Postup statistického zpracování výsledků stopové analýzy při použití transformace dat

Milan Meloun
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice
milan.meloun@upce.cz
a
Jiří Militký
Katedra textilních materiálů, Technická univerzita Liberec, 461 17 Liberec
jiri.militky@vslib.cz

Summary: *Exploratory Data Analysis provides the first contact with the data and serves to uncover unexpected departures from familiar (Gaussian) models. When the data does not fulfil all assumption about the sample i. e. the sample distribution differs from the Gaussian, normal one the user is faced with the problem how to analyze the data. Power or Box-Cox transformation involves finding a scale that can clarify the analysis of the data or simplify the distribution of the data. It help to promote symmetry, constancy of variability, linearity, or additivity of effect, depending on the structure of the data. The proper transformation leads to symmetric distribution of data, stabilizes the variance, or makes the distribution closer to normal. The method and software are presented on the illustrative study case of a trace analysis of the determination of cobalt in potatoes.*

Souhrn: *Průzkumová analýza dat provádí první kontakt s daty a slouží k odhalení všech statistických zvláštností výběru, asymetrie rozdělení výběru a vybočujících hodnot. Když data nesplňují požadavky, kladené na výběr, nevykazují Gaussovo rozdělení a navíc obsahují vybočující hodnoty je uživatel vystaven problému jak vyčíslit odhad střední hodnoty. Mocninná a Box-Coxova transformace pak slouží k nalezení objektivního odhadu střední hodnoty. Zaručuje uživateli spolehlivý odhad střední hodnoty i v takovém případě, jako jsou data stopové analýzy, která mívají vždy silně sešikmené rozdělení. Navržená metoda s doprovodným software je dokumentována na úlohách vyčíslení bodového a intervalového odhadu střední hodnoty u stanovované stopového kadmia v bramborách. Oba datové výběry vykazují silně sešikmené, asymetrické rozdělení.*

ÚVOD

Účelem průzkumové (exploratorní) analýzy dat EDA výsledků analytické metody je odhalit statistické zvláštnosti a ověřit předpoklady o datech pro následné statistické zpracování. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí.

Před standardní statistickou analýzou je nezbytné vyšetřit platnost základních předpokladů, tj. nezávislost, homogenitu a normalitu prvků výběru. Reprezentativní náhodný výběr je popsán základními vlastnostmi: prvky výběru x_i jsou vzájemně nezávislé a dostatečné četnosti, výběr je homogenní a pochází z normálního rozdělení pravděpodobnosti, všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru. Vychází se z *pořádkových statistik*, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, a *pořadové pravděpodobnosti* $P_i = i/(n+1)$, pro kterou platí, že $100P_i$ procentní výběrový kvantil je hodnota, pod kterou leží $100P_i$ procent prvků výběru. Vynesením hodnot $x_{(i)}$ proti P_i, $i = 1, \ldots, n$, se získá hrubý odhad *kvantilové funkce Q(P).* Ta je inverzní k funkci distribuční a jednoznačně charakterizuje rozdělení výběru.

METODICKÁ ČÁST

1. Postup analýzy dat

Experimentální data se v analytické laboratoři často vyznačují asymetrickým rozdělením a porušením dalších předpokladů, kladených na výběr. Uveďme proto nejprve obecnou osnovu analýzy dat.

A. V průzkumové analýze dat se vyšetřují *statistické zvláštnosti,* jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost podezřelých hodnot. Odhalí se také anomálie a odchylky rozdělení výběru od typického rozdělení, obyčejně Gaussova. Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického software totiž nabízí řadu diagnostických grafů a diagramů. Pokud je rozdělení dat nevhodné pro standardní statistickou analýzu (tj. většinou asymetrické), provádí se často vhodná transformační úprava dat. Pokud bylo indikováno sešikmené rozdělení nebo rozdělení s dlouhými konci, vede často ke zlepšení mocninná a Boxova-Coxova transformace. Transformace je vhodná především při asymetrii rozdělení původních dat, resp. nekonstantnosti rozptylu.

B. Pro případ rutinních měření se ověří *základní předpoklady,* kladené na výběr, jako jsou nezávislost prvků, homogenita výběru, dostatečný rozsah výběru a rozdělení výběru. Jsou-li závěry tohoto kroku optimistické, následuje vyčíslení odhadů polohy a rozptýlení, tj. obyčejně aritmetického průměru a rozptylu. Dále se vyčíslí intervaly spolehlivosti následované testováním statistických hypotéz. V případě pesimistickém následuje pokus o úpravu dat.

C. V konfirmatorní analýze je nabízena paleta rozličných odhadů polohy, rozptýlení a tvaru. Základní jsou *klasické odhady* a *robustní odhady* (necitlivé na odlehlé prvky výběru, resp. další předpoklady o datech). Z dalších lze pak uvést např. adaptivní. Z nabídky odhadů parametrů vybírá uživatel ty, jež odpovídají závěrům průzkumové analýzy dat a ověření předpokladů o výběru.

A. Průzkumová (exploratorní) analýza dat (EDA)
- Odhalení stupně symetrie a špičatosti výběrového rozdělení;
- Indikace lokální koncentrace dat;
- Nalezení vybočujících a podezřelých prvků ve výběru;
- Porovnání výběrového rozdělení dat s typickými rozděleními;
- Mocninná transformace dat;
- Box-Coxova transformace dat.

