Chem. Listy *115*, 300–304 (2021) Referát

IDENTIFIKACE ODLEHLÝCH HODNOT V CHEMICKÝCH MĚŘENÍCH S UVÁŽENÍM NESYMETRIE JEJICH ROZDĚLENÍ PRAVDĚPODOBNOSTI

JOSEF TOŠENOVSKÝ a FILIP TOŠENOVSKÝ

VŠB-Technická univerzita Ostrava, Fakulta materiálově technologická, 17. listopadu 2172/15, 708 00 Ostrava - Poruba
josef.tosenovsky@vsb.cz

Došlo 11.1.21, přijato: 19.2.21.

Klíčová slova: odlehlé hodnoty, kladná a záporná šikmost, kvartily, krabicový graf

Obsah

1. Úvod

Výsledky chemických měření velmi často obsahují ojedinělá data, která jsou výrazně větší (menší) než většina ostatních, tedy odlehlá pozorování (outliers). Jsou-li zařazena do dalšího zpracování, mohou výrazně ovlivnit výsledky. Jsou-li vyloučena, může dojít ke ztrátě cenných informací. Není-li možné jejich věcné vyhodnocení resp. opakované měření, nezbývá, než tato data vyhodnotit statistickými prostředky, zejména testy. Ty však také pracují s těmito daty a jsou jimi ovlivněny a navíc je většina běžně dostupných testů vázána na splnění konkrétních předpokladů, např. normální rozdělení dat. To často není splněno a právě u dat s kladnou resp. zápornou šikmostí nemusí být výskyt odlehlostí chybnou hodnotou a vyskytuje se mnohem častěji. Proto je vhodné brát v úvahu tuto šikmost při hodnocení extrémních hodnot.

Nejprve se však musí šikmost rozdělení zjistit. Běžně se posuzuje koeficientem šikmosti (skewness), který je také odlehlými daty ovlivněn. Dobrou cestou při hodnocení extrémních dat může být kombinace robustního ukazatele šikmosti[1] a grafické metody, která není vázána na splnění specifických předpokladů. Často je používán klasický krabicový graf[2,3] (boxplot). Ten ale u souborů s kladnou nebo zápornou šikmostí vylučuje nadbytek dat. Proto byl klasický krabicový graf několikrát vylepšován, výrazně např. v r. 1990 (cit.[4]) a naposledy v r. 2018, kde je nejvýraznější úprava pro případy nesymetrických rozdělení dat[5].

Cílem této studie je otestovat účinnost nových krabicových grafů[4,5], objasnit uživatelům princip jejich konstrukce a poskytnou chemikům jednoduchý návod na jejich praktické použití pouze s všeobecně dostupným programem Excel Microsoft.

Vlastnosti nových krabicových grafů jsou vyzkoušeny a ilustrovány jednak na simulovaných datech, kde je možné vytvořit uměle problémové situace, ale také na reálných datech[6] z chemického průmyslu.

2. Základní vzorce a jejich analýza

2.1. Klasický krabicový graf (Tukey)

Krabicový graf tvoří krabička s levým a pravým vousem. Body q_1, q_2 a q_3 jsou první, druhý a třetí kvartil daného souboru. Vzdálenost mezi krajními body vousů tvoří interval, dále označovaný (A, B), kde body mimo tento interval jsou považovány za odlehlé. Jsou to tedy body pod hranicí A resp. nad hranicí B (obr. 1). Je zřejmé, že čím delší je vous, tím méně je vyloučených hodnot. Nové úpravy základního krabicového grafu prodlužují pravý (levý) vous ve směru kladné (záporné) šikmosti dat.

