Communications of the Korean Statistical Society
2012, Vol. 19, No. 4, 591–595

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.4.591

Skew Normal Boxplot and Outliers

Myung-Hoe Huh[1,a], Yonggoo Lee[b]

[a]Department of Statistics, Korea University
[b]Department of Applied Statistics, Chung-Ang University

Abstract

We frequently use Tukey's boxplot to identify outliers in the batch of observations of the continuous variable. In doing so, we implicitly assume that the underlying distribution belongs to the family of normal distributions. Such a practice of data handling is often superficial and improper, since in reality too many variables manifest the skewness. In this short paper, we build a modified boxplot and set the outlier identification procedure by assuming that the observations are generated from the skew normal distribution (Azzalini, 1985), which is an extension of the normal distribution. Statistical performance of the proposed procedure is examined with simulated datasets.

Keywords: Boxplot, outlier, skew normal distribution.

1. Background and Aim

For the case of continuous observations $x_1, \ldots, x_n$, Tukey's boxplots are frequently used for describing the location and the spread along with outliers. In the plot, the observations outside the 'fence' are declared as outliers, where the fence is set at

$$q_1(x_1, \ldots, x_n) - 1.5 \cdot \mathrm{IQR}, \qquad q_3(x_1, \ldots, x_n) + 1.5 \cdot \mathrm{IQR}, \tag{1.1}$$

where $q_1(x_1, \ldots, x_n)$ and $q_3(x_1, \ldots, x_n)$ are the first and the third quartile of the observations and $\mathrm{IQR} = q_3(x_1, \ldots, x_n) - q_1(x_1, \ldots, x_n)$.

When the observations are generated from $N(\mu, \sigma)$, the proportion of data points outside the fence is about 0.7%. When it is not the case, the proportion of outliers may vary from situation to situation. When the underlying distribution is skewed, Tukey's boxplot may show too many 'outliers' in one side. If we transform the variable in some way to correct the skewness, the 'outliers' may disappear. It means that such data points are not genuine outliers. Hence, for the benefit of exploratory data analysts, we need to modify Tukey's boxplot that accepts moderately skewed distributions.

Hubert and Vandervieren (2008) proposed an adjusted boxplot and the associated outlier identification procedure. They considered moderately skewed cases arising from the gamma, chi-square, F, Pareto, and G_g-distributions and proposed a modification of (1.1) by fitting the 0.35 tail percentiles using a robust measure of the skewness. Their methodological base is rather empirical than theoretical.

In this study, we consider the skew normal distribution proposed by Azzalini (1985) as underlying distributions for the observations. The skew normal distribution $SN(\mu, \sigma, \alpha)$ has the density

$$f(x\,;\mu,\sigma,\alpha) = \frac{2}{\sigma}\,\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right), \quad -\infty < x < \infty.$$

[1] Corresponding author: Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Seoul 136-701, Korea. E-mail: stat420@korea.ac.kr

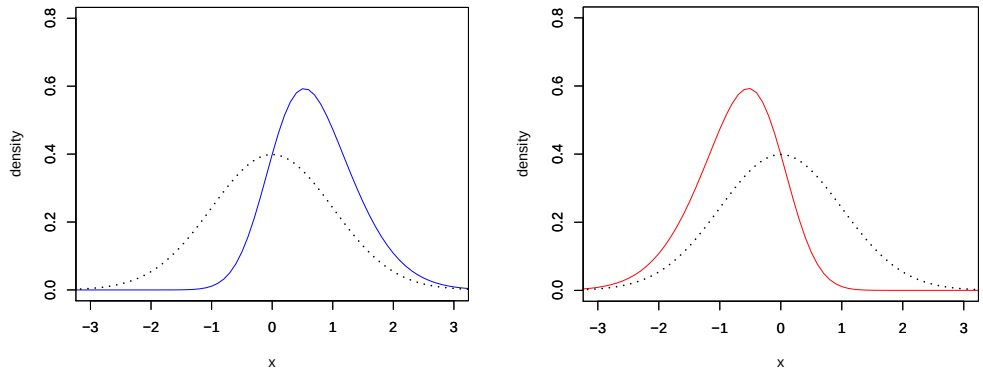

Figure 1: *Densities of skew normal distributions for* $\alpha = 2$ *[Left] and* $\alpha = -2$ *[Right] with the* $N(0,1)$ *density superimposed*

When $\alpha = 0$, the skew normal distribution is identical to $N(\mu, \sigma)$. When α is positive (negative), $SN(\mu, \sigma, \alpha)$ is skewed to the right (left). Figure 1 shows three cases of the skew normal distribution: $SN(0, 1, 2)$, $SN(0, 1, 0)$ and $SN(0, 1, -2)$. As such, the skew normal distribution extends the normal distribution.

