COMPSTAT'2004 Symposium

AN ADJUSTED BOXPLOT FOR SKEWED DISTRIBUTIONS

Ellen Vanderviere and Mia Huber

Key words: Boxplot, skewness, medcouple.
COMPSTAT 2004 section: Graphics.

Abstract: The boxplot is a very popular graphical tool to visualize the distribution of continuous univariate data. First of all, it 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 whiskers are pinpointed as outliers, hence the boxplot also includes information from the tails. However, when the data are skewed, usually too many points are classified as outliers. This is because the outlier rule is solely based on measures of location and scale, and the cutoff values are derived from the normal distribution. We present a generalization of the boxplot that includes a robust measure of skewness in the determination of the whiskers. We show with several simulation results that this adjusted boxplot gives a more accurate representation of the data and of possible outliers.

1 Introduction

One of the most frequently used graphical techniques for analyzing a univariate data set is the *boxplot*, proposed by Tukey [6].
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

$$[Q_1 - 1.5 \text{ IQR} \; ; \; Q_3 + 1.5 \text{ IQR}] \qquad (\text{with IQR } = Q_3 - Q_1)$$

  as outlier and marking them on the plot
- drawing the whiskers (i.e. the lines that go from the ends of the box to the most remote points that are no outliers)

This construction implies that a boxplot gives information about the location, spread, skewness and tails of the data.

However, in some cases the information about the tails that is given by the boxplot, is not reliable. As was already mentioned in Hoaglin, Mosteller

and Tukey [3], too many points are classified as outlier when the data are sampled from a skewed distribution. Consider e.g. the chi-squared distribution with *df* degrees of freedom, which is very skewed for small *df*. Table 1 lists the theoretical lower whisker and the upper whisker for $df = 1, 5$ and 20. The last column gives the probability of a type I error, which we define as the probability to exceed the whiskers, or equivalently, to mark regular observations as outliers. We see that this probability is about 7.6% for χ^2_1 which is very high, and decreases with *df*. For the normal distribution on the other hand, the expected percentage of outliers is only 0.7%.

Distr.	Lower outlier cutoff	Upper outlier cutoff	Total % outliers
χ^2_1	-1.730	3.155	7.58
χ^2_5	-3.252	12.552	2.80
χ^2_{20}	2.888	36.392	1.39
N(0,1)	-2.698	2.698	0.70

Table 1: Theoretical lower and upper outlier cutoff values for several distributions, and the expected percentage of classified outliers according to the boxplot rule.

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

2 Generalization of the boxplot

2.1 A robust measure of skewness

To measure the skewness of a continuous distribution F, we will use the *medcouple*, which we denote as MC. It is defined as

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

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

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

From the definition we see 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 it becomes negative at a left skewed distribution. Finally, a symmetric distribution has a zero medcouple. As shown in Brys, Hubert and Struyf [1] this robust measure of skewness has a bounded influence function and a breakdown value of 25%. Besides, the MC turned out to be the overall winner when comparing it with two other robust skewness measures which are solely based on quantiles, namely the QS (quartile skewness) and the OS (octile skewness). 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 MC, a fast algorithm of $O(n \log n)$ time has been constructed, and Matlab and S-PLUS functions are available.

2.2 Possible models

We generalize the original boxplot by introducing functions $h_l(\mathrm{MC})$ and $h_r(\mathrm{MC})$ in the cutoff values to classify the outliers. Thus instead of using the interval

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

for the regular observations, we propose the boundaries of the interval to be defined as

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

We additionally require that $h_l(0) = h_r(0) = 1.5$ in order to obtain the original boxplot at symmetric distributions. Note that by using different functions h_l and h_r we allow to obtain whiskers of different length. Moreover the boundaries are location and scale equivariant due to the location and scale invariancy of the medcouple.

We studied three different models, which are easy and which do not contain too many parameters, namely a

(1). *linear model*: $h_l(\mathrm{MC}) = 1.5 + a \text{ MC}$, $h_r(\mathrm{MC}) = 1.5 + b \text{ MC}$.

