vodní hospodářství®

www.vodnihospodarstvi.cz

ročník 56
9
2006

Příští generace:

FDO 700 IQ

optické měření rozpuštěného kyslíku s IQ

FDO

IQ SENSOR NET

O_2

- Maximální přesnost - a to hned od začátku
- Nejvyšší stabilita
- Nejmenší nároky na údržbu

WTW

25 YEARS
EWA
1981 - 2006

MĚSTSKÉ VODY

MĚSTSKÉ VODY 2006
Břeclav, 5. - 6. října
Další info na adrese:
www.ardec.cz
info@ardec.cz
Poslední možnost!

VRV
CVTVHS
POVODÍ LABE

Organizátoři si Vás dovolují pozvat na konferenci
VODNÍ TOKY 2006
Přihláška s podrobnostmi je vložena v časopise

PŘÍLOHA
· ČL ·

Statistické zpracování vodohospodářských dat

1. Interaktivní analýza jednorozměrných dat

Milan Meloun

Klíčová slova

exploratorní analýza - mocninná transformace - Box-Coxova transformace - test správnosti - interval spolehlivosti - střední hodnota - průměr - směrodatná odchylka - medián - kvantily - písmenové hodnoty - kvantilový graf - diagram rozptýlení - krabicový graf - graf polosum - graf symetrie - graf rozptýlení s kvantily - jádrový odhad hustoty pravděpodobnosti - histogram - kvantil-kvantilový graf - Hornův postup pivotů

Souhrn

Při posouzení správnosti naměřených výsledků tvoří exploratorní analýza dat důležitou pomůcku, která využívá kvantilových diagnostik ke sledování stupně symetrie a špičatosti rozdělení výběru, lokální koncentrace dat a přítomnosti vybočujících hodnot. Mezi nejdůležitější patří vedle kvantilového grafu a grafu rozptýlení s kvantily také krabicový graf, vrubový krabicový graf, graf polosum a graf symetrie, kvantil-kvantilový graf, jádrový odhad hustoty pravděpodobnosti a histogram. Intervalový odhad míry polohy a Studentův t-test správnosti představují diagnostiky k testování správnosti n alezené střední hodnoty. U malých výběrů 4 ≤ n ≤ 20 se doporučuje Hornův postup pivotů, který je také vhodný pro svou dostatečnou robustnost vůči asymetrii rozdělení a vůči vybočujícím hodnotám.

◆

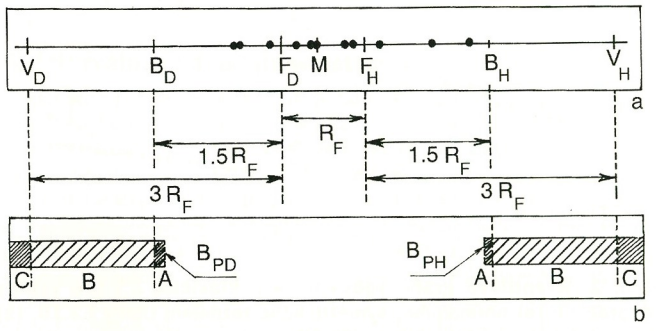
1. Úvod

Otázka spolehlivosti a správného vyhodnocení experimentálních dat se v době osobních počítačů ocitá u každého měření dat na prvním místě. V kontrolní laboratoři, ať už vodohospodářské, chemické, biologické, fyzikální či jakékoliv jiné, tvoří základ experimentální práce měření na přístroji. V laboratořích dnes představují instrumentální metody spojovací článek mezi přírodovědnými a technickými obory, protože moderní počítačem řízené přístroje používá každá laboratoř. Navíc na každém psacím stole laboratoře nacházíme počítač, většinou nejvyšší kvality, kapacity a rychlosti, vybavený moderním software. Je proto neomluvitelné vyhodnocovat naměřená data zjednodušenými, aproximativními postupy pozůstalými z dob kalkulaček. Kontrolní orgány, komisaři akreditačních komisí ale především konkurenční pracoviště v zahraničí se předhánějí při vyhodnocování dat v užívání špičkového software s rigorózními matematickými postupy, ve kterých není žádného zjednodušení či zanedbání nějakých statistických předpokladů. Výsledky dosažené těmito náročnějšími postupy se pak berou za validní a jedině správné a přijatelné třeba v okružním testu. Ukažme si zde proto jeden z novějších postupů interaktivní statistické analýzy dat, který je založen na diagnostikování v dialogu s osobním počítačem čili na interaktivní analýze a který nabízí uživateli hlubší pohled do všech tajemství, ukrytých v experimentálních datech. S tímto problémem souvisí obvykle i vhodný software, který zajistí bezproblémové a přátelské prostředí a "nechá data promluvit". Nezapomeňme přitom na důležité pravidlo, že "úroveň užívaného software dnes prozrazuje úroveň celého pracoviště".

2. Postup interaktivní analýzy dat

Obecný postup náročnější statistické analýzy jednorozměrných dat lze vyjádřit následujícím schématem. Interaktivní přístup uvedený postup ulehčuje, protože většina statistického software obsahuje uvedené statistické diagnostiky a testy.

