

Některé robustní postupy určení střední hodnoty a rozptýlení souboru výsledků a jejich použití

Dempír J.¹, Dohnal L.²

¹Pracoviště Hlízov

²Referenční laboratoř pro klinickou biochemii, ÚKBLD VFN a 1. LF UK, Praha

SOUHRN

Autoři popisují většinou neparametrické postupy odhadů střední hodnoty a rozptýlení výsledků v souborech dat. Také jsou uvedeny postupy a použití kvantilů, mediánů, vážení výsledků a jejich transformace. Při volbě vhodného hodnotícího postupu lze obecně doporučit postup jednoduchý, např. použití kvartilů či mediánu. U souborů s významnou šikmostí dat lze použít transformace, i když např. lambda transformace bez použití počítačového programu je pracná a znamená vždy zásah do experimentálních dat.

Klíčová slova: střední hodnota, rozptyl, robustní statistiky.

SUMMARY

Dempír J., Dohnal L.: Application of Some Robust Procedures to Evaluation of Set of Results

In the present article some robust procedures estimating the mean values and their variances of a set of results are dealt with. Non-parametric methods using quartiles, various kinds of medians, weighted results and dominant cluster are mentioned. For data with significant skewness log-transformation or lambda-transformation are recommended, although the latter is time-consuming and both transformations have influence upon the shape of original data. From practical point of view approaches using quartiles and/or medians are to be preferred because of their simplicity.

Key words: mean value, variance, robust statistics.

Úkolem hodnocení souboru výsledků bude v našem případě určení charakteristiky střední hodnoty a charakteristiky rozptýlení výsledků. Způsob určení střední hodnoty i rozptýlení výsledků závisí na typu jejich rozložení (distribuci). Pro normální rozložení platí, že střední hodnotu charakterizuje aritmetický průměr a rozptýlení směrodatná odchylka reprodukovatelnosti (nebo její druhá mocnina, tedy rozptyl reprodukovatelnosti). Pro střední hodnotu lognormálního rozložení použijeme analogicky geometrický průměr. V obecném případě blíže neurčeného rozdělení používáme tzv. robustní (též neparametrické) postupy. Pro střední hodnotu např. medián a pro rozptýlení např. rozpětí nebo maximální diferenci dvou výsledků. V této práci se pokoušíme seznámit čtenáře s některými robustními postupy ke zjištění střední hodnoty a rozptýlení, které nejsou v klinických laboratořích příliš známé, ale přitom jsou poměrně jednoduché a v analytice se již delší dobu s úspěchem používají.

Robustní postupy hodnocení typicky používáme v případech, když:

- hodnocený soubor má odlehlé nebo z odlehlosti podezřelé výsledky, které nemohly být opraveny a není vhodné je vyloučit;
- hodnocený soubor nemá normální rozložení výsledků, např. vykazuje významnou šikmost;
- hodnocený soubor má velké rozptýlení dat.

Pokud bude řeč o významnosti nebo intervalu spolehlivosti, v celém sdělení pracujeme s hladinou významnosti 0,05.

Správnost střední hodnoty \bar{x} se prokazuje nejlépe současnou analýzou vhodného referenčního, respektive certifikovaného referenčního materiálu s udanou

střední hodnotou μ a její rozšířenou nejistotou U . Nevýznamnost rozdílu $(\bar{x} - \mu)$ a tedy nevýznamnost vychýlení (bias) střední hodnoty \bar{x} je prokázána, jestliže je splněna nerovnost [1]

$$-2\sigma < (\bar{x} - \mu) < 2\sigma, \quad (1)$$

$$\sigma = \sqrt{u_{CRM}^2 + s_r^2 + \frac{s_r^2}{n}} \quad (2),$$

kde $u_{CRM} = \frac{U_{CRM}}{2}$, s_r je směrodatná odchylka reprodukovatelnosti a s_r je směrodatná odchylka opakovatelnosti.