B. Ověření předpokladů o datech:
- Ověření nezávislosti prvků dat;
- Ověření homogenity rozdělení dat;
- Určení minimálního rozsahu dat.
- Ověření normality rozdělení dat.

C. Konfirmatorní analýza dat (CDA) - odhady parametrů (polohy, rozptýlení a tvaru)
1. Klasické odhady (bodové a intervalové) parametrů;
2. Robustní odhady (bodové a intervalové) parametrů;

2. Transformace dat

Pokud se na základě analýzy reálného výběru zjistí, že rozdělení dat se příliš odlišuje od normálního, vzniká problém, jak data vůbec vyhodnotit. Pak již není módus totožný s mediánem ani střední hodnotou a vlastní interpretace parametru polohy je ztížena. Efektivní odhad parametru polohy je možný jen při znalosti rozdělení pravděpodobnosti. Běžné statistické testy předpokládají symetrické rozdělení dat. Běžné robustní metody odhadu parametrů polohy a rozptýlení zde nefungují dobře, protože opět předpokládají, že symetricky rozdělená data obsahují vybočující hodnoty. Je zřejmé, že již symetrizační transformace bude v analýze dat velmi užitečná. Část lze nalézt *vhodnou transformaci*, která vede ke

stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě. Vychází se z představy, že zpracovávaná data jsou nelineární transformací normálně rozdělené náhodné veličiny x a hledá se k nim inverzní transformace $g(x)$.

(a) Transformace stabilizující rozptyl: nekonstantnost rozptylu je původním jevem u řady měření v instrumentálních metodách. Indikuje buď neplatnost aditivního modelu měření $x_i = \mu + \varepsilon_i$, kde ε_i jsou náhodné chyby s nulovou střední hodnotou a konstantním rozptylem, nebo indikuje nenormalitu rozdělení výběru. Stabilizace rozptylu vyžaduje nalezení transformace $y = g(x)$, ve které je již rozptyl $\sigma^2(y)$ konstantní. Pokud je rozptyl původní proměnné x funkcí typu $\sigma^2(x) = f_1(x)$, lze rozptyl $\sigma^2(y)$ určit z Taylorova rozvoje funkce $g(x)$

$$\sigma^2(y) \approx \left[\frac{dg(x)}{dx}\right]^2 f_1(x) = C$$

kde C je konstanta. Hledaná transformace $g(x)$ je pak řešením diferenciální rovnice

$$g(x) \approx C \int \frac{dx}{\sqrt{f_1(x)}}$$

U řady instrumentálních metod je zajištěna konstantnost relativní chyby měření $\delta(x)$. To znamená, že rozptyl $\sigma^2(x)$ je dán funkcí $f_1(x) = \delta^2(x)\, x^2 = \text{konst}\, x^2$. Po dosazení vyjde $g(x) = \ln x$. Optimální je pro tento případ logaritmická transformace původních dat. Z toho vyplývá také vhodnost použití geometrického průměru. Pokud je závislost $\sigma^2(x) = f_1(x)$ mocninná, bude optimální transformace $g(x)$ také mocninná. Jelikož pro normální rozdělení je střední hodnota na rozptylu nezávislá, bude transformace stabilizující rozptyl také zajišťovat přiblížení k normalitě.

(b) Symetrizující tranformace: zesymetričtění rozdělení výběru se provede *jednoduchou mocninnou transformací*

$$y = g(x) = \begin{cases} x^{\lambda} & \lambda > 0 \\ \ln x & \text{pro} \quad \lambda = 0 \\ -x^{-\lambda} & \lambda < 0 \end{cases}$$

Tato transformace však nezachovává měřítko, není vzhledem k hodnotě λ všude spojitá, zachovává však pořadí dat ve výběru a hodí se pouze pro kladná data. Optimální odhad $\hat{\lambda}$ se hledá s ohledem na minimalizaci vhodných charakteristik asymetrie. Kromě šikmosti $\bar{g}_1(y)$ je možné užít i robustní verzi šikmosti definovanou výrazem

$$\hat{g}_{1R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{\tilde{y}_{0.75} - \tilde{y}_{0.25}} ,$$

kde y_P je P%ní kvantil transformovaného výběru. Stejně jednoduché je sledovat rozdíl mezi střední hodnotou y a mediánem $\tilde{y}_{0.5}$ pomocí statistiky šikmosti

$$g_P = \frac{\bar{y} - \tilde{y}_{0.5}}{\sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n - 1}}}$$

Pro symetrická rozdělení je statistika $\hat{g}_p(y)$ rovna nule. Stejně tak jsou rovny nule i statistiky $\hat{g}_1(y)$ a $\hat{g}_{1R}(y)$. Hodnotu $\hat{\lambda}$ lze hledat pomocí rankitového grafu, kde pro optimální $\bar{\lambda}$ budou kvantily $y_{(i)}$ ležet přibližně na přímce.