1. Průzkumová (exploratorní) analýza dat (EDA) vyšetřuje především stupeň symetrie a špičatosti rozdělení, lokální koncentraci dat a odhaluje také vybočující a podezřelá data.
2. Ověření základních předpokladů o výběru dat se týká ověření normality, ověření nezávislosti, ověření homogenity a konečně i určení minimální četnosti analyzovaných dat.
3. Transformace dat následuje v případě porušení některého z předpokladů o výběru. Patří sem mocninná, exponenciální transformace a Boxova-Coxova transformace.
4. Vyčíslení nejlepších odhadů parametrů polohy, rozptýlení a tvaru se týká vyčíslení jednak klasických odhadů (aritmetický průměr a rozptyl), jednak robustních odhadů (medián, uřezané průměry, winsorizovaný rozptyl) a konečně i adaptivních M-odhadů. Retransformovaný průměr po transformaci dat se přesto obvykle jeví jako nejlepší odhad střední hodnoty.



Obr. 1. Konstrukce barierově-číslicového schématu indikujícího vybočující hodnoty: a) diagram rozptýlení s mediánem *M*, kvartily F_D (dolní) a F_H (horní), vnitřní hradby B_D (dolní) a B_H (horní), vnější hradby V_D (dolní) a V_H (horní); b) oblast vybočujících hodnot: A přilehlé (B_{PD} je blízké B_D a B_{PH} je blízké B_H), B značí oblast vnějších a C vzdálených bodů.

3. Exploratorní diagnostiky v analýze jednorozměrných dat

Prvním krokem v analýze jednorozměrných dat je průzkumová, exploratorní analýza. Jejím cílem je odhalit statistické zvláštnosti v datech a ověřit předpoklady o výběru pro následné rigorózní statistické zpracování. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí. Z různých typů výběru se v laboratoři nejvíce uplatňuje **reprezentativní náhodný výběr**, $\{x_i\}$, $i = 1, \ldots, n$, který má čtyři základní vlastnosti: (1) Jednotlivé prvky výběru x_i jsou vzájemně nezávislé. (2) Výběr je homogenní, tj. všechna x_i pocházejí ze stejného rozdělení pravděpodobnosti s konstantním rozptylem. (3) Předpokládá se také, že jde o normální rozdělení pravděpodobnosti. (4) Všechny prvky souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.

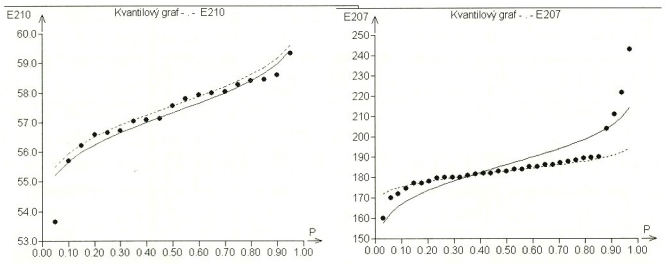
Před vlastní analýzou je vždy nezbytné ověřit platnost základních předpokladů, tj. nezávislost, homogenitu a normalitu výběru **Obr. 1.** Využívá se k tomu robustních kvantilových charakteristik, které umožňují sledování lokálního chování dat a které jsou vhodné pro malé nebo středně velké výběry. Vychází se z **pořádkových statistik** výběru $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Platí, že střední hodnota i-té pořádkové statistiky je rovna $100P_i$ procentnímu kvantilu výběrového rozdělení $F^1(P_i) = Q(P_i)$, kde $F(x)$ označuje distribuční funkci a $Q(P_i)$ kvantilovou funkci výběru. Symbol $P_i = i/(n + 1)$ označuje **pořadovou pravděpodobnost**. Připomeňme, že $100P_i$ procentní výběrový kvantil je hodnota, pod kterou leží $100P_i$ procent prvků výběru. Optimální hodnoty P_i závisí na předpokládaném rozdělení výběru. Pro normální rozdělení se doporučuje volba $P_i = (i - 3/8)/(n + 1/4)$. 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. V průzkumové analýze se často používá speciálních **kvantilů** L pro pořadové pravděpodobnosti $P_i = 2^{-i}$, $i = 1, 2, \ldots$, které se také nazývají **písmenové hodnoty**, viz. **Tabulka 1.**

Symbol u_{P_i} označuje kvantil normovaného normálního rozdělení $N(0, 1)$. Kromě **mediánu** ($i = 1$) existují pro každé $i > 1$ dvojice kvantilů, a to dolní a horní písmenová hodnota L_D a L_H. Dolní písmenová hodnota je pro pořadovou pravděpodobnost $P_i = 2^{-i}$, zatímco horní je pro $P_i = 1 - 2^{-i}$. Počet písmenových hodnot závisí na rozsahu výběru. Pro velikost výběru n lze určit n_L písmenových hodnot včetně mediánu dle vztahu $n_L = 1.44 \ln(n + 1)$.