(2). *quadratic model*: $h_l(\mathrm{MC}) = 1.5 + a_1 \text{ MC} + a_2 \text{ MC}^2$, $h_r(\mathrm{MC}) = 1.5 + b_1 \text{ MC} + b_2 \text{ MC}^2$

(3). *exponential model*: $h_l(\mathrm{MC}) = 1.5 \; e^{a\mathrm{MC}}$, $h_r(\mathrm{MC}) = 1.5 \; e^{b\mathrm{MC}}$.

2.3 Defining the constants

In order to determine the constants in the models mentioned above, we require that the expected percentage of marked outliers is 0.7%, which coincides with the outlier rule of the original boxplot at the normal distribution. If we use for example the linear model, this implies that the constants a and b should satisfy $Q_1 - (1.5 + a \text{ MC}) \text{ IQR} = Q_\alpha$ and $Q_3 + (1.5 + b \text{ MC}) \text{ IQR} = Q_\beta$ 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 $\frac{Q_1 - Q_\alpha}{\text{IQR}} - 1.5 = a$ MC, and $\frac{Q_\beta - Q_3}{\text{IQR}} - 1.5 = b$ MC. Linear regression without intercept can then be used to obtain estimates of the parameters a and b. The parameter derivation in case of the quadratic or the exponential model is analogous to that of the linear case. For the exponential model for example, we obtain the linear system

$$\begin{cases} \ln\left(\frac{2}{3}\,\frac{Q_1 - Q_\alpha}{\text{IQR}}\right) = a \text{ MC} \\ \ln\left(\frac{2}{3}\,\frac{Q_\beta - Q_3}{\text{IQR}}\right) = b \text{ MC} \end{cases}$$

To derive the constants we used 12,605 distributions from the family of Γ, χ^2, F, Pareto and G_g-distributions [4]. 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 always selected such that the medcouple did not exceed 0.6. Doing so, we retain 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 currently 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 1(a) we show for the parameter b the fitted regression curves, after applying (robust) reweighted LTS regression [5] for each model. Note that the regression results are based on the whole set of distributions we considered. On the vertical axis we have set the response value for the exponential model. Figure 1(b) only displays the G_g distributions (with the same fits superimposed).

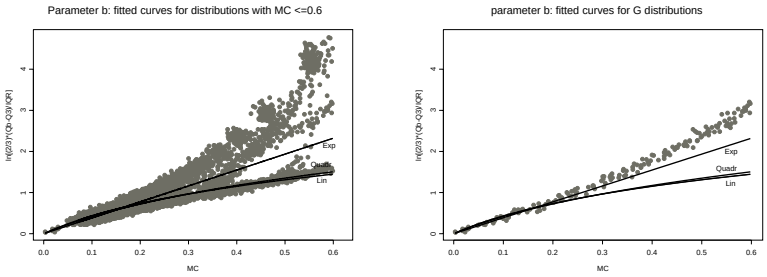

Figure 1: Regression curves for the linear, quadratic and exponential model.

2.4 The adjusted boxplot

From Figure 1 we see that the exponential model is the most appropriate one. Although the fitted line will produce an underestimate of Q_β for some distributions, the same quantile will be overestimated for others. So it gives a good compromise for the whole set of distributions we considered. The linear and quadratic model on the other hand give underestimates for a large group of distributions. For the estimate of the left tail, we have not included the figures because of lack of space. Here it could be seen that the linear model fails completely to estimate Q_α, whereas the exponential and quadratic model perform much better. As the exponential model is appropriate for both the left and the right tail and as it only includes one parameter (on each side), we will use the *exponential model* in the definition of our adjusted boxplot. To make the model easier, we rounded off the estimated values of $a = -3.79$ and $b = 3.87$ to $a = -3.5$ and $b = 4$. We will investigate other possibilities such as $-a = b = 4$ and $-a = b = 3$ in further research.