Základem pro výpočet hraničních bodů krabicového grafu ve všech uvažovaných vzorcích jsou kvartily q_1, q_2 a q_3. Pomocí kvartilů je počítáno:

kvartilové rozpětí $QR = q_3 - q_1$ *(1)*

horní kvartilové rozpětí $UQR = q_3 - q_2$ *(2)*

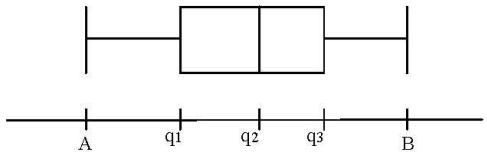
dolní kvartilové rozpětí $LQR = q_2 - q_1$ *(3)*

Dále bude použita robustní charakteristika šikmosti (Bowle), která má tvar:

$$Bc = \frac{UQR - LQR}{QR} = \frac{(q_3 - q_2) - (q_2 - q_1)}{q_3 - q_1} = \frac{q_3 - 2q_2 + q_1}{q_3 - q_1} \quad (4)$$

Pro symetrické rozdělení je LQR = UQR, takže Bc = 0; pro kladnou šikmost bude UQR > LQR a Bc > 0; pro zápornou šikmost je UQR < LQR a Bc < 0.

Bc má tedy stejná znaménka jako momentová charakteristika šikmosti:



Obr. 1 **Klasický krabicový graf a jeho klíčové body**

$$\mathrm{Sk} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} \qquad (5)$$

Základní krabicový graf [2,3] počítá hranice A, B, které vymezují interval (A, B) pro přijetí dat, takto:

$$A = q_1 - c \cdot QR,\ B = q_3 + c \cdot QR \qquad (6)$$

kde nejčastěji c = 1,5.

2.2. Nové krabicové grafy a princip jejich fungování

Kimber[4] upravil vzorec (*6*) na tvar:

$$A = q_1 - c \cdot 2LQR,\ B = q_3 + c \cdot 2UQR \qquad (7)$$

proto, aby prodloužil hranice překročení ve směru šikmosti, tj. rozšířil interval (A, B) a tím snížil počet vyloučených bodů u nesymetrických rozdělení. Logika úpravy tohoto krabicového grafu je lépe vidět v jiném vyjádření:

$$A = q_1 - c \cdot QR \cdot (1 - Bc),\ B = q_3 + c \cdot QR \cdot (1 + Bc) \qquad (8)$$

Pro kladnou šikmost je Bc > 0, takže 1 + Bc > 1 a položka c·QR(1+Bc), která se přičítá ke q_3, je větší, než původní c·QR. Tím se sníží počet bodů, překračujících B. Pro symetrická rozdělení s Bc = 0 je potom vzorec (*8*) shodný se vzorcem klasickým (*6*). Pro zápornou šikmost, pokud LQR > UQR (jak se očekává), je Bc < 0, takže 1 – Bc > 1 a od q_1 se odečítá ve vzorci (8) položka větší než c.QR ve vzorci (*6*); proto pro dolní hranice grafů K a T platí K < T, tedy opět méně hodnot je menších (vyřazených) u K než u klasické dolní hranice grafu T v (*6*). Vzorce pro Kimberův krabicový graf (*7*) a jeho upravenou verzi (*8*) (adjusted boxplot) jsou stejné: dosadíme-li do vzorce (*8*) pro bod A:

$$1 - Bc = 1 - \frac{q_3 - 2q_2 + q_1}{q_3 - q_1} = \frac{2LQR}{QR} \qquad (8a)$$

dostáváme:

$$A = q_1 - c \cdot QR \cdot \frac{2LQR}{QR} = q_1 - c \cdot 2LQR \qquad (8b)$$

Podobně pro hraniční bod B bude po dosazení za

$$1 + Bc = 1 + \frac{q_3 - 2q_2 + q_1}{q_3 - q_1} = \frac{2UQR}{QR} \qquad (8c)$$

horní hranice:

$$B = q_3 + c \cdot QR \frac{2UQR}{QR} = q_3 + c \cdot 2UQR \qquad (8d)$$

Nejnovější verze krabicového grafu[5], která je označená jako RS (Ratio Skewnes), upravuje hranice A, B v základním vzorci (*6*) tak, že místo Kimberova posílení položkou (1 + B) a (1 – B) používá účinnější podíl kvartilových rozpětí. Tak vzniknou vzorce pro dolní hranici:

