S E C T I O N O F S T A T I S T I C S

DEPARTMENT OF MATHEMATICS

KATHOLIEKE UNIVERSITEIT LEUVEN

T E C H N I C A L R E P O R T

TR-06-11

AN ADJUSTED BOXPLOT FOR SKEWED DISTRIBUTIONS

Hubert, M., Vandervieren, E.

http://wis.kuleuven.be/stat/

An Adjusted Boxplot for Skewed Distributions

Mia Hubert
Department of Mathematics - UCS,
Katholieke Universiteit Leuven,
W. de Croylaan 54, B-3001 Leuven, Belgium
E-mail: mia.hubert@wis.kuleuven.be

Ellen Vandervieren
Department of Mathematics & Computer Science,
University of Antwerp,
Middelheimlaan 1, B-2020 Antwerp, Belgium
E-mail: ellen.vandervieren@ua.ac.be

10 November, 2006

Abstract: The boxplot is a very popular graphical tool to visualize the distribution of continuous unimodal data. It shows information about the location, spread, skewness as well as the tails of the data. However, since the fences are derived from the normal distribution, usually too many points are classified as outliers when the data are skewed. We present an adjustment of the boxplot that includes a robust measure of skewness in the determination of the whiskers. We show with several examples and simulation results that this adjusted boxplot gives a more accurate representation of the data and of possible outliers.

Key words: Boxplot, Skewness, Medcouple

1. INTRODUCTION

One of the most frequently used graphical techniques for analyzing a univariate data set is the *boxplot*, proposed by Tukey (1977). If $X_n = \{x_1, x_2, \ldots, x_n\}$ is a univariate data set, the

boxplot is constructed by

- putting a line at the height of the sample median Q_2
- drawing a box from the first quartile Q_1 to the third quartile Q_3
- classifying all points outside the interval (the fence)

$$[Q_1 - 1.5 \text{ IQR} \; ; \; Q_3 + 1.5 \text{ IQR}] \tag{1}$$

  as outlier and marking them on the plot
  (where $\text{IQR} = Q_3 - Q_1$)

- drawing the whiskers (i.e. the lines that go from the ends of the box to the most remote points that are no outliers).

The boxplot thus shows information about the location and the spread of the data by means of the median and the interquartile range. The length of the whiskers on both sides of the box and the position of the median within the box are helpful to detect possible skewness in the data. Finally, observations that fall outside the fences are pinpointed as outliers, hence the boxplot also includes information from the tails. However, in some cases the information about the tails given by the boxplot, is not reliable.

As an example we consider the time intervals between coal mining disasters (Jarret 1979). This data set contains 190 time intervals, measured in days, between explosions in coal mines from 15th March 1851 to 22nd March 1962 inclusive. From the boxplot of these data, shown in Figure 1, it can be seen that the underlying distribution of the data set is skewed to the right. However, the upper whisker is rather short, which causes a lot of observations (more precisely 6.84%) to be classified as 'right' outlier. Clearly, this percentage of right outliers is not realistic. A boxplot with a longer upper whisker and less potential right outliers would give a more accurate representation of the data.

This phenomenon, which often occurs at skewed distributions, was already mentioned in Hoaglin, Mosteller and Tukey (1983). If the data come for example from a χ^2_1-distribution, the probability to exceed the lower fence is zero, whereas it can be expected that 7.56% of the (regular) data exceed the upper fence. Similarly, we can expect a 7.76% upper exceedance probability at the lognormal distribution. For the normal distribution on the other hand, the expected exceedance percentage is only 0.7%, i.e. 0.35% on both sides of the distribution.

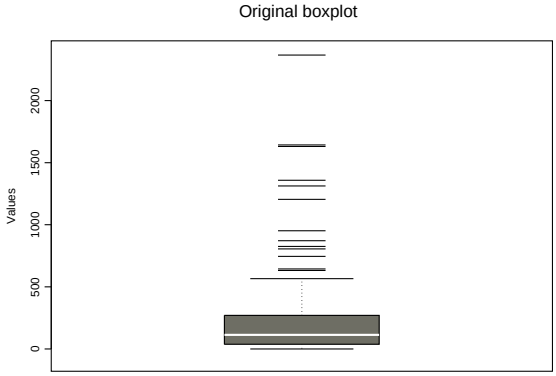


Figure 1: Original boxplot of the time intervals between coal mining disasters.

The large discrepancy between these percentages is caused by the fact that the fences of the boxplot are solely based on measures of location and scale, and that the cutoff values are derived from the normal distribution.

To circumvent these problems several modifications to the boxplot have been proposed in the literature. Some authors have adjusted the fences towards skewed data by use of the semi-interquartile range $Q_2 - Q_1$ and $Q_3 - Q_2$, i.e. they define the fences

$$[Q_1 - k_1(Q_2 - Q_1); Q_3 + k_2(Q_3 - Q_2)] \tag{2}$$

for some constants k_1 and k_2. In Kimber (1990) both constants are set to 1.5, while Aucremanne et al. (2004) use $k_1 = k_2 = 3$. However, no justification for these choices can be found. Schwertman, Owens and Adnan (2004) present a method to determine k_1 and k_2 based on the required exceedance percentage. Their method has the strong limitation that the resulting constants are based on the expected value of the quartiles, which are in general not known in advance. Only for normal or almost normal data, they provide fixed values.

In Carling (2000), the quartiles Q_1 and Q_3 in (1) are replaced with the median Q_2, and the constant 1.5 is changed by a formula that includes the third and fourth moment of the presumed distribution. This is an interesting approach, but it has the serious drawback that the skewness and kurtosis of the distribution need to exist and should be known in advance. In exploratory data analysis these moments have to be estimated, but this issue is not discussed by the author.

We propose an adjustment to the boxplot that can be applied to all distributions, even without finite moments. Moreover, we estimate the underlying skewness with a robust

measure, to avoid masking the real outliers. The structure of this paper is as follows. In Section 2, we present our generalization of the boxplot that includes a robust measure of skewness in the determination of the whiskers. To construct this adjusted boxplot we will derive new outlier rules at the *population* level. To draw the boxplot at a particular data set, we then just need to plug in the finite-sample estimates. In Section 3, we apply the new boxplot to several real data sets, whereas in Section 4 a simulation study is performed at uncontaminated as well as contaminated data sets. Section 5 concludes and gives directions for future research.