To summarize, the adjusted boxplot marks the observations that fall outside the interval

$$[Q_1 - 1.5\ e^{-3.5\ \mathrm{MC}}\ \mathrm{IQR}\ ;\ Q_3 + 1.5\ e^{4\ \mathrm{MC}}\ \mathrm{IQR}] \tag{1}$$

3 Simulation study

To compare our adjusted boxplot with the original one, the percentage of left and right outliers (observations that fall outside the boundaries defined by (1)), together with the total percentage of outliers, is computed at samples of different size n from several distributions. For each distribution, 100 samples of size n were considered. The average percentages of outliers are reported in Table 2. Standard errors are below 0.2% for most of the entries. The superscript $*$ means a standard error between 0.2% and 0.5%, $**$ between 0.5% and 0.9%.

At the normal distribution, we notice that, slightly remarkably, the adjusted boxplot 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. At larger sample sizes, we see that the total percentage of outliers is again close to 0.7%.

Much more pronounced differences can be seen at the skewed distributions. At the χ^2_1 distribution for example the average number of marked outliers is less than 0.18%. The adjusted boxplot of the Pareto(3,1) distribution now yields at most 1.23% outliers, as opposed to more than 8% at the original boxplot. We also included two distributions that were not used in the calibration of the exponential model, namely the Pareto(1,3) and the G_3 distribution. 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 right

		original boxplot			adjusted boxplot		
Distr	n	% L	% R	Tot %	% L	% R	Tot %
$N(0,1)$	100	0.600	0.700	1.300	1.180	0.800	1.980*
	500	0.358	0.402	0.760*	0.462*	0.634*	1.096*
	1000	0.335	0.362*	0.697*	0.484*	0.445*	0.929*
χ^2_1	100	0.000	7.350*	7.350*	0.000	0.180	0.180
	500	0.000	7.940**	7.940**	0.000	0.032	0.032
	1000	0.000	7.726**	7.726**	0.000	0.015	0.015
χ^2_{20}	100	0.060	1.360	1.420	0.880	0.780	1.660
	500	0.002	1.478*	1.480*	0.400*	0.392*	0.792*
	1000	0.002	1.456**	1.458**	0.382*	0.311*	0.693*
$\Gamma(0.1, 0.5)$	100	0.000	7.960*	7.960*	0.000	0.410	0.410
	500	0.000	7.716**	7.716**	0.000	0.030	0.030
	1000	0.000	7.708**	7.708**	0.000	0.019	0.019
Pareto(3,1)	100	0.000	8.130*	8.130*	0.280	0.950	1.230
	500	0.000	8.350**	8.350**	0.034	0.620*	0.654*
	1000	0.000	7.943**	7.943**	0.000	0.558*	0.558*
$F(90, 10)$	100	0.000	5.210*	5.210*	1.480*	0.960	2.440*
	500	0.000	5.000**	5.000**	0.584*	0.636*	1.220*
	1000	0.000	5.230**	5.230**	0.485**	0.714*	1.199**
Pareto(1,3)	100	0.000	12.250*	12.250*	0.710*	2.490	3.200*
	500	0.000	12.338**	12.338**	0.000	2.314*	2.314*
	1000	0.000	12.461**	12.461**	0.000	2.166*	2.166*
G_3	100	0.000	16.300*	16.300*	0.000	3.290	3.290
	500	0.000	16.516*	16.516*	0.000	2.966*	2.966*
	1000	0.000	16.408**	16.408**	0.000	3.028**	3.028**

Table 2: For different distributions and samples sizes, the mean percentage of left outliers (% L), right outliers (% R) and the mean total percentage of outliers (Tot %) are reported, resulting from the original boxplot and the adjusted boxplot. The superscript $*$ means a standard error between 0.2% and 0.5%, $**$ between 0.5% and 0.9%. No superscript is set if the standard error is smaller than 0.2%.

whiskers. The lower whiskers are often still somewhat too small, yielding zero percentages of marked outliers. We don't consider the latter as a too serious problem as it is mainly the outlyingness to the right which is of importance at right-skewed distributions. However, to improve the fits, several modifications could be made (we thank a referee for pointing out several of them). The most natural one is to include tail information of the distributions 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 et al. [2]. We see however several disadvantages of such a procedure. First of all, the model will become more complex with more estimators and parameters. The robustness will decrease as the tail measures have a lower breakdown value, and the variability of the whisker's length will increase, due to the variability of the tail measures.