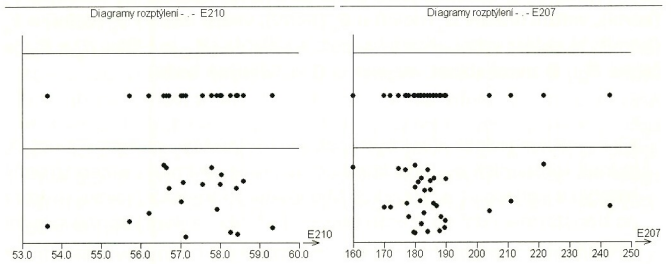
Kvantilový graf Obr. 2a, 2b (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika $x_{(i)}$) umožňuje přehledně znázornit data a snadněji rozlišit tvar rozdělení, který může být symetrický, sešikmený k vyšším nebo nižším hodnotám. Ke snadnějšímu porovnání s normálním rozdělením se do tohoto grafu zakreslují i kvantilové funkce normálního rozdělení $N_{P_i} = \hat{\mu} + \hat{\sigma} u_{P_i}$, pro $0 \le P_i \le 1$, a to: (1) klasických odhadů parametrů polohy a rozptýlení $\hat{\mu} = \bar{x}$ a $\hat{\sigma} = s$, a (2) robustních odhadů $\hat{\mu} = \tilde{x}_{0.5}$ a $\hat{\sigma} = R_F/1.349$.

Tabulka 1. Označení písmenových hodnot

i	i-tý kvantil	Pořadová pravděpodobnost P_i	Symbol písmenové hodnoty L	Hodnota kvantilu u_{Pi}
1	Medián	$2^{-1} = 1/2$	M	0
2	Kvantity	$2^{-2} = 1/4$	F	-0.674
3	Oktily	$2^{-3} = 1/8$	E	-1.15
4	Sedecily	$2^{-4} = 1/16$	D	-1.53



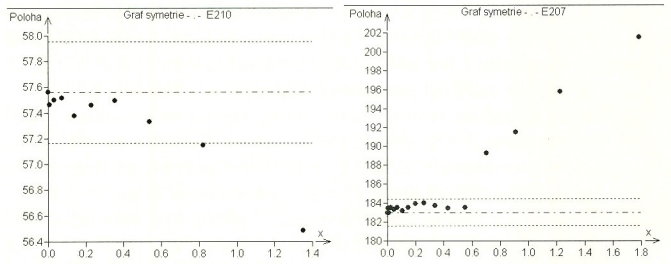
Obr. 2. Kvantilové grafy (robustní --- a klasické ...) pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.



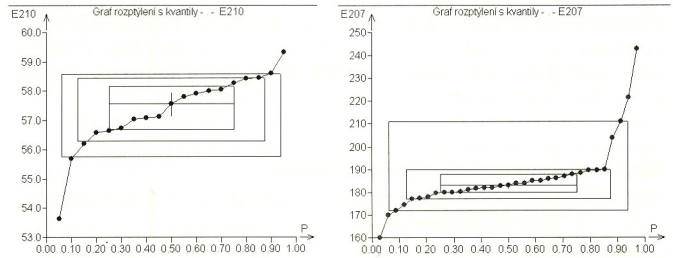
Obr. 3. Konstrukce (a) diagramu rozptýlení a (b) rozmítnutého diagramu rozptýlení pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.

Diagram rozptýlení Obr. 3a, 3b (osa *x*: hodnoty *x*, osa *y*: libovolná úroveň, obyčejně *y* = 0) představuje jednorozměrnou projekci kvantilového grafu do osy *x*, zatímco **rozmítnutý diagram rozptýlení** představuje týž graf, ale body jsou vhodně rozmítnuté ve směru *y*-nové osy. I při své jednoduchosti tento diagram názorně ukazuje na lokální koncentraci dat a indikuje i podezřelá a vybočující měření.

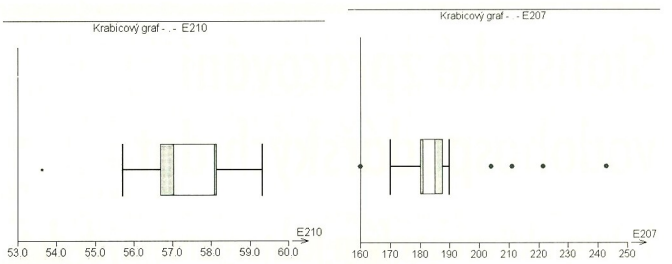
Krabicový graf Obr. 4a, 4b (osa *x*: úměrná hodnotám *x*, osa *y*: libovolný interval) umožňuje vedle znázornění robustního odhadu polohy, mediánu *M* také posouzení symetrie v okolí kvantilů a posouzení symetrie u konců rozdělení a často i identifikaci odlehlých dat. Jde o obdélník délky $R_F = F_H - F_D = x_{0.75}^{\%} - x_{0.25}^{\%}$ s vhodně zvolenou šířkou, která je úměrná hodnotě $\sqrt{n}$. V místě mediánu je vertikální čára. Od obou protilehlých stran tohoto obdélníku pokračují úsečky. Ty jsou ukončeny *přilehlými hodnotami* B_{PH} a B_{PD}, ležícími uvnitř *vnitřních hradeb* nejblíže k jejich hranicím BH, BD, tj. $B_H = F_H + 1.5R_F$ a $B_D = F_D - 1.5R_F$. Pro data pocházející z normálního rozdělení platí $B_H - B_D = 4.2$. Prvky výběru mimo vnitřní hradby jsou považovány za podezřelá měření (kroužky). Obdobou je **vrubový krabicový graf**, který umožňuje i posouzení variability mediánu, vyjádřenou robustním intervalem spolehlivosti $I_D \le M \le I_H$.