2. SKEWNESS ADJUSTMENT TO THE BOXPLOT

2.1 A robust measure of skewness

To measure the skewness of a continuous unimodal distribution F, we use the *medcouple* (MC), introduced in Brys, Hubert and Struyf (2003). It is defined as

$$\mathrm{MC}(F) = \underset{x_1<Q_2<x_2}{\mathrm{med}} h(x_1, x_2)$$

with x_1 and x_2 sampled independently from F, Q_2 the median of F and the kernel function h given by

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i}.$$

This definition is inspired by the quartile skewness (QS), introduced in Bowley (1920) and Moors et al. (1996), defined as

$$\mathrm{QS} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

with as before $Q_1 = F^{-1}(0.25)$ and $Q_3 = F^{-1}(0.75)$.

It clearly follows from its definition that the medcouple always lies between -1 and 1. A distribution that is skewed to the right has a positive value for the medcouple, whereas the MC becomes negative at a left skewed distribution. Finally, a symmetric distribution has a zero medcouple. As shown in Brys et al. (2003), this robust measure of skewness has a bounded influence function and a breakdown value of 25%, which means that at least 25% outliers are needed to obtain a medcouple of +1 (or -1). Besides, the MC turned out to be the overall winner when comparing it to two other robust skewness measures which are

solely based on quantiles, namely the QS and the octile skewness (OS), given by

$$\mathrm{OS} = \frac{(Q_{0.875} - Q_2) - (Q_2 - Q_{0.125})}{Q_{0.875} - Q_{0.125}}$$

with $Q_{0.875} = F^{-1}(0.875)$ and $Q_{0.125} = F^{-1}(0.125)$.

The MC combines the strengths of OS and QS: it has the sensitivity of OS to detect skewness and the robustness of QS towards outliers. For the computation of the medcouple, a fast algorithm of $\mathrm{O}(n \log n)$ time has been constructed, and Matlab and S-PLUS functions are available from the authors website.

2.2 Incorporating skewness into the boxplot

In order to make the original boxplot skewness-adjusted, we incorporate the medcouple into the definition of the whiskers. This can be done by introducing some functions $h_l(\mathrm{MC})$ and $h_u(\mathrm{MC})$ into the outlier cutoff values. Instead of using the fence

$$[Q_1 - 1.5 \text{ IQR} \; ; \; Q_3 + 1.5 \text{ IQR}],$$

we propose the boundaries of the interval to be defined as

$$[Q_1 - h_l(\mathrm{MC}) \text{ IQR} \; ; \; Q_3 + h_u(\mathrm{MC}) \text{ IQR}].$$

Additionally, we require that $h_l(0) = h_u(0) = 1.5$ in order to obtain the original boxplot at symmetric distributions. As the medcouple is location and scale invariant, this interval is location and scale equivariant. Note that by using different functions h_l and h_u, we allow the fences to be asymmetric around the box, so that adjustment for skewness is indeed possible. Also remark that at the population level, no distinction should be made between the whiskers and the fences. At finite samples on the other hand, the whiskers are drawn up to the most remote points before the fences.

Three different models have been studied, namely a

1. *linear model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5 + a \text{ MC} \\ h_u(\mathrm{MC}) &= 1.5 + b \text{ MC} \end{aligned} \tag{3}$$

2. *quadratic model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5 + a_1\ \mathrm{MC} + a_2\ \mathrm{MC}^2 \\ h_u(\mathrm{MC}) &= 1.5 + b_1\ \mathrm{MC} + b_2\ \mathrm{MC}^2 \end{aligned} \tag{4}$$

3. *exponential model*:

$$\begin{aligned} h_l(\mathrm{MC}) &= 1.5\ e^{a\mathrm{MC}} \\ h_u(\mathrm{MC}) &= 1.5\ e^{b\mathrm{MC}} \end{aligned} \tag{5}$$

with $a, a_1, a_2, b, b_1, b_2 \in \mathbb{R}$. Note that each of these models is simple and contains only a few parameters. This is very important for exploratory data analysis.

2.3 Determination of the constants

In order to find good values for a, a_1, a_2, b, b_1 and b_2, we fit a whole range of distributions and try to define the fences such that the expected percentage of marked outliers is close to 0.7%, which coincides with the outlier rule of the original boxplot at the normal distribution. At the linear model (3), this implies that the constants a and b should satisfy

$$\begin{cases} Q_1 - (1.5 + a\ \mathrm{MC})\ \mathrm{IQR} \approx Q_\alpha \\ Q_3 + (1.5 + b\ \mathrm{MC})\ \mathrm{IQR} \approx Q_\beta \end{cases}$$

where in general Q_p denotes the pth quantile of the distribution, $\alpha = 0.0035$ and $\beta = 0.9965$. The previous system can be rewritten as

$$\begin{cases} \frac{Q_1 - Q_\alpha}{\mathrm{IQR}} - 1.5 \approx a\ \mathrm{MC} \\ \frac{Q_\beta - Q_3}{\mathrm{IQR}} - 1.5 \approx b\ \mathrm{MC}. \end{cases}$$

Linear regression without intercept can then be used to obtain estimates of the parameters a and b. The parameter determination at the quadratic and at the exponential model is analogous to that of the linear case. At the exponential model we obtain the linear system

$$\begin{cases} \ln(\frac{2}{3}\ \frac{Q_1 - Q_\alpha}{\mathrm{IQR}}) \approx a\ \mathrm{MC} \\ \ln(\frac{2}{3}\ \frac{Q_\beta - Q_3}{\mathrm{IQR}}) \approx b\ \mathrm{MC} \end{cases}$$

so that again linear regression without intercept can be applied.

