Resistant outlier rules and the non-Gaussian case[*]

Kenneth Carling[†]

September, 1998

Abstract

The techniques of exploratory data analysis include a resistant rule, based on a linear combination of quartiles, for identification of outliers. This paper shows that the substitution of the quartiles with the median leads to a better performance in the non-Gaussian case. The improvement occurs in terms of resistance and efficiency, and an outside rate that is less affected by the sample size.

The paper also studies issues of practical importance in the spirit of robustness by considering moderately skewed and fat tail distributions obtained as special cases of the Generalized Lambda Distribution

Keywords: Asymptotic efficiency, Generalized Lambda Distribution, Kurtosis, Outside rate, Resistance, Skewness, Small-sample bias

[*] Reprinted from Computational Statistics and Data Analysis, Vol 33, 2000, pp 249–258. Homepage http://www.elsevier.com/locate/csda or ScienceDirectTM at http://www.sciencedirect.com

[†] Office of Labor Market Policy Evaluation, Box 513, S-75120 Uppsala, Sweden. This research was completed when the author visited the Department of Statistics at Yale University. The visit was made possible by The Sasakawa Young Leaders' Fellowship fund.

1 Introduction

The ideas and techniques of exploratory data analysis (EDA) have received considerable attention in the statistical literature (see, for example, Tukey, 1977 and the two books edited by Hoaglin, Mosteller, and Tukey, 1983 and 1985) and they are a cornerstone in applied statistical analysis. Among other techniques, EDA provides a rule to screen for potential outliers in symmetric univariate data (cf. the Boxplot rule). The rule is based on selected order-statistics, and the emphasis is on resistance of the rule.

The notion of outliers is vague, and identification of them is not necessarily of interest *per se*. There exists a vast literature on this topic, e.g., Barnett (1978), Barnett and Lewis (1995), Beckman and Cook (1983), Hampel et al. (1986), and aforementioned books by Tukey (1977) and Hoaglin et al. (1983, 1985). Rather than reviewing the contents of these works and the centuries of work they refer to, I will present three examples calling for simple and resistant outlier rules. First, in data obtained from physical measurements, outliers may point at recording or measurement errors, and even suggest calibrating problems with the measuring device or discordance between observers. Second, the existence of outliers may caution the analyst to search for statistical procedures that accommodate outliers, e.g. the analyst may prefer the Kruskal-Wallis procedure rather than the classical F-test procedure for Analysis of Variance. Third, statistics derived from very large data sets of moderate quality, for which a careful examination of each observation is prohibitively expensive, may be stabilized by automatic trimming of outliers.

In this paper I propose an improvement of the Boxplot rule by including the sample *median* in the linear combination of order statistics, rather than only the first and the third quartiles. It will be shown by theoretical reasoning and by simulations that such a slight modification of the rule leads to higher precision and resistance, and a reduction of the bias of the rule in small samples.

This comparative study extends beyond Gaussian data 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.

The paper proceeds as follows. In section 2, the rules are outlined and their resistance and efficiency compared. The outside rate, i.e. the expected proportion of labelled outliers in a non-contaminated sample, has been found in small samples to deviate considerably from its asymptotic counterpart. Section 3 is devoted to this problem. In section 4, a compact representation relates the outside rate to the sample size and other variables in the interest of clarifying miscellaneous issues of practical relevance. A short summary ends the paper.

2 Resistance and efficiency

To focus ideas and simplify notation, consider a sample of size n with observations $x_1, ..., x_m, y_1, ..y_{n-m}$, where $X \sim F$ and F is the distribution of interest. Note that the outliers, i.e. $y_1, ..y_{n-m}$, do not belong to F and they are not necessarily identically distributed. In the following, contamination is taken to mean a situation where $m < n$ and the degree of contamination is $p = (n-m)/n$.

Given a potentially contaminated sample, the conventional way of defining the Boxplot rule, hereafter referred to as Tukey's rule, is

$$c^U = q_3 + k_1 (q_3 - q_1), \tag{1}$$

where q_1 and q_3 are the sample quartiles, of which a precise definition applicable for small samples will be deferred to section 3, and k_1 is a constant selected to meet a pre-specified outside rate under some model. Outsiders are the observations in the sample which exceed in value the upper cut-off point c^U (For simplicity, attention is restricted to outsiders in the right tail). The outside rate in non-contaminated samples, i.e. $p = 0$, of size n will be denoted r_n and the population or asymptotic outside rate r_∞ in the following. In general, the outside rate is inversely related to k_1, whereas the relation to n will be examined below. A slight modification of (1) is

$$c^U = q_2 + k_2 (q_3 - q_1), \tag{2}$$

where q_2 is the sample median. The Median Rule, as it hereafter will be referred to, is a natural competing definition.