V analytické chemii se pro výpočet obecné maximální (přípustné) reprodukovatelnosti pro danou koncentraci často používá Horwitzův vztah [2], který byl autorem definován pro koncentraci vyjádřenou v procentech nebo v mg/kg. Tento vztah je empirický a byl odvozen na základě statistického zpracování velkého množství analytických dat z mnoha různých oborů. Přípustné hodnoty reprodukovatelnosti pro konkrétní případ mohou být často mnohem přísnější. Dá se říci, že hodnota podle Horwitz je dobrá pomůcka, nemáme-li k dispozici nic sofistikovanějšího. Pokud tedy není k dispozici jiné vhodné srovnání (např. literární údaj o maximální přípustné reprodukovatelnosti vyplývající z biologické variability), můžeme experimentálně zjištěnou směrodatnou odchylku reprodukovatelnosti srovnat se směrodatnou odchylkou reprodukovatelnosti s_r vypočtenou podle Horwitzova vztahu.

$$s_r = 0,04 \times c^{0,8495} \quad (3),$$

kde c je koncentrace složky vyjádřená v procentech nebo

$$s_R = 0,16 \times c^{0,8495} \quad (4)$$

c je koncentrace složky vyjádřená v mg/kg (ppm). Pořád většinou pracujeme s koncentracemi vztaženými na jednotku objemu, za c můžeme dosadit z následujícího vztahu

$$c = c_v \times \frac{1000}{h} \quad (4a),$$

kde c_v je koncentrace složky vyjádřená v gramech na litr roztoku a h je specifická hmotnost roztoku vyjádřená rovněž v gramech na litr roztoku.

Pro posouzení hodnoty opakovatelnosti s_r můžeme jako hrubý odhad maximální přípustné opakovatelnosti vzít třetinu nebo polovinu s_R .

$$s_r \leq \frac{1}{3} s_R \quad (5)$$

V současnosti se prováděné analýzy zaměřují často na stanovení stopových obsahů sloučenin či prvků, např. toxikologické analýzy, antidopingová vyšetření, stopové biogenní prvky. Získané soubory mohou mít velké rozptýlení výsledků, nemají normální rozložení, ale významnou šikmost, a obsahují odlehlé nebo z odlehlosti podezřelé výsledky, které nelze z různých důvodů opravit a není žádoucí je vylučovat. V těchto případech je vhodné použít k odvození statistických charakteristik polohy a rozptýlení robustních postupů. Některé z níže uvedených postupů (např. medián, použití lognormálního rozdělení) jsou v klinických laboratořích již dlouho používané a uvádíme je zde pro porovnání. Jiné postupy (např. Gastwirthův medián, winsorizace dat, metoda převažujícího shluku, transformace lambda) se ještě v této oblasti nevžily, a proto se pokoušíme na ně upozornit. Použití kvartilů či mediánu lze doporučit jako nejjednodušší postupy, hlavně pokud chceme potlačit vliv extrémních výsledků. Postup s Gastwirthovým mediánem a metoda dvojího mediánu jsou účinnější (a složitější) než předcházející. Vážení výsledků přichází v úvahu tehdy, jestliže máme (byť nestatistické, ale dobré) důvody při zjišťování střední hodnoty a rozptýlení posílit vliv některých výsledků a vliv jiných oslabit. Winsorizace je jednoduchá metoda k potlačení extrémů. Metoda převažujícího shluku jednoduchým způsobem určuje hlavní cluster v datech, jehož střední hodnotu a směrodatnou odchylku lze spočítat klasickým postupem, a může být užitečná, pokud se jedná např. o bimodální data. Logaritmická transformace a transformace lambda se uplatní při odstraňování šikmosti dat.

1. Použití kvartilů \tilde{x}_r [3]

Vyloučí-li se ze souboru výsledků uspořádaného od nejmenšího výsledku k největšímu dolní čtvrtina výsledků, pak první (nejnižší) zbylý výsledek je dolní kvartil označený $\tilde{x}_{0,25}$. Podobně při vyloučení horní čtvrtiny výsledků je poslední (nejvyšší) zbylý výsledek horní kvartil $\tilde{x}_{0,75}$. Podobně při vyloučení třetiny výsledků obdržíme

dolní tercil $\tilde{x}_{0,33}$ a horní tercil $\tilde{x}_{0,67}$. Medián (odhad střední hodnoty) má označení $\tilde{x}_{0,5}$ nebo krátce \tilde{x} .