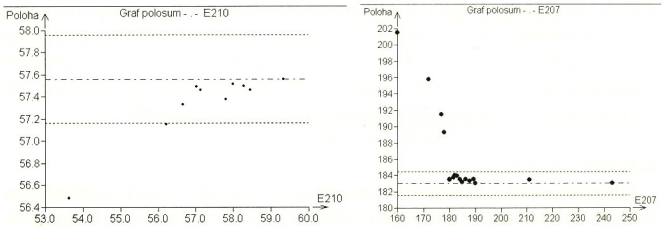
Obr. 6. Grafy symetrie pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.



Obr. 7. Konstrukce grafu rozptýlení s kvantily pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.



Obr. 4. Konstrukce (a) krabicového grafu, a (b) vrubového krabicového grafu z dat diagramu rozptýlení pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07. Prázdná kolečka indikují vybočující hodnoty.



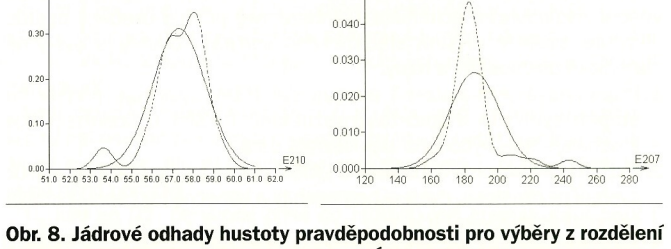
Obr. 5. Grafy polosum pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07.

Graf polosum Obr. 5a, 5b (osa *x*: pořádkové statistiky $x_{(i)}$, osa *y*: $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$ diagnostikuje tak, že pro symetrické rozdělení je grafem horizontální přímka, určená rovnicí $\tilde{x}_{0.5} = M$.

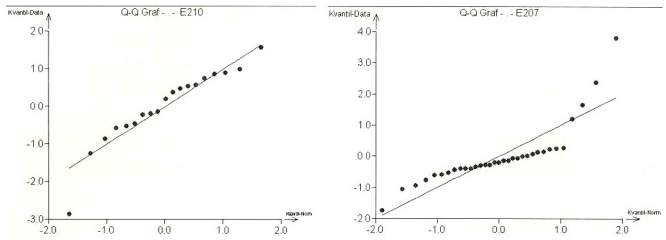
Graf symetrie Obr. 6a, 6b (osa *x*: $u_{P_i}^2/2$ pro $P_i = i/(n+1)$, osa *y*: $Z_i = 0.5(x_{(n+1-i)} + x_{(i)})$ je obdobou předešlého grafu, u kterého symetrická rozdělení vykazují horizontální přímku $y = \tilde{x}_{0.5} = M$. Pokud tato přímka nemá nulovou směrnici, je směrnice odhadem parametru šikmosti, asymetrie.

Graf rozptýlení s kvantity Obr. 7a, 7b (osa x: P_i, osa y: $x_{(i)}$) představuje vlastně kvantilový graf, který se získá spojením bodů $\{x_{(i)}, P_i\}$ lineárními úseky a pro symetrická rozdělení nabývá tato kvantilová funkce sigmoidálního tvaru. Pro rozdělení sešikmená k vyšším hodnotám je konvexně rostoucí a pro rozdělení sešikmená k nižším hodnotám konkávně rostoucí. Do kvantilového grafu se zakreslují tři obdélníky *F*, *E* a *D*: (1) *Kvartilový obdélník F:* na ose *x* pravděpodobnosti $P_2 = 2^{-2} = 0.25$ a $1 - 2^{-2} = 0.75$. (2) *Oktilový obdélník E:* na *y* oktily E_D a E_H a na ose *x* $P_3 = 2^{-3} = 0.125$ a $1 - 2^{-3} = 0.875$. (3) *Sedecilový obdélník D:* na *y* sedecily D_D, D_H a na *x* $P_4 = 2^{-4} = 0.0625$ a $1 - 2^{-4} = 0.9375$. Tato pomůcka může diagnostikovat i určité anomálie: (a) Symetrické unimodální rozdělení výběru obsahuje obdélníky symetricky uvnitř sebe. (b) Nesymetrická rozdělení mají pro rozdělení sešikmené k vyšším hodnotám vzdálenosti mezi dolními hranami obdélníků *F*, *E* a *D* výrazně kratší než mezi jejich horními hranami. (c) Odlehlá pozorování jsou indikována tím, že na kvantilové funkci mimo obdélník *F* se objeví náhlý vzrůst, kdy hodnota směrnice roste nade všechny meze. (d) Vícemodální rozdělení jsou indikována tím, že na kvantilové funkci uvnitř obdélníku *F* je několik úseků s téměř nulovými směrnicemi.