To derive the constants we used 12,605 distributions from the family of Γ, χ^2, F, Pareto and G_g-distributions (Hoaglin et al. 1985). More precisely, we used $\Gamma(\beta, \gamma)$ distributions

with scale parameter $\beta = 0.1$ and shape parameter $\gamma \in [0.1; 10]$, χ^2_{df} distributions with $df \in [1; 30]$, F_{m_1,m_2} distributions with $(m_1, m_2) \in [1; 100] \times [1; 100]$, Pareto distributions $Par(\alpha, c)$ with $c = 1$ and $\alpha \in [0.1; 20]$, and G_g-distributions with $g \in [0; 1]$.

The parameters of the distributions were selected such that the medcouple did not exceed 0.6. Doing so, we retained a large collection of distributions that are not extremely skewed. It appeared that constructing one good and easy model that also includes the cases with MC > 0.6 is hard to find, hence we only concentrated on the more common distributions with moderate skewness. Note that we only considered symmetric and right skewed distributions, as the boundaries just need to be switched for left skewed distributions.

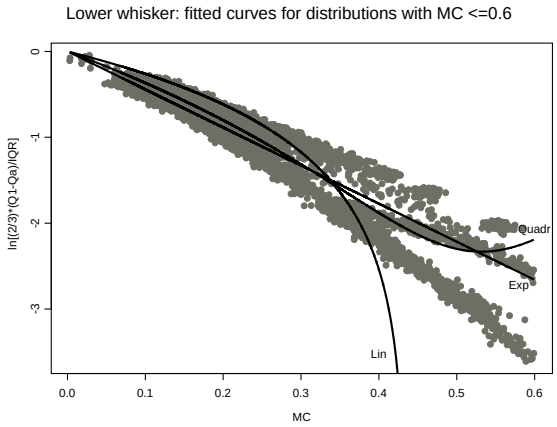
To obtain the population values of the medcouple and the quartiles at all these distributions, we generated 10,000 observations from each of them, and used their finite-sample estimates as the true values.

In Figure 2(a) we show the fitted regression curves for the lower whisker, after applying LS regression for the linear, quadratic and exponential model and based on the whole set of distributions we considered. For reasons of clarity, we have set on the vertical axis the response value of the exponential model, which is $\ln(\frac{2}{3}\frac{Q_1 - Q_\alpha}{IQR})$. Hence, only the exponential fit is presented by a straight line. Figure 2(b) only displays the G_g distributions (with the same fits superimposed). Figures 3(a) and 3(b) show analogous results for the upper whisker.

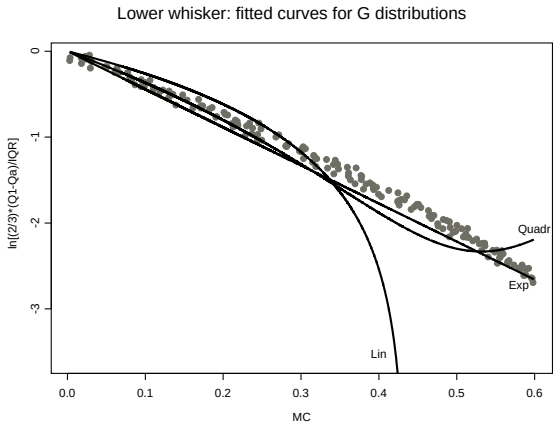
2.4 The adjusted boxplot

From Figure 2 and Figure 3 it can be seen that the linear model completely fails to determine accurate lower whiskers, whereas the exponential and quadratic model perform much better. For the upper whiskers, the quadratic model gives less accurate estimates than the exponential model. Consequently, as the exponential model is appropriate for both the left and the right tail, we will use the *exponential model* in the definition of our adjusted boxplot, rather than the quadratic model. Also remark that the exponential model only includes one parameter (on each side), which makes it more simple than the quadratic model.

Although the exponential fit will produce an underestimate of Q_α (respectively Q_β) for some distributions, the same quantile will be overestimated for others. Consequently it gives a good compromise for the whole set of distributions we considered. If we would have a priori information of the distribution, for example, we would know that it belongs to the class of G_g distributions, it is clear from Figure 2(b) and Figure 3(b) that a more specific



(a)



(b)

Figure 2: Lower whisker: regression curves for the linear, quadratic and exponential model.

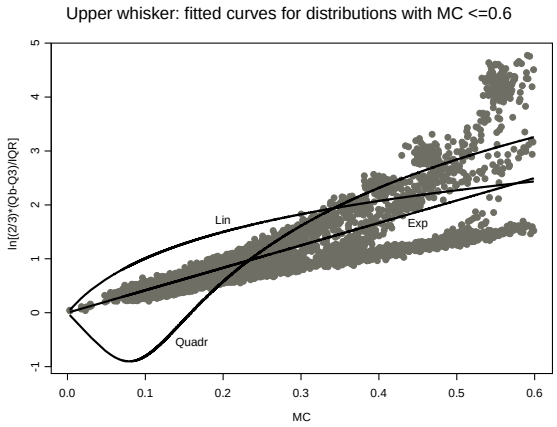
model could be constructed, for example by including results from extreme value theory (see e.g. Vandewalle, Beirlant, and Hubert 2004).

To ease the model and for robustness reasons, we rounded off the estimated values of the exponential model $a = -3.79$ and $b = 3.87$ to $a = -4$ and $b = 3$. To summarize we thus can say that when $\mathrm{MC} \geq 0$, all observations outside the interval

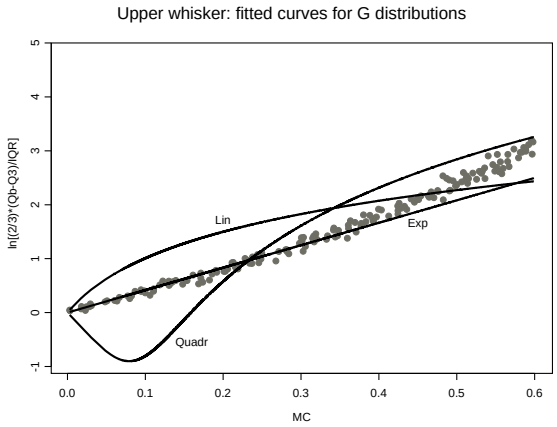
$$[Q_1 - 1.5\ e^{-4\mathrm{MC}}\ \mathrm{IQR}\ ;\ Q_3 + 1.5\ e^{3\mathrm{MC}}\ \mathrm{IQR}] \tag{6}$$

will be marked as potential outlier. For $\mathrm{MC} < 0$, the interval becomes

$$[Q_1 - 1.5\ e^{-3\mathrm{MC}}\mathrm{IQR}\ ;\ Q_3 + 1.5\ e^{4\mathrm{MC}}\mathrm{IQR}].$$



(a)



(b)

Figure 3: Upper whisker: regression curves for the linear, quadratic and exponential model.