Not surprisingly, the Tukey and the Median Rule share several properties. They are both invariant to change of location and scale, and have a breakdown point of roughly 25% (the breakdown point gives the fraction of outliers the estimator can cope with, see Huber 1981). Secondly, asymptotically, they are equivalent for symmetric batches of data if $k_2 =$

$k_1 + 0.5$. Thirdly, the labelling of outliers or the outside rate in small non-contaminated samples is highly dependent on the quantile estimation technique employed and, presumably, the distribution from which the data derives. The latter undesired properties will be addressed in the next two sections.

The first comparison concerns the resistance of the rules. Hampel et al. (1986) discuss the magnitude of the fraction of outliers in batches of data, and make the point that one should not be surprised to find 10 percent of the observations to be erroneous. This insight has implications for the choice of an outlier detection rule, as it is to be expected that small to moderately large samples contain not a single outlier, but rather multiples of outliers. Hence, there is a need for methods for which the presence of an outlier is not masking the existence of itself and other outliers in the same batch. Resistance is taken to mean a rule that does not suffer from the aforementioned problem.

As an operational definition of resistance I consider the deviation in upper cut-off point obtained from a p-contaminated sample with respect to the non-contaminated population counterpart. Formally, I consider the deviation $\left|c^U\left(r_\infty, p\right) - C^U\left(r_\infty, p = 0\right)\right|$ to be the parameter of interest, where $c^U\left(\cdot\right)$ denotes the sample estimate of the upper cut-off point and $C^U\left(\cdot\right)$ the population upper cut-off point. Ideally, the deviation should be small for any distribution and a realistic degree of contamination.

To explore the relative deviation for a broad class of distributions, the Generalized Lambda Distribution, GLD(α_3, α_4), is handy (Ramberg et al., 1979). It permits the skewness and the kurtosis to be varied one at the time. Figure 1 and 2 provide a subset of obtained relative deviation of the Tukey and the Median Rule for varying degrees of kurtosis (figure 1) and skewness (figure 2). In each of the two figures, three lines are provided to show the results for three levels of contamination, $p = 0.01, 0.05, 0.1$. Furthermore, results were obtained for various large sample outside rates, $r_\infty = 0.01(0.01)0.2$, although only the case $r_\infty = 0.01$ is shown in the figure. Figure 1 gives the ratio as a function of kurtosis, $\alpha_4 = 2(1)10$, setting the skewness $\alpha_3 = 0$ and figure 2 gives the ratio as a function of skewness, $\alpha_3 = 0(0.25)2$, for $\alpha_4 = 9$[1]. The general impression is that the Median Rule is quite more resistant than Tukey's rule.

Just as a high resistance ensures that the upper cut-off point c^U be

[1] The choice of $\alpha_4 = 9$ is driven by the desire to show the results for a sequence of values of skewness, yet being within the admissible combination of skewness of kurtosis.

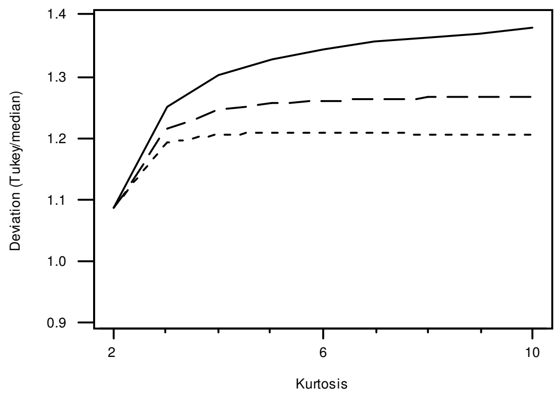

Fig 1. Deviation (Tukey/median). The ratio as a function of kurtosis for three levels of contamination, p. The top function refers p=0.01, below is p=0.05,0.1.