Pozice kvantilu \tilde{x}_r je dána vztahem

$$x_r = (n + 1) \times \left(r - \frac{i}{n+1}\right) \times (x_{i+1} - x_i) + x_i \quad (6),$$

kde r je označení kvantilu, n je počet výsledků v souboru a i se určí z nerovnosti

$$i \leq r(n + 1) \leq (i + 1) \quad (7)$$

2. Robustní odhady směrodatné odchylky a střední hodnoty pomocí kvartilů

Tyto robustní odhady jsou dány následujícími vztahy [3]

$$s_{rob} = \frac{\tilde{x}_{0,75} - \tilde{x}_{0,25}}{1,3490} \quad (8)$$

$$x_{rob} = \frac{\tilde{x}_{0,75} + \tilde{x}_{0,25}}{2} \quad (9)$$

Příklad:

Stanovením glukózy v krevním séru byly nalezeny tyto výsledky (mmol/l):

x_i : 1 2 4 5,3 7,1 8 8,1 13 24

Tento soubor výsledků má následující charakteristiky (za předpokladu normality):

$n = 9$ (různých laboratoří), $\bar{x} = 8,06$, $s_R = 6,98$

Uvedený soubor výsledků, který je velmi špatné kvality, výsledky nedosahují ani řádové shody, byl zvolen záměrně, aby vynikly rozdíly mezi středními hodnotami a rozptýlenými určenými různými postupy. Ve skutečnosti by správně neměl být vůbec hodnocen, místo toho by použitá metoda měla být podrobena zkoumání s cílem odstranit velké rozptýlení výsledků. Soubor má významnou šikmost $sk = 1,97$ (momentový odhad koeficientu šikmosti) proti tabelované hodnotě $sk(0,95;9) = 0,946$. Měření je málo přesné, což se projevuje ve velké hodnotě reprodukovatelnosti $s_R = 6,98$. Má-li se experimentální s_R srovnat se s_R podle Horwitze, je nutno koncentraci glukózy vyjádřenou v mmol/l přepočítat na mg/l, tj. přibližně na mg/kg.

Relativní molekulová hmotnost glukózy je 180,1572; 1 mmol/l glukózy je tedy 180,1572 mg/l, tj. 180,1572 mg/kg glukózy. 8,06 mmol/l glukózy je 1452 mg/kg glukózy.

Pro tuto koncentraci udává Horwitzův vztah pro s_R :

$$s_R = 0,16 \times 1452^{0,8495} = 77,66 \text{ mg/kg} = 0,43 \text{ mmol/l glukózy}$$

Horwitzův vztah tedy udává hodnotu zhruba 16krát menší, než jaká byla vypočtena z experimentálních dat za předpokladu normality. Je proto vhodné použít robustní postupy. Výpočet statistických parametrů provedeme pomocí kvartilů.

Horní kvartil: $i \leq (0,75 \times 10) \leq (i + 1) \quad i = 7$

$$\tilde{x}_{0,75} = 10 \times (0,75 - \frac{7}{10}) \times (13 - 8,1) + 8,1 = 10,55$$

Dolní kvartil: $i \leq (0,25 \times 10) \leq (i + 1)$ $i = 2$

$$\tilde{x}_{0,25} = 10 \times (0,25 - \frac{2}{10}) \times (4 - 2) + 2 = 3,00$$

Robustní odhad směrodatné odchylky:

$$s_{rob} = \frac{10,55 - 3,00}{1,3490} = 5,60$$

Robustní odhad střední hodnoty:

$$\bar{x}_{rob} = \frac{10,55 + 3,00}{2} = 6,78$$

Robustní odhad střední hodnoty 6,78 je výrazně menší než aritmetický průměr 8,06.

Robustní odhad směrodatné odchylky 5,60 podobně jako klasická směrodatná odchylka 6,98 mnohokrát převyšuje Horwitzovu hodnotu.

