[3] James W., Stein C.: *Estimation with Quadratic Loss*, Proceed. 4th Berkeley Sym on Math. Statist., p. 361, 1961.
[4] Guanadeskian R., Kettenring J. R.: *Biometrics* **28**, 80 (1972).
[5] Campbell N. A.: *Appl. Statist.*, 29, 231 (1980).
[6] Hu J., Skrabal P., Zollinger H.: *Dyes and Pigments*, **8**, 189 (1987).
[7] Chambers J. M., Cleveland W. S., Kleiner B., Tukey P. A.: *Graphical Methods Data Analysis*. Duxburg Press, Belmont, California 1983.
[8] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap.
[9] Jolliffe I. T.: *Principal Component Analysis*. Springer Verlag, New York 1986.
[10] Barnett V., (Edit.): *Interpreting Multivariate Data*. Wiley, Chichester 1981, kap. 1
[11] Everitt B. S.: *Graphical Techniques for Multivariate Data*. London 1978.
[12] Andrews D. F.: *Biometrics*, **28**, 125 (1972).
[13] Kulkarni S. R., Paranjape S. R.: *Commun. Statist.*, **13**, 2511 (1984).
[14] Guanadeskian R.: *Methods for Statistical Data Analysis of Multivariate Observatio* Wiley, New York 1977.
[15] Kleiner B., Hartigan J. A., *J. Amer. Statist. Assoc.*, **76**, 260 (1981).
[16] Kres H.: *Statistical Tables for Multivariate Analysis*. Springer, New York 1983.
[17] Seber G. A. F.: *Multivariate Observations*. Wiley, New York 1984.
[18] Stryjewska E., Rubel S., Henrion A., Henrion G.: *Z. Anal. Chem.*, **327**, 679 (198
[19] Mudholkar G. S., Trivedi M. S., Lin T. C.: *Technometrics*, **24**, 139 (1982).
[20] Johnson R.A., Wichern D.W.: *Applied Multivariate Statistical Analysis*, Prentice Ha 1982
[21] Ajvjazin S., Bežajeva Z., Staroverov O.: *Metody vícerozměrné analýzy*, SNTL Pra 1981
[22] Meloun M., Militký J. , Forina M.: *Chemometrics for Analytical Chemistry, Volu 1. PC-Aided Statistical Data Analysis*, Ellis Horwood, Chichester 1992.
[23] Brereton R. G. *Multivariate Pattern Recognition in Chemometrics, Illustrated by Ca Studies*, Elsevier 1992,
[24] Krzanowski W. J.: *Principles of Multivariate Analysis, A User's Perspective*, Oxfo Science Publications 1988,
[25] Jeffers J. N. R., *Applied Statistician*, **16**, 225 (1967).
[26] Meloun M. , Militký J., *Statistické zpracování experimentálních dat*, Plus Praha 199
[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 Yo
[30] Meloun M. , Militký J., *Sbírka úloh - Statistické zpracování experimentálních da* Univerzita Pardubice, 1996.

Zpracování dat s ohledem na limitu detekce

Jiří Militký[1], Milan Meloun[2]

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

[2] *Katedra analytické chemie, Universita Pardubice, Katedra analytické chemie, nám. ČS.legií 565, 532 10 Pardubice, e-mail: milan.meloun@upce.cz*

Motto: *I nula nese informaci*

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. Jsou vybrány především metody, které umožňují zpracovat zešikmené 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 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 $(x_1, \ldots, x_N)$ stanoví vhodný odhad střední hodnoty μ a porovná se s povolenou úrovní μ_0. S ohledem na variabilitu měření je vhodné ověřit, zda μ_0 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.

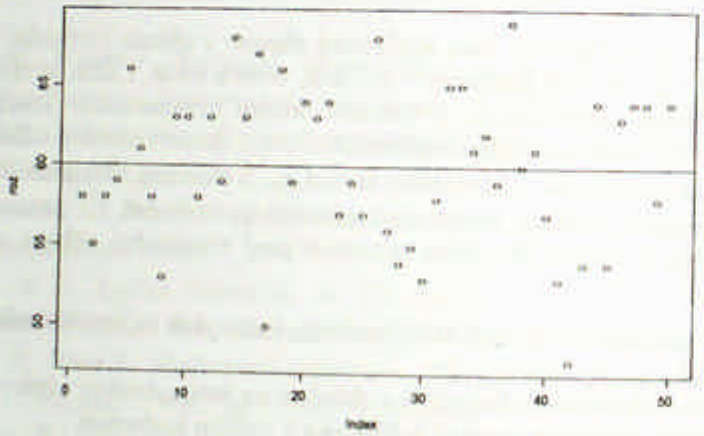
Tyto zvláštnosti pak omezují použití různých technik založených na průzkumové analýze a identifikaci vybočujících měření. Také robustní techniky obyčejně selhávají, protože eliminují extrémy, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