Hines - Hinesův selekční graf (osa x: $\bar{x}_{0.5}/x_{1-P_i}$, osa y: $\tilde{x}_{Pi}/\tilde{x}_{0.5}$): diagnostickou pomůckou pro odhad optimálního parametru λ je selekční graf dle Hinese a Hinesové, obr. 1.
Vychází z požadavků symetrie jednotlivých kvantilů kolem mediánu

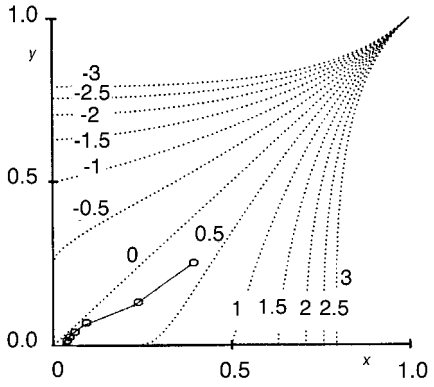
$$\left(\frac{\tilde{x}_{P_i}}{\tilde{x}_{0.5}}\right)^{\lambda} + \left(\frac{\tilde{x}_{0.5}}{\tilde{x}_{1-P_i}}\right)^{-\lambda} = 2$$

kde jako pořadové pravděpodobnosti jsou obvykle voleny hodnoty, $P_i = 2^{-i}$, $i = 2, 3$. K porovnání průběhu experimentálního bodů s ideálním (teoretickým) pro zvolené λ se do grafu zakreslují i řešení rovnice $y^{\lambda} + x^{-\lambda} = 2$ pro $0 \le x \le 1$ a $0 \le y \le 1$:

a) pro $\lambda = 0$ je řešením přímka $y = x$,

b) pro $\lambda < 0$ je řešením vztah $y = (2 - x^{-\lambda})^{1/\lambda}$,

c) pro $\lambda > 0$ je řešením vztah $x = (2 - y^{\lambda})^{-1/\lambda}$.



Obr. 1 Ukázka selekčního grafu pro výběr, vykazující téměř lognormální rozdělení (stopová analýza).

Podle umístění experimentálních bodů na teoretických křivkách selekčního grafu lze odhadovat velikost λ a posuzovat kvalitu transformace v různých vzdálenostech od mediánu.

(c) Normalizační transformace: pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti se užívá rodiny *Boxovy-Coxovy transformace*

$$y = g(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln x & (\lambda = 0) \end{cases}$$

Boxova-Coxova transformace má tyto vlastnosti:

1. Transformace $g(x)$ jsou vzhledem k veličině λ spojité, protože v okolí nuly platí

$$\lim_{\lambda \to 0} \frac{x^{\lambda} - 1}{\lambda} = \lim_{\lambda \to 0} x^{\lambda} \cdot \ln x = \ln x$$

2. Všechny transformace procházejí bodem $[y = 0; x = 1]$ a mají v tomto bodě společnou směrnici, jsou zde co do průběhu totožné.

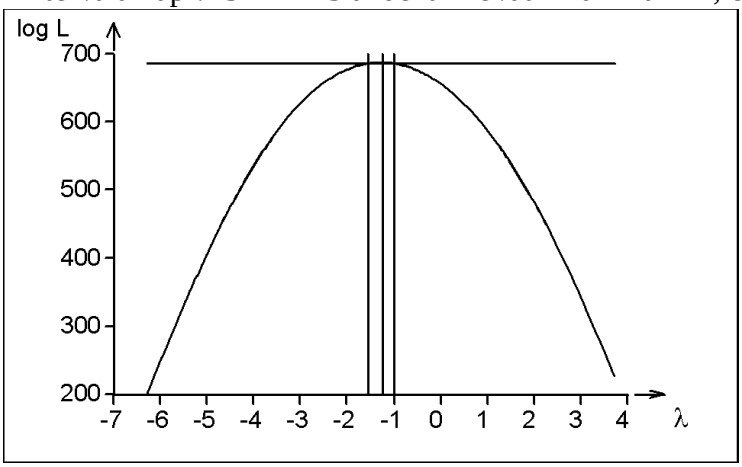
3. Mocninné transformace s exponenty -2; -3/2; -1; -1/2; 0; 1/2; 1; 3/2; 2 jsou co do křivosti rovnoměrně rozmístěné.

Boxova-Coxova transformace je použitelná pouze pro kladná data. Rozšíření této transformace na oblast, kdy rozdělení dat začíná od prahové hodnoty x_0, spočívá v náhradě x rozdílem $(x - x_0)$, který je vždy kladný.

Graf logaritmu věrohodnostní funkce (osa x: λ, osa y: ln L): pro odhad parametru λ v Boxově-Coxově transformaci lze užít metodu maximální věrohodnosti s tím, že pro $\lambda = \hat{\lambda}$ je rozdělení transformované veličiny y normální, $N(\mu_y, \sigma^2(y))$. Po úpravách bude logaritmus věrohodnostní funkce ve tvaru

$$\ln L(\lambda) \;=\; -\frac{n}{2}\,\ln s^2(y) \;+\; (\lambda - 1)\sum_{i=1}^{n}\ln x_i$$

kde $s^2(y)$ je výběrový rozptyl transformovaných dat y. Průběh věrohodnostní funkce $\ln L = f(\lambda)$ lze znázornit ve zvoleném intervalu např. $-3 \le \lambda \le 3$ a identifikovat i maximum $\hat{\lambda}$, obr. 2.



Obr. 2 Graf logaritmu věrohodnostní funkce pro výběr z lognormálního rozdělení (stopová analýza).

