Využití exploratorní analýzy v analýze mikroelementů

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Univerzita Pardubice, 532 10 Pardubice

Souhrn: *Účelem průzkumové analýzy dat (EDA), jako prvního kroku v analýze jednorozměrných dat, je odhalení jejich statistických zvláštností pomocí 12 grafických diagnostik, ověření základních předpokladů a případná mocninná transformace. Při průzkumové analýze složitějších, nákladných nebo unikátních měření je totiž nutné posoudit zvláštnosti chování dat ještě před vlastní, rutinní statistickou analýzou. Jedině tak lze zabránit provádění numerických výpočtů bez hlubších statistických souvislostí. EDA je pak obzvláště důležitá ve stopové analýze mikroelelemntů a kontrole kvality.*

1. Postup analýzy dat ve stopové analýze

V prvním kroku se v průzkumové (exploratorní) analýze dat vyšetřují *statistické zvláštnosti,* jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnosti podezřelých hodnot. Odhalí se tak anomálie a odchylky rozdělení výběru od předpokládaného rozdělení Gaussova. Interaktivní statistická analýza na počítači tento postup ulehčuje, většina statistického software totiž nabízí řadu diagnostických grafů a diagramů.

V druhém kroku 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 klasických odhadů polohy a rozptýlení, tj. aritmetického průměru a rozptylu ve kroku čtvrtém. Sem patří i konstrukce intervalů spolehlivosti a případně testování hypotéz. V opačném případě následuje pokus o třetí krok, obsahující symetrizující transformaci dat.

Ve třetím kroku se provádí mocninná a Boxova-Coxova transformace, které mohou vést k symetričtějšímu rozdělení výběru a umožňují provedení korek-tnějšího odhadu polohy a rozptýlení. Transformace je vhodná především při nekontantnosti rozptylu a při asymetrii rozdělení původních dat.

Ve čtvrtém kroku se v konfirmatorní analýze nabízí paleta rozličných odhadů polohy, rozptýlení a tvaru, jež lze rozdělit do skupin: *klasické odhady* a *robustní odhady* (necitlivé na odlehlé prvky výběru, resp. další předpoklady o datech). Z nabídky odhadů parametrů vybírá uživatel uvážlivě ty, jež mají statistický smysl a odpovídají závěrům průzkumové analýzy dat a závěrům ověření předpokladů o výběru.

Postup statistické analýzy jednorozměrných dat lze shrnout do čtyř kroků:

A. Průzkumová (exploratorní) analýza dat (EDA)

1. Zkoumání zvláštností dat:

Odhalení stupně symetrie a špičatosti rozdělení;
Indikace lokální koncentrace dat a rozdělení výběru;
Nalezení vybočujících a podezřelých prvků ve výběru.

2. Ověření předpokladů o datech:

Ověření normality rozdělení; Ověření nezávislosti prvků výběru;
Ověření homogenity rozdělení výběru; Určení minimálního rozsahu výběru.

3. Transformace dat:

Mocninná transformace; Box-Coxova transformace.

B. Konfirmatorní analýza dat (CDA) - odhady parametrů
(polohy, rozptýlení a tvaru)

1. Klasické odhady (bodové a intervalové): - momentové;
2. Robustní odhady (bodové a intervalové): - kvantilové, - uřezané, - winsorizované.

A. Postup analýzy rutinních dat

Při zpracování rutinních výsledků měření předpokládáme, že známe rozdělení souboru dat. Předpokládá se, že rozdělení dat je normální a data asi splňují předpoklady nezávislosti a homogenity. Účelem je

a) Test nezávislosti: testování nezávislosti prvků výběru - autokorelace,
b) Test homogenity: testování homogenity výběru,
c) Test normality: testování normality rozdělení výběru.

Z grafických metod se k předběžné analýze rutinních dat nejčastěji užívá *rankitových grafů* a *grafů rozptýlení s kvantily*. Nejsou-li však o rozdělení dat dostupné žádné informace nebo očekává-li se výrazně nenormální rozdělení, je vhodné provést

a) průzkumovou analýzu dat s využitím grafických diagnostik,
b) určení výběrového rozdělení a jeho konstrukci.

