

A ROBUSTIFICATION OF THE JARQUE-BERA TEST OF NORMALITY

Guy Brys, Mia Hubert and Anja Struyf

Key words: Robustness, skewness, tail weight, goodness-of-fit tests.

COMPSTAT 2004 section: Robustness.

Abstract: Many statistical tests have been proposed to find out whether a sample is drawn from a normal distribution or not. Here we discuss the Jarque-Bera test [1] which is based on the classical measures of skewness and kurtosis. As these measures are based on moments of the data, this test has a zero breakdown value. In other words, a single outlier can make the test worthless. In this paper we propose normality tests based on robust measures of skewness and tail weight. We investigate their power and their robustness by means of simulations and examples. We also outline how this approach can be extended to test for other distributions than the normal.

1 Introduction

The third and fourth moments of a distribution are called the skewness and kurtosis. For any distribution F with finite central moments μ_k ($k \leq 3$), the *skewness* is defined as

$$\gamma_1(F) = \frac{\mu_3(F)}{\mu_2(F)^{3/2}}.$$

Skewness describes the asymmetry of a distribution. 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.

For any distribution F with finite central moments μ_k ($k \leq 4$), the *kurtosis* is defined as

$$\gamma_2(F) = \frac{\mu_4(F)}{\mu_2(F)^2}.$$

There is no agreement on what it really measures. Strictly speaking, kurtosis measures both peakedness and tail heaviness of a distribution relative to that of the normal distribution. Consequently, its use is restricted to symmetric distributions. Finite-sample versions of γ_1 and γ_2 will be denoted by b_1 and b_2 .

The classical skewness and kurtosis coefficient have some common disadvantages. They both have a zero breakdown value, and so they are very sensitive to outlying values. One single outlier can make the estimate become very large or small, making it hard to interpret. Another disadvantage is that they are only defined on distributions having finite moments.

In Section 2 we propose several measures of skewness and of left and right tail weight for univariate continuous distributions. Their interpretation

is clear and they are robust against outlying values. Contrary to the kurtosis coefficient, the tail weight measures can be applied to symmetric as well as asymmetric distributions. In Section 3 we describe the Jarque-Bera test and introduce some robust normality tests. Section 4 includes simulation results while Section 5 applies the tests on real data. Finally, Section 6 concludes and gives outlines for future research.

2 Robust measures of skewness and tail weight

Assume we have independently sampled n observations $X_n = \{x_1, x_2, \dots, x_n\}$ from a continuous univariate distribution F . We will consider the *medcouple* (MC), a robust skewness measure, proposed in Brys et al. [2] and extensively discussed in Brys et al. [3]. It is defined as

$$MC(F) = \operatorname{med}_{x_1 \leq m_F \leq x_2} h(x_1, x_2)$$

with x_1 and x_2 sampled from F , $m_F = F^{-1}(0.5)$ and the kernel function h given by

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

Furthermore, we consider the *left medcouple* (LMC) and *right medcouple* (RMC), respectively the left and right tail weight measure, as defined in Brys et al. [4]. To construct these measures we have applied the medcouple to respectively the left and right half of the samples:

$$\text{LMC}(F) = -\text{MC}(x < m_F) \quad \text{and} \quad \text{RMC}(F) = \text{MC}(x > m_F).$$

Finite sample versions will be denoted by MC_n , LMC_n and RMC_n . These measures can be computed at any distribution, even when finite moments do not exist. Their computation can be performed in $O(n \log n)$ time due to the fast algorithm described in Brys et al. [3]. They satisfy all natural requirements of skewness or tail weight measures including location and scale invariance. Moreover, they have good robustness properties. More details can be found in the cited references.

3 Normality tests

In this section we discuss four normality tests for the following null and alternative hypothesis:

$$\begin{cases} H_0 : \text{The data are sampled from a normal distribution} \\ H_1 : \text{The data are not sampled from a normal distribution} \end{cases}$$

First, the Jarque-Bera normality test (JB) uses the classical skewness and kurtosis coefficient. As been stated in Moors [7], under the normality assumption ($\gamma_1 = 0$ and $\gamma_2 = 3$) we can write:

	JB	MC1	MC2	MC3
ω	(0 3)	(0)	(0.199 0.199)	(0 0.199 0.199)
Σ_k	(6 0) (0 24)	(1.25)	(-2.62 -0.0123) (-0.0123 2.62)	(1.25 0.323 -0.323) (0.323 2.62 -0.0123) (-0.323 -0.0123 2.62)