$$A = q_1 - c \cdot QR \cdot \frac{LQR}{UQR} \qquad (9)$$

a pro horní hranici:

$$B = q_3 + c \cdot QR \cdot \frac{UQR}{LQR} \qquad (10)$$

3. Testování nových grafických metod

Cílem dále uvedených úloh je ukázat, že nová konstrukce krabicového grafu je provedena tak, aby graf respektoval kladnou resp. zápornou šikmost. Prakticky to znamená, že ve směru zjištěné šikmosti je hranice překročení A resp. B obtížnější (delší) než u klasického grafu. Je tedy u nových grafů pro kladnou šikmost (Bc > 0) vzdálenost bodu B od kvartilu q_3 větší, podobně u záporné šikmosti (Bc < 0) je bod A vzdálenější od q_1. Tím se prodlouží interval (A, B) *ve směru šikmosti* a pouze ve směru šikmosti.

Všechny úlohy, které jsou uvedené v kapitolách 3.1. až 3.3. mají jednotné zadání: vypočítat charakteristiky šikmosti, zjistit dolní resp. horní hranici klasického a nového krabicového grafu a porovnat počty vyřazených hodnot. Analýzu provést ve směru zjištěné šikmosti. U každého souboru je uvedeno posledních 5 největších hodnot pro kladnou šikmost nebo nejmenších 5 hodnot pro zápornou šikmost. Je-li potřeba, pak se uvádí i 10 hodnot.

3.1. Data s kladnou šikmostí

a) data (n = 100) jsou generována z rozdělení chí-kvadrát (1) (χ_1^2) a doplněna odlehlou hodnotou. Charakteristiky šikmosti: Sk = 14,48 a Bc = 0,50, tedy kladná šikmost.

Vzhledem ke kladné šikmosti sledujeme horní hranici B krabicového grafu.

Výsledky výpočtů, potřebných k nalezení hranice B:

q_1 = 0,086; q_2 = 0,39; q_3 = 1,30; QR = 1,214; LQR = 0,304; UQR = 0,91.

Největších 10 náhodných čísel včetně odlehlé hodnoty (tučně):

2,77; 2,93; 2,95; 3,18; 3,38; 3,54; 3,73; 3,79; 4,83; **10,27288**.
Odlehlá hodnota 10,27288 je generovaná jako kritická hodnota $\chi^2_1(0{,}00135)$.
Závěry: Nejlépe hodnotí data krabicový graf RS, který vynechal jen odlehlou hodnotu. Klasický krabicový graf vyřadil kromě odlehlé hodnoty ještě 6 platných (tab. I).

b) data (cit.[6], s. 91, úloha C2.02; n = 90) představují obsah acetonu v surovém isopropylaminu [%]. Charakteristiky šikmosti: Sk = 8,97 a Bc = 0,41.
Pozornost bude tedy soustředěna na horní hranici B krabicového grafu.
Největších 10 hodnot obsahu acetonu [%]:
0,86; 0,90; 0,96; 0,97; 1,12; 1,14; 1,15; 1,58; 1,61; 2,15.
Pro klasický krabicový graf[3] a pro dvě nové konstrukce[4,5] (K, RS) byly vypočítány hranice intervalu přijetí (A, B). Předmětem zájmu je, vzhledem ke kladné šikmosti dat, počet vyřazených měření v důsledku překročení *horní hranice* B u klasického grafu (T) a nových grafů (K, RS).
Závěry: Z výsledku je patrné, že krabicový graf RS je *ve směru kladné šikmosti* „šetrnější“, respektuje kladnou šikmost dat a vyřazuje nejméně měření (tab. I).

