu\text{g/g}$ (data byla publikována v [2]). Získané koncentrace v $\mu\text{g/g}$ jsou

DL DL 1.24 1.49 1.50 1.56 1.61 1.78, kde DL označuje hodnoty pod limitou detekce.

Limita detekce přístroje je $\text{limd} = 1 \mu\text{g/g}$. Účelem je odhad střední hodnoty rozptylu a intervalu spolehlivosti za předpokladu normálního rozdělení.

Nevhodný postup s vynecháním hodnot pod mezí detekce.

Průměr = 1.53 a výběrová směrodatná odchylka = 0.177. Kvantil t rozdělení $t_{0.975}(5) = 2.571$

95 % ní interval spolehlivosti $UC = 1.72$ $LC = 1.34$

Maximalizace věrohodnostní funkce

Parametr $h = 0.25$, parametr $g = 0.11$ a tabelovaná hodnota $\lambda = 0.3387$

Průměr = 1.35 a výběrová směrodatná odchylka = 0.355

95 % ní interval spolehlivosti $UC = 1.72$ $LC = 0.98$

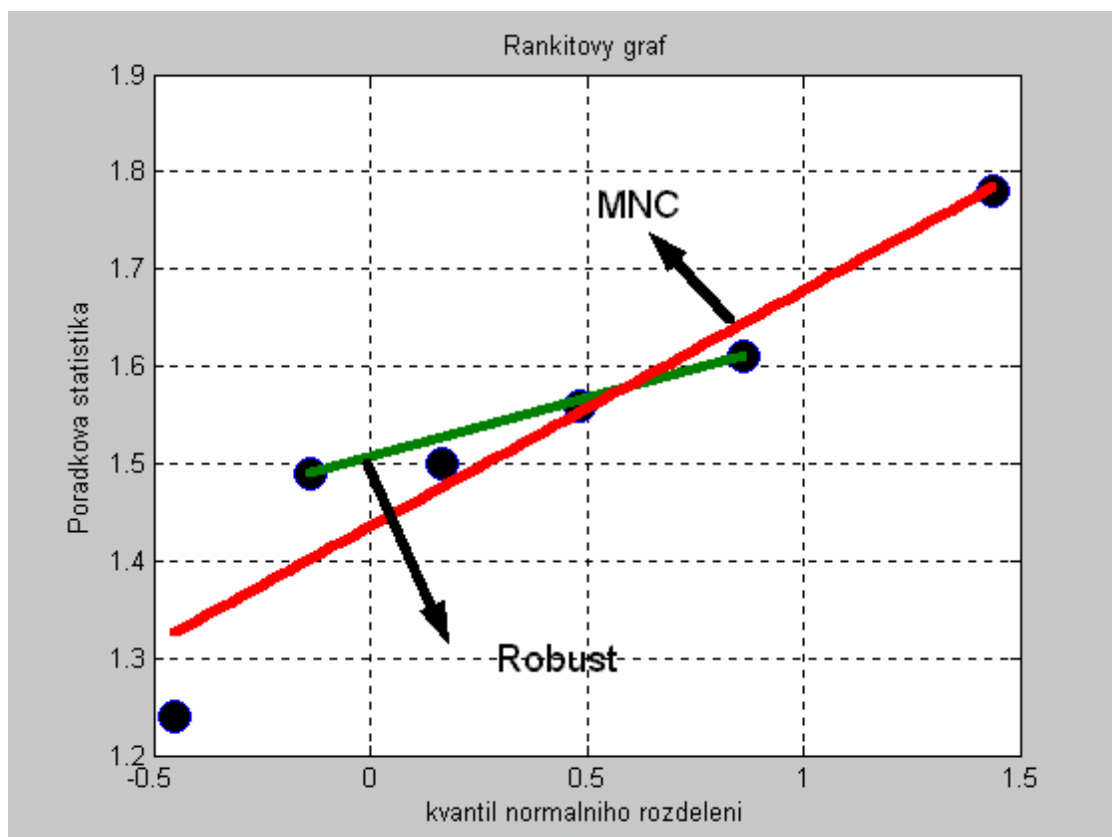
Jednokroková aproximace maximalizace věrohodnostní funkce

Průměr = 1.46 a výběrová směrodatná odchylka = 0.2

95 % ní interval spolehlivosti $UC = 1.67$ $LC = 1.25$

Pořádkové statistiky

Na obr 1. je rankitový graf spolu s regresními přímkami pro klasickou a robustní MNC.



Obr 1. Rankitový graf

Ze směrnice a úseku určených klasickou MNČ vyšlo
Průměr = 1.43 a výběrová směrodatná odchylka = 0.24 .
95 % ní interval spolehlivosti $UC = 1.68$ $LC = 1.18$

Je patrné, že postupy beroucí v úvahu limitu detekce vedou k výrazně nižší dolní mezi intervalu spolehlivosti.

8. Závěr

Je patrné, že statistické zpracování dat v oblasti stopové analýzy má celou řadu specifických zvláštností. V řadě případů je třeba budovat i pro zdánlivě jednoduché situace poměrně komplikované modely. Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy bez hlubšího rozboru zde může vést ke katastrofickým závěrům.

Poděkování:

Tato práce vznikla s podporou výzkumného centra Textil LN00B090

9. Literatura

- [1] Meloun M., Militký J.: *Zpracování experimentálních dat*, East Publishing Praha 1998
- [2] Cohen, A.C.: *Technometrics* **3**, 535 (1961)
- [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. : *Environmetrics* **6**, 181 (1995)
- [6] Chen L.: *J. Appl. Statist.* **25**, 739 (1998)
- [7] Smothers D.D. a kol.: *Veterinary Parasitology* **81**,211 (1999)
- [8] May J. a kol. : *Decision Sciences* **31**, 129 (2000)
- [9] Munro A. H., Wixley R. A.J. : *J. Amer. Statist. Assoc.* **65**, 212 (1970)
- [10] Huybrechts T., a kol. : *J. of Chromatography* **A975**, 123 (2002)
- [11] Haas C.N., Sheff P. A.: *Environ. Sci. Technol.* **24**, 912 (1990)
- [12] Kuttatharmmakul S. a kol. : *Trends in analytical chemistry* **19**, 215 (2000)
- [13] Tiku M.L.: *J. Amer. Statist. Assoc.* **63**, 134 (1968)
- [14] Shumway R. H. a kol.: *Technometrics*, **31**, 347-356 (1989)