Note that while this adjusted boxplot accounts for skewness, it does not yet account for tail heaviness. This could be done by including tail information of the distribution as well. We could for example try to construct a model which includes robust measures of left and right tail, such as those proposed in Brys, Hubert and Struyf (2006). We see however several disadvantages of such a procedure. First of all, the model would become more complex with more estimators and parameters. The robustness would decrease as the tail measures have a lower breakdown value, and the variability of the whisker's length would increase, due to the variability of the tail measures.

3. EXAMPLES

3.1 Coal mine data

We recall the coal mine data from the introduction where we have illustrated that the original boxplot marks too many observations as outlier. Figure 4 shows both the original and the

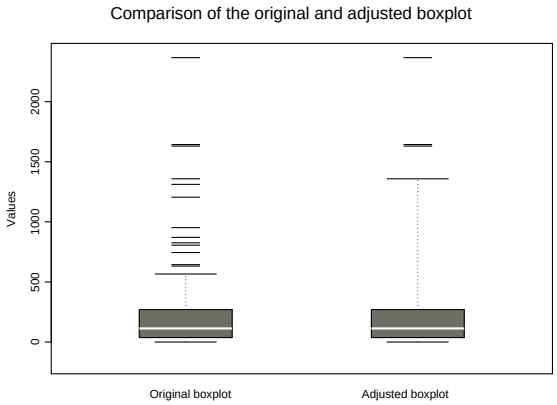


Figure 4: Coal mine data: a comparison of the original and the adjusted boxplot.

adjusted boxplot, obtained using our S-PLUS code. It can be seen that the adjusted boxplot yields a more accurate representation of the data. The upper whisker has become larger and now better reflects the skewness of the underlying distribution. Besides, it causes less observations to be marked as upper outlier.

3.2 Condroz data

The Condroz data (Goegebeur, Planchon, Beirlant and Oger 2005) contain the pH-value and the Calcium (Ca) content in soil samples, collected in different communities of the Condroz region in Belgium. As in Vandewalle et al. (2004), we focus on the subset of 428 samples with a pH-value between 7.0 and 7.5.

From the normal quantile plot in Figure 5 it can be seen that the distribution of Calcium is right skewed. This also follows from the MC value, which equals 0.16. Besides, we notice 6 upper outliers and 3 lower outliers in Figure 5. In Vandewalle et al. (2004), they were also identified as such, based on a robust estimator of the tail index. The outliers appeared to be measurements from communities at the boundary of the Condroz region and hence, can be considered to be sampled from another distribution.

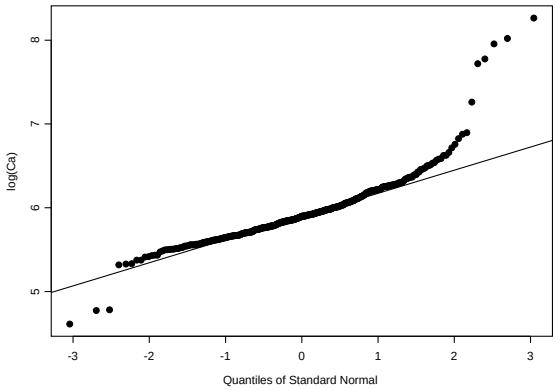


Figure 5: QQ normal quantile plot of the Condroz data with pH between 7.0 and 7.5.

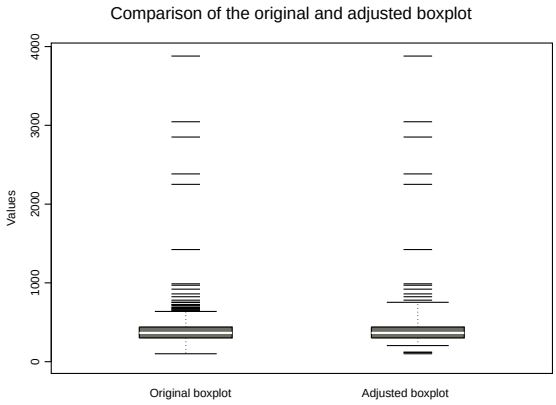


Figure 6: Original and adjusted boxplot of the Condroz data.

We see from the original boxplot in Figure 6 that a substantial number of observations are exceeding the upper whisker, leading to a black box in which the 'outlying' observations can no longer be recognized. Moreover, none of the cases is indicated as a lower outlier. Another visualisation of the data is given in the index plot in Figure 7. Full lines were drawn at the median, the first and third quartile of the data. The dotted lines refer to the whiskers of the original boxplot. We see that 20 of the regular data points are marked as upper outliers. The adjusted boxplot, shown in Figure 6, has a longer upper whisker and a shorter lower whisker. The skewness of the underlying distribution is more pronounced and the shorter left tail is better reflected. Now, less observations exceed the upper cutoff, whereas the three smallest cases are marked as lower outliers. Note that only two marks are visible, as two of the three smallest observations almost coincide as can be seen on the index

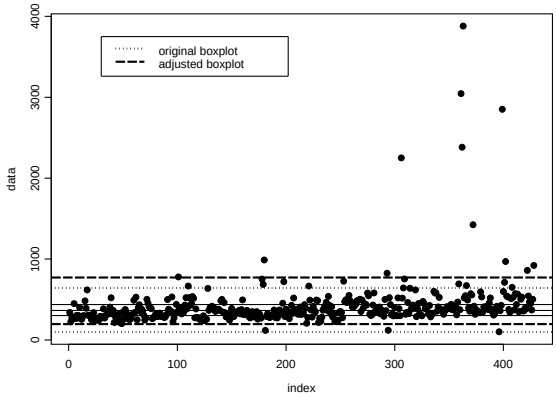


Figure 7: Plot of the Ca measurements versus their index. Full lines were drawn at the median, the first and third quartile. Dashed lines were used to indicate the boundaries of the adjusted boxplot. The dotted lines refer to the boundaries of the original boxplot.