Pro asymptotický 100(1 - α)%ní interval spolehlivosti parametru λ platí

$$2\left[\ln L(\hat{\lambda}) \;-\; \ln L(\lambda)\right] \;\le\; \chi^2_{1-\alpha}(1)$$

kde $\chi^2_{1-\alpha}(1)$ je kvantil χ^2-rozdělení s jedním stupněm volnosti. V tomto intervalu spolehlivosti leží všechna λ, pro která je $\ln L(\lambda)$ větší nebo roven $\ln L(\hat{\lambda}) - 0.5\chi^2_{1-\alpha}(1)$. Výhodně lze do grafu logaritmu věrohodnostní funkce $\ln L(\lambda)$ na λ zakreslit obyčejně 95% interval spolehlivosti. Z tohoto grafu lze snadno odhadnout jak kvalitu transformace, odhad exponentu $\hat{\lambda}$, tak i posoudit, v jakých mezích se může hodnota λ pohybovat. Platí totiž, že čím je interval spolehlivosti exponentu $\hat{\lambda}$ tj. $\langle L_D, L_H \rangle$ širší, tím je transformace méně výhodná. Pokus tento interval obsahuje i hodnotu $\lambda = 1$, není transformace ze statistického hlediska přínosem.

3. Zpětná transformace

Pokud se podaří nalézt vhodnou transformaci, která vede k přibližné normalitě, lze určit $\bar{y}$, $s^2(y)$, interval spolehlivosti $\bar{y} \pm t_{1-\alpha/2}(n-1).\, s(y)/\sqrt{n}$ a provádět i statistické testování. Problém však spočívá v tom, že všechny statistické charakteristiky a jejich intervaly spolehlivosti je třeba určit pro původní proměnné.

1. **Nekorektní (naivní) přístup** spočívá v pouhé zpětné transformaci $\bar{x}_R = g^{-1}(\bar{y})$. Pro jednoduchou mocninnou transformaci vede zpětná transformace na obecný průměr definovaný vztahem

$$\bar{x}_R \;=\; \bar{x}_\lambda \;=\; \left[\frac{\sum_{i=1}^{n} x_i^{\lambda}}{n}\right]^{1/\lambda}$$

Pro $\lambda = 0$ se místo x^λ používá $\ln x$ a místo $x^{1/\lambda}$ pak e^x. Hodnota $\bar{x}_R = \bar{x}_{-1}$ představuje *harmonický průměr*, $\bar{x}_R = \bar{x}_0$ *geometrický průměr*, $\bar{x}_R = \bar{x}_1$ *aritmetický průměr* a $\bar{x}_R = \bar{x}_2$ *kvadratický průměr*. Tento způsob zpětné transformace nebere v úvahu variabilitu střední hodnoty.

2. **Správnější přístup** zpětné transformace vychází z Taylorova rozvoje funkce $y = g(x)$ v okolí $\bar{y}$. Pro retransformovaný průměr $\bar{x}_R$ lze pak odvodit přibližný vztah

$$\bar{x}_R \approx g^{-1}\left[\bar{y} - \frac{1}{2}\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)\right]$$

Pro rozptyl vyjde $$s^2(x_R) \approx \left(\frac{dg(x)}{dx}\right)^{-2} s^2(y) \ .$$

Zde jednotlivé derivace jsou vyčísleny v bodě $x = \bar{x}_R$. Pro 100(1 - α)%ní interval spolehlivosti střední hodnoty původního souboru dat *x* platí

$$I_D \leq \mu \leq I_H$$

kde

$$I_D = g^{-1}\left(\bar{y} + G - t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$I_H = g^{-1}\left(\bar{y} + G + t_{1-\alpha/2}(n-1)\frac{s(y)}{\sqrt{n}}\right)$$

$$G = -0.5\frac{d^2 g(x)}{dx^2}\left(\frac{dg(x)}{dx}\right)^{-2} s^2(y)$$

Symbolem $t_{1-\alpha/2}(n-1)$ je označen 100(1 - α/2)%ní kvantil Studentova rozdělení s (n - 1) stupni volnosti. Při znalosti hodnot konkrétní transformace $y = g(x)$ a odhadů $\bar{y}$, $s^2(y)$ je snadné vyčíslit hodnoty $\bar{x}_R$ a $s^2(x_R)$:

a) Pro speciální případ λ = 0, tzn. logaritmickou transformaci typu $g(x) = \ln x$, bude

$$\bar{x}_R \approx \exp\left[\bar{y} + 0.5\, s^2(y)\right]$$

Rozptyl se určí vztahem $$s^2(x_R) \approx \bar{x}_R^2\, s^2(y).$$

b) Pro případ λ ≠ 0 a Boxovy-Coxovy transformace bude $\bar{x}_R$ jedním z kořenů kvadratické rovnice, pro které platí

$$\bar{x}_{R,1,2} = \left[0.5(1 + \lambda\bar{y}) \pm 0.5\sqrt{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2\, s^2(y))}\right]^{1/\lambda}$$

Jako odhad x_R se pak bere kořen $\bar{x}_{R,i}$, který je nejbližší mediánu $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. Při znalosti retransformovaného průměru $\bar{x}_R$ lze z vyčíslit i odpovídající rozptyl

$$s^2(x) = \bar{x}_R^{-2\lambda+2}\ s^2(y) \ .$$

4. Ilustrační úloha

Úloha I. *Odhad střední hodnoty obsahu kadmia v bramborách*
Je třeba určit spolehlivý odhad střední hodnoty obsahu kadmia v bramborách v oblastech jižní Moravy a jižních Čech. Předem je třeba vyšetřit rozdělení a určit počet odlehlých hodnot výběru o celkové četnosti n = 28. K vyhodnocení je třeba užít průzkumovou analýzu dat, ověření předpokladů o náhodném výběru, event. transformaci dat.
Data: z úlohy E219, ref. [2]: obsah kadmia 100.x v bramborách [mg/kg]

8.0	1.8	1.5	5.2	8.0	2.0	3.6	1.3	3.2	6.0	7.0	1.9	1.5	5.5
1.5	1.0	1.0	3.3	3.2	1.1	11.5	3.3	6.0	2.0	0.9	0.5	2.7	5.5

Řešení:

(1) Přehled popisných statistik: software NCSS2000 vyčíslil parametry polohy, rozptýlení a tvaru, z nichž nejdůležitější jsou zde uvedeny. Všechny odhady je třeba vynásobit 10^{-2} [mg/kg].