If we have a priori information of the distribution, for example, we know that it belongs to the class of G_g distributions, it is clear from Figure 1

that a more specific model could be constructed, for example by including results from extreme value theory. Another possibility is to vary the quantiles $\alpha = 0.0035$ and $\beta = 0.9965$ within one (or all) distributions to see if each of its tails is being modeled appropriately. Or different functional forms could be considered for the two tails. This will certainly be considered in our future research.

4 Example

In this section we consider a sample of size $n = 200$ from a G_1-distribution (which is exactly the lognormal distribution) and apply both the original and our adjusted boxplot. As we did not yet implement a graphical representation, we summarize the results as in Figure 2. On the plot with the data versus their index, we have drawn full lines at the median, and at the first and third quartile of the data. Next, we have drawn dashed lines for the original boxplot, and dash-dotted lines at the boundaries of the adjusted boxplot.

We see that by introducing the medcouple in our definition, both the left and the right boundary have shifted upwards. We notice a significant decrease from 10% to 3.5% of right outliers. The lower bound lies much closer to the data points. This might yield more left outliers, but it also better reflects the shorter left tail.

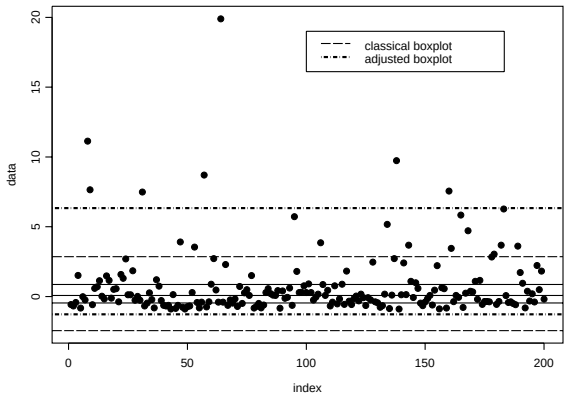

Figure 2: A sample of 200 observations from a lognormal distribution with the boundaries to classify outliers based on the original and the adjusted boxplot.

5 Conclusion

A frequently used graphical tool to analyze a univariate data set is the boxplot. Unfortunately, when drawing this boxplot to a skewed distribution, the tail information is not reliable. Therefore, we have presented a generalization 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. Finally, some simulation results and a graphical representation have been given to indicate the gain of accuracy, achieved by using the adjusted boxplot at skewed distributions.

Note that in this paper we have studied the adjusted boxplot with respect to its type I error, which we have defined as the probability to wrongly declare regular observations as outliers. In the future we will also study its behavior at data sets which contain real outliers.

References

[1] Brys G., Hubert M., Struyf A. (2003). *A robust measure of skewness.* Journal of Computational and Graphical Statistics, to appear.

[2] Brys G., Hubert M., Struyf A. (2004). *Robust measures of tail weight.* Submitted.

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

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

[5] Rousseeuw P.J. (1984). *Least median of squares regression.* Journal of the American Statistical Association **79**, 871 – 880.

[6] Tukey, J.W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Address: E. Vanderviere, University of Antwerp, Department of Mathematics and Computer Science, Middelheimlaan 1, B-2020 Antwerp, Belgium
M. Huber, Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, Belgium

E-mail: `ellen.vandervieren@ua.ac.be`, `mia.hubert@wis.kuleuven.ac.be`