3. Medián [5]

Medián $\tilde{x}_{0,5}$ je prostřední hodnota uspořádaného souboru, v našem případě

$$\tilde{x}_{0,5} = x_5 = 7,1$$

Lze jej vypočíst s použitím rovnic (6) a (7)

$$i \leq (0,5 \times 10) \leq (i + 1) \quad i = 4$$

$$\tilde{x}_{0,5} = 10 \times (0,5 - \frac{4}{10}) \times (7,1 - 5,3) + 5,3 = 7,1$$

Interval spolehlivosti mediánu $I(\tilde{x}_{0,5})$ se odhadne z interkvartilového rozpětí [3]:

$$I(\tilde{x}_{0,5}) = \tilde{x}_{0,5} \pm 1,57 \times \frac{(\tilde{x}_{0,75} - \tilde{x}_{0,25})}{\sqrt{n}} =$$

$$= 7,1 \pm 1,57 \frac{10,55 - 3,0}{\sqrt{9}} = \begin{cases} 11,1 \\ 3,1 \end{cases} \quad (10)$$

4. Gastwirthův medián [6]

Gastwirthův medián \tilde{x}_{Gst} patří mezi robustní odhady střední hodnoty. Počítá se pomocí klasického mediánu, dolního a horního tercilu. Odhaduje se ze vztahu

$$\tilde{x}_{Gst} = 0,4 \times \tilde{x}_{0,5} + 0,3 \times (\tilde{x}_{0,67} + \tilde{x}_{0,33}) \quad (11)$$

Příklad:

Glukóza v séru (viz výše).

Horní tercil: $i \leq (0,67 \times 10) \leq (i + 1)$ $i = 6$

$$\tilde{x}_{0,67} = 10 \times (0,67 - \frac{6}{10}) \times (8,1 - 8) + 8 = 8,07$$

Dolní tercil: $i \leq (0,33 \times 10) \leq (i + 1)$ $i = 3$

$$\tilde{x}_{0,33} = 10 \times (0,33 - \frac{3}{10}) \times (5,3 - 4) + 4 = 4,39$$

Pro Gastwirthův medián dostaneme hodnotu

$$\tilde{x}_{Gst} = 0,4 \times 7,1 + 0,3 \times (8,07 + 4,39) = 6,58$$

Pro interval spolehlivosti $I(\tilde{x}_{Gst})$ je možno použít interkvartilového rozpětí podle rovnice (10).

5. Metoda dvojího mediánu

Při aplikaci této metody se postupuje následovně:

- zjistí se medián \tilde{x}_1 uspořádaných výsledků;
- vypočtou se absolutní rozdíly $|x_i - \tilde{x}_1|$;
- zjistí se medián absolutních rozdílů MAD_1 ;
- vypočte se faktor f_1 :

$$f_1 = (0,7722 + \frac{1,604}{n}) \times t(P, n - 1)$$

kde $t(P, n-1)$ je Studentův kvantil pro pravděpodobnost P (typicky $P = 0,95$) a počet stupňů volnosti $(n-1)$;

– pro každé x_i se vypočte testovací kritérium TK_i

$$TK_i = \frac{|x_i - \tilde{x}_1|}{f_1 \times MAD_1} \quad (12)$$

– výsledky, jejichž $TK_i > 2$ se vyloučí a celý postup se opakuje.

Příklad:

Glukóza v séru (viz výše).

Table 1. Dual median, data (x_i) and enumerations; definitions of symbols in text

x_i	Step 1			Step 2		
	$(x_i - 7,1)$	$ x_i - 7,1 $	TK_1	$(x_i - 6,2)$	$ x_i - 6,2 $	TK_2
1	-6.1	0.0	0.0	-5.2	0.9	0.19
2	-5.1	0.9	0.13	-4.2	0.9	0.19
4	-3.1	1.0	0.15	-2.2	1.8	0.38
5,3	-1.8	1.8	0.26	-0.9	1.9	0.40
7,1	0.0	3.1	0.46	0.9	2.2	0.47
8	0.9	5.1	0.75	1.8	4.2	0.89
8,1	1.0	5.9	0.87	1.9	5.2	1.10
13	5.9	6.1	0.90	6.8	6.8	1.44
24	16.9	16.9	2.49	-	-	-

Výpočty k prvnímu kroku:

$$MAD_1 = 3,1$$

$$f_1 = (0,7722 + \frac{1,604}{9}) \times 2,3060 = 2,1917$$

$$TK_i = \frac{|x_i - 7,1|}{2,1917 \times 3,1} = \frac{|x_i - 7,1|}{6,7942}$$

Vylučuje se výsledek $x_9 = 24$, jehož $TK_9 = 2,49$ je větší než 2.