Skewed distributions of some kinds have received increased interest by the Korean statistical community. Seo *et al.* (2009) studied the circular distribution derived from the skew normal distribution, and Kim (2010) applied a skew normal distribution to the calibration problem.

This study aims to build a boxplot and to set the outlier identification procedure by assuming that the observations are distributed to the skew normal distribution (Azzalini, 1985), In Section 2, we propose a modified boxplot with marked outliers, called the skew normal boxplot; subsequently, Tukey's boxplot will be called the normal boxplot. In Section 3, statistical performance of the proposed procedure is examined with simulated datasets.

2. Proposed Procedure

To identify outliers, the underlying distribution $SN(\mu, \sigma, \alpha)$ should be fitted robustly. For that purpose, we propose the following estimation procedure.

First, compute the sample skewness which we define as

$$\text{skewness}(x_1, \ldots, x_n) = \frac{q_3(x_1, \ldots, x_n) - q_2(x_1, \ldots, x_n)}{q_2(x_1, \ldots, x_n) - q_1(x_1, \cdots, x_n)},$$

where $q_2(x_1, \ldots, x_n)$ is the median. Then, α is found by matching to the corresponding quantity of $SN(0, 1, \alpha)$. Such estimate $\hat{\alpha}$ of α is safe from 25% contamination.

Second, compute the quartiles $\dot{q}_1$ and $\dot{q}_3$ of $SN(0, 1, \hat{\alpha})$ and estimate σ by

$$\hat{\sigma} = \frac{q_3(x_1, \ldots, x_n) - q_1(x_1, \ldots, x_n)}{\dot{q}_3 - \dot{q}_1}.$$

Third, compute the median $\dot{\mu}$ of $SN(0, \hat{\sigma}, \hat{\alpha})$ and estimate μ by

$$\hat{\mu} = q_2(x_1, \ldots, x_n) - \dot{\mu}.$$

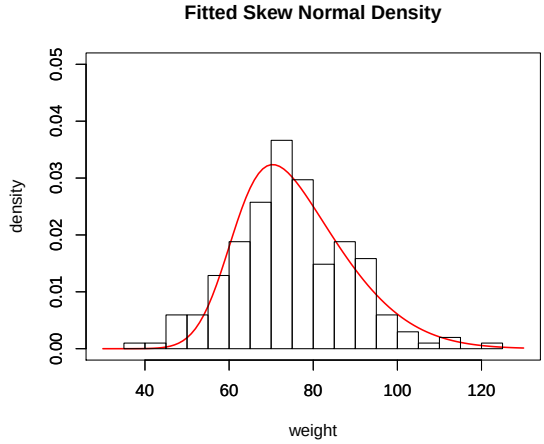

Figure 2: *Histogram of AIS weight with the superimposed skew normal density*

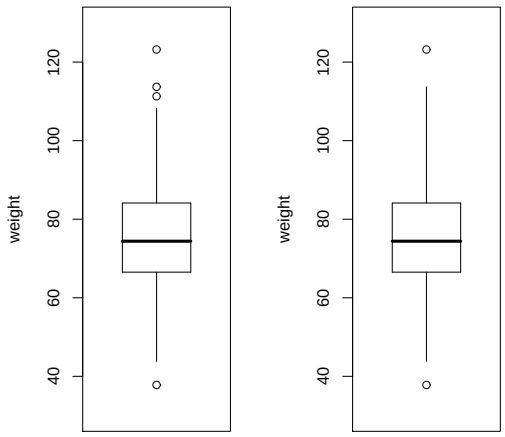

Figure 3: *The normal boxplot [Left] and the skew normal boxplot [Right]*

We define the outliers as the observations below the minimum of $SN(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$'s 0.35 percentile and the lower value of (1.1) or above the maximum of $SN(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$'s 99.65 percentile and the upper value of (1.1). Therefore, for the case in which observations are generated from the skew normal distribution, less than 0.7%, or roughly 1% at maximum, of the observations are expected to be declared as outliers.

As a numerical illustration, we apply the procedures to the weight variable of Australian Institute of Sports dataset ($n = 202$), as given in the `R` package `sn` (Azzalini, 2011). The fence values of Tukey's boxplot are 40.1(kg) and 110.5(kg), and, accordingly, one small observation with the data value 37.8 and three large observations with data values 111.3, 113.7, 123.2 are declared as outliers. See Figure 3 [Left].

