A Comparison of Some New Measures of Skewness

G. Brys, M. Hubert[⋆], and A. Struyf[⋆⋆]

Department of Mathematics and Computer Science, University of Antwerp (UIA), Belgium
E-mail: guy.brys, mia.hubert, anja.struyf@ua.ac.be

Summary. Asymmetry of a univariate continuous distribution is commonly described as skewness. The well-known classical skewness coefficient is based on the first three moments of the data set, and hence it is strongly affected by the presence of one or more outliers. In this paper we propose several new measures of skewness which are more robust against outlying values. Their properties are compared using both real and simulated data.

1 Introduction

Statistical models often assume symmetric distributions, and when the data are asymmetric we try to apply a symmetrizing transformation first. The latter is not always possible, however. Sometimes the asymmetry is an inherent factor. Asymmetry is described by *skewness*. A symmetric distribution has zero skewness, an asymmetric distribution with the largest tail to the right has positive skewness, and a distribution with a longer left tail has negative skewness. To measure the skewness of a univariate data set $X_n = \{x_1, x_2, \ldots, x_n\}$ sampled from a continuous distribution one typically uses the classical skewness coefficient b_1. It is defined as

$$b_1(X_n) = \frac{m_3(X_n)}{m_2(X_n)^{3/2}}$$

where m_3 and m_2 denote the third and second empirical moments of the data. However, b_1 may be strongly affected by even a single outlier. Therefore, we will investigate several measures of skewness which are less sensitive to outlying values. We introduce these measures and we discuss their properties in Section 2. In Section 3 we look at their performance at symmetric and asymmetric distributions. Their robustness towards outlying values is studied in Section 4. The limiting distribution of these skewness measures is studied in Section 5. Finally, Section 6 contains some conclusions and directions for further research.

[⋆] Postdoctoral Fellow of the Fund for Scientific Research - Flanders (Belgium)

[⋆⋆] Research Assistant of the Fund for Scientific Research - Flanders (Belgium)

2 New Measures of Skewness

2.1 Definitions

We first investigate skewness measures based on certain quantiles of the data. This is in analogy with the median which estimates the center of the data, and with the interquartile range which estimates its scale in a robust way. Hinkley (1975) suggested to use the following class of skewness measures

$$\frac{(Q_{1-p} - Q_{0.5}) - (Q_{0.5} - Q_p)}{Q_{1-p} - Q_p} \tag{1}$$

where Q_p $(0 < p < 1)$ is the p-th quantile of X_n. We will investigate two measures that belong to this class. The *quartile skewness* corresponds with $p = 0.25$ in (1):

$$QS = \frac{(Q_{0.75} - Q_{0.5}) - (Q_{0.5} - Q_{0.25})}{Q_{0.75} - Q_{0.25}}$$

and it is also known as the Bowley coefficient (Bowley, 1920; Moors et al., 1996). Next, we will consider the *octile skewness*, which takes $p = 0.125$ in (1), yielding

$$OS = \frac{(Q_{0.875} - Q_{0.5}) - (Q_{0.5} - Q_{0.125})}{Q_{0.875} - Q_{0.125}}.$$

From these definitions it is clear that QS is less sensitive to outliers than OS. But on the other hand, OS uses more information from the tails of the distribution and thus will be more appropriate to detect asymmetry in the data. This will become apparent in Sections 3 and 4.

By replacing some of the quantiles in (1) with actual data points, we get another measure of skewness which we call *medcouple*. For all $x_i \neq x_j$ let

$$h_1(x_i, x_j) = \frac{(x_{(j)} - Q_{0.5}) - (Q_{0.5} - x_{(i)})}{x_{(j)} - x_{(i)}}$$

with $x_{(i)} < x_{(j)}$ the sorted arguments of h_1. In the special case that $x_i = x_j = Q_{0.5}$, we set

$$h_1(x_i, x_j) = \begin{cases} +1 & i > j \\ 0 & i = j \\ -1 & i < j \end{cases}$$

The medcouple is then defined as

$$MC = \text{med}_{x_i \leq Q_{0.5} \leq x_j} h_1(x_i, x_j).$$

Next, we also replace the median by an observation. Let h_2 be given by

$$h_2(x_i, x_j, x_k) = \frac{(x_{(k)} - x_{(j)}) - (x_{(j)} - x_{(i)})}{x_{(k)} - x_{(i)}}$$

where we assume $\{x_{(i)}, x_{(j)}, x_{(k)}\}$ to be sorted in ascending order. In the special case where $x_{(i)} = x_{(k)}$ we set $h_2(x_i, x_j, x_k) = 0$. We now define the *medtriple* as

$$MT = \text{med}_{i<j<k} h_2(x_i, x_j, x_k).$$

Instead of taking the median over all couples or triples of data points, we can also use a repeated median. In this way, we obtain two other estimators which are computationally more complex. The *repeated medcouple* is defined as

$$RMC = \text{med}_i \text{med}_{\substack{x_i \leq Q_{0.5} \leq x_j \\ \text{or} \\ x_j \leq Q_{0.5} \leq x_i}} h_1(x_i, x_j)$$

and the *repeated medtriple* as

$$RMT = \text{med}_i \text{med}_{j \neq i} \text{med}_{k \notin \{i,j\}} h_2(x_i, x_j, x_k).$$