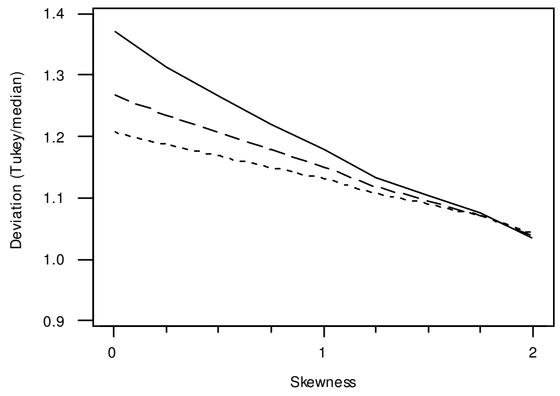

Fig 2. The three lines shows the ratio of deviation (Tukey/median) for three levels of contamination, 0.01, 0.05, 0.1 starting from the top of the figure

insensitive to outliers, thereby avoiding masking, it is desired that c^U be precisely determined from the bulk of the data. Such a stability of c^U carries over to the trimmed mean or whatever other statistic is sought for the trimmed data. As the second comparison of the rules, the asymptotic variation of the cut-off point, c^U, is derived by using the following result

(David, 1981),

$$\begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \sim N \left\{ \begin{pmatrix} Q_1 \\ Q_2 \\ Q_3 \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \frac{3}{16f_1^2} & \frac{2}{16f_1f_2} & \frac{1}{16f_1f_3} \\ \frac{2}{16f_1f_2} & \frac{4}{16f_2^2} & \frac{2}{16f_2f_3} \\ \frac{1}{16f_1f_3} & \frac{2}{16f_2f_3} & \frac{3}{16f_3^2} \end{pmatrix} \right\}. \tag{3}$$

Here, upper-case letters are used to indicate population parameters, and lower-case letters to indicate sample estimates of the parameters, and $f_j, (j = 1, .., 3)$, is the value of the density function evaluated at Q_j. As it stands, it is possible to derive the asymptotic relative efficiency, $ARE = V\left(c_2^u\right)/V\left(c_1^u\right)$, of the cutoffs for the two rules. Let C_1^U denote the population upper cutoff provided by the first rule, and C_2^U the population upper cutoff provided by the second. Since c_1^U and c_2^U are linear functions of the sample quartiles, they are also consistent estimates of the population parameters and follow a normal distribution asymptotically. Hence, the comparison can focus on their asymptotic variance. The asymptotic variances are

$$V\left(c_1^U\right) = \frac{1}{16n}\left\{\frac{3\left(1+k_1\right)^2}{f_3^2} + \frac{3k_1^2}{f_1^2} - \frac{2k_1\left(1+k_1\right)}{f_1f_3}\right\} \tag{4}$$

and

$$V\left(c_2^U\right) = \frac{1}{16n}\left\{\frac{4}{f_2^2} + \frac{3k_2^2}{f_3^2} + \frac{3k_2^2}{f_1^2} + \frac{4k_2}{f_2f_3} - \frac{4k_2}{f_1f_2} - \frac{2k_2^2}{f_1f_3}\right\}. \tag{5}$$

These expressions can be evaluated given a density function, f, and by considering a population outside rate, r_∞, being deterministically and inversely related to k_2 and k_1.

Once again, the ARE is studied by use of the Generalized Lambda Distribution. Focusing first on the impact of kurtosis, figure 3 gives ARE as a function of kurtosis, $\alpha_4 = 2(1)10$, setting $\alpha_3 = 0$. The figure clearly indicates that the rules are equally efficient for symmetric batches of data. This holds true for $r_\infty \in (0, 0.2)$, although only the natural choice of an outside rate $r_\infty = 0.01$ is shown in the figure.

Shifting the focus to the impact of skewness in the data, figure 3 also shows by two lines ARE as a function of skewness, $\alpha_3 = 0(0.25)2$, setting $\alpha_4 = 9$. As might be expected, the relative efficiency of Tukey's rule decreases as the skewness in the data becomes more pronounced. The figure