Table 1: Asymptotic mean ω and covariance matrix Σ_k of the (joint) distribution of several measures of skewness and tail weight.

$$\sqrt{n} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \rightarrow_{\mathcal{D}} N_2 \left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix} \right)$$

which leads to the Jarque-Bera test statistic:

$$T = n \left(\frac{b_1^2}{6} + \frac{(b_2 - 3)^2}{24} \right) \approx \chi_2^2.$$

This test can be viewed as a special case of the following generalization. Let $w = (w_1, w_2, \dots, w_k)$ be estimators of $\omega = (\omega_1, \omega_2, \dots, \omega_k)$, such that

$$\sqrt{n} (w_1 \ \dots \ w_k) \rightarrow_{\mathcal{D}} N_k (\omega, \Sigma_k)$$

then the generalized test statistic T

$$T = n(w - \omega)^t \Sigma_k^{-1} (w - \omega) \approx \chi_k^2.$$

We can thus easily expand the number of goodness-of-fit tests. Taking $k = 2$, $w_1 = b_1$ and $w_2 = b_2$ leads to the JB test with $\omega_1 = \gamma_1 = 0$ and $\omega_2 = \gamma_2 = 3$. A test based on the medcouple (MC1) given in Brys et al. [3] has $k = 1$ and $w_1 = MC$. Tests solely based on one robust tail weight are proposed in Brys et al. [4]. In the latter paper also a comparison is included with a robust test proposed in Schmid and Trede [8] based solely on quantiles of the data. However as the power of these tests appeared to be rather low, we here construct tests that are based on the joint distribution of several robust measures of skewness and tail weight. First, we define a test based on the left and right medcouple (MC2) with $k = 2$, $w_1 = LMC$ and $w_2 = RMC$. Next, we propose a normality test based on MC, LMC and RMC (MC3) where $k = 3$, $w_1 = MC$, $w_2 = LMC$ and $w_3 = RMC$. Table 1 shows the values of ω and Σ_k under the null hypothesis of normality for the proposed normality tests. They are derived from the influence function of the estimators, as described in Brys et al. [3] and in Brys et al. [4].

Note that a robust test of normality could also be obtained by removing the outliers from the data, using an outlier detection rule such as provided by the boxplot or a rule based on robust estimators of location and scale. When the majority of the data are indeed normally distributed, this is a valuable alternative to our robust tests as both the boxplot and the most popular robust estimators of location and scale (such as M-estimators) are based on this

normal assumption and thus will indicate the correct set of outliers. However it becomes more complicated when even the majority of the data points do not come from a normal distribution. A boxplot for example then typically shows too many outliers, and to construct an outlier rule one should have knowledge about the underlying distribution of the regular observations. Robust tests, such as the ones presented in Schmid and Trede [8], Moors [7] and in this paper, are based on characteristics of the majority of the data points, and hence they do not require an outlier detection procedure. Consequently, these tests are less powerful to detect non-normality than classical tests, but they are less sensitive to outlying values. This will be demonstrated in the next sections.

4 Simulation study

We investigate the performance of the four normality tests at Tukey's class of gh-distributions [6]. When a random variable Z is standard gaussian distributed, then

$$Y_{g,h} = \begin{cases} \frac{(e^{gZ}-1)e^{\frac{hZ^2}{2}}}{g} & g \neq 0 \\ Ze^{\frac{hZ^2}{2}} & g = 0 \end{cases}$$

is said to follow a gh-distribution $G_{g,h}$ with parameters $g \in \mathbb{R}$ and $h \geq 0$. The parameter g controls the skewness of the distribution, whereas h effects the tail weight. We generated 1000 samples of size 1000 from some $G_{g,h}$ distributions, and summarized the results in Table 2 by calculating the fraction of 1000 samples on which the null hypothesis of normality was rejected at the 5% significance level. In this way, the first column depicts the actual size of the tests, while the other columns represent their power.

It is straightforward to see that the JB test outperforms the other normality tests, followed by MC3 which is much more conservative. The poor performance of MC1 at fat tailed but symmetric distributions $G_{0,h}$, is due to the fact this test is based solely on the skewness of the distribution. Although MC2 is based on tail weight only, it also detects deviations from symmetry, which is a consequence of considering both the left and the right tail weight. Nevertheless, the power values of MC1 and MC2 are mostly lower than those of their combined test MC3.