Tabulka 1. Přehled odhadů parametrů polohy a rozptýlení (NCSS2000 a ADSTAT): odhady je třeba vynásobit 10^{-2} [mg/kg].

Střední hodnota	Bodový odhad	Dolní mez	Horní mez	Užito
Aritmetický průměr	3.571	2.512	4.631	28
Geometrický průměr	2.677	---	---	28
Harmonický průměr	1.983	---	---	28
Medián	2.95	1.433	4.467	28
Módus	1.20	---	---	28
5%ní uřezaný průměr	3.351	2.344	4.357	25
(jeho winsorizovaný odhad)	3.461	---	---	25
10%ní uřezaný průměr	3.210	2.117	4.303	22
(jeho winsorizovaný odhad)	3.468	---	---	22
40%ní uřezaný průměr	2.739	1.247	4.232	6
(jeho winsorizovaný odhad)	2.668	---	---	6
M-odhad	3.264	2.210	4.317	28
Hoggův *M*-odhad	3.571	2.512	4.631	28
Směrodatná odchylka	2.733	2.161	3.72	28
Rozpětí	11.0	---	---	28
Interkvartilové rozpětí	4.0	---	---	28

Z těchto odhadů si má uživatel vybrat správný. Pro $n = 28$ bylo vyčísleno minimum 0.50 a maximum 11.0 a dále z parametrů polohy aritmetický průměr $\bar{x} = 3.57$ s 95%ním intervalovým odhadem $L_L = 2.51$ a $L_U = 4.63$, medián $\hat{x}_{0.5} = 2.95$ s 95%ním intervalovým odhadem $L_L = 1.43$ a $L_U = 4.47$.

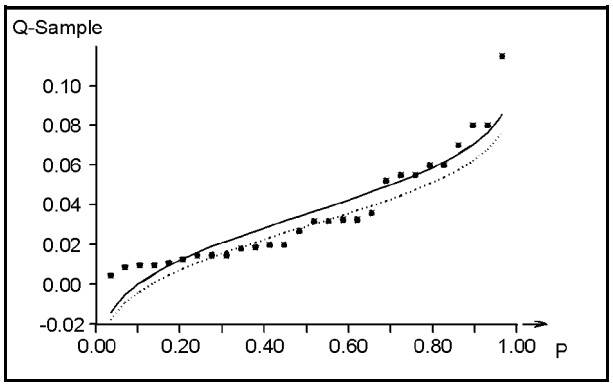
Dále je geometrický průměr $x_g = 2.68$, harmonický průměr $x_h = 1.98$, modus $x_M = 1.20$, a následující uřezané průměry (v závorce je u každé hodnoty uvedena winsorizovaná hodnota) $\bar{x}(5\%) = 3.35$ (3.46) s $s(5\%) = 2.67$ (2.45) a pro $n(5\%) = 25$, $\bar{x}(10\%) = 3.21$ (3.47) s $s(10\%) = 2.81$ (2.54) a pro $n(10\%) = 22$, $\bar{x}(40\%) = 2.74$ (2.67) s $s(40\%) = 3.44$ (1.36) pro $n(10\%) = 6$.

Robustní *M*-odhad polohy je $\hat{\mu}_M = 3.26$ a rozptýlení $\sigma_M = 2.60$ s intervalovým odhadem $L_L = 2.21$ a $L_U = 4.32$ a dále robustní Hoogův *M*-odhad polohy je $\hat{\mu}_M = 3.57$ a rozptýlení $\sigma_M = 2.73$ s intervalovým odhadem $L_L = 2.51$ a $L_U = 4.63$.

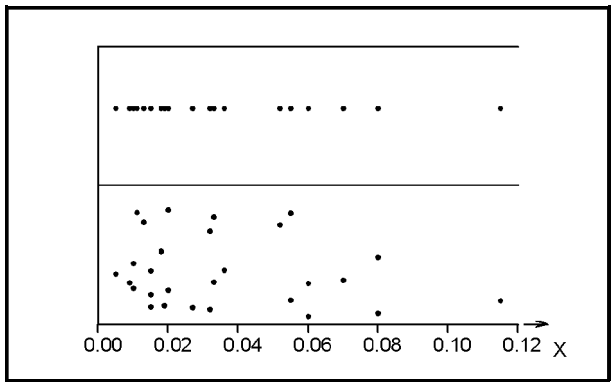
Z parametrů rozptýlení jsou to směrodatná odchylka $s = 2.73$, rozpětí $R = 11.0$, interkvartilové rozpětí $R_F = 4.0$ a z parametrů tvaru je to šikmost $g_1 = 1.12$ (test ukazuje, že odchylka od 0 je statisticky významná a jde o nenormální rozdělení) a špičatost $g_2 = 3.67$.