Výpočty k druhému kroku:

$$MAD_2 = 2,05$$

$$f_2 = (0,7722 + \frac{1,604}{8}) \times 2,3646 = 2,3000$$

$$TK_2 = \frac{|x_i - 6,2|}{2,3000 \times 2,05} = \frac{|x_i - 6,2|}{4,7151}$$

Jelikož v druhém kroku jsou všechna testovací kritéria menší než 2, nedochází k vylučování výsledků. Střední hodnotou je druhý medián $\tilde{x} = 6,2$.

Jeho interval spolehlivosti se odhadne z interkvartilového rozpětí redukovaného souboru podle rovnice (10)

$$\begin{aligned} \bar{x}_{0,75} &= 8,0 & \bar{x}_{0,25} &= 4,0 \\ I(6,2) &= 6,2 \pm 1,57 \times \frac{(8-4)}{\sqrt{8}} \end{aligned} \begin{array}{l} \nearrow 8,4 \\ \searrow 4,0 \end{array}$$

6. Vážení výsledků

Omezení vlivu výsledků na obou koncích rozdělení lze docílit jednoduchou metodou, kterou lze nazvat vážení výsledků. Jednotlivým výsledkům sledovaného souboru se přidělí váhy w_i a sice tak, že výsledky kolem průměru dostanou váhy nejvyšší a výsledky směrem k nižším a vyšším hodnotám váhy o to nižší, o co jsou od průměru v absolutní hodnotě vzdálenější. Výsledky kolem průměru mají váhy nejvyšší, poněvadž jsou nejpravděpodobnější. To je rozdíl proti běžným metodám vyhodnocování, kdy se výsledky sčítají všechny se stejnou (typicky jednotkovou) vahou. Při metodě vážení výsledků se pro výpočet průměru a směrodatné odchylky použijí všechny výsledky, nedochází k jejich vylučování. Ovšem výsledky okrajové se berou s malými vahami. Mírou váhy je hustota pravděpodobnosti plynoucí z Gaussova rozdělení.

Při metodě vážení výsledků se vyjde z klasického aritmetického průměru \bar{x} a klasické směrodatné odchylky hodnoceného souboru. Jednotlivým výsledkům x_i se přidělí váhy w_i podle vzorce

$$w_i = e^{-\frac{1}{2} \left(\frac{x_i - \bar{x}}{s} \right)^2} \quad (13)$$

a vypočtou se $\sum w_i, \sum x_i w_i, \sum x_i^2 w_i$.

Pro vážený průměr \bar{x}_w platí

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i} \quad (14)$$

Pro druhou mocninu vážené směrodatné odchylky s_w^2 platí

$$s_w^2 = \frac{1}{\sum w_i} \left[\sum x_i^2 w_i - \frac{(\sum x_i w_i)^2}{\sum w_i} \right] \quad (15)$$

Příklad:

Glukóza v séru (viz výše).

Table 2. Weighting of results, data (x_i) and enumerations; definitions of symbols in text

i	x_i	w_i	$x_i w_i$	$x_i^2 w_i$
1	1	0.5996	0.5996	0.5996
2	2;	0.6860	1.3720	2.7440
3	4	0.8444	3.3775	13.5099
4	5,3	0.9248	4.9014	25.9777
5	7,1	0.9906	7.0332	49.9355
6	8	1.0000	7.9997	63.9976
7	8,1	1.0000	8.0999	65.6089
8	13	0.7785	10.1199	131.5585
9	24	0.0737	1.7691	42.4590
	$\bar{x} = 8.06$	$\sum w_i =$	$\sum x_i w_i =$	$\sum x_i^2 w_i =$
	$s = 6.98$	6.8976	45.2723	396.3909

Vážený průměr \bar{x}_w je

$$\bar{x}_w = \frac{45,2723}{6,8976} = 6,56$$

Druhá mocnina vážené směrodatné odchylky s_w^2 a vážená směrodatná odchylka s_w jsou

$$s_w^2 = \frac{1}{6,8976} \times [396,3909 - \frac{45,2723^2}{6,8976}] = 14,3886$$

$$s_w = 3,79$$

Interval spolehlivosti váženého průměru $I(\bar{x}_w)$ je

$$I(\bar{x}_w) = 6,56 \pm \frac{2,3060 \times 3,79}{\sqrt{9}} \begin{array}{l} \nearrow 9,47 \\ \searrow 3,65 \end{array}$$

7. Winsorizace dat [7]

Tato metoda spočívá v náhradě maximálního a minimálního výsledku v uspořádaném souboru dat výsledkem sousedním, tzn. první výsledek se nahradí druhým a n-tý výsledek (n-1)-ním. Tímto postupem se do určité míry omezí odlehlost dat, ale zachová se jejich trend – velký výsledek se nahradí opět velkým a malý malým. Obecný postup winsorizace lze nalézt v literatuře [3].