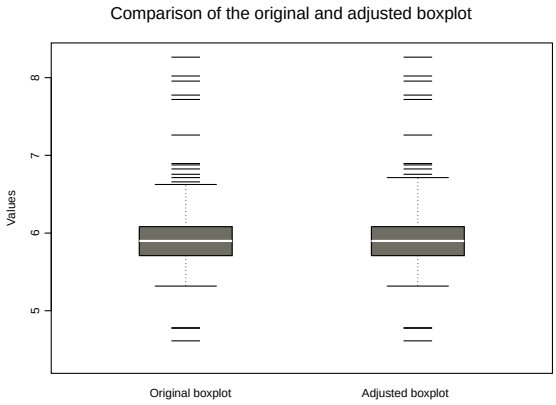


Figure 8: Original and adjusted boxplot of the log-transformed Condroz data.

plot in Figure 7. Here, the dashed lines refer to the whiskers of the adjusted boxplot.

As the data are highly skewed, it is common practice to apply a log transformation to the data to make them more symmetric. The resulting boxplots are shown in Figure 8. We see that the original and adjusted boxplot do not differ very much as the MC of the log transformed Calcium values equals only 0.044. We also notice that the original boxplot applied to the transformed data now shows the same outliers as the adjusted boxplot of the raw data. From this example, one could conclude that the adjusted boxplot is not needed at all and that alternatively, first a transformation could be applied, after which the whiskers could be retransformed to the original unit scale. This approach would certainly work out

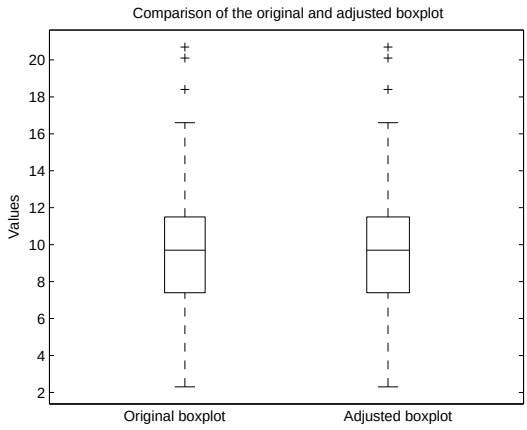


Figure 9: Air data: a comparison of the original and the adjusted boxplot in Matlab.

well in many examples, but it has the drawback that first an appropriate symmetrizing transformation has to be found.

3.3 Air data

In Figure 8 we already noticed that the adjusted boxplot is very similar to the original boxplot when data are nearly symmetric. Here, we consider another example, the wind speed variable from the air data (Chambers and Hastie 1992), measuring the wind speed (in miles per hour) for 111 consecutive days. We now use our Matlab implementation to compare the original and the adjusted boxplot, depicted in Figure 9. As MC = 0.012, we see that both boxplots are equal. Note that the small value of MC slightly effects the fences, as defined in (6), but here does not effect the whiskers as they are drawn to the largest (smallest) non-outlier. This example thus again illustrates that we can consider the adjusted boxplot as a generalization of the original boxplot towards skewed distributions.

3.4 Length of Stay data

Our next example is concerned with data of 201 patients, who stayed in the University Hospital of Lausanne in the year 2000. The data are kindly provided by A. Marazzi (Institute of Social and Preventive Medicine, Lausanne). One of the main objectives is to estimate and predict the total resource consumption of this group of patients (Ruffieux, Paccaud and Marazzi 2000). For this purpose, one can focus on the variable 'length of stay' (LOS) in days, which is an easily available indicator of hospital activity and is used for various

purposes, such as management of hospital care, quality control, appropriateness of hospital use and hospital planning. The most natural way to compute an estimate of the expected LOS, is to use the arithmetic mean. However, the underlying distribution of the LOS data has two features, which make the use of this simple statistic questionable. First of all, the distribution of the LOS data is skewly distributed to the right, as can be seen on the histogram in Figure 10. Besides, three observations are clearly isolated from the majority of the data and may therefore be regarded as outlying values for the length of stay.

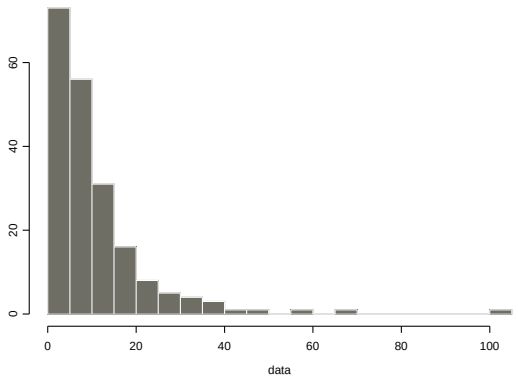


Figure 10: Histogram of the 'Length of Stay' data.

The adjusted boxplot of the LOS data is shown in Figure 11. Due to the upward shift of the upper whisker, only the three largest observations are marked as outliers. At the original boxplot on the other hand, seventeen observations are detected as potential upper outlier. This illustrates again that the adjusted boxplot accounts for skewness. Consequently, one could use the adjusted boxplot as a trimming rule which accounts for skewness. Taking the mean of all observations within the lower and upper fence of the adjusted boxplot, then gives a more realistic estimate of the expected LOS.

3.5 Consumer Expenditure Survey data

Boxplots are often used to compare the distribution of a variable within several groups. Our adjusted boxplot can be used for this purpose as well. To illustrate, we consider data derived from the Consumer Expenditure Survey (CES) of 1995, collected by the Bureau of Labor Statistics, U.S Department of Labor, and available at
http://econ.lse.ac.uk/courses/ec220/G/iedata/ces/.