c) výběr z dat (cit.[6], s. 91, úloha C2.02; n = 15); z dat 3.1b byla vybrána část pro porovnání výsledků u souborů o různém rozsahu. Charakteristiky šikmosti jsou:
Sk = 4,11 a Bc = 0,76. Pozornost bude soustředěna na největších 5 hodnot: 0,12; 0,36; 0,36; 0,86; 1,61. Hranice B horního vousu a vyloučené hodnoty jsou v tabulce I.
Závěry: Vzhledem k malému rozsahu souboru ve srovnání s 3.1b je také méně vyloučených hodnot. Větší „citlivost“ k šikmosti je patrná u grafu RS, grafy T
a K vylučují jednu hodnotu.

3.2. Data se zápornou šikmostí

a) data (n = 100) jsou generována z Weibullova rozdělení W (5; 400) a doplněna o 3 odlehlé hodnoty (průměr – k.s, k = 3; 3,5 a 4). Charakteristiky šikmosti jsou: Sk = –3,29 a Bc = –0,17, tedy záporná šikmost. Výsledky výpočtů, potřebných k nalezení hranice A: q_1 = 308,56; q_2 = 369,82; q_3 = 412,93; QR = 104,37; LQR = 61,26; UQR = 43,11.
Vzhledem k záporné šikmosti sledujeme dolní hranici krabicového grafu A. Nejmenších 5 hodnot souboru se třemi odlehlými čísly (tučně): **66,105; 103,355; 140,605**; 142,523; 187,436.
Závěr: Zápornou šikmost zohledňuje nejvýrazněji graf RS, nejméně graf T (tab. II).

b) data (cit.[6], s. 99, úloha C2.31B ; n = 14): obsah modrých látek ve filtrátech [%]. Jedná se o kompletní výsledky měření malého rozsahu. Přesto, že data mají normální rozdělení (test chí-kvadrát, p-hodnota = 0,10), jsou empirické charakteristiky šikmosti záporné: Sk = –1,99 a Bc = –0,54. Pozornost bude proto soustředěna na nejmenších 5 hodnot souboru: 1,18; 2,22; 2,23; 2,24; 2,33.
Závěr: V intervalu (průměr – 3s; průměr + 3s), zde (1,02; 4,08), se u normálního rozdělení nachází 99,73 % hodnot, tedy drtivá většina; pravděpodobnost výskytu hodnoty mimo tento interval je 0,0027. Všechna data daného souboru jsou skutečně v tomto intervalu (x_{min} = 1,18, x_{max} = 3,38). Přesto je krabicovým grafem T vyloučena jedna hodnota, u nových grafů žádná (tab. II). Vzhledem k záporné šikmosti prodlužují nové grafy délku dolního vousu; jednotlivé délky jsou: graf T: 0,84; graf K: 1,29 a graf RS: 2,78. Horní vous se naopak zkracuje; graf T: 0,84; graf K: 0,39; graf RS: 0,25. Nové grafy jsou tedy orientovány *jen ve směru šikmosti.*

3.3. Data s nulovou šikmostí

a) data (n = 100) jsou generovaná ze Studentova rozdělení t_{10} a doplněna třemi odlehlými hodnotami: 4,297 = průměr + 3s; 4,39 a 4,58 jsou kritické hodnoty rozdělení t_{10} pro hladinu významnosti 0,00135 a 0,001.
U dat se Studentovým rozdělením se očekává symetrie, avšak empirické charakteristiky šikmosti vychází nenulové: Sk = 2,91 a Bc = 0,21, tedy kladná šikmost. Pozornost