The approach to define estimators based on pairs or triples of observations is not new. In the location setting we have the Hodges-Lehmann estimator (Hodges and Lehmann, 1963), given by

$$HL_n = \text{med}_{i<j} \frac{x_i + x_j}{2}.$$

As a robust measure of scale we mention the Q_n estimator (Rousseeuw and Croux, 1993) that is defined as the first quartile of the set of pairwise distances

$$\{x_j - x_i; 1 \leq i, j \leq n \text{ and } x_i \leq x_j\}.$$

In simple regression the Theil-Sen estimator (Theil, 1950; Sen, 1968) and the repeated median line (Siegel, 1982) are based on all pairwise slopes through two data points. Moreover, Rousseeuw and Hubert (1996) have constructed scale estimators based on the vertical height of the triangle formed by three data points.

2.2 Mathematical and Statistical Properties

Let γ denote any of the six skewness measures defined in Section 2.1, and X_n a univariate sample from a continuous distribution. Then the following properties hold:

Property 1. *γ is location and scale invariant, i.e.*

$$\gamma(aX_n + b) = \gamma(X_n)$$

for any $a > 0$ and $b \in \mathbb{R}$.

Property 2. *If we invert a data set, its skewness is inverted as well:*

$$\gamma(-X_n) = -\gamma(X_n).$$

Property 3. *If X_n is symmetric around its median, then* $\gamma(X_n) = 0$.

Properties 1 and 2 follow immediately from the definitions, and imply Property 3. All of them express natural requirements of any skewness measure. It is also straightforward to show that all new measures are bounded.

Property 4. $\gamma(X_n) \in [-1, 1]$.

Finally let us compare the robustness of the skewness measures towards contamination. For this, we use the breakdown value ε^* which roughly measures the maximum proportion of outliers an estimator can resist without achieving its extreme values (Rousseeuw and Leroy, 1987).

Table 1 lists the breakdown values of all the new estimators. We see that they all have a positive breakdown value, ranging from 12.5% up to even 50%. The classical skewness coefficient b_1 on the other hand is based on moments of the data set, and thus it has zero breakdown value. To reduce space we do not include the proofs here. The results for QS and OS are trivial. To obtain the breakdown values of MC, MT, RMC and RMT one can use similar arguments as in Rousseeuw and Hubert (1996). From these proofs it appears that breakdown occurs when all the outliers are placed at the extreme side of the data set and when the distance between adjacent outliers increases when we move from the innermost to the outermost one.

Note that we also could have considered other skewness measures from the class (1). But here we see that, from the robustness point of view, we can only compete with the other measures by choosing p around 25%.

Table 1. Breakdown value of the new skewness measures

estimator	ϵ^*
QS	25%
OS	12.5%
MC	25%
MT	20.6%
RMC	25%
RMT	50%

3 Performance at Non-contaminated Distributions

In this section we will compare the performance of b_1 and the six new estimators on simulated data from a symmetric and from several skewed distributions. For this we consider Tukey's class of g-distributions (Hoaglin et al., 1985). When a random variable Z is gaussian distributed, then

$$Y_g = \frac{(e^{gZ} - 1)}{g}$$

is said to follow a g-distribution G_g with parameter $g \in I\!R$. For $g = 0$ we set $Y_0 \equiv Z$ and thus we have zero skewness. For $g = 1$ we obtain the shifted lognormal distribution. Negative resp. positive values of g yield left-tailed resp. right-tailed distributions. Note that Y_{-g} has the same distribution as $-Y_g$ so we can restrict ourselves to $g \geq 0$. In Figure 1 the density of the g-distribution is depicted for several values of the parameter g.