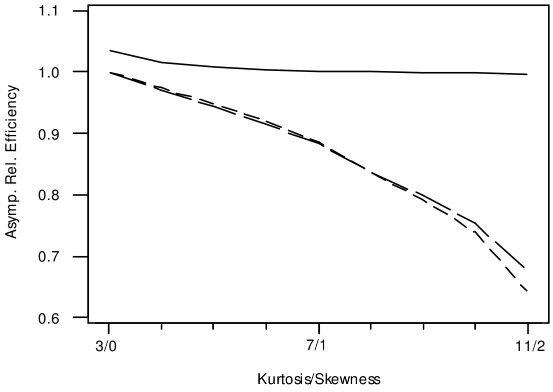

Fig 3. The solid line shows ARE as a function of kurtosis. The dashed/dotted lines show ARE as a function of skewness for r=0.01,0.05 resp.

shows the case where $r_\infty = 0.01, 0.05$. The asymptotic variance calculations provide a very good insight as to what to expect in finite samples, which is of more direct interest. I pursued small sample simulations to confirm this claim. In short, the Median Rule had an equal or smaller variance than the Tukey rule, with the notable exception of symmetric distributions with a low value of the kurtosis for which case the Tukey rule was slightly more efficient. However, in finite samples the issue of bias arises and it will be the topic of the next section.

3 A comparison of the small sample outside rate

It is well known, in the application of Tukey's rule, that if $k_1 = 1.5$ and the non-contaminated sample comes from a Gaussian distribution, then about 0.7 percent will be labelled outliers (Hoaglin, Iglewicz, and, Tukey, 1986), provided that the sample is large. However, they have also shown that for small sample sizes the outside rate may be as high as 10 percent. As another example, Kimber (1990) finds empirically that the upper outside rate per observation, r_n, is approximately bounded by

$$100r_n \leq 4.8 + 17.5n^{-1} \tag{6}$$

for exponential data and a choice of k being 1.5. This finding is quite noteworthy since it implies that even for a sample size being as large as

100 observations, one can not hope large sample results to hold particularly well. It goes without saying that it is an unattractive feature of the outlier rule, and there is a need for a substantial reduction of the dependence on the sample size to make the rule useful. Fortunately, this may be possible in part as will be discussed in section 4.

The dependence on the sample size for the outside rate will also be the subject for a comparison of the two rules. Particularly, the rules in (1) and (2) will be compared by an approximation to the *upper outside rate per observation*, r_n, for a given r_∞ and sample size n. The statistic, which is the expected rate of outsiders found in a non-contaminated random sample of size n, was first used by Kimber (1990). However, the outside rate per observation has often been used in studies of symmetric distributions (Brant, 1990, Hoaglin et al., 1986). Before proceeding, however, the important question on how to define sample quartiles must be addressed.

Equations (1) and (2) may look simple, but the definition of sample quartiles has long been debated (see Cleveland, 1985, Freund and Perles, 1987, Frigge, Hoaglin, and Iglewicz, 1989, Harrell and Davis, 1982, Hoaglin et al., 1983, Hoaglin and Iglewicz, 1987, Hyndman and Fan, 1996). Frigge, Hoaglin, and Iglewicz, (1989) give eight definitions which have been used in various statistical softwares and similar contexts. Most of these definitions can be represented as

$$q_1 = (1-g)\,x_{(j)} + gx_{(j+1)}, \tag{7}$$

where $x_{(j)}$ and $x_{(j+1)}$ are the j:th and the $(j+1)$:th ordered observations. The controversy is in the choice of j and g, i.e. if and how interpolation should be done. The Boxplot originally used $j+g = [(n+3)/2]/2$ ($[x]$ denotes the largest integer that does not exceed x), where j equals the integer part of the ratio and g the remaining fraction (Tukey, 1977). However, Frigge et al. (1989) recommend the use of the *ideal* or *machine fourth*, for which $j+g = n/4 + 5/12$. As an example, figure 4 shows the r_n obtained by the original definition of fourth and by the ideal fourth. The introduction of the ideal fourth was an important improvement since the bias is decreased and the remaining bias is a smooth function of n. Hence, the following results will build on the ideal fourth.