Jádrový odhad hustoty pravděpodobnosti **Obr. 8a, 8b** (osa *x: x*, osa *y:* hustota pravděpodobnosti) a histogram patří k nejužívanějším pomůckám a histogram pak k nejstarším diagramům hustoty pravděpodobnosti. U histogramu jde o obrys sloupcového grafu, kde jsou na ose x jednotlivé třídy, definující šířky sloupců, a výšky sloupců odpovídají empirickým



Obr. 8. Jádrové odhady hustoty pravděpodobnosti pro výběry z rozdělení (a) normálního, symetrického rozdělení Úlohy E2.10, (b) asymetrického rozdělení Úlohy E2.07. Čárkovaně je znázorněna hustota Gaussova rozdělení s parametry $\bar{x}$ a s^2 a plnou čarou jádrový odhad hustoty pravděpodobnosti empirického rozdělení výběru.



Obr. 9. Grafy *Q-Q* pro porovnání rozdělení výběru normálního rozdělení s teoretickým rozdělením.

hustotám pravděpodobnosti. Kvalitu histogramu ovlivňuje ve značné míře volba počtu tříd L a všech délek intervalů Δx_j. Pro přibližně symetrická rozdělení výběru lze vyčíslit L podle vztahu $L = \mathrm{int}(2\sqrt{n})$, kde funkce int($x$) označuje celočíselnou část čísla x nebo je možné užít výraz $L = \mathrm{int}(2.46(n-1)^{0.4})$.

Kvantil-kvantilový graf (graf Q-Q) Obr. 9a, 9b (osa x: $Q_T(P_i)$, osa y: $x_{(i)}$) umožňuje posoudit shodu výběrového rozdělení, charakterizovaného kvantilovou funkcí $Q_E(P)$ s kvantilovou funkcí zvoleného teoretického rozdělení $Q_T(P)$. Za odhad kvantilové funkce výběru se užívají pořádkové statistiky $x_{(i)}$. Při shodě výběrového rozdělení se zvoleným teoretickým rozdělením musí platit přibližná rovnost kvantilů $x_{(i)} = Q_T(P_i)$, kde P_i je pořadová pravděpodobnost. Pokud je rozdělení výběru shodné se zvoleným teoretickým rozdělením, je závislost $x_{(i)}$ na $Q_T(P_i)$ lineární a výsledná závislost se nazývá graf *Q-Q*. Těsnost lineární závislosti experimentálními body lze posoudit korelačním koeficientem a využít ho jako rozhodčí kritérium při hledání typu rozdělení.

4. Mocninná a Boxova-Coxova transformace dat

Pokud se na základě analýzy dat zjistí, že rozdělení výběru dat se systematicky odlišuje od rozdělení normálního, vzniká problém, jak data vůbec vyhodnotit. Často je pak nejlepším řešením vhodná transformace dat, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení. Zesymetričtění rozdělení výběru je možné provést užitím prosté **mocninné transformace**

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

která však nezachovává měřítko a není vzhledem k exponentu λ všude spojitá a proto se hodí pouze pro kladná data. Optimální odhad exponentu λ se hledá s ohledem na optimalizaci charakteristik asymetrie (šikmosti) a špičatosti. Pro přiblížení rozdělení výběru k rozdělení normálnímu vzhledem k šikmosti a špičatosti je vhodná **Boxova-Coxova transformace**

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

která je použitelná rovněž 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 $L(\lambda)$ lze znázornit ve zvoleném intervalu, např. $-3 \leq \lambda \leq 3$, a identifikovat maximum křivky, jejíž souřadnice x indikuje odhad $\hat{\lambda}$. Dva průsečíky křivky ln $L(\lambda)$ s rovnoběžkou s osou x indikují 100(1 - α)% interval spolehlivosti parametru λ. Čím bude interval spolehlivosti λ_D, λ_H širší, tím je mocninná nebo Boxova-Coxova transformace méně výhodná. Pokud obsahuje interval λ_D, λ_H i hodnotu λ = 1, není transformace ze statistického hlediska přínosem.

Zpětná transformace: Po vhodné transformaci se vyčíslí $\bar{y}$, $s^2(y)$ a potom pomocí zpětné transformace využitím Taylorova rozvoje v okolí $\bar{y}$ se odhadnou retransformované parametry polohy a rozptýlení $\bar{x}_R$ a $s^2(\bar{x}_R)$ původních dat. Uvedený postup vede vesměs k nejlepším odhadům polohy $\bar{x}_R$ a rozptýlení $s^2(\bar{x}_R)$ a je zvláště vhodný v případech asymetrického rozdělení výběru.