Příklad:

Glukóza v séru (viz výše).

Po naznačené operaci dostaneme korigovaný soubor dat

x_i : 2 2 4 5,3 7,1 8 8,1 13 13

$$n = 9 \quad \bar{x}_{corr} = 6,94 \quad s_{corr} = 4,12$$

Pro interval spolehlivosti korigovaného průměru $I(\bar{x}_{corr})$ dostaneme

$$I(\bar{x}_{corr}) = 6,94 \pm \frac{2,3060 \times 4,12}{\sqrt{9}} \begin{array}{l} \nearrow 10,11 \\ \searrow 3,77 \end{array}$$

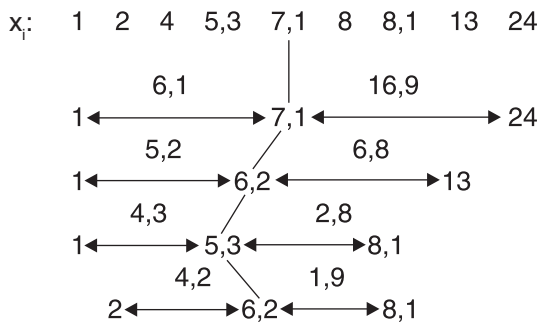
8. Metoda převažujícího shluku (dominant cluster method) [8]

Postup odvození střední hodnoty uspořádaného souboru je tento:

- zjistí se medián \bar{x}_1 ;
- zjistí se rozdíly $(x_n - \bar{x}_1)$ a $(\bar{x}_1 - x_1)$;
- vyloučí se buď x_n , nebo x_1 podle toho, který rozdíl je větší;
- zjistí se medián redukovaného souboru \bar{x}_2 a zjišťují se obdobně rozdíly $(x_{\max} - \bar{x}_2)$ a $(\bar{x}_2 - x_{\min})$ a vyloučí se x_{\max} nebo x_{\min} podle toho, který rozdíl je větší;
- ve vylučování dat se pokračuje tak dlouho, až zůstane pouze 5 výsledků. Jejich aritmetický průměr se považuje za střední hodnotu.

Příklad:

Glukóza v séru (viz výše).



Vyloučeny byly výsledky: 1, 2, 13, 24; zbylo pět výsledků 4, 5,3, 7,1, 8,8,1 – tedy pět dat. Jejich aritmetický průměr je střední hodnotou.

$$\bar{x} = 6,5$$

9. Lognormální rozložení dat [5, 9]

Významná šikmost experimentálních dat x_i se většinou odstraní transformací na dekadické logaritmy X_i .

$$X_i = \log x_i \quad (16)$$

Transformované veličiny se testují na normalitu rozložení. Prokáže-li se, že X_i jsou rozloženy normálně, znamená to, že původní data x_i jsou rozložena lognormálně.

U transformovaných hodnot X_i se vypočte aritmetický průměr \bar{X} , směrodatná odchylka S a interval spolehlivosti $I(\bar{X})$. Provede se retransformace \bar{X} na geometrický průměr \bar{x}_g a retransformace intervalu spolehlivosti $I(\bar{X})$ na interval spolehlivosti geometrického průměru $I(\bar{x}_g)$.

$$\bar{x}_g = \text{antilog } \bar{X} \quad (17)$$

$$I(\bar{x}_g) = \text{antilog } I(\bar{X}) \quad (18)$$

Interval spolehlivosti $I(\bar{x}_g)$ je asymetrický vůči \bar{x}_g , nedosahuje však nikdy záporných hodnot. Jeho použití se proto doporučuje pro soubory dat kolem meze stanovitelnosti.

Příklad:

Glukóza v séru (viz výše).