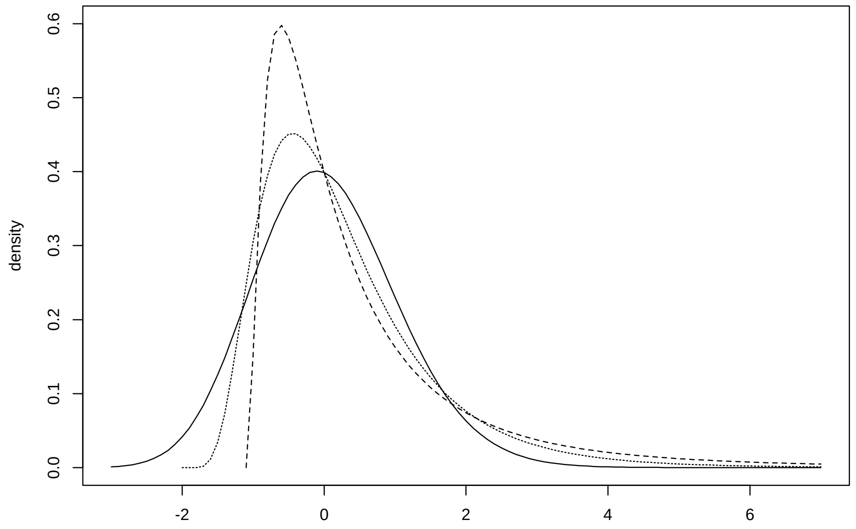

Fig. 1. Density of the g-distribution for $g = 0.1$ (full line), $g = 0.5$ (dotted line), and $g = 0.9$ (dashed line).

First, we have generated 1000 samples of each $n = 100$ observations from G_g with g ranging from 0 to 2. The average estimated skewness obtained with QS, OS, MC, MT, RMC and RMT versus the value of g is depicted in Figure 2. From this figure the monotone relationship between the parameter of the population distribution and the estimated skewness is obvious. Moreover we see that MT, MC, and RMT behave similarly, while QS and RMC give on average smaller and OS on average larger values for the skewness. This gap between the three groupes of curves widens when g becomes larger.

Figure 3 shows the relationship between the parameter g and the average estimated classical skewness coefficient over 1000 samples of size $n = 100$ and $n = 1000$. Also here we observe a monotone relationship between g and b_1. More-

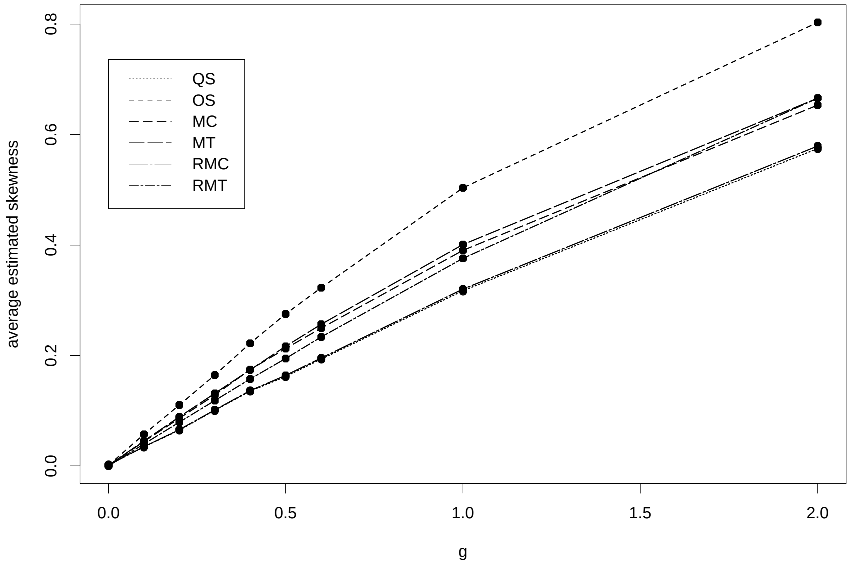

Fig. 2. Average of six skewness measures over 1000 samples of size $n = 100$ for several values of g.

over it is interesting to see how both curves become more and more discrepant when the skewness increases. This is due to the fact that samples of size 1000 are more likely to contain very large values as g increases. Figure 3 thus shows how b_1 explodes in the presence of this kind of outliers. We will investigate this behaviour in more detail in Section 4.

Let us now concentrate on the behaviour of the estimators at a *symmetric* distribution. For this, we consider the simulation for $g = 0$ and $n = 100$. In Table 2 we have listed the average estimated skewness and the standard error of the different estimators. We see that the average estimate is close to zero for all of them, and that MT and RMT have the smallest variation. The classical skewness b_1 on the other hand displays more variation. This is not surprising because of its greater range. This behaviour appeared to be similar at the larger samples with $n = 1000$.

At *right-tailed* distributions we expect the estimated skewness to be positive. Therefore we focus now on the previous simulations for distributions with $g > 0$. Tables 3 and 4 give the frequency of strictly positive values for all the skewness measures for several values of g. We want this frequency to be close to 1. From the two tables we conclude that the classical skewness b_1, the medtriple MT and the repeated medtriple RMT are the best estimators to detect small positive skewness. The quartile skewness QS and the repeated medcouple RMC perform much worse than the others.