5. Intervalový odhad parametrů

Představuje interval, ve kterém se bude se zadanou pravděpodobností či statistickou jistotou (1 - α) nacházet skutečná hodnota čili „pravda" daného parametru μ. Neznámý parametr μ odhadujeme dvěma číselnými hodnotami L_D a L_H, které tvoří **meze** tzv. *intervalu spolehlivosti* čili konfidenčního intervalu. Interval spolehlivosti pokryje parametr μ s předem zvolenou, statistickou jistotou čili dostatečně velikou pravděpodobností $P = (1 - \alpha)$, což lze vyjádřit vztahem $P(L_D < \mu < L_H) = 1 - \alpha$, nazvanou *koeficient spolehlivosti* (čili konfidenční koeficient, statistická jistota). Je obyčejně roven 0.95 nebo 0.99. Parametr α se nazývá *hladina významnosti*. Interval spolehlivosti vyjadřuje tvrzení: *"Statistická jistota, s jakou bude „pravda" μ ležet v náhodných mezích L_D, L_H je rovna právě 1 - α"*. Vlastnosti intervalu spolehlivosti: (1) Čím je rozsah výběru n větší, tím je interval spolehlivosti užší. (2) Čím je odhad přesnější a má menší rozptyl, tím je interval spolehlivosti užší. (3) Čím je vyšší statistická jistota (1 - α), tím je interval spolehlivosti širší.

Konstrukce intervalových odhadů: Postup konstrukce intervalu spolehlivosti střední hodnoty μ normálního rozdělení $N(\mu, \sigma^2)$:

1. **Velký výběr $n \geq 30$**: Když nejlepším bodovým odhadem střední hodnoty μ je výběrový průměr s rozdělením , pak v intervalu $N(\mu, \sigma^2/n)$ leží přibližně 95% hodnot náhodných veličin výběru o rozsahu n a *100(1 - α)%ní interval spolehlivosti střední hodnoty* μ bude vyčíslen vztahem

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}},$$

kde hodnota 1.96 je 100(1 - 0.05/2) = 97.5%ní kvantil normovaného normálního rozdělení $u_{0.975}$.

2. **Malý výběr $n \leq 30$:** v praxi obvykle neznáme směrodatnou odchylku σ ale pouze její odhad s a je-li $t_{1-\alpha/2}(n-1)$ je 100(1 - α/2)%ní kvantil Studentova rozdělení bude *100(1 - α)%ní interval spolehlivosti střední hodnoty* μ roven

$$\bar{x} - t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}}$$

Meze intervalu spolehlivosti závisí vedle chyby s i na rozsahu výběru n. Pro větší rozsahy výběru ($n > 30$) lze použít místo kvantilu $t_{1-\alpha/2}$ kvantilu normovaného normálního rozdělení $u_{1-\alpha/2}$ a *100(1 - α)%ní oboustranný interval spolehlivosti rozptylu* σ^2 se vypočte dle

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)},$$

kde $\chi^2_{1-\alpha/2}(n-1)$ je horní a $\chi^2_{\alpha/2}(n-1)$ dolní kvantil rozdělení χ^2. Robustní *interval spolehlivosti mediánu* se přibližně vyčíslí

$$\tilde{x}_{0.5} - u_{1-\alpha/2}\frac{0.707s}{\sqrt{n}} \leq med \leq \tilde{x}_{0.5} + u_{1-\alpha/2}\frac{0.707s}{\sqrt{n}}$$

6. Analýza malých výběrů

Předem je třeba si uvědomit, že závěry z malých výběrů jsou vždy zatíženy značnou mírou nejistoty. Malých rozsahů proto užijeme jen tam, kde skutečně není možné zvýšit počet měření.

Hornův postup pro malé výběry, $4 \leq n \leq 20$ je založený na pořádkových statistikách. Nejprve se určí hloubka pivotu je H = (int((n + 1)/2))/2 nebo H = (int((n + 1)/2 + 1)/2, pak dolní pivot jako $x_D = x_{(H)}$ a horní pivot dle $x_H = x_{(n+1-H)}$. Odhadem parametru polohy je potom *pivotová polosuma* $P_L = (x_D + x_H)/2$ a a odhadem parametru rozptýlení je *pivotové rozpětí* $R_L = x_H - x_D$. Lze definovat i náhodnou veličinu k testování $T_L = P_L/R_L$, která má přibližně symetrické rozdělení, jehož vybrané kvantily jsou dostupné v tabulce [1]. Potom se *95%ní interval spolehlivosti střední hodnoty* vypočte vztahem

$$P_L - R_L t_{L,0.975}(n) \leq \mu \leq P_L + R_L t_{L,0.975}(n).$$

7. Test správnosti výsledku

Testy hypotéz o parametrech μ a σ^2 normálního rozdělení: soubor s $N(\mu, \sigma^2)$, výběr rozsahu n a vypočteme průměr a směrodatnou odchylku s. Testy správnosti výsledku měření lze provést pomocí intervalu spolehlivosti dle pravidla: pokud 100(1 - α) %ní interval spolehlivosti parametru μ obsahuje zadanou hodnotu μ_0, nelze na hladině významnosti α zamítnout hypotézu $H_0 : \mu = \mu_0$.