Standardní statistická analýza zde vede k přehnaně optimistickým závěrům. Pro pozitivně zešikmená rozdělení platí, že i když je aritmetický průměr x_A asymptoticky nevychýleným odhadem je s velkou pravděpodobností menší než skutečná hodnota parametru polohy μ. To je dobře patrné z obr 1 v příkladu 1.

Příklad 1

Na obr 1 jsou výsledky jednoduché simulace spočívající v opakované generaci (n = 100) výběrů velikosti 30 z lognormálního rozdělení s parametry $\mu = 0.8$ a $\sigma = 1.4$. Pro každý výběr byl určen aritmetický průměr x_a a vyjádřena odchylka $ro = \left(x_a - \exp\left(\mu + 0.5\sigma^2\right)\right)$ Ze všech 100 výsledků bylo určeno procento případů, že je ro záporné (aritmetický průměr je nižší než střední hodnota), které je označeno jako roz. Tento postup byl opakován 50krát a byl stanoven medián roz (ten je vynesen jako horizontální čára. Je patrné, že mediánová čára je na úrovni 60 % a většina bodů leží nad úrovní 50 %

Tedy aritmetický průměr je systematicky nižší než střední hodnota.



Obr 1. Procento případů, kdy je aritmetický průměr nižší než střední hodnota (ognormální rozdělení)

Analogické výsledky byly získány i pro jiné velikosti výběrů.

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 omezení dílčích problémů spojených s velikostí výběru a zešikmením rozdělení dat. Je navržen postup pro případ, že některá data jsou pod limitou detekce (cenzurované výběry). Závislost v datech 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 $(<\infty)$ 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 $\alpha = 0.05$ 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

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

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

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

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

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

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 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

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

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 nižší dolní mez *CID* než odpovídá reálnému rozdělení statistiky t. Interval spolehlivosti definovaný rov. (5) je tedy celý *posunut doleva* 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.**

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

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

kde $g_2(x)$ je špičatost 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 pro:

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

V případech (A) a (C) jde o použití vhodné transformace vedoucí ke zlepšení statistických vlastností testovacích statistik (A), resp. původních dat (C). 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.

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

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.

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

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

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

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

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

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

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

a

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

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 možnosti 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

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

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

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

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

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

Pro data z příkladu 2. vyšlo opravené alfa = .034 a opravené $t_{(1-alfa/2)}$ = 2.1374

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

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

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

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

kde d se počítá podle vztahu

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

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

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

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

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

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.

5. Výběry obsahující nulové prvky

Model vychází z předpokladu, že se hodnoty pod limitou detekce považují za nulové. Předpokládá se, že studovaný soubor obsahuje nulové hodnoty s pravděpodobností $(1-p)$ a nenulové hodnoty s pravděpodobností p, charakterizované hustotou pravděpodobnosti $f_1(x)$. Hustota pravděpodobnosti náhodné veličiny x je pak [12]

$$\begin{aligned} f(x) &= p * f_1(x, \mu, a_1, \ldots a_m) && \text{pro } x \neq 0 \\ f(x) &= 1 - p && \text{pro } x = 0 \end{aligned} \qquad (25)$$

Zde μ je střední hodnota rozdělení f_1 a a_1, a_2... jsou další parametry rozdělení. Účelem je určit interval spolehlivosti pro celkovou (populační) střední hodnotu $\tau = p * \mu$ tj průměr celé populace. K dispozici je náhodný výběr x_1, x_2 x_n obsahující k nul a n-k ostatních hodnot. Pro odhady parametrů lze použít věrohodnostní funkce a pro stanovení intervalu spolehlivosti populační střední hodnoty τ věrohodnostní poměr.