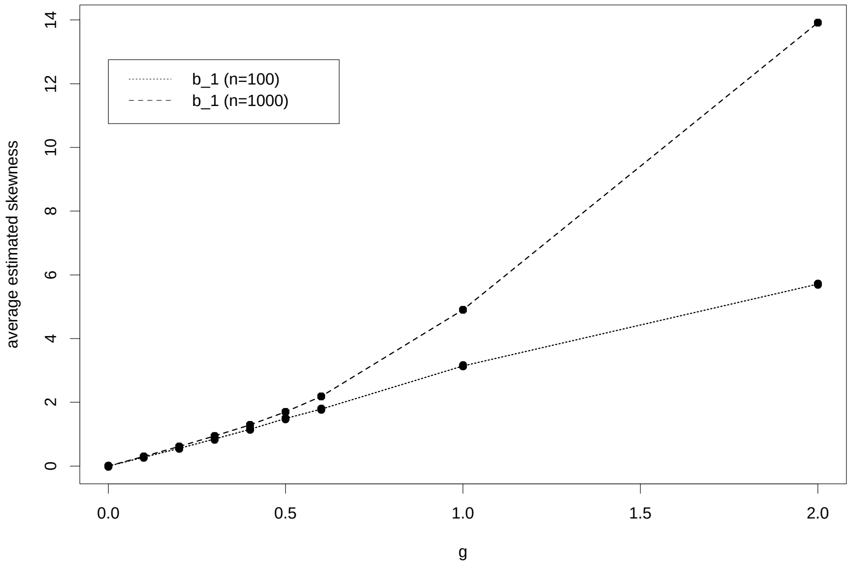

Fig. 3. Average of the classical skewness coefficient b_1 over 1000 samples of size $n = 100$ and $n = 1000$ for several values of g.

4 Performance Under Contamination

As mentioned before, the classical skewness b_1 can be highly influenced by a few outliers. Let us first illustrate this on a real data set.

The speed of light data set (available at the Data and Story Library at `http://lib.stat.cmu.edu/DASL/`) measures the time required for light to travel from a laboratory to a mirror and back, over a total distance of 7400m. This data set contains 66 observations. From the boxplot in Figure 4 it is clear that they are sampled from a symmetric distribution. Moreover, there are two clear outliers. The skewness estimates for this data set are given in Table 5. We see that b_1 is heavily influenced by the outlying observations and that it suggests a left-tailed distribution. The QS and OS are slightly positive, hence they are not attracted by the outliers and even detect a very small right tail. The other estimators reflect very well the symmetry of the regular data points.

Next, we have performed several simulations. As before, data sets of size $n = 100$ were drawn from a G_g distribution with g varying between 0 and 0.5. For $g = 0.3$ the boxplots of the skewness estimates on 1000 random data sets are shown in Figure 5. We see again that the distribution of b_1 has a long tail to the right, due to samples containing points far in the tail of the distribution. For the other values of g we obtained comparable results.

Table 2. Average estimated skewness and standard error at the symmetric distribution G_0.

estimator	ave	st.error
b_1	0.00089	0.0073
QS	0.00200	0.0040
OS	0.00111	0.0033
MC	0.00064	0.0033
MT	0.00081	0.0014
RMC	0.00168	0.0040
RMT	0.00006	0.0019

Table 3. Fraction of skewness estimates strictly positive for 1000 samples of $n = 100$ observations.

estimator	$g = 0.1$	$g = 0.2$	$g = 0.3$	$g = 0.4$	$g = 0.5$	$g = 0.6$
b_1	0.879	0.991	0.999	1.000	1.000	1.000
QS	0.626	0.677	0.761	0.839	0.888	0.918
OS	0.718	0.845	0.946	0.979	0.997	0.998
MC	0.675	0.789	0.890	0.936	0.976	0.974
MT	0.840	0.973	0.999	1.000	1.000	1.000
RMC	0.625	0.680	0.759	0.846	0.895	0.922
RMT	0.738	0.894	0.969	0.994	0.999	0.999

Table 4. Fraction of skewness estimates strictly positive for 1000 samples of $n = 1000$ observations.

estimator	$g = 0.1$	$g = 0.2$	$g = 0.3$
b_1	1.000	1.000	1.000
QS	0.793	0.951	0.994
OS	0.957	0.999	1.000
MC	0.889	0.995	0.999
MT	1.000	1.000	1.000
RMC	0.795	0.950	0.994
RMT	1.000	1.000	1.000

Table 5. Skewness of the speed of light data.

b_1	QS	OS	MC	MT	RMC	RMT
-4.39	0.11	0.09	0	0	0	0

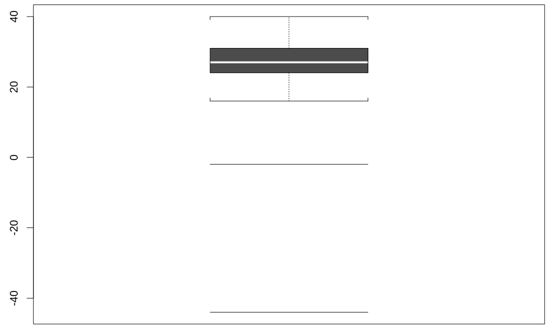

Fig. 4. Boxplot of the speed of light data.

