Rigozní přístup k analýze jednorozměrných dat

Rigorous Approach to the Univariate Data Analysis

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 řady grafických diagnostik, ověření základních předpokladů o výběru a především rigorózní mocninná a Box-Coxova transformace dat. 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 a kontrole kvality.*

Summary: The purpose of the proposed Exploratory Data Analysis (EDA), as the first step in the analysis of the univariate data analysis is to reveal their statistical peculiarities of data elements by using a series of graphical diagnostics, validating basic assumptions, and rigorous power and Box-Cox data transformation. In the exploratory analysis of more complex, costly or unique measurements, it is necessary to assess the peculiarities of data behavior prior to the actual, routine statistical analysis. Only then the numerical calculations can be performed without deeper statistical connections. The EDA is then particularly important in trace analysis and quality control.

Procedure of the univariate data analysis

1. In the first step, statistical peculiarities such as local data concentration, shape specificity of data distribution, and the presence of suspicious or outliers values are investigated in exploratory data analysis. Anomalies and variations in the distribution of the selection from the predicted Gaussian distribution are thus revealed. Interactive statistical analysis on the computer enables this process easier, because most statistical software offers a number of diagnostic graphs and diagrams. In addition, the basic assumptions of sample, such as elements independence, sample homogeneity, sufficient number of elements in a sample and sample distribution, are verified. If the conclusions of this step are optimistic, then the classical estimates of position and spread, i.e. the arithmetic mean and the variance could be calculated. This step also includes the construction of confidence intervals and possibly hypothesis testing. Otherwise, an attempt is performed to symmetrize the data transformation.
2. In the second step, the power and Box-Cox transformations are performed in a confirmatory analysis, which can lead to a more symmetrical distribution of the sample elements and allow for a more correct estimate of location and spread. The transformation is suitable especially for the sample heteroscedasticity and for the asymmetry of the distribution of the original data sample. There is also a variety of different location, dispersion and shape estimates that can be here calculated: classic estimates and robust estimates (insensitive to outlying elements of the selection, or other assumptions about data). From the parameter estimation menu, the user chooses carefully those that have a statistical meaning and correspond to the conclusions of the exploratory data analysis and to the conclusions of the selection assumptions.

Procedure:

1. **Exploratory Data Analysis (EDA) - data peculiarities and verification of data assumptions:**
   Uncovering the degree of symmetry, skewness and curtosis of sample.
   Indication of local data concentration and sample distribution.
   Finding outliers and suspicious elements in the sample.
   Verification of distribution symmetry and normality.
   Verifying the independence of sample elements.
   Verification of homogeneity of selection distribution.
   Determine a minimum number of the sample elements.
2. **Confirmative Data Analysis (CDA) - Estimates of Position, Spread and Shape:**
   Rigorous Data Transformation: Power Transformation; Box-Cox transformation.
   Classical estimates (point and interval)**:** - momentary;
   Robust estimates (point and interval)**:** - quantile, - trimmed, - winsorized.

1. Postup analýzy jednorozměrných dat

V prvním kroku se v průzkumové (exploratorní) analýze dat vyšetřují *statistické zvláštnosti,* jako je lokální koncentrace dat, tvarové zvláštnosti rozdělení dat a přítomnost 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. Dále jsou ověřeny *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 v kroku druhém. Sem patří i konstrukce intervalů spolehlivosti a případně testování hypotéz. V opačném případě následuje pokus o symetrizující transformaci dat.

Ve druhém kroku se v konfirmatorní analýze provádí mocninná a Boxova-Coxova transformace, které mohou vést k symetričtějšímu rozdělení výběru a umožňují provedení správnějšího odhadu polohy a rozptýlení. Transformace je vhodná především při nekonstantnosti rozptylu a při asymetrii rozdělení původních dat. Nabízí se také paleta rozličných odhadů polohy, rozptýlení a tvaru, které 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

1. Průzkumová analýza dat (EDA) - z**koumání zvláštností dat a ověření předpokladů o datech:**
- Odhalení stupně symetrie a špičatosti rozdělení.
- Indikace lokální koncentrace dat a rozdělení výběru.
- Nalezení vybočujících a podezřelých prvků ve výběru.
- Ověření normality rozdělení.
- Ověření nezávislosti prvků výběru.
- Ověření homogenity rozdělení výběru.
- Určení minimálního rozsahu výběru.

2. Konfirmatorní analýza dat (CDA) - odhady parametrů polohy, rozptýlení a tvaru:
- **Rigorozní transformace dat:** Mocninná transformace. Box-Coxova transformace.
- **Klasické odhady** (bodové a intervalové): - momentové.
- **Robustní odhady** (bodové a intervalové): - kvantilové, - uřezané, - winsorizované.

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.