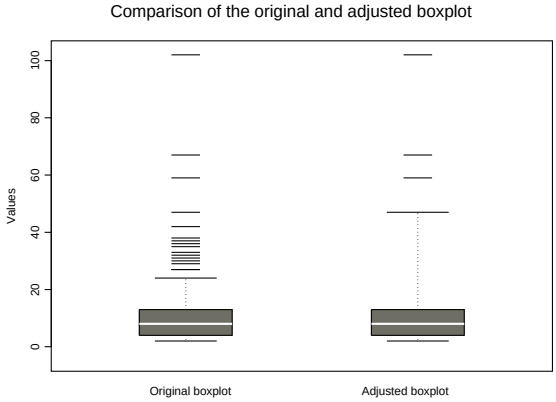


Figure 11: The original and the adjusted boxplot of the 'Length of Stay' data.

In this paper we focus on the variables 'EXP' and 'REFRACE', which represent the total household expenditure and the ethnicity of the reference person respectively.

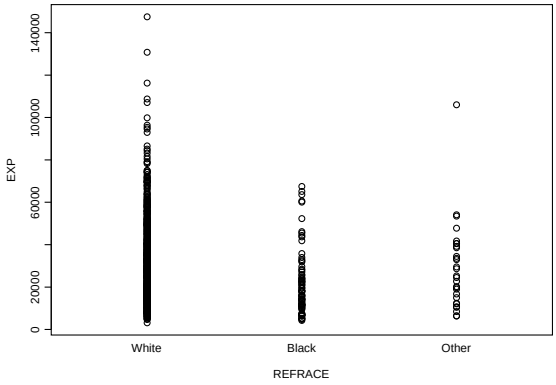


Figure 12: Scatterplot of 'EXP' versus 'REFRACE'.

In Figure 12 the variable 'EXP' is plotted versus 'REFRACE', which can be either 'white', 'black' or 'other' (e.g. American Indian, Aleut, Eskimo, Asian, Pacific Islander etc.). This plot shows that the underlying distribution of the variable 'EXP', conditional on the ethnicity factor, is right skewed and long tailed.

To continue exploration of the data set, we apply the adjusted boxplot to each of the ethnicity subgroups. The result is shown in Figure 13. On the adjusted boxplots, we have superimposed dotted lines which refer to the whiskers of the original boxplot. At the 'white' and 'black' group, the upper whisker has been shifted upwards, yielding less right outliers and emphasizing the skewness of the underlying distribution. Furthermore, at the 'black'

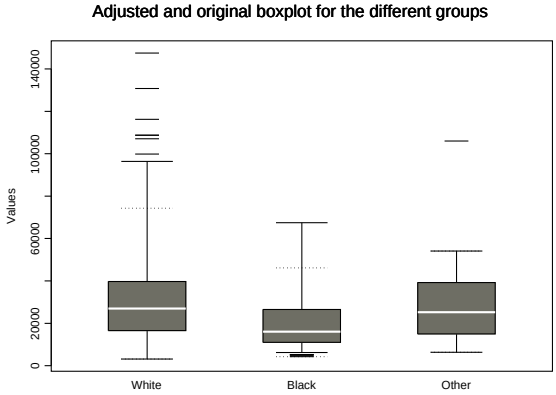


Figure 13: Boxplots of the different groups in the Ces data. The adjusted boxplots are plotted, with dotted lines at the height of the original whiskers superimposed.

group the lower whisker has shifted too, now better reflecting the shorter left tail. Finally, both the original and the adjusted boxplot give the same result at the 'other' group. This is caused by the fact that the majority of the observations in this subgroup come from a symmetric distribution. Hence, the adjusted whiskers give the same result as the original ones.

4. SIMULATION STUDY

To compare our adjusted boxplot with the original one, a simulation study has been done. We focussed on several right skewed distributions, such as the normal, G_g, χ^2, Γ, Pareto and F-distribution. More detailed information can be found in Table 1.

4.1 Performance at uncontaminated data sets

For each of the considered distributions, we generated 100 samples of size 1000 and computed the percentage of left and right outliers (observations that fall outside the boundaries defined by (1) resp. (6)). The average percentages of left and right outliers for the original boxplot (crosses) and for the adjusted boxplot (black dots) are reported in Figure 14. Figure 14(a) gives the result for the right tail, whereas Figure 14(b) concentrates on the left tail.

At the normal distribution, we notice that, slightly remarkable, the adjusted boxplot

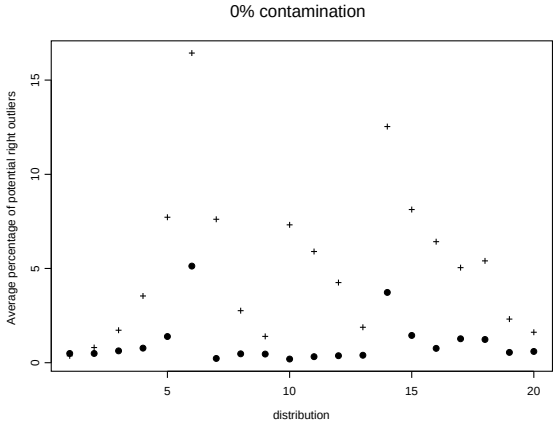
no.	Distribution	no.	Distribution
1	N(0,1)	11	$\Gamma(0.1, 0.75)$
2	$G_{0.1}$	12	$\Gamma(0.1, 1.25)$
3	$G_{0.25}$	13	$\Gamma(0.1, 5)$
4	$G_{0.5}$	14	Pareto(1,1)
5	G_1	15	Pareto(3,1)
6	G_3	16	Pareto(6,1)
7	χ^2_1	17	F(90,10)
8	χ^2_5	18	F(10,10)
9	χ^2_{20}	19	F(10,90)
10	$\Gamma(0.1, 0.5)$	20	F(80,80)

Table 1: The 20 different distributions that are used in the simulation study.