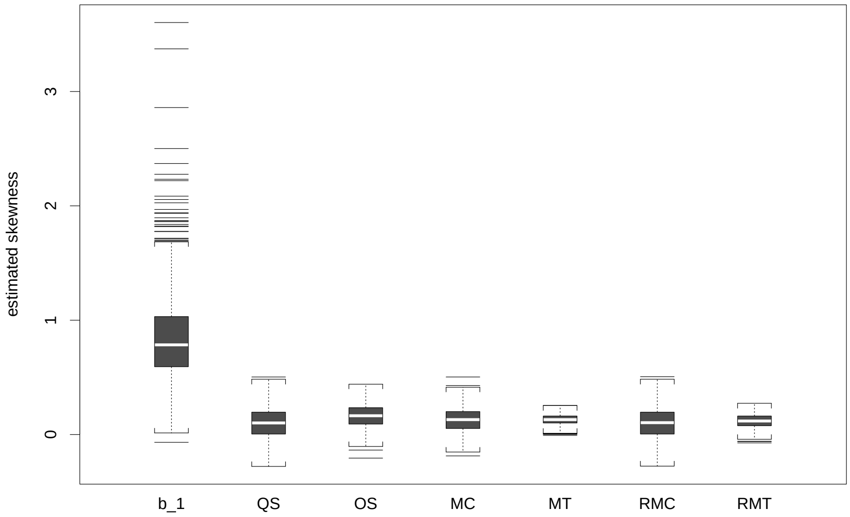

Fig. 5. Boxplots of skewness estimates on 1000 random samples of $n = 100$ observations drawn from $G_{0.3}$.

Then we have replaced 5% of the data (obtained with $g = 0.3$) with outliers spread out far in the right tail of the samples. The boxplots of the estimates on these contaminated samples are shown in Figure 6. We see that the median value and the dispersion of b_1 have increased considerably compared to the uncontaminated situation in Figure 5 (notice the different scales of the vertical axes). On the other hand, the median value and the dispersion of the other estimators have changed very little.

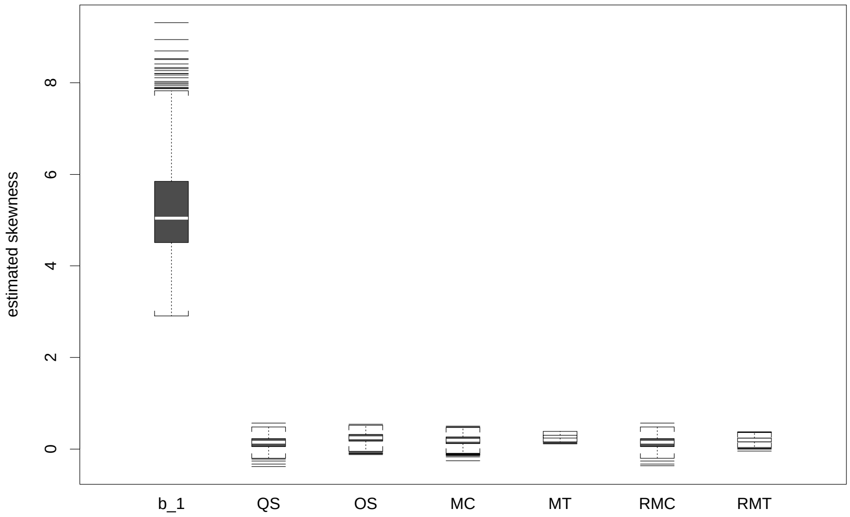

Fig. 6. Boxplots of skewness estimates on 1000 random samples of $n = 100$ observations drawn from $G_{0.3}$ with 5% contamination.

To compare the sensitivity of the six new estimators QS, OS, MC, MT, RMC and RMT in more detail, we have plotted in Figure 7 for each measure and for several values of g the difference between the average estimated value at the contaminated and at the original data sets. From this figure it is seen that with a contamination of 5% RMC and QS are less influenced than MC, OS and RMT, while MT is the most influenced by a small fraction of contamination. The differences between the measures remain quite stable when we vary the skewness parameter g of the underlying distribution.