Because, in small samples, the outside rate is not amenable to analytic calculation, simulations will be employed. Let M be the number of replicates of size n drawn from the GLD(α_3, α_4), then the estimated upper

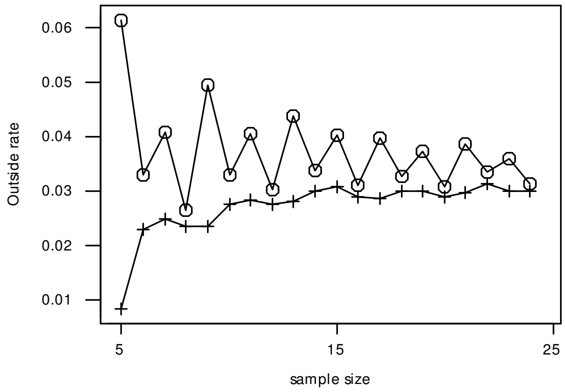

Figure 4: An example of the small sample outside rate's, r_n, relation to the sample size for two definitions of the sample quartiles. The circles refer to Tukey's definition of the fourth and the plus-symbols refer to the *ideal* fourth. For the example, GLD(0,3), which comes very close to the standard normal distribution, was used and the asymptotic outside rate, r_∞, was taken to be 0.03.

outside rate is

$$\widehat{r}_n = \frac{\sum_{m=1}^{M} \sum_{i=1}^{n} I\left\{x_{i,m} > c_m^U\left(r_\infty\right)\right\}}{nM}, \tag{8}$$

where $x_{i,m}$ refers to observation i in replicate m, and $I\left\{\cdot\right\}$ is an indicator function taking on unity if true, and zero otherwise. In the experiments, $M = [180000/n]$ samples are drawn and the number of outside observations in each sample is determined. For each sample and sample size $n = 6(1)25(5)50(10)100(50)300$, the finite sample outside rate, r_n, is estimated for $r_\infty = 0.01(0.01)0.2$, $\alpha_3 = 0(0.25)2$, and $\alpha_4 = 2(1)10$, for admissible combinations of skewness and kurtosis. The range of parameters to control skewness and kurtosis is sufficient for encompassing many of the commonly used probability distributions, or at least to closely approximate them.

The aim is to find a reasonably simple function that approximates $r_n -$

r_∞. In exploring the possibilities of expanding $[r_n - r_\infty]$ by linearization, along the lines of David (1981), I found it reasonable to consider the bias to be inversely related to n, yet the impact of the higher order terms was difficult to determine analytically. Hence, I relied on regression analysis to approximate the bias. The following functions provides the main results, as a percentage of bias, where superscript t refers to Tukey's rule and m to the Median Rule

$$100\frac{r_n^t - r_\infty}{r_\infty} = -\frac{215.3}{n} \tag{9}$$

$$100\frac{r_n^m - r_\infty}{r_\infty} = -\frac{162.2}{n}. \tag{10}$$

In either case R^2 was found to be 90%. The fit can be improved by including the terms r_∞ and r_∞^2, in which case $R^2 = 0.97$. The inclusion will not affect the comparison of the two rules and the terms are excluded in the interest of simplicity. In comparing the two rules, it is worth noting that the bias-percentage does not depend on the distribution. I was unable to find any relation between the bias-percentage and the skewness and the kurtosis of the distributions, despite a considerable effort to seek such relations.

4 What determines r_n and what should k_2 be?

Equations (1) and (2) include the constants k_1 and k_2 on which I have been silent. Although none of the above cited works has claimed a particular choice of k_1 to be optimal in some sense, one often sees $k_1 = 1.5$ being employed. The rationale is that under the Gaussian model, $k_1 = 1.5$ implies that $r_\infty \approx 0.035$, i.e. a very small fraction of observations in a large non-contaminated sample would be erroneously labelled outliers. By studying the relationship between r_n and the broad class of distributions encompassed by GLD(α_3, α_4), perhaps a more compelling argument could be given for a suitable choice of k_2. As a side remark, recall that under symmetry it is natural to take $k_2 = k_1 + 0.5$.

Relying on the simulated data in the previous section, I obtained the reasonably good and stable fit

$$\begin{aligned} 100r_n &= -8.07 + \frac{3.71}{n} + \frac{17.63}{k_2} - \frac{23.64}{nk_2} + 0.83\alpha_3 + 0.48\alpha_3^2 \\ &\quad +0.48\left(\alpha_4 - 3\right) - 0.04\left(\alpha_4 - 3\right)^2, \end{aligned} \tag{11}$$

for which $R^2 = 99.2\%$. Moreover, the distribution of the relative error, i.e. the ratio of the residual and the outside rate, is accurately and compactly summarized by $N(0, \sigma = 0.004)$. Hence, the regression curve gives a predicted outside rate which rarely errors by more than one percent, in either direction. On the other hand, the curve may predict negative outside rates for large values of k_2, a spurious result due to the fact that $r_\infty < 0.01$ was not included in the experimental setting.

