DALARNA
UNIVERSITY

**Working papers in transport, tourism, information technology and microdata analysis**

# A reflection on outlier detection and skewed distributions

**Kenneth Carling**
**Editor: Hasan Fleyeh**

**Nr: 2017:02**

Working papers in transport, tourism, information technology and microdata analysis
ISSN: 1650-5581

# A reflection on outlier detection and skewed distributions

Authors[♦]: Kenneth Carling

This version: 2017-10-20

**Abstract**: It seems that a paper of mine appearing in this journal (Carling, 2000) has prompted the development of outlier detection methods for highly skewed data. However, I wrote the paper in the spirit of Exploratory Data Analysis (Tukey, 1977) and I shared Tukey's opinion, and I still hold it, that skewed data are better to be transformed for approximate symmetry prior to detection of outliers (or other data analyses).

**Key words**: Box-plot, Non-Normality, Outlier rules

The Boxplot rule for detection of outliers is efficient, elegant, simple and well-known in most scientific domains where data analysis is regularly conducted. It has emerged from the work of Tukey (1977) into a rule that labels observations greater than $q_3 + 1.5(q_3 - q_1)$ (or less than $q_1 - 1.5(q_3 - q_1)$) as outliers, where $q_1, q_3$ are the first and third sample quartiles. In Carling (2000), I argued that the rule could be improved by defining the cut-off values as $q_2 \pm c(n)(q_3 - q_1)$, where $q_2$ is the sample median and the constant $c(n) = \frac{(17.63n - 23.64)}{(7.74n - 3.71)}$ which is approximately 2.3 for large samples (i.e. $n$ large).

The recommendation of the value of the (sample-sized adjusted) constant in the rule was deliberately made to accommodate the situation when the parent distribution of the sample (absent outliers) observations deviated to some extent from the Normal distribution. As stated in the paper: "This comparative study extends beyond Gaussian by considering batches of data which moderately deviate therefrom. This is a desirable extension since perfect normality is rarely obtained in applied work, even after application of a suitable transformation." (Carling, 2000, p. 250). Subsequent research, frequently appearing in CSDA, has argued for additional modifications of the outlier rule

[♦] Kenneth Carling is a professor in Statistics and Micro-data Analysis at the School of Technology and Business Studies, Dalarna university, SE-791 88 Falun, Sweden. E-mail: kca@du.se. Phone: +46-23-778967.

to adjust for skewed data. Forerunners were Schwertman, Owens, and Adnan (2004) (see also Carter, Schwertman, and Kiser, 2009) who discussed the idea of the lower and the upper cut-offs to depend on the distance between the median and the first and third quartiles, respectively. Hubert and Vandervieren (2008) furthered this development by letting the asymmetry of the cut-offs be contingent on a robust measure of skewness. They have been followed by many others such as, to mention some recent instances, Bruffaerts, Verardi, Vermandele (2014) and Dovoedo and Chakraborti (2015).

In particular the paper of Hubert and Vandervieren (2008) has diffused broadly to other areas where data analysis is being conducted. Hubert and Vandervieren (2008) stated with reference to Carling (2000): "It is not clear how the method performs when these shape parameters first need to be estimated.". Their writing was in reference to equation (4.1) that appeared in Carling (2000), an expression that showed how the outside rate was a function of the skewness and the kurtosis of the parent distribution of the data. As I am receiving (from practitioners) a substantial number of comments and questions regarding estimation of the shape parameters I would want to stress that, in the context of exploratory data analysis, I find the idea of estimating higher order moments prior to performing outlier detection bizarre. It was certainly not the message I intended to convey to practitioners of exploratory data analysis. Nor do I find it appropriate to suggest practitioners to apply skewness-adjusted outlier rules whenever they encounter skewed data as these inevitably require some prior estimation based on the potentially contaminated sample.

Instead I would urge the practitioner to transform data for approximate symmetry before outlier detection. Church (1979, p. 435) stated in his summarizing review of Tukey (1977): "Very often it is more convenient to look at some transform of the original variable. If the distribution is far from symmetrical…, one end of the distribution will be too crowded to permit careful inspection.". Further, Hoaglin, Mosteller, and Tukey (1983) in their book on Exploratory Data Analysis dedicated an entire chapter to the arguments in favor for transformation prior to analysis and the outline of simple and robust transformation methods. So in conclusion, I would urge the data analyst encountering skewed data to apply a simple transformation to achieve approximate symmetry and then use the outlier rule presented in Carling (2000).

## References

Bruffaerts C, Verardi V, Vermandele C (2014). A generalized boxplot for skewed and heavy-tailed distributions, Statistics & Probability Letters, 95, 110-117.

Carling K, (2000). Resistant outlier rules and the non-Gaussian case, Computational Statistics & Data Analysis, 33, 249-258.

Carter N J, Schwertman N C, Kiser T L, (2009). A comparison of two boxplot methods for detecting univariate outliers which adjust for sample size and asymmetry, Statistical Methodology, 6, 604-621.

Church R M, (1979). How to look at data: A review of John W. Tukey's Exploratory Data Analysis, Journal of Experimental Analysis of Behavior, 31, 433-440.

Dovoedo Y H, Chakraborti S, (2015). Boxplot-Based Outlier Detection
for the Location-Scale Family, Communication in Statistics – Simulation and Computation, 44, 1492-1513.

Hoaglin D C, Mosteller F, Tukey J W, (1983). Understanding Robust and Exploratory Data Analysis, Wiley, New York.

Hubert M, Vandervieren E, (2008). An adjusted boxplot for skewed distributions, Computational Statistics & Data Analysis, 52, 5186-5201.

Schwertman N C, Owens M A, Adnan R, (2004). A simple more general boxplot method for identifying outliers, Computational Statistics & Data Analysis, 47, 165-174.

Tukey J W, (1977). Exploratory Data Analysis, Addison-Wesley, Reading, MA.