Tabulka I
Vyloučené hodnoty u jednotlivých grafů pro data 3.1

	3.1a	
Tukey	Kimber	RS
B = 3,12	B = 4,03	B = 6,75
Nad B: 3,18; 3,39; 3,54; 3,72; 3,79; 4,82; 10,27	Nad B: 4,83; 10,27	Nad B: 10,27
	3.1b	
Tukey	Kimber	RS
Interval (A, B): (–0,34; 0,94)	Interval (A, B): (–0,15; 1,14)	Interval (A, B): (–0,06; 1,60)
Nad B: 0,96; 0,97; 1,12; 1,14; 1,15; 1,58; 1,61; 2,15	Nad B: 1,15; 1,58; 1,61; 2,15	Nad B: 1,61; 2,15
	3.1c	
Tukey	Kimber	RS
B = 0,86	B = 1,23	B = 3,95
Nad B: 1,61	Nad B: 1,61	Nad B: nic

Tabulka II
Vyloučené hodnoty u jednotlivých grafů pro data 3.2

	3.2a	
Tukey	Kimber	RS
A: 152	A: 124,78	A: 86,09
Pod A: 66,11; 103,36; 140,61; 142,52	Pod A: 66,11; 103,36;	Pod A: 66,11
	3.2b	
Tukey	Kimber	RS
A:1,4	A: 0,98	A: –0,28
Pod A: 1,18	Pod A: nic	Pod A: nic

je proto zaměřena na největších 5 hodnot, z nichž poslední tři jsou odlehlé (tučně):
2,906; 3,774; **4,397**; **4,393; 4,587**.
Závěry: Také u teoreticky symetrických rozdělení je vhodné použít nové krabicové grafy, pokud je zjištěna šikmost. Pozornost je soustředěna na data ve směru šikmosti.
Zde jsou u klasického grafu vyloučeny 4 hodnoty, u RS grafu žádná (tab. III).

b) data (cit.[6], s. 93, úloha C2.08; n = 60) představují obsah plynného dusíku, který vzniká jako vedlejší produkt při výrobě čpavku a je oddělován v čistící jednotce [%].
Charakteristiky šikmosti: Sk = –0,006 a Bc = 0,03, tedy prakticky symetrie.
Nejmenších 5 hodnot: 19,6; 20,0; 20,6; 20,7; 21,0.
Největších 5 hodnot: 29,0; 29,4; 29,6; 29,8; 30,9.
Závěry: U dat se potvrdila normalita (test chí-kvadrát, p-hodnota = 0,23) a minimální šikmost. Pozornost byla proto soustředěna v obou směrech: pod hranici A a nad hranici B, avšak žádná hodnota daného měření nevybočuje mimo interval (A, B) (tab. III). Grafy zde fungují rovnocenně. Také zde je ale vidět, že i při minimální šikmosti (Bc = 0,03) dochází k mírnému prodloužení hranice B (délky horního vousu);
pro T graf: B – q_3 = 6,53 a pro RS graf: B – q_3 = 6,99.

c) výběr z dat (cit.[6], s. 93, úloha C2.08; n = 15): z dat 3.3b byla vybrána část pro porovnání výsledků u souborů o různém rozsahu. Charakteristiky šikmosti vychází: Sk = –0,62; robustní charakteristika Bc = –0,03. Data mají normální rozdělení (test Shapiro-Wilk, p-hodnota = 0,94).
Závěry: Ani jeden z krabicových grafů nevyloučil žádnou hodnotu. Grafy zde fungovaly rovnocenně (tab. III).

4. Postup při aplikaci nových grafů

Doporučené kroky při aplikaci upravených krabicových grafů:

a) výpočet charakteristik šikmosti, nejlépe klasické i robustní podle (*4*) a (*5*); b) výpočet kvartilů q_1,q_2 a q_3; c) výpočet kvartilových rozpětí QR,LQR,UQR a poměrů LQR/UQR a UQR/LQR; d) výpočet hranic A a B podle vzorců (*7*), (*9*) a (*10*). Všechny výpočty lze provést s programem Microsoft Excel a postup lze vyzkoušet na úlohách 3.1a resp. 3.2a, kde jsou k tomuto účelu uvedené potřebné hodnoty kvartilů.