(2) Základní diagnostické grafy EDA jsou užity ke grafickému znázornění datového výběru: *kvantilový graf* (obr. 3) vykazuje řadu odlehlých hodnot a asymetrické rozdělení, klasická a empirická křivka se totiž od sebe výrazně liší. Oba *diagramy rozptýlení* (obr. 4) indikují asymetrické rozdělení a řadu odlehlých hodnot jak v horní části, tak i v dolní části výběru. *Graf polosum* (obr. 5) a *graf symetrie* (obr. 6) vykazují asymetrické rozdělení, protože značné množství bodů leží vně konfidenčního intervalu

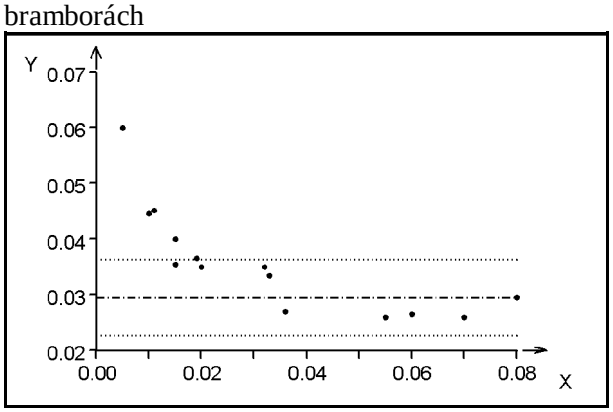
mediánové přímky. *Graf rozptýlení s kvantily* (obr. 7) ukazuje na řadu odlehlých bodů, které leží vně sedecilového obdélníku. Poloha mediánu *M* je vyznačena krátkou mediánovou úsečkou ve střední části kvartilového grafu pro $P_1 = 0.5$. V *kruhovém grafu* (obr. 8) se liší obě kruhové křivky, teoretická elipsa pro normální rozdělení a empirická zborcená elipsa pro výběrové rozdělení.



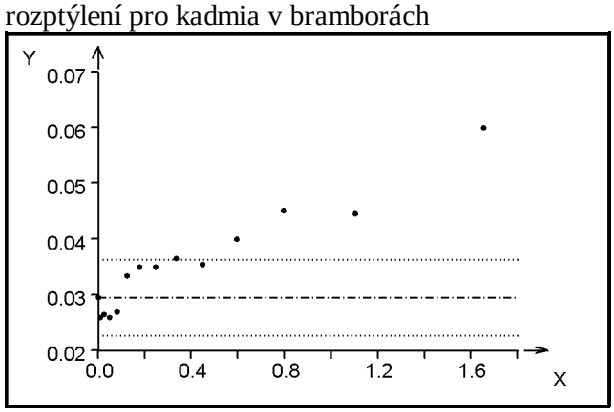
Obr. 3 Kvantilový graf pro obsah kadmia v bramborách



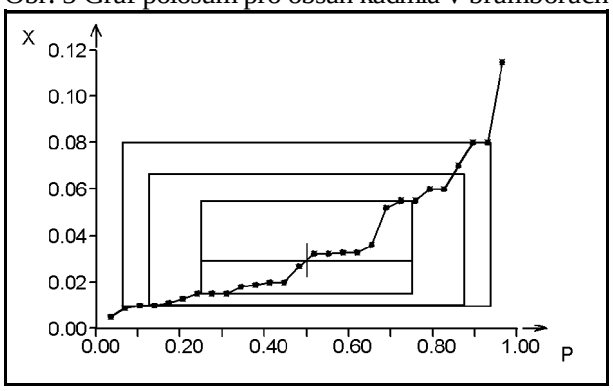
Obr. 4 Diagram rozptýlení a rozmítnutý diagram rozptýlení pro kadmia v bramborách



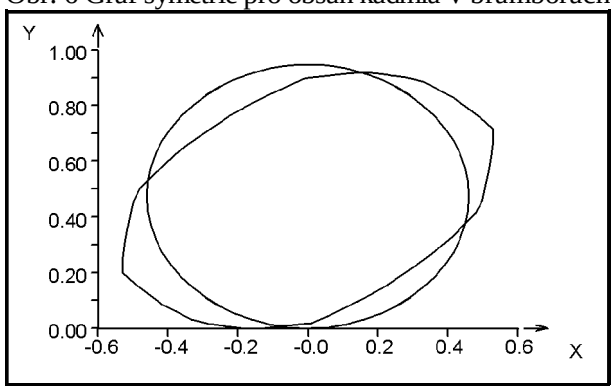
Obr. 5 Graf polosum pro obsah kadmia v bramborách