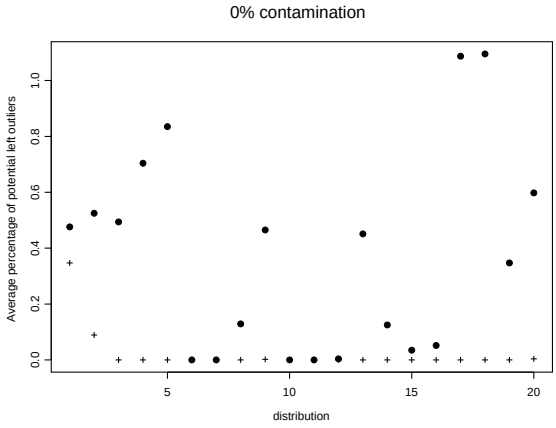
classifies more observations as outliers than before. This is because the finite-sample medcouple is not exactly zero, hence the adjusted whiskers are slightly different from the original ones. Though, the discrepancy is rather small (the total percentage of outliers, classified by the adjusted boxplot is about 0.96% as opposed to about 0.7% at the original boxplot).

Much more pronounced differences can be seen at the skewed distributions. At the χ^2_5 distribution for example, the average number of marked outliers is less than 0.6% at the adjusted boxplot as opposed to more than 2.7% at the original boxplot. The adjusted boxplot of the Pareto(3,1) distribution now yields on average at most 1.48% outliers, whereas on average more than 8% of the observations are marked as outlier at the original boxplot. Note that the G_3 distribution was not used in the calibration of the exponential model, but also here we see that our model highlights much fewer outliers than before.

As we see, the improvements differ somewhat over the distributions. The overall improvement is mainly due to a substantial increase of the upper whiskers. This causes the adjusted boxplot to mark less observations as right outlier than before. Besides, due to the exponential factor we added, the lower whiskers now also yield small percentages of marked outliers for many distributions. This is in accordance with the lower whisker of the original boxplot at the normal distribution, and better allows to detect real left outliers at skewed distributions.



(a)



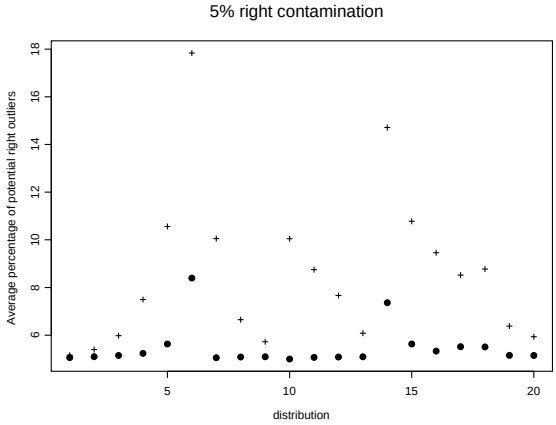
(b)

Figure 14: For different distributions, the average percentage of outliers is reported, resulting from the original boxplot (plus symbols) and the adjusted boxplot (black dots). (a) shows the result for the right tail, whereas (b) concentrates on the left tail.

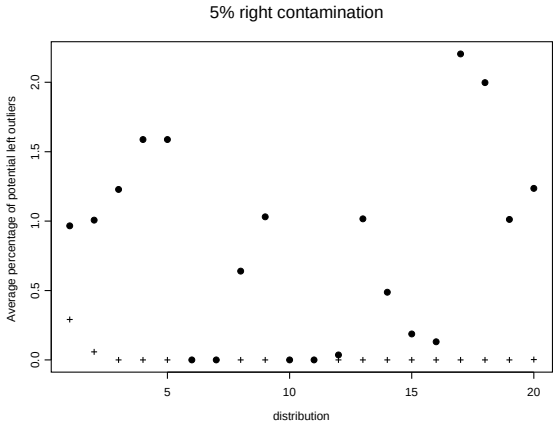
4.2 Performance at contaminated data sets

To get an idea of the robustness of the skewness-adjusted boxplot, we looked at its performance when applied to contaminated data sets. We generated again 100 samples of size 1000 for each distribution, but now replaced 5% of the data by right (respectively left) outliers, coming from a normal distribution. The results when adding 5% of right contamination are depicted in Figure 15.

Figure 15(a) clearly shows that the adjusted boxplot detects the expected 5% of right



(a)

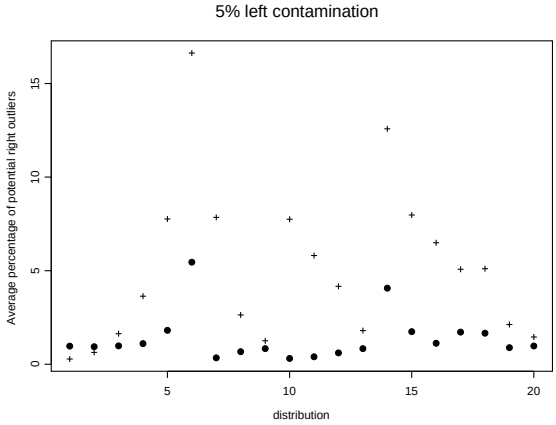


(b)

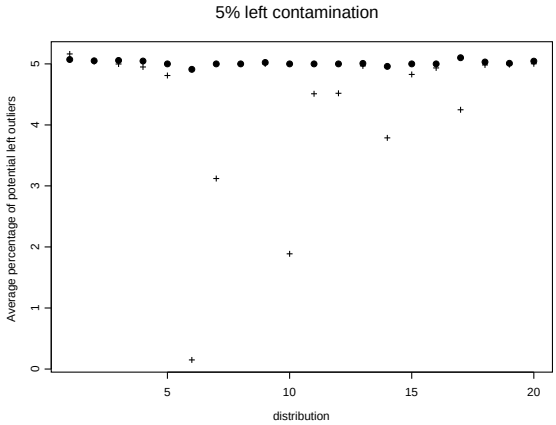
Figure 15: Simulation results, when adding 5% of right contamination. (a) focusses on the right tail, whereas (b) concentrates on the left tail.

outliers, without marking too many observations as potential right outlier. This is not always the case when we apply the original boxplot. For example at the G_1-distribution (which is in fact a lognormal distribution), more than 10% of the observations were marked as right outlier. Furthermore, the adjusted boxplot gives again a more balanced view of potential left outliers, which can be seen from Figure 15(b).