Next, we have repeated the simulation with 15% of contamination. The corresponding sensitivity curves are shown in Figure 8. Compared to Figure 7 we see that QS, RMC, RMT and MT behave approximately the same with respect to each other. The medcouple MC shows somewhat more sensitivity and comes close to RMT. But the largest change is due to the octile skewness OS which is clearly heavily influenced by the outlying values. This is caused by its low breakdown value of only 12.5%.

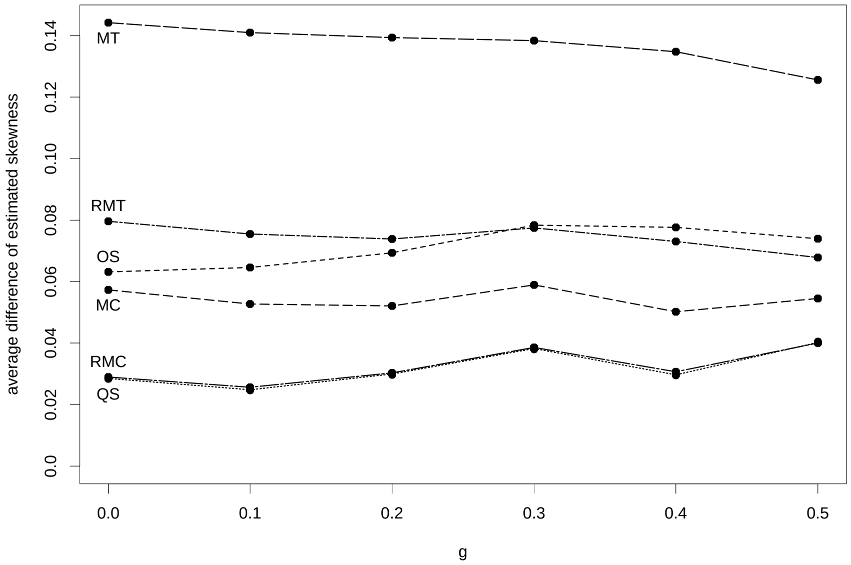

Fig. 7. Difference between average skewness estimate at contaminated and at uncontaminated data, for different values of g and 5% contamination.

Finally we were interested to see how the different skewness measures react when contamination is added in the right tail of a symmetric distribution. For this we have made boxplots of the estimated skewness measures on 1000 samples from G_0 with 5% contamination (Figure 9). Here, we see that MC, RMC, QS and OS are the most robust estimators, while MT and RMT very often estimate positive skewness. As in Figure 6 the classical measure b_1 always yielded very large values (with median 5.05) and was therefore omitted from this plot.

5 Limiting Distributions

To study the limiting distribution of the different estimators, we have made normal QQ-plots of the estimated skewness based on samples of size $n = 1000$ from G_0 and G_2 (see Figure 10 to Figure 15). The QQ-plot of the RMT could not be made due to its computational complexity.

In Moors et al. (1996) it is shown that b_1, QS and similarly OS are asymptotically normal distributed. Figure 10 shows that the rate of convergence of b_1 is very slow at asymmetric distributions, whereas QS (Figure 11) and OS (Figure 12) converge much faster. For the other estimators the limiting distribution is still unknown, but Figures 13, 14 and 15 suggest that normality is indeed satisfied and that it is reached at a faster rate than for b_1.

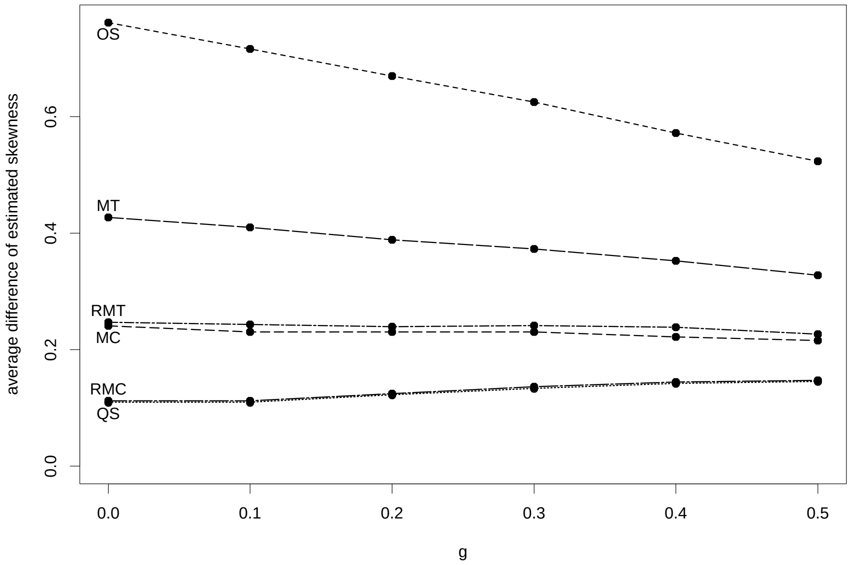

Fig. 8. Difference between average skewness estimate at contaminated and at uncontaminated data, for different values of g and 15% contamination.