Obr. 6 Graf symetrie pro obsah kadmia v bramborách

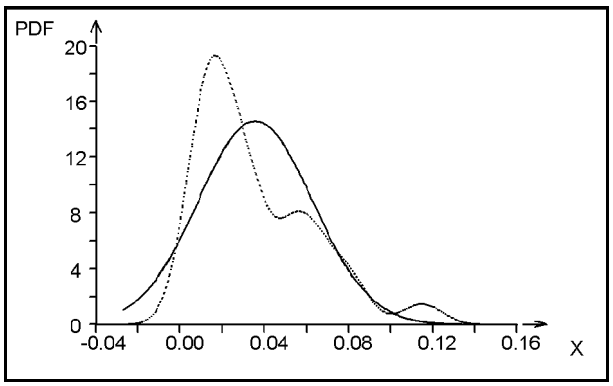


Obr. 7 Graf rozptýlení s kvantily pro obsah kadmia v bramborách

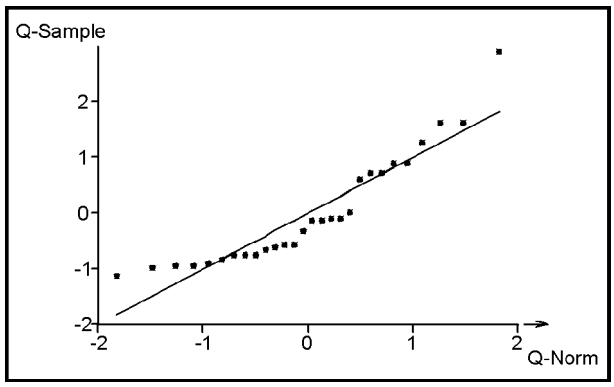


Obr. 8 Kruhový graf pro obsah kadmia v bramborách

(3) Určení výběrového rozdělení (EDA): výběrové rozdělení je definováno svou symetrií, šikmostí a špičatostí a lze ho indikovat pomocí čtyř grafů: *Jádrový odhad hustoty pravděpodobnosti* (obr. 9) ukazuje nenormální rozdělení, protože obě křivky, teoretická aproximující normální rozdělení a empirická pro výběrové rozdělení, se významně odlišují. V *rankitovém Q-Q grafu* (obr. 10) většina bodů neleží na přímce normálního rozdělení, což je důkaz, že výběrové rozdělení není normálního charakteru. Korelační koeficient *Q-Q* grafu $r_{xy} = 0.9879$ ukazuje na log.-normální nebo sešikmené exponenciální rozdělení.



Obr. 9 Jádrový odhad hustoty pravděpodobnosti pro obsah kadmia v bramborách



Obr. 10 Rankitový Q-Q graf pro obsah kadmia v bramborách

Tabulka 2. Kvantilové míry polohy, rozptýlení a tvaru pro obsah kadmia v bramborách [mg/kg]

Kvantil	P	Dolní kvantil Q_D	Horní kvantil Q_H	Rozsah R_Q	Polosuma Z_Q	Šikmost S_Q	Délka konců T_Q
Median	0.5	2.95	2.95	-			
Kvartil	0.25	1.50	5.50	4.00	3.50	-0.004	0.000
Oktil	0.125	1.04	6.63	5.59	3.83	0.093	0.334
Sedecil	0.0625	0.97	8.00	7.03	4.48	0.093	0.564

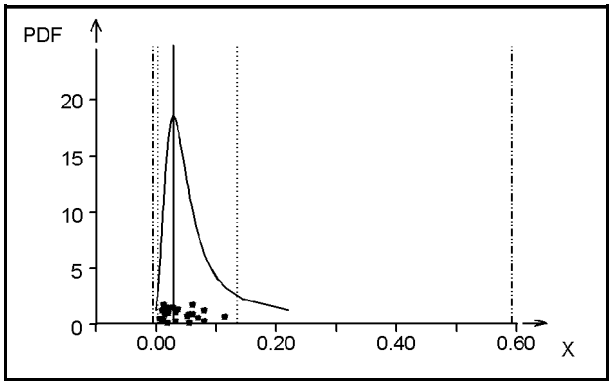
Délka oktilových konců $T_E = 0.334$ se liší od tabulované hodnoty pro normální rozdělení $T_E = 0.534$ a také sedecilových konců $T_D = 0.564$ se liší od tabulované hodnoty pro normální rozdělení $T_D = 0.822$. Bodový odhad šikmosti $g_1 = 1.12$ a bodový odhad špičatosti $g_2 = 3.67$ ukazují, že výběrové rozdělení je sešikmené a nedá se aproximovat normálním.

***(4) Ověření základních předpokladů o reprezentativním náhodném výběru*:** vyšetřením základních předpokladů, kladených na reprezentativní, náhodný výběr bylo dosaženo těchto závěrů:

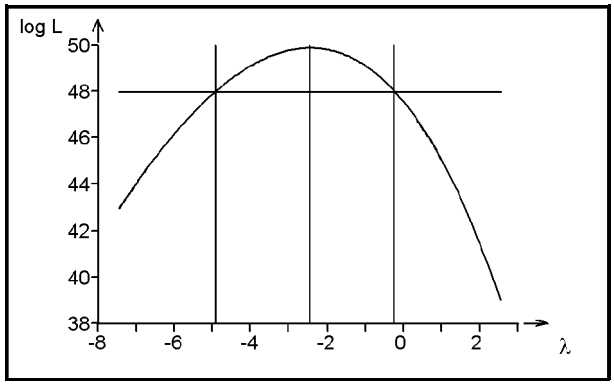
(a) *Vyšetření nezávislosti prvků výběru*: von Neumannův test nezávislosti prvků ve výběru dospěl k hodnotě testačního kritéria $t_{17} = 0.083 < t_{0.975}(28+1) = 2.045$, a proto je nezávislost přijata.

(b) *Vyšetření normality výběrového rozdělení:* Jarque-Berrův test kombinované šikmosti a špičatosti vede k testační statistice $C_1 = 8.737 > \chi^2(0.95, 2) = 5.992$, což dokazuje, že předpoklad normality je zamítnut.