8. Ilustrativní úlohy dle cit [2]

Úloha E2.10 *Symetrie výběrového rozdělení a správnost obsahu dusičnanů v pitné vodě*

Fotometrickou metodou se salicylanem sodným byl stanoven obsah dusičnanů [mg.l^{-1}] v reálném vzorku pitné vody o deklarované hodnotě

57.0 mg.l^{-1}. Exploratorní analýzou je třeba učinit závěry o typu a symetrii rozdělení a výskytu odlehlých hodnot. Je třeba ověřit správnost střední hodnoty obsahu dusičnanů v pitné vodě vůči deklarované hodnotě.

Data: Nalezený obsah dusičnanů v pitné vodě [mg. l^{-1}]: 57.56 57.80 58.59 56.72 59.33 58.27 56.65 57.03 56.58 55.71 58.00 57.08 58.41 53.64 57.13 58.04 58.45 57.92 56.21.

Úloha E2.07 *Symetrie rozdělení a správnost hodnot chloridů v přírodních vodách*

Přírodní vody představují zdroje pitné vody v různých lokalitách, ve kterých se sleduje řada složek, např. chloridy s obsahem 190 mg. l^{-1}. Je třeba provést průzkum tvaru a symetrie rozdělení na základě grafu polosum, symetrie a kvantilového grafu a dále ověřit správnost obsahu chloridů vůči deklarované hodnotě.

Data: Nalezený obsah chloridů v přírodních vodách [mg. l^{-1}] v okružním testu: 184.0 183.0 221.6 182.9 180.0 189.5 181.0 174.6 243.0 189.7 180.1 178.0 181.6 172.0 180.0 177.0 182.0 170.0 179.6 188.5 204.0 184.0 160.0 185.0 187.0 185.0 190.0 186.2 187.9 186.0 182.0 211.0 177.3.

Řešení:

Na statistickém vyhodnocení a testu správnosti intervalem spolehlivosti dvou výběrů, a to z rozdělení symetrického u Úlohy E2.10 a asymetrického u Úlohy E2.07 budou vysvětleny nejprve užívané diagnostické grafy exploratorní analýzy: kvantilový graf (**Obr. 2.**) odhaluje asymetrické rozdělení a odlehlé body. Diagram rozptýlení a rozmítnutý diagram rozptýlení (**Obr. 3.**) odhaluje míru rozptýlení prvků výběru. Kolečka v krabicovém grafu (**Obr. 4.**) značí odlehlé body. Body mezi horním a dolním konfidenčním pásem v grafu polosum (**Obr. 5.**) a v grafu symetrie (**Obr. 6.**) se týkají symetrického rozdělení, zatímco body vně pásu rozdělení asymetrického. Asymetrický tvar grafu roztýlení s kvantily (**Obr. 7.**) potvrzuje, že rozdělení na **Obr. 7b.** je silně asymetrické. Jádrové odhady hustoty pravděpodobnosti pro oba výběry (**Obr. 8.**) ukazují, které rozdělení je asymetrické a ne-normální. Leží-li body těsně na přímce kvantil-kvantilového grafu, jde o symetrické a obvykle normální rozdělení. Body vzdálené od konců přímky značí odlehlé hodnoty. Exploratorní analýza předurčí volbu, zda k testu správnosti využijeme intervalový odhad aritmetického průměru nebo retarnsformovaného průměru po předešlé transformaci dat. Studentův *t*-test správnosti analytického výsledku využívá intervalu spolehlivosti: nachází-li se totiž hodnota μ_0 (tj. "pravda", správná hodnota, norma, standard) ve vyčísleném intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Z **tabulky 2** je zřejmé, že výběr úlohy E2.10 vykazuje symetrické normální rozdělení a k testování lze proto použít aritmetický průměr. protože leží je deklarovaná hodnota $\mu_0 = 57.0$ v intervalu spolehlivosti $[L_D = 56.70; L_H = 57.94]$, je nalezený odhad aritmetického průměru = 57.32 správný. Protože je normalita rozdělení výběru úlohy E2.07 zamítnuta, nelze použít aritmetický průměr se svým intervalem spolehlivosti, protože oba vedou k falešným závěrům. Data E2.07 je třeba nejprve transformovat exponenciální nebo Boxovou-Coxovou transformací a k retransformovanému průměru $\bar{x}$ = 183.72 nebo 183.64 vyčíslit také interval spolehlivosti $[L_D = 179.80; L_H = 188.24]$ nebo $[L_D = 179.42; L_H = 187.86]$. Protože deklarovaná hodnota $\mu_0 = 190.0$ neleží ani v jednom retransformovaném intervalu spolehlivosti, je retransformovaný průměr nesprávný. **Viz tabulka 2.**

Tabulka 2. Mocninná, exponenciální a Boxova-Coxova transformace u výběru Úlohy E2.10 a E2.07 (QC-Expert)