6 Conclusions

In this paper we have proposed several new measures of skewness that are not based on moments of the data. Therefore they are not as vulnerable towards outliers as the classical skewness b_1.

When we compare the performance of these measures at uncontaminated symmetric and asymmetric distributions, our preference goes to MT, RMT and b_1 followed by OS and MC. The QS and RMC measures on the other hand do not detect asymmetry adequately.

At contaminated data, we see that b_1, MT and RMT on average give no precise estimates, and that MC outperforms OS.

The medcouple MC is thus the overall winner. With a naive algorithm MC can be computed in O(n^2) time. This is still reasonable compared to the O(n^3) computation of MT and RMT, but it is too slow for large data sets. Therefore we will focus our further research on constructing a faster algorithm for MC. In the meantime, we recommend to use the OS estimator as a faster alternative to MC.

Apart from these computational aspects, we will also study the influence function and limiting distribution of the medcouple.

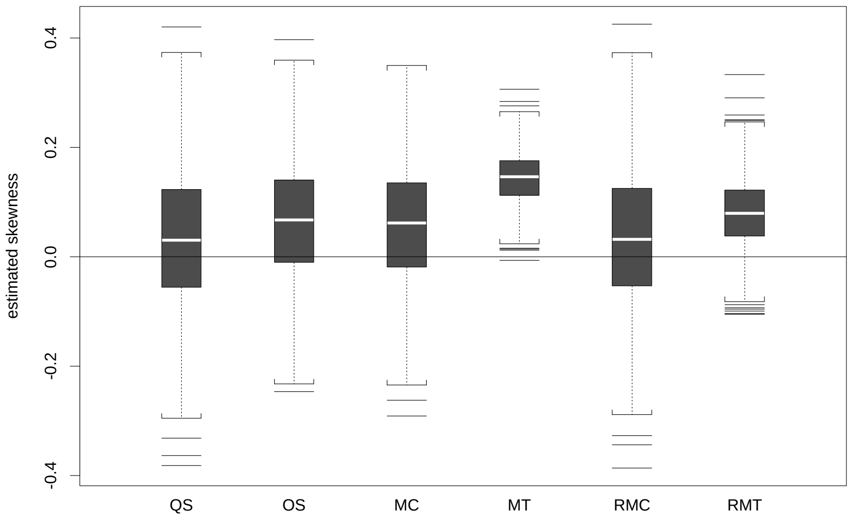

Fig. 9. Boxplots of skewness estimates on 1000 random samples of $n = 100$ observations drawn from G_0 with 5% contamination.

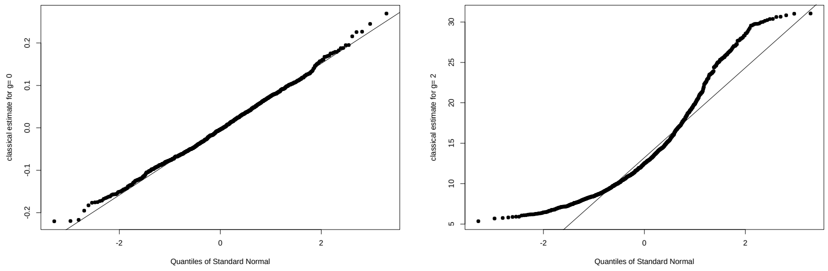

Fig. 10. Normal QQ-plots of b_1 based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

Acknowledgement

We thank both referees for their constructive remarks which helped to improve the presentation of this paper.

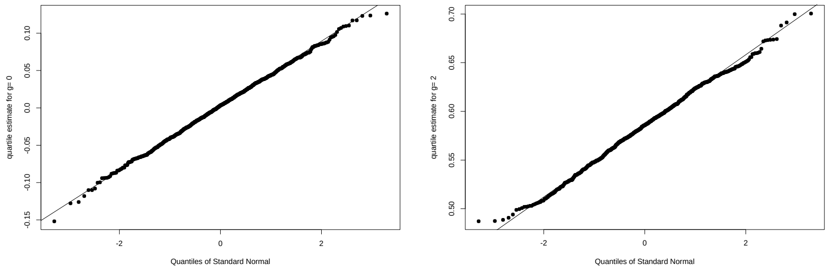

Fig. 11. Normal QQ-plots of QS based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