Tabulka III
Vyloučené hodnoty u jednotlivých grafů pro data 3.3

	3.3a	
Tukey	Kimber	RS
B: 3,41	B: 3,95	B: 4,76
Nad B: 3,77; 4,29; 4,39; 4,58	Nad B: 4,39; 4,58	Nad B: nic
	3.3b	
Tukey	Kimber	RS
(A; B): (16,53; 33,93)	(A; B): (16,75; 34,15)	(A; B): (16,96; 34,39)
Mimo (A;B): nic	Mimo (A;B): nic	Mimo (A;B): nic
	3.3c	
Tukey	Kimber	RS
(A; B): (20,35;33,55)	(A; B): (20,20; 33,40)	(A; B): (20,04; 33,26)
Mimo (A;B): nic	Mimo (A;B): nic	Mimo (A;B): nic

5. Závěry

Klasické krabicové grafy jsou poměrně známý prostředek k identifikaci odlehlých hodnot. Méně známá bude asi jejich poslední úprava z r. 2018, která směřuje k takovému výpočtu hranic grafu, aby byl dobře použitelný pro nesymetrická rozdělení. Cílem tohoto příspěvku bylo seznámit čtenáře s novými krabicovými grafy, vyzkoušet jejich funkčnost a ukázat možné použití. Aplikace je snadná i s programem Microsoft Excel, takže každý praktik může prezentované poznatky ihned použít. Výhodou těchto grafických metod je také skutečnost, že jsou neparametrické a lze je tedy použít např. bez častého předpokladu normality dat. To umožňuje jejich použití pro data vykazující kladnou nebo zápornou šikmost. Nejnovější verze krabicových grafů byly zkoumány v 10 000 simulacích na souborech od velikosti $n = 10$ (cit.[5], str. 351). Tento rozsah lze proto považovat za minimální velikost pro jejich použití. S rostoucím počtem dat se zvyšuje spolehlivost závěrů, podobně jako u jiných grafických či numerických metod.

V našich simulacích se ukázalo, že u symetrických rozdělení postačují klasické krabicové grafy. Naopak pro data s velkou šikmostí jsou lepší nové typy grafů. Pro menší soubory, např. o rozsahu $n = 3$ až $n = 10$ v analytické chemii, lze doporučit pro vyloučení odlehlých hodnot např. neparametrický Dean-Dixonův Q test [7].

Vypracováno s finanční podporou VŠB – Technická univerzita Ostrava jako projekt SP2020/61.

Seznam zkratek

QR kvartilové rozpětí
UQR horní kvartilové rozpětí
LQR dolní kvartilové rozpětí
Bc robustní koeficient šikmosti (Bowle)
K označení pro Kimberův krabicový graf
T označení pro Tukeyův krabicový graf
RS označení pro poměr šikmosti krabicového grafu (boxplot Ratio Skewnes)
s výběrová směrodatná odchylka

LITERATURA

1. Bowley A. L.: *Elements of Statistics.* 4.vyd. Charles Scribner´s Sons, New York 1920.
2. Spear M. E.: *Charting Statistics*. McGraw-Hill, New York 1952.
3. Tukey J. W., McGill R., Larsen W. A.: American Statistician *32*, 12 (1987).
4. Kimber A. C.: Applied Statistics *39*, 21 (1990).
5. Walker M. L., Dovoedo Y. H., Chakraborti S., Hilton C. W.: American Statistician *72*, 348 (2018).
6. Meloun M., Militký J.: *Kompendium statistického zpracování dat*. Academia, Praha 2002.
7. Dean R. B., Dixon W. J.: Anal. Chem. *23*, 636 (1951).

J. Tošenovský and F. Tošenovský *(Faculty of Materials Science and Technology, VSB-Technical University of Ostrava):* **Identification of Outliers in Chemical Measurements Allowing for Asymmetry in their Probability Distribution**

The aim of the paper is to acquaint the reader with new box plots, clarify the principles they are based on, try out their functionality and show their application in detecting outliers in data samples with positive or negative skewness.

Keywords: outliers, skewness, percentiles, boxplot