Asian Journal of Chemistry Vol. 14, Nos. 3–4 (2002), 1772–1774

NOTE

The Boxplot: A Robust Exploratory Data Analysis Tool for the Definition of the Threshold for Outlier Data

M. BOUNESSAH
Department of Chemistry, Riyadh College of Technology
P.O. Box 42826, Riyadh 11551, Saudi Arabia
E-mail: bouna@rct.edu.sa

Data from chemical analyses are subject to variability due to sampling, sample preparation, analytical method, variation in the phenomena under study, etc. In classical statistics, prior to the analysis of data, it is required to transform it in order to satisfy the Gaussian distribution assumption. However, this is not always possible owing to the presence of "wild" values, a phenomenon omnipresent in studies such as in environmental pollution and many other specialised fields. This note introduces an unconventional technique for the analysis and interpretation of univariate data.

Key Words: Robust, Data, Analysis, Threshold, Outlier data.

Chemical analysis data from environmental studies and many other specialized fields are subject to many factors that affect their variability including sampling, analytical procedure, physico-chemical environment, man-made pollution, etc. These factors are reflected in the empirical data distribution by inconsistencies, polymodal behaviour, the presence of "wild" data, etc. These extreme data are of utmost importance for instance in environmental studies as they may indicate a potential pollution, contamination, etc. In classical statistics, the threshold for defining such data is based on the x + 2sd (or 3sd) method; x being the mean and sd the standard deviation. In cases where the data distribution is skewed, data is transformed (log, square root, etc.) prior to analysis in order to satisfy the Gaussian distribution prerequisite; however this is not always successful because of the presence of extreme data.

An unconventional approach to understanding single-data distribution and defining a threshold for outliers is the use of Exploratory Data Analysis (EDA) as developed by Tukey[1] and Hoaglin *et al.*[2]. This technique has been applied elsewhere using the boxplot by many workers such as Kürzl[3]; O'Connor and Reimann[4]; O'Connor *et al.*[5]; Reimann *et al.*[6]; Reimann[7]. This note depicts the boxplot, its properties and its advantages in analysing univariate data.

Definition, properties and applications of the boxplot: The boxplot displays the characteristics of the empirical distribution for single data at a glance: location, spread, skewness, tail lengths and outliers ("wild" values). The box represents 50% of ordered data stretching between the lower hinge and the upper hinge which represent the lower and the upper quartile of the data respectively (Fig. 1). The bar in this box indicates the median, which by its position depicts the symmetry or skewness of the data. The two hinges, the median and the two extreme values (lowest and highest) are known as the 5-number summary. The box also describes the spread of the data symbolised by the h-spread, whereas its width shows the sample size. The whiskers include all data, from the hinges up to the lower and the upper fences which define the outliers ("wild data") cutoffs. The cutoffs are found by subtracting or adding a step (1.5 of the h-spread) to the lower and to the upper fences respectively. This means, in a given data set, 25% of the data can be arbitrarily wild without significantly affecting the median and hinges; and since the outlier cutoffs are detined by the h-spread, they are not affected by outlier data and therefore can resist disturbances due the data.

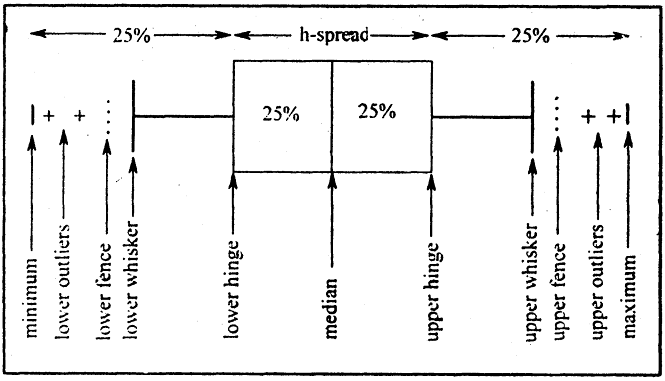

Fig. 1. Definition of the boxplot and its properties (see text for details)

This graphical technique has been applied successfully by many workers in the past (Yusta *et al.*[8]; Bounessah and Atkin[9]; O'Connor and Reimann[4]; Kurzl[3]; O'Connor *et al.*[5]; Reimann, *et al.*[6]). Fig. 2 shows an example of boxplot-compar-

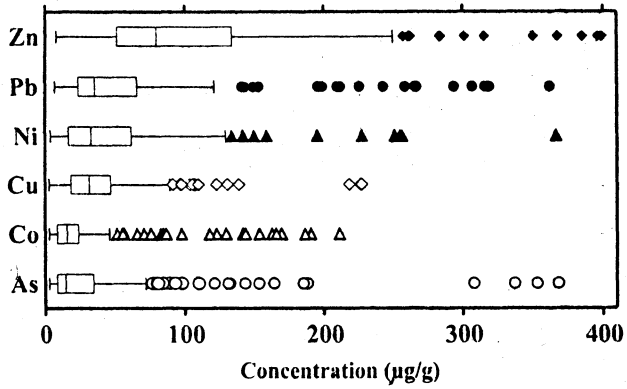

Fig. 2. Boxplot-comparison between As, Co, Cu, Ni, Pb and Zn in a single geochemical group; for graphical clarity, extreme outliers lying to the right where omitted (data from Bounessah and Atkin[9]).

ison between As, Co, Cu, Ni, Pb and Zn for the geochemical data taken from Bounessah and Atkin[9]. Visual comparison is enhanced and from the empirical distribution of the data, one can easily notice that all data are skewed to the right, and that all elements analysed are characterized by outlier values. In this example, the log-transformation of the data did not approach the normal distribution and these were analysed and interpreted without prior transformation.

In conclusion, the boxplot is a very useful tool for graphically portraying the empirical distribution of data. It has shown its effectiveness for the analysis of univariate data where the variability of measurements may be affected by many factors. The boxplot shows the following advantages:

(1) It gives a quick insight into the empirical distribution of data and its statistics: central location, skewness, outliners, etc.

(2) It is very robust when defining the cutoff for outlier data that may affect statistical parameters in classical analysis (mean, standard deviation, etc.).

(3) No particular model is assumed or fitted to the data when using the boxplot, which avoids the need to transform data.

Finally, the boxplot is incorporated in statistical softwares such as Statistica and SPSS and should become an integral part in the analysis and interpretation of data.

REFERENCES

1. J.W. Tukey, Addison-Wesley, Reading (1977).
2. D.C. Hoaglin, F. Mosteller and W. Tukey, John Wiley & Sons, New York (1983).
3. H. Kürzl, *J. Geochem. Explor.*, **30**, 309 (1988).
4. P.J. O'Connor and C. Reimann, *J. Geochem. Explor.*, **47**, 63 (1992).
5. P.J. O'Connor and C. Reimann and H. Kürzl, *Can. Inst. Min. Metall.*, 449 (1988).
6. C. Reimann, H. Kurzl and F. Wurzer, *Sci. Rev.*, **21** (1987).
7. C. Reimann, *J. Geochem. Explor.*, **31**, 75 (1987).
8. I. Yusta, F. Velasco and J.M. Herrero, *App. Geochem.*, **13**, 421 (1998).
9. M. Bounessah and B.P. Atkin, *J. Afr. Earth. Sci.*, **19**, 51 (1994).

(*Received*: 18 February 2002; *Accepted*: 10 June 2002) AJC-2764