Table 3. Lognormal distribution, data (x_i) and enumerations; definitions of symbols in text

i	x_i	X_i
1	1	0.0000
2	2	0.3010
3	4	0.6021
4	5.3	0.7243
5	7.1	0.8513
6	8	0.9031
7	8.1	0.9085
8	13	1.1139
9	24	1.3802
		$\bar{X} = 0.7538$
		$S = 0.4146$

Transformovaná data jsou normálně rozložena a mají pouze nevýznamnou šikmost. Jejich retransformací dostaneme:

$$x_g = \text{antilog } 0,7538 = 5,67$$

$$I(\bar{x}_g) = 0,7538 \pm \frac{2,3060 \times 0,4146}{\sqrt{9}} \begin{matrix} \nearrow 1,0725 \\ \searrow 0,4351 \end{matrix}$$

$$I(\bar{x}_g) = \begin{matrix} \nearrow \text{antilog } 1,0725 = 11,8 \\ \searrow \text{antilog } 0,4351 = 2,7 \end{matrix}$$

10. Transformace lambda [10]

Tato transformace, ačkoliv je bez použití počítačového programu pracná, je považována za nejlepší pro hodnocení výsledků s významnou šikmostí.

Experimentální hodnoty x_i se transformují na X_i podle rovnice

$$X_i = \frac{x_i^\lambda - 1}{\lambda} \quad (19),$$

kde λ se volí tak, aby šikmost X_i byla zanedbatelná, např. menší než 0,05. U transformovaných hodnot se vypočte aritmetický průměr \bar{X} a směrodatná odchylka S . Pro šikmost $s\hat{k}$ platí vztah [9, 10]

$$s\hat{k} = \frac{(X_i - \bar{X})^3}{S^3} \times \frac{\sqrt{n}}{\sqrt{(n-1)^3}} \quad (20)$$

Retransformací obdržíme lambda průměr x_λ a jeho interval spolehlivosti $I(x_\lambda)$

$$\bar{x}_\lambda = (\bar{X} \times \lambda + 1)^{\frac{1}{\lambda}} \quad (21)$$

$$I(\bar{x}_\lambda) = \begin{matrix} \nearrow (\bar{X}_2 \times \lambda + 1)^{\frac{1}{\lambda}} \\ \searrow (\bar{X}_1 \times \lambda + 1)^{\frac{1}{\lambda}} \end{matrix} \quad (22)$$

Příklad:

Glukóza v séru (viz výše).

Vhodné λ bylo nalezeno zkusem ($\lambda = 0,22$), takže transformace byla provedena podle vztahu

$$X_i = \frac{x_i^{0,22} - 1}{0,22}$$

Table 4. Lambda transformation, data (x_i) and enumerations; definitions of symbols in text

i	x_i	X_i
1	1	0.0000
2	2	0.7488
3	4	1.6209
4	5.3	2.0148
5	7.1	2.4506
6	8	2.6367
7	8.1	2.6564
8	13	3.4464
9	24	4.6004
		$\bar{X} = 2.2417$
		$S = 1.3728$

Transformované hodnoty X_i jsou normálně rozloženy a mají nevýznamnou šikmost: $s\hat{K} = 0,008$ proti tabulované $sk(0,95,9) = 0,946$ [11]. Interval spolehlivosti transformovaného průměru $l(\bar{X})$ je

$$l(\bar{X}) = 2,2417 \pm \frac{2,3060 \times 1,3728}{\sqrt{9}} \begin{cases} \rightarrow 3,2969 \\ \rightarrow 1,1865 \end{cases}$$

Provedeme retransformaci

$$\bar{x}_k = (2,2417 \times 0,22 + 1)^{\frac{1}{0,22}} = 6,19$$

$$l(\bar{x}_k) = \begin{cases} \rightarrow (3,2969 \times 0,22 + 1)^{\frac{1}{0,22}} = 11,93 \\ \rightarrow (1,1865 \times 0,22 + 1)^{\frac{1}{0,22}} = 2,87 \end{cases}$$

11. Přehled výsledků

V následujícím přehledu jsou použity tyto zkratky: SH = střední hodnota, SD = směrodatná odchylka, IS = interval spolehlivosti.