Three parameters of the skew normal distribution are estimated as $\hat{\mu} = 60.74$, $\hat{\sigma} = 20.32$, $\hat{\alpha} = 3.00$. Figure 2 shows the histogram with the fitted skew normal density superimposed.

With reference to $SN(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$, 0.35 percentile and 99.65 percentiles are 49.0 and 120.1. Therefore,

Table 1: Monte-Carlo outcome for the simulated 1,000 observations from the skew normal distribution $SN(0, 1, \alpha)$

alpha	Normal boxplot		Skew normal boxplot	
	#below (se)	#above (se)	#below (se)	#above (se)
0	3.70 (0.07)	3.66 (0.07)	3.07 (0.07)	3.03 (0.07)
1	2.38 (0.05)	5.88 (0.08)	2.07 (0.05)	4.44 (0.09)
2	0.65 (0.03)	11.30 (0.12)	0.63 (0.03)	5.96 (0.12)
3	0.06 (0.01)	14.46 (0.14)	0.06 (0.01)	5.87 (0.11)
4	0.00 (0.00)	16.04 (0.15)	0.00 (0.00)	5.86 (0.11)
5	0.00 (0.00)	16.59 (0.15)	0.00 (0.00)	5.99 (0.10)
10	0.00 (0.00)	16.68 (0.16)	0.00 (0.00)	5.93 (0.10)
100	0.00 (0.00)	16.75 (0.16)	0.00 (0.00)	5.88 (0.10)

* '#below' ('#above') is the average number of nominated outliers below the lower fence (above the upper fence).

the fence values distinguishing inner and outer points are min(40.1, 49.0) and max(110.5, 120.1). Thus, only two observations, the smallest 37.8 and the largest 123.2, are declared as outliers.

Figure 3 [Right] shows the skew normal boxplot. Whisker lengths are different in the skew normal boxplot, contrasting the 'almost' same lengths in the normal boxplot.

3. Simulation Study

What are the primary differences between the outlier identification procedure based on the normal boxplot and the one based on the skew normal boxplot? To answer the question, we designed a Monte-Carlo study in which 1,000 observations are generated from $SN(0, 1, \alpha)$ for $\alpha = 0, 1, 2, 3, 4, 5, 10, 100$ and two types of outlier identification procedures are applied. The number of repetitions are 1,000 $(= R)$. Table 1 summarizes the outcomes by the average numbers of nominated outliers and their standard errors (in parenthesis).

For the case $\alpha = 0$, in which the observations are generated from the normal distribution, normal boxplot procedure's '#below' and '#above' are around 3.7. Skew normal boxplot's '#below' and '#above' are slightly smaller, around 3.1. As α increases to 2, normal boxplot procedure's '#below' falls less than 1.0, while '#above' exceeds 10. Furthermore, for the case $\alpha = 5$, normal boxplot procedure's '#below' becomes 0, while '#above' exceeds 16. Even in these cases and the cases of $\alpha = 10$ or 100, skew normal boxplot procedure's '#above' does not exceed 7. As such, the normal boxplot often shows too many outliers in the skewed case, while the skew normal boxplot's list of outliers is more compact.

4. Concluding Remarks

From the viewpoint of exploratory data analysis(EDA), it is always better for data analysts to assume less on the underlying distribution for observations. The principle can be applied to the boxplot and the associated procedure for outlier identification. The normal boxplot assumes the normal distribution, which is quite often inappropriate. In comparison, the skew normal boxplot proposed here assumes the skew normal distribution that allows moderate skewness and extends the normal distribution in a natural way. It is more flexible and reliable.

However, the skew normal boxplot is not a panacea for all kinds of the skewness. Its performance may be poor for heavily skewed cases like the log-normal distribution, since the skewness of $SN(\mu, \sigma, \alpha)$ does not exceed 1.34, as measure by the ratio of the lengths of the quartiles from the median. Hence, it is wise to think of a variable transformation to correct for the skewness for such cases.

References

Azzalini, A. (1985). A class of distribution which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. (2011). R package sn (Version 0.4). http://www.r-project.org.

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

Kim, S. (2010). New calibration methods with asymmetric data, *Korean Journal of Applied Statistics*, **23**, 759–765.

Seo, H. S., Shin, J. K. and Kim, H. M. (2009). Projected circular and *l*-axial skew-normal distributions, *Korean Journal of Applied Statistics*, **22**, 879–891.

Received February 28, 2012; Revised March 27, 2012; Accepted March 28, 2012