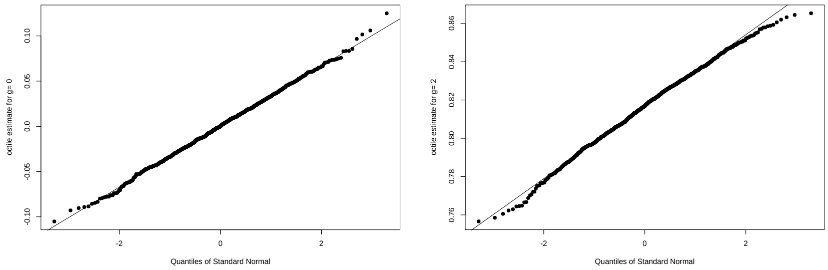

Fig. 12. Normal QQ-plots of OS based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

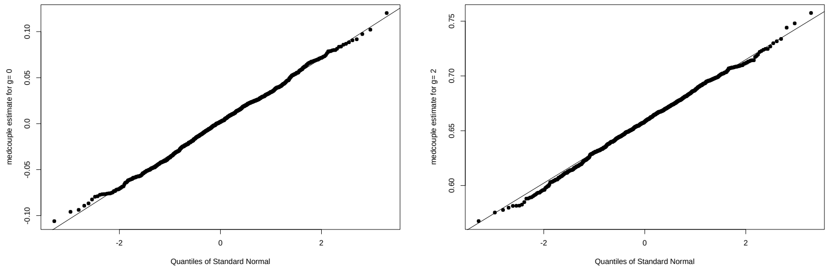

Fig. 13. Normal QQ-plots of MC based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

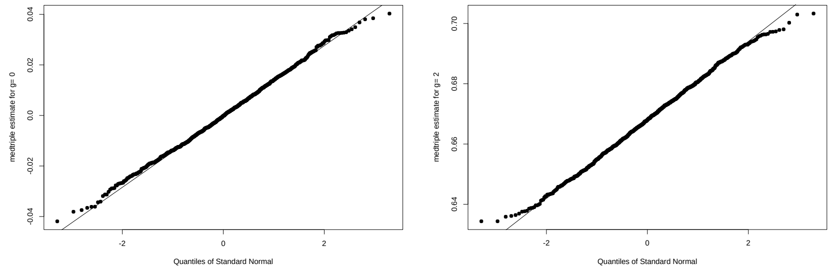

Fig. 14. Normal QQ-plots of MT based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

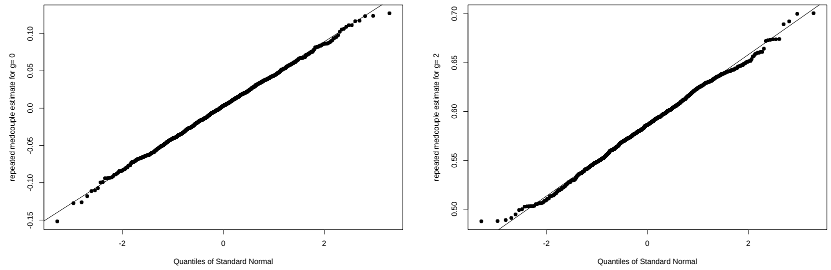

Fig. 15. Normal QQ-plots of RMC based on samples of size $n = 1000$ for $g = 0$ (left panel) and $g = 2$ (right panel).

References

A.L. Bowley. *Elements of statistics*. Charles Scribner's Sons, New York, 1920.

D.V. Hinkley. On power transformations to symmetry. *Biometrika*, 62:101–111, 1975.

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

J.L. Hodges and E.L. Lehmann. Estimates of location based on rank tests. *Ann. Math. Statist.*, 34:598–611, 1963.

J.J.A. Moors, R.Th.A. Wagemakers, V.M.J. Coenen, R.M.J. Heuts, and M.J.B.T. Janssens. Characterizing systems of distributions by quantile measures. *Statistica Neerlandica*, 50: 417–430, 1996.

P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. Am. Statist. Assoc.*, 88:1273–1283, 1993.

P.J. Rousseeuw and M. Hubert. Regression-free and robust estimation of scale for bivariate data. *Computational Statistics and Data Analysis*, 21:67–85, 1996.

P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection.* Wiley, New York, 1987.

P.K. Sen. Estimates of the regression coefficient based on Kendall's Tau. *J. Am. Statist. Assoc.*, 63:1379–1389, 1968.

A.F. Siegel. Robust regression using repeated medians. *Biometrika*, 69:242–244, 1982.

H. Theil. A rank-invariant method of linear and polynomial regression analysis (Parts 1-3). *Ned. Akad. Wetensch. Proc., Ser. A*, 53:386–392, 521–525, 1397–1412, 1950.