Věrohodnostní funkce má pro model (25) tvar

$$L = \left(1 - \frac{\tau}{\mu}\right)^k * \left(\frac{\tau}{\mu}\right)^{n-k} * \prod f_1(x_i, \mu, \ldots) \qquad (26)$$

Maximalizací L se obecně určí maximálně věrohodné odhady všech parametrů (s indexem v).

Pro pevné τ_0 je věrohodnostní funkce ve tvaru

$$L_0 = \left(1 - \frac{\tau_0}{\mu}\right)^k * \left(\frac{\tau_0}{\mu}\right)^{n-k} * \prod f_1(x_i, \mu, \ldots) \qquad (27)$$

Maximalizací L_0 se pak určí podmíněné odhady (s indexem 0), které závisí na velikosti τ_0

Věrohodnostní poměr je dán výrazem $-2 * \ln\left(R(\tau_0)\right)$ kde

$$R(\tau_0) = \frac{\max L_0}{\max L} \qquad (28)$$

Funkce $R(\tau)$ se označuje jako věrohodnostní profil. Věrohodnostní poměr má přibližně $\chi^2(1)$ rozdělení. Pro oboustranný 100*(1-α) procentní interval spolehlivosti tedy platí, že jde o množinu všech τ pro, které je splněna nerovnost

$$R(\tau) > \exp\left(-\chi^2_{1-\alpha}(1)/2\right) \qquad (29)$$

Předpokládejme, že f_1 je hustota pravděpodobnosti normálního rozdělení $N(\mu,\sigma^2)$. Po dosazení do rov. (26) a analytické minimalizaci resultují maximálně věrohodné odhady

$$\mu_v = \frac{U}{n-k} \qquad \tau_v = \frac{U}{n} \qquad \sigma_v^2 = \frac{C}{n-k} - \frac{U^2}{(n-k)^2}$$

kde $U = \sum_i x_i \quad C = \sum_i x_i^2$

Opakováním téže procedury pro pevné τ_0 se určí podmíněný odhad rozptylu

$$\sigma_0^2 = \frac{C - 2 * \mu_0 * T + (n-k) * \mu_0^{\ 2}}{n-k} \tag{30}$$

Odhad střední hodnoty je reálný kořen kubické rovnice

$$\mu^3 - A\mu^2 + B\mu - C = 0$$

kde

$$A = \frac{(2n-k)\tau_0 + 3U}{2(n-k)} \qquad B = \frac{C(n-k) + (3n-k)U\tau_0}{2(n-k)^2} \qquad C = \frac{C * n * \tau_0}{2(n-k)^2}$$

Pro věrohodnostní profil pak platí vztah

$$R(\tau) = \frac{k_0 * \exp\left(-\sum_{i-1}^{n-k} (x_i - \mu_0)^2 / \left(2 * \sigma_v^2\right)\right)}{k_v * \exp\left(-\sum_{i-1}^{n-k} (x_i - \mu_v)^2 / \left(2 * \sigma_v^2\right)\right)} \tag{31}$$

kde

$$k_0 = \left(1 - \frac{\tau}{\mu_0}\right)^k * \left(\frac{\tau}{\mu_0}\right)^{n-k} * \left(\frac{1}{2 * \pi * \sigma_0^2}\right)^{\frac{n-k}{2}} \quad \text{a}$$

$$k_v = \left(1 - \frac{\tau}{\mu_v}\right)^k * \left(\frac{\tau}{\mu_v}\right)^{n-k} * \left(\frac{1}{2 * \pi * \sigma_v^2}\right)^{\frac{n-k}{2}}$$

Pro konstrukci intervalu spolehlivosti odhadu populační střední hodnoty t byl na základě těchto vztahů sestaven program CIZERO v jazyce MATLAB, který hledá meze intervalu spolehlivosti pro zadané (1-α). Tento program by použit při řešení příkladu 3.