- předpoklad normality SH = 8,06 SD = 6,98
- pomocí kvartilů SH = 6,78 SD = 5,60
- medián a interkvartilové rozpětí SH = 7,1
IS = 3,1...11,1
- Gastwirthův medián a interkvartilové rozpětí SH = 6,58 IS = 3,1.11,1
- metoda dvojího mediánu a interkvartilové rozpětí redukovaného souboru SH = 6,2 IS = 4,0.8,4
- vážených výsledků SH = 6,56 SD = 3,79 IS = 3,65.9,47
- winsorizace dat SH = 6,94 SD = 4,12
IS = 3,77...10,11
- metoda převažujícího shluku (dominant cluster method) SH = 6,5
- lognormální rozložení dat SH = 5,67 IS = 2,7.11,8
- transformace lambda SH = 6,19 IS = 2,87..11,93

12. Diskuse a závěr

Interval spolehlivosti vypočtený za předpokladů normality (a) a pomocí kvartilů (b) je tak široký, že zasahuje do záporných hodnot, a nemá tudíž fyzikální smysl. Výsledky metodami (a) a (b) jsou tedy prakticky nepoužitelné. Všechny ostatní metody, kromě metody převažujícího shluku (h), která samotná interval spolehlivosti nepočítá, už poskytují odhady intervalu spolehlivosti, které jsou reálné.

Z nich nejvyšší odhad střední hodnoty 7,1 je dán klasickým mediánem – metoda (c) – a nejnižší střední hodnotu 5,67 dává geometrický průměr – metoda (i). Přitom obě tyto metody poskytují přibližně stejný interval spolehlivosti – asi 2,9 ... 11,5.

Nejmenší interval spolehlivosti 4,0 ... 8,4 a prakticky druhou nejmenší střední hodnotu 6,2 dává metoda dvojího mediánu a interkvartilového rozpětí redukovaného souboru (e).

Metody (d), (i) a (j) dávají interval spolehlivosti, který je asymetricky rozložen kolem střední hodnoty a střední hodnota je blíže jeho dolní mezi. Ostatní intervaly spolehlivosti jsou vzhledem ke střední hodnotě symetrické.

Pro data v našem příkladu se jeví jako nejrealističtější výsledky metody (d) a metody (j). Obě udávají střední hodnotu kolem 6,3 a značně široký interval spolehlivosti cca 3 ... 11, který je asymetrický vzhledem ke střední hodnotě.

Literatura

- Roper, P., Burke, S., Lawn, R., Barwick, V., Walker, R. *Applications of Reference Materials in Analytical Chemistry*. Cambridge : Royal Society of Chemistry 2001.
- Horwitz, W., Lamps, L. R., Boyer, K. W. The Empirical Relationship between Standard Deviation and Concentration. *J. Ass. Off. Anal. Chem.*, 1980, 63, p. 1344–1347.
- Meloun, M., Militký, J. *Statistické zpracování experimentálních dat*. Praha : PLUS spol. s r. o. 1994, ISBN 80–85297–56–6.
- Govindaraju, K., Rubeška, I., Paukert, T. 1944 Report on Zinnwaldite ZW-C Analysed by Ninety-Two IWG-Member-Laboratories. *Geostandards Newsletter*, 1994, 18, p. 1–42.
- Doerffel, K. *Statistik in der analytischen Chemie*. Leipzig : VEB Deutscher Verlag für Grundstoffindustrie 1966.
- Gastwirth, J. On Robust Procedures. *J. Amer. Stat. Assoc.*, 1966, 61, p. 929–935.
- Winsor, C. P. The Future of Data Analysis. *Ann. Math. Statist.*, 1962, 33, p. 1–67.
- Ellis, P. J., Copelowitz, I., Steele, T. W. Estimation of the Mode by the Dominant Cluster Method. *Geostandards Newsletter*, 1977, 1, p. 123–130.
- Sachs, L. *Angewandte Statistik*. Berlin : Springer Verlag 1974.
- Christie, O. H. J., Alfsen, K. H. Data Transformation as a Means to Obtain Reliable Consensus Values for Reference Materials. *Geostandards Newsletter*, 1977, 1, p. 47–49.
- ČSN 01 2841: Referenční materiály. Praha : Vydavatelství ÚNM 1988.

Do redakce došlo 17. 8. 2004.

Adresa pro korespondenci:
Dr. Ing. Josef Dempír, CSc.
285 32 Hlízov 22