The results of the simulation study with 5% of left contamination are reported in Figure 16. Figure 16(b) clearly shows that the adjusted boxplot succeeds in marking the 5% of left outliers that were added. This is not always the case at the original boxplot. For example at the $\Gamma(\beta, \gamma)$ distribution with $\beta = 0.1$ and $\gamma = 0.5$, less than 2% of the left outliers



(a)



(b)

Figure 16: Simulation results, when adding 5% of left contamination. (a) shows the results for the right tail, whereas (b) concentrates on the left tail.

were detected. Also in the right tail, the adjusted boxplot performs well, as can be seen from Figure 16(a). For most distributions the percentage of observations that have been marked as potential right outlier is rather small. Only at the G_3-distribution more than 5% of right outliers were detected, which is due to the extreme skewness of the underlying distribution.

5. DISCUSSION AND CONCLUSION

A frequently used graphical tool to analyse a univariate data set is the boxplot. Unfortunately, when drawing the boxplot of a skewed distribution, the tail information is not reliable.

Therefore, we have presented an adjustment of the boxplot, that takes the skewness factor into account. To measure skewness of the data, the medcouple has been used and different models for generalizing the original boxplot have been studied. The overall winner seems to be an exponential model.

In order to compare the skewness-adjusted boxplot to the original boxplot, we have applied both boxplots to some real data sets and a graphical representation has been obtained from S-Plus and Matlab functions.

Finally, a simulation study has been performed, based on uncontaminated as well as contaminated data sets. The results indicate the gain of accuracy, achieved by using the adjusted boxplot at skewed distributions.

Our S-PLUS and Matlab functions are available from
http://wis.kuleuven.be/stat/robust.html, the latter being part of the LIBRA toolbox (Verboven and Hubert 2005).

While in this paper we focussed on the detection of univariate outliers, the idea of the adjusted boxplot has been applied to detect multivariate outliers in the context of independent component analysis (Brys, Hubert and Rousseeuw 2005). This methodology can be extended to find outlying observations at multivariate skewed distributions. This could lead to an extension of the adjusted boxplot for bivariate data, such as the bagplot (Rousseeuw, Ruts and Tukey 1999). Also skewness-adjusted modifications of the robust PCA method ROBPCA (Hubert, Rousseeuw and Vanden Branden 2005) will be studied in further research.

References

[1] Aucremanne, L., Brys, G., Hubert, M., Rousseeuw, P.J., Struyf, A. (2004), "A Study of Belgian Inflation, Relative Prices and Nominal Rigidities using New Robust Measures of Skewness and Tail Weight," *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, pp. 13 – 25.

[2] Bowley, A.L. (1920), *Elements of Statistics*, New York: Charles Scribner's Sons.

[3] Brys, G., Hubert, M., Struyf, A. (2003), "A Robust Measure of Skewness," *Journal of Computational and Graphical Statistics*, **13**, pp. 996 – 1017.

[4] Brys, G., Hubert, M., Struyf, A. (2006), "Robust Measures of Tail Weight," *Computational Statistics and Data Analysis*, **50**, pp. 733 – 759.

[5] Brys, G., Hubert, M., Rousseeuw, P.J. (2005), "A Robustification of Independent Component Analysis", *Journal of Chemometrics*, **19**, pp. 364 – 375.

[6] Carling, K. (2000), "Resistant Outlier Rules and the Non-Gaussian Case," *Computational Statistics and Data Analysis*, **33**, pp. 249 – 258.

[7] Chambers, J.M., Hastie, T.J. (1992), *Statistical Models in S*, Wadsworth and Brooks, Pacific Grove, pp. 348 – 351.

[8] Goegebeur, Y., Planchon, V., Beirlant, J., Oger, R. (2005), "Quality Assesment of Pedochemical Data Using Extreme Value Methodology," *Journal of Applied Science*, **5**, pp 1092 – 1102.

[9] Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis*, New York: Wiley, pp. 58 – 77.

[10] Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1985), *Exploring Data Tables, Trends and Shapes*, New York: Wiley, pp. 463 – 478.

[11] Hubert, M., Rousseeuw, P.J., Vanden Branden, K. (2005), "ROBPCA: a New Approach to Robust Principal Components Analysis", *Technometrics*, **47**, pp. 64 – 79.

[12] Jarret, R.G. (1979), "A Note on the Intervals Between Coal Mining Disasters," *Biometrika*, **66**, pp. 191 – 193.

[13] Kimber, A.C. (1990), "Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions ," *Applied Statistics*, **39**, pp. 21 – 30.

[14] Moors, J.J.A., Wagemakers, R.Th.A., Coenen, V.M.J., Heuts, R.M.J. and Janssens, M.J.B.T. (1996), "Characterizing Systems of Distributions by Quantile Measures", *Statistica Neerlandica*, **50**, pp. 417 – 430.

[15] Rousseeuw, P.J., Ruts, I., Tukey, J.W. (1999), "The Bagplot: a Bivariate Boxplot", *The American Statistician*, **53**, pp. 382 – 387.

[16] Ruffieux, C., Paccaud F., Marazzi A. (2000)," Comparing Rules for Truncating Hospital Length of Stay," *Casemix Quaterly,* **2** n.1.

[17] Schwertman, N.C., Owens, M.A., Adnan, R. (2004), "A Simple More General Boxplot Method for Identifying Outliers," *Computational Statistics and Data Analysis*, **47**, pp. 165 – 174.

[18] Tukey, J.W. (1977), "Exploratory Data Analysis," Massachusetts: Reading (Addison-Wesley), pp. 39 – 49.

[19] Vandewalle, B., Beirlant, J., Hubert, M. (2004), "A robust estimator of the tail index based on an exponential regression model", *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel, pp. 367-376.

[20] Verboven, S., Hubert, M. (2005), "LIBRA: a MATLAB Library for Robust Analysis", *Chemometrics and Intelligent Laboratory Systems*, **75**, pp. 127 – 136.