6. Praktické zpracování dat

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 $(x_1, x_2, \ldots x_n)$ z (ne)známého rozdělení $f(x)$. Základní problém je nenulová šikmost (g_1 # 0) a špičatost odpovídající nenormálnímu rozdělení. (g2 # 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

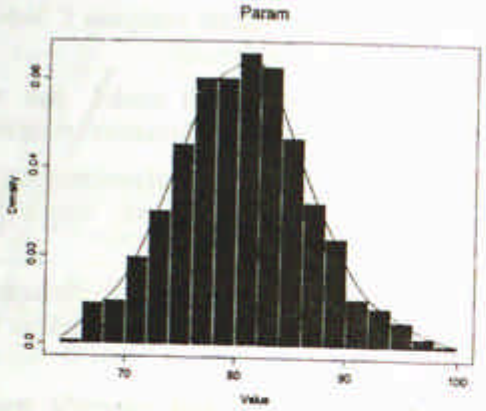
Jak tyto tak i další postupy konstrukce intervalu spolehlivosti střední hodnoty jsou založeny na nějakých předpokladech a nejsou universální pro všechny situace. Většina postupů je uvedena v knize [1].

Příklad 2. Určení koncentrace pregnenolonu v pupeční krvi novorozenců

Byla sledována úroveň steroidů v krvi 100 novorozenců. Účelem je odhadnout střední úroveň a interval spolehlivosti obsahu těchto steroidů Na základě průzkumové analýzy bylo zjištěno, že v logaritmické transformaci není třeba vylučovat odlehlé hodnoty. Rozdělení je systematicky zešikmené k vyšším hodnotám i když šikmost 0,985 není příliš vysoká.. Výsledky metody Bootstrap pro průměr (histogram pro průměry z 1000 simulací) jsou znázorněny na obr. 1. Intervaly spolehlivosti počítané různými metodami jsou uvedeny v tabulce 1.

Je patrné, že vhodnost intervalu spolehlivosti závisí na přijatém předpokladu o datech. Metoda Bootstrap poskytuje interval, který se příliš neliší od klasického intervalu počítaného ze všech bodů. Také speciální postupy pro zešikmená rozdělení zde nepřinášejí výrazné změny . Je to způsobeno poměrně nevýraznou šikmostí. Eliminace podezřelých bodů a logaritmická transformace vedou k výrazné změně mezi

intervalu spolehlivosti. Pro malé odchylky od nulové šikmosti tedy techniky založené na zešikmených rozděleních nemají větší význam.



Obr. 1. Histogram dílčích průměrů (Bootstrap)

Tabulka 1. Souhrn 95% nich intervalů spolehlivosti střední hodnoty

TYP	LC	UC
Log transformace	53.1822	71.8021
Klasická normalita	68.631	92.635
Normalita bez 2 vyb. bodů	65.61	86.209
Hall Edgew. rov(14)	69.69	91.57 (t^{-1}=1.809)
Reduk. T rov (19a)	67.708	93.55 (t=2.1374)
Bootstrap mean	68.298	92.582

Příklad 3. Určení koncentrace ethyl parathionu v ovzduší

V rámci monitorování toxických látek byl monitorován toxický ethyl parathion v ovzduší u Herber Station v Californii (data byla publikována v [11]). Získané koncentrace v $\mu g/m^3$ jsou

0.0090 0.0090 0.0090 0.0090 0.0180 0.0320 0.0120 0.0150 0.0090 0.0780 0.0920 0.0230 0.0180 0.0100

Limita detekce přístroje je limd = 0.01 a hodnoty 0.09 jsou tedy pouze dosazeny Místo nich mohou být nuly, či jiná čísla od 0 do 0.01.Účelem je stanovit 90 procentní interval spolehlivosti střední hodnoty .

Cenzurování na mezi detekce (Postup z kap.5).

Průměr populační = 0.0245, průměr MLE = 0.0469
a výběrový rozptyl MLE = 0.000583

90 % ní interval spolehlivosti UC = 0.0354 LC = 0.0136

Klasický postup s vynecháním hodnot pod mezí detekce.

Průměr = 0.0381 a výběrový rozptyl = 0.000505

90 % ní interval spolehlivosti UC = 0.0504 LC = 0.0258

Je patrné, že postup berouci v úvahu limitu detekce vede k výrazně nižší horní mezi intervalu spolehlivosti a vynechání hodnot pod mezí detekce nevede ke zlepšení.

8. Závěr

Je patrné, že statistické zpracování dat v oblasti životního prostředí 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 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] 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)

[10] Militký J,. Meloun M.: Sborník přednášek z konference „Statistika a řízení jakosti", Liberec, listopad 2000

[11] Shumway R. H. a kol.: Technometrics, **31**, 347-356 (1989)

[12] Kvanli A. H. a kol. : J. of Business anf Econ. Statist. 16, 362-368 (1998)