Now, because of the choice of designing the experiment to include only symmetric and right skewed distributions, one has that $2r_n^U \geq r_n^U + r_n^L$, where the superscripts U and L refer to upper and lower outside rates respectively, and strict equality holds under symmetry. It makes sense to specify the desired asymptotic outside rate over both tails and derive the appropriate value of k_2, and then perform a calculation to adjust for the sample size. Consider as an example the choices $\alpha_3 = 0.5$, $\alpha_4 = 5$, and a target outside rate in both tails of 0.02. In this case, k_2 is found to equal about 2.3 and the outside rate has a lower bound for the Gaussian case at 0.2%. To preserve this property for a given sample size n, one takes

$$k_2 = \frac{17.63n - 23.64}{7.74n - 3.71}.$$

Obviously, the function in (11) lends itself to consider an arbitrary distribution appropriate for a specific application, yet, in the absence of compelling subject-matter information, I find it reasonable to let $k_2 = 2.3$ be the default choice with the suitable adjustment for the sample size.

5 Summary

The results of the paper show that the Median Rule performs better than Tukey's rule with respect to the criteria being studied. I have been unable to find any other arguments in favor of Tukey's rule, and hence an implementation of the Median Rule seems warranted. Moreover, I have been able to provide additional light on a suitable choice for the constant, k_2, in the rule, which ought to be of practical help. Finally, the suggestion made by Hoaglin et al. (1987) to consider the ideal fourths as the sample estimate of the quartiles leads to a dramatic improvement in both of the rules.

References

[1] Barnett, V., (1978), "The study of outliers: Purpose and model", *Applied Statistics*, 27, 242-250.

[2] Barnett, V., and Lewis, T., (1995) "Outliers in statistical data, 3ed.", Wiley, Chicester, England.

[3] Beckman, R.J., and Cook, R.D., (1983), "Outliers.......s", *Technometrics*, 25, 119-163.

[4] Brant, R., (1990), "Comparing classical and resistant outlier rules", *Journal of the American Statistical Association*, 85, 1083-1090.

[5] Cleveland, W.S., (1985), "The elements of graphing data", Hobart Press, Summit, New Jersey.

[6] David, H.A., (1981), "Order Statistics", 2nd ed., Wiley, New York.

[7] Freund, J.E., and Perles, B.M., (1987), "A new look at quartiles of ungrouped data", *The American Statistician*, 41, 200-203.

[8] Frigge, M., Hoaglin, D.C., and Iglewicz, B., (1989), "Some implementations of the Boxplot", *The American Statistician*, 43, 50-54.

[9] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A., (1986), "Robust Statistics: The Approach Based on Influence Functions", Wiley, New York.

[10] Harrell, F.E., and Davis, C.E., (1982), "A new distribution-free quantile estimator", *Biometrika*, 69, 635-640.

[11] Hoaglin, D.C., and Iglewicz, B., (1987), "Fine-tuning some resistant rules for outlier labelling", *Journal of the American Statistical Association*, 82, 1147-1149.

[12] Hoaglin, D.C, Iglewicz, B., and Tukey, J.W., (1986), "Performance of some resistant rules for outlier labelling", *Journal of the American Statistical Association*, 81, 991-999.

[13] Hoaglin, D.C., Mosteller, F., and Tukey, J.W., (1983), "Understanding robust and exploratory data analysis", Wiley, New York.

[14] Hoaglin, D.C., Mosteller, F., and Tukey, J.W., (1985), "Exploring data tables, trends and shapes", Wiley, New York.

[15] Huber, P.J., (1981), "Robust statistics", Wiley, New York.

[16] Hyndman, R.B., and Fan, Y., (1996) "Sample quantiles in Statistical Packages", *The American Statistician*, 50, 361-365.

[17] Kimber, A.C., (1990), "Exploratory data analysis for possibly censored data from skewed distribution", *Applied Statistics*, 39, 21-30.

[18] Ramberg, J.S., Tadikamalla, P.R., Dudewicz, E.J., and Mykytka, E.F., (1979), "A Probability Distribution and Its Uses in Fitting Data", Technometrics, 21, 201-214.

[19] Tukey, J.W., (1977), "Exploratory data analysis", Addison-Wesley.