Pokud nebylo nalezeno vhodné aproximující rozdělení, provádí se *mocninná transformace*, která by měla zlepšit rozdělení dat. Kombinace metod závisí na konkrétních datech a konkrétních požadavcích analýzy.

B. Postup při nesplnění předpokladů o datech

1. Nesplnění předpokladu nezávislosti prvků

Pokud prvky měření nejsou nezávislé, vzrůstá nebezpečí, že odhady budou systematicky vychýleny a nadhodnoceny pro pozitivní hodnotu autokorelačního koeficientu. Nezbývá, než hlouběji analyzovat logické příčiny a snažit se o jejich odstranění, zkontrolovat celý měřicí řetězec a provést nová měření.

2. Nesplnění předpokladu normality výběru

Rozdělení dat je buď jiné než normální, nebo jsou v datech vybočující měření. V případě nenormálního rozdělení dat může jít o odchylky pouze v délce konců, nebo se jedná *o sešikmená rozdělení.* V případě symetrických rozdělení lišících se od normálního délkou konců lze použít pro odhad parametrů polohy a rozptýlení jednoduché robustní techniky. U sešikmených rozdělení je vždy výhodné začít hledáním mocninné transformace. Pokud byla mocninná transformace úspěšná a byl nalezen optimální exponent λ, provádí se další analýza v této transformaci a nakonec se vyčíslí zpětná transformace do původních proměnných. Nebyla-li mocninná transformace úspěšná, je možné pomocí technik průzkumové analýzy dat nalézt vhodné teoretické aproximující rozdělení a realizovat další postup na základě obecných vztahů pro rozptyl, resp. střední hodnotu. S využitím tohoto konfidenčního intervalu pro střední hodnotu sešikmených rozdělení lze také provádět testy významnosti parametru polohy.

3. Přítomnost vybočujících hodnot

Na základě logické analýzy je třeba nejdříve zvážit, zda nejde o sešikmené rozdělení. Body, které se jeví vybočující pro symetrické (speciálně normální) rozdělení, mohou být pro sešikmená rozdělení naopak přijatelné. Pokud jde o vybočující pozorování, lze použít dvou alternativ.

První alternativa spočívá v jejich vyloučení z další analýzy, což však není vždy zcela nejvhodnější. Pokud jsou vybočující měření výsledkem řídce se vyskytujících jevů, může tím totiž

dojít ke ztrátě informace úplně. Proto lze tyto hodnoty vyloučit jedině při doplnění o nová experimentální data.

Druhá alternativa spočívá v použití robustních metod. Tento postup však nemusí být vždy korektní. Robustnost spočívá v přiblížení se k přijatému modelu měření bez ohledu na jeho platnost. Pokud se analýzy vybočujících měření účastní experimentátor, měl by rozhodnout, která měření jsou evidentní hrubé chyby (jako je selhání přístroje, špatný zápis dat), a která jsou jen podezřelá. Evidentní hrubé chyby je vhodné z další analýzy vyloučit, ale podezřelá měření je lépe ponechat. Robustními metodami se jejich vliv na odhady parametrů výrazně oslabí.

4. Nedostatečný rozsah výběru

Nejjednodušší je v tomto případě provést dodatečná měření. Platí, že čím jsou data méně rozptýlená, tím menší počet jich stačí k zajištění dostatečné přesnosti odhadu. Pokud nelze provést dodatečné experimenty, je možné použít techniky vhodné pro malé výběry - viz Hornův postup.

Tento postup je vhodný zejména pro analýzu rutinních měření, kde jsou o chování dat předběžné informace. Když se analyzují výsledky nových měření nebo neznámé výběry, je vždy třeba začít průzkumovou analýzou dat a stanovit statistické zvláštnosti výběru.

Doporučená 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] Meloun M., Militký J.: *Statistická analýza experimentálních dat*, Academia Praha 2004, ISBN 80-200-1254-0.

[5] Meloun M., Militký J.: *Kompendium statistického zpracování dat*, Academia Praha 2006, ISBN 80-200-1396-2.

[6] Meloun M., Militký J., Hill M: *Počítačová analýza vícerozměrných dat v příkladech*, Academia Praha 2005, ISBN 80-200-1335-0.