(c) *Vyšetření homogenity výběru:* vně intervalu Hoaglinových mezí $[B^*_L = -6.98; B^*_U = 13.99]$ nejsou žádné odlehlé hodnoty.

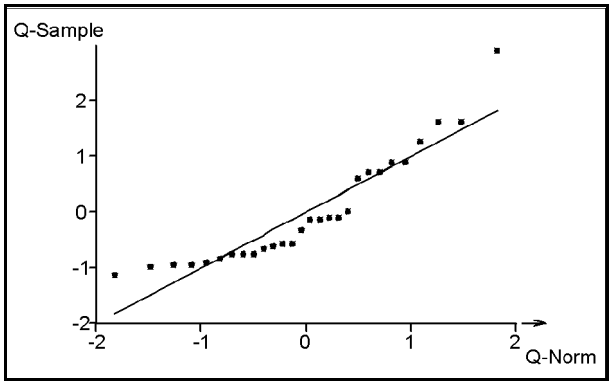


Obr. 11 Jádrový odhad hustoty pravděpodobnosti pro obsah kadmia v bramborách



Obr. 12 Graf logaritmu maximální věrohodnosti pro obsah kadmia v bramborách po Box-Coxově transformaci

(5) Transformace dat: asymetrické rozdělení výběru původních dat vyžaduje transformaci dat. Z grafu logaritmu maximální věrohodnosti plyne, že Box-Coxova transformace je statisticky významná, protože pod segmentem v tomto grafu neleží hodnota +1.



Obr. 13 Rankitový Q-Q graf pro obsah kadmia v bramborách po Box-Cox transformaci

Klasický odhad parametru polohy pro původní data aritmetický průměr $\bar{x} = 3.57$ je nepoužitelný, protože není splněn předpoklad symetrického a normálního rozdělení. Symetrizující mocninná transformace (ADSTAT 1.25, $\hat{\lambda} = 0.13$ čili číslo blízké nule indikující tak log.-normální rozdělení) vede na opravený průměr $\bar{x}_R = 2.79$ s intervalem spolehlivosti $L_D = 2.04$ a $L_H = 3.77$ a normalizační Box-Coxova transformace (ADSTAT 1.25, $\lambda = 0.13$) vede na stejný opravený průměr $\bar{x}_R = 2.79$ se stejným intervalem spolehlivosti jako mocninná transformace.

(6) ***Závěr úlohy*****:** diagnostiky průzkumové analýzy dat vedou k závěru, že 28 hodnot původních dat vykazuje asymetrické, silně sešikmené rozdělení. Nelze proto použít klasické odhady parametrů polohy a rozptýlení, platící pouze pro symetrické rozdělení a data je třeba nejprve transformovat mocninnou nebo Box-Coxovou transformací. Re-transformovaný průměr pak představuje odhad parametru polohy $\bar{x}_R = 2.79 \times 10^{-2}$ [mg/kg] s intervalem spolehlivosti $L_D = 2.04 \times 10^{-2}$ [mg/kg] a $L_H = 3.77 \times 10^{-2}$ [mg/kg]. Rozdělení dat lze považovat za logaritmicko-normální. Podobné výsledky o střední hodnotě přináší 40%ní uřezaný průměr.

ZÁVĚRY

Symetrizující mocninná transformace a normalizující Boxova-Coxova transformace dat slouží k určení parametrů polohy pro případ nesymetrického rozdělení dat. Vlastní výpočet má postup:

1. Pro mocninnou transformaci se počítají různé míry symetrie a výběrová špičatost v rozmezí $-3 \le \lambda \le 3$ s krokem 0.1. Je možno kreslit Hinesův-Hinesové selekční graf k určení optimální hodnoty λ. Na základě těchto informací se zadává zvolená hodnota λ. V této transformaci se pak vyčíslí $\bar{y}$, $s^2(y)$, šikmost $g_1(y)$ a špičatost $g_2(y)$.

2. Pro Box-Coxovu transformaci se počítá $\ln L(\lambda)$, různé míry symetrie a výběrová špičatost v rozmezí $-3 \le \lambda \le 3$ s krokem 0.1. Jsou tištěny optimální hodnoty těchto měr. Je kreslen graf závislosti $\ln L$ na λ spolu s 95%ním intervalem spolehlivosti a na základě těchto informací se zvolí hodnota λ. V této transformaci se počítá $\bar{y}$, $s^2(y)$, šikmost $g_1(y)$ a špičatost $g_2(y)$. Jsou určeny i retransformované hodnoty $\bar{x}_R$ a 95%ní interval spolehlivosti pro retransformovanou střední hodnotu μ.

Poděkování

Autoři děkují za finační podporu Grantové agentury ČR, č. 303/00/1559.

Literatura:

[1] Meloun M., Militký J.: *Statistické zpracování experimentálních dat*, PLUS Praha 1994, ISBN 80-85297-56-6.

[2] Meloun M., Militký J.: *Statistické zpracování experimentálních dat - Sbírka úloh s disketou*, Univerzita Pardubice 1997, ISBN 80-7194-075-5.

[3] Kupka K.: *Statistické řízení jakosti*, Trilobyte Pardubice 1998, ISBN 80-238-1818-X.

[4] Militký J.: *Moderní statistické metody pro životní prostředí*, PHARE, Svazek 15, Vysoká škola báňská, Ostrava 1996, ISBN 80-7078-360-5.