(1) Odhady klasických parametrů	E2.10	E2.07
Deklarovaná hodnota	57.0	190.0
Odhad aritmetického průměru	57.32	186.16
Dolní mez intervalu spolehlivosti	56.70	180.83
Horní mez intervalu spolehlivosti	57.94	191.50
Odhad směrodatné odchylky	1.28	15.05
Odhad šikmosti	-1.13	2.04
Odhad špičatosti	4.70	8.24
Jarque-Berrův test normality: Normality je	přijata	zamítnuta
Test správnosti vůči deklarované hodnotě: Správnost je	přijata	přijata
(2) Prostá mocninná transformace, exponenciální transformace:		
Odhad optimálního exponentu	-0.42	0.50
Opravený odhad průměru původních dat	57.53	183.72
Dolní mez intervalu spolehlivosti	56.93	179.80
Horní mez intervalu spolehlivosti	58.03	188.24
Test správnosti s deklarovanou hodnotou: Správnost je	přijata	zamítnuta
(3) Boxova-Coxova transformace:		
Odhad optimálního exponentu	2.50	-1.26
Opravený odhad průměru původních dat	57.49	183.64
Dolní mez intervalu spolehlivosti	56.44	179.42
Horní mez intervalu spolehlivosti původních dat	58.54	187.86
Test správnosti s deklarovanou hodnotou: Správnost je	přijata	zamítnuta

9. Závěr

V postupu statistického vyhodnocení výsledků měření slouží průzkumová analýza dat EDA jako výhodná pomůcka k vyšetření zvláštností statistického chování dat. Z nejdůležitějších pomůcek jsou to vedle kvantilového grafu a grafu rozptýlení s kvantily i diagram rozptýlení a rozmítnutý diagram rozptýlení, krabicový graf, vrubový krabicový graf, graf polosum a symetrie, kvantil-kvantilový graf, jádrový odhad hustoty pravděpodobnosti a histogram k určení tvaru rozdělení. U malých výběrů $4 \le n \le 20$ poskytuje správné odhady střední hodnoty Hornův postup pivotů. Pivotová polosuma a pivotové rozpětí umožňují vyčíslit i intervalový odhad střední hodnoty a navíc jsou oba odhady dostatečně robustní vůči asymetrii rozdělení malého výběru a i vůči odlehlým hodnotám. Studentův *t*-test správnosti analytického výsledku je ekvivalentní vůči intervalu spolehlivosti. Nachází-li se totiž hodnota μ_0 (tj. "pravda", správná hodnota, norma, standard) v intervalu spolehlivosti $[L_D; L_H]$, je stanovení správné. Exploratorní analýza předurčí volbu, zda k testu správnosti využijeme intervalový odhad aritmetického průměru v případě symetrického rozdělení nebo retransformovaného průměru v případě asymetrického rozdělení. Interaktivní statistická analýza při užití vhodného software umožňuje jednoznačně vyšetřit správnost analytického výsledku.

Literatura

[1] M. Meloun; J. Militký: *Statistické zpracování experimentálních dat,* Plus Praha 1994 (1. vydání), East Publishing 1996 (2. vydání), Academia Praha 2004 (3. vydání).

[2] M. Meloun; J. Militký: *Kompendium statistického zpracování dat,* Academia Praha 2002.

[3] ADSTAT, TriloByte Statistical Software s. r. o., Pardubice 1990.

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie
Chemicko-technologická fakulta
Univerzita Pardubice,
nám. Čs. Legií 565, 532 10 Pardubice,
http://meloun.upce.cz, email: milan.meloun@upce.cz,
telefon: 466037026, fax: 466037068, ICQ: 224-001-003

Computer-Assisted Statistical Data Analysis: 1. Interactive analysis of the univariate data (Meloun, M.)

Key Words

exploratory data analysis - power transformation - Box-Cox transformation - accuracy test - confidence interval - mean value - arithmetic mean - standard deviation - median - quantiles - letter values - quantile plot - jittered dot diagram - Box-and-whisker plot - midsum plot - symmetry plot - quantile-box plot - kernel estimation of probability density - histogram - quantile-quantile plot - Horn procedure of pivots

The first step of the univariate data analysis called an exploratory data analysis (EDA) isolates certain basic statistical features and patterns of data. This is a set of various descriptive graphically oriented techniques which are typically free of strict statistical assumptions about data. The EDA-techniques are often called „distribution-free" and isolates certain statistical features and patterns of data. For graphical visualization of data the EDA uses the quantile plot, the dot and jittered dot diagrams, the (notched) box-and-whisker plot while the sample distribution is investigated by the midsum plot, the symmetry plot and the quantile-box plot. The construction of sample distribution i.e. the estimation of probability density function is done by the kernel estimation of probability density function, the histogram, the bar chart, the quantile-quantile plot. When an exploratory data analysis (EDA) finds that the sample distribution differs from a normal one or when a confirmatory data analysis (CDA) does not prove a sample independence and a sample homogeneity, the original data should be transformed. The power transformation and the Box-Cox transformation improves a sample symmetry and stabilizes a sample variance. The plot of logarithm of maximal likelihood function enables to find an optimum power transformation. According to results of an examination about sample assumptions the classical, robust or adaptive estimates of location and spread are calculated.