We now compare the robustness of the normality tests using contaminated normal samples. Contaminated samples were created by taking normal samples of size $1000(1-\varepsilon)$, and adding a normal sample of outliers $N(a, \sigma^2=1)$ of size 1000ε with $a = 7$ (right contamination, RC) and $a = -7$ (left contamination, LC). With central contamination (CC) a normal sample $N(0, \sigma^2=0.05)$ of size 1000ε was added. We have also studied a more dispersed symmetric contamination (SC) by adding a normal sample $N(0, \sigma^2=5)$ of size 1000ε . Here, we take ε equal to 1% and 5%. In Table 3 the fraction of 1000 samples on which the null hypothesis of normality was rejected is given. These values should remain close to the prescribed significance level of 5%.

	$G_{0,0,0}$	$G_{0,0,1}$	$G_{0,0,2}$	$G_{0,0,3}$	$G_{0,1,0,0}$	$G_{0,1,0,1}$	$G_{0,3,0,0}$	$G_{0,3,0,1}$
JB	0.038	0.999	1.000	1.000	1.000	1.000	1.000	1.000
MC1	0.038	0.058	0.063	0.076	0.225	0.235	0.965	0.941
MC2	0.036	0.218	0.688	0.939	0.064	0.254	0.425	0.560
MC3	0.030	0.196	0.617	0.914	0.223	0.383	0.986	0.991

Table 2: Fraction of 1000 samples of data size 1000 from several distributions $G_{g,h}$ on which the null hypothesis of normality was rejected at the 5% significance level.

	RC(1%)	RC(5%)	LC(1%)	LC(5%)	SC(1%)	SC(5%)	CC(1%)	CC(5%)
JB	1.000	1.000	1.000	1.000	0.991	1.000	0.058	0.162
MC1	0.061	0.319	0.067	0.338	0.039	0.072	0.028	0.033
MC2	0.056	0.359	0.044	0.362	0.058	0.094	0.041	0.046
MC3	0.050	0.588	0.052	0.598	0.058	0.092	0.031	0.047

Table 3: Fraction of 1000 samples of data size 1000 from several contaminated normal samples on which the null hypothesis of normality was rejected.

We notice that the JB test is very sensitive to outlying values, especially at right, left and symmetric contamination. The robust normality tests perform better in all cases. Their performance is very comparable except at 5% left and right contamination where MC1 and MC2 are more robust than MC3.

5 Examples

In this section we analyse two data sets which illustrate the robustness of our tests compared to the JB test.

The first data set comes from the Associated Examining Board in Guilford [5] and contains a sample of 1000 scores of students on the writing of a paper. From the normal QQ-plot of Figure 1(a) and the boxplot in Figure 1(b) the assumption of normality seems appropriate. Only a few minor outliers are visible on the boxplot. In Table 4 the non-robustness of the JB test is illustrated. Normality is rejected at the 5% significance level when the outliers from the boxplot are included, but is accepted when they are excluded. On the contrary, the robust normality tests are based on the majority of the data and so they behave the same in both situations. As could be expected, they all detect normality in this data set. Note also the higher significance values of the robust tests compared to the JB test.

Our second example is the speed of light data set [9] which 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, which is low compared with our simulation study, but we will see that also in this situation the Jarque-Bera test may fail. From the normal QQ-plot and the boxplot in Figure 2 we have the impression that these data come from a normal

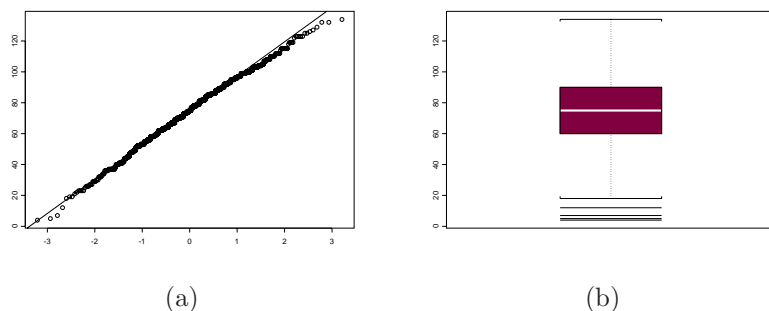


Figure 1: The Guilford data: (a) normal QQ-plot; (b) boxplot.

	JB	MC1	MC2	MC3
Guilford, outliers included	0.039	1.000	0.901	0.975
Guilford, outliers excluded	0.087	1.000	0.967	0.995
Speed of light, outliers included	0.000	1.000	0.308	0.492
Speed of light, outliers excluded	0.855	1.000	0.957	0.992

Table 4: Significance of the tests JB, MC1, MC2 and MC3 at the Guilford and speed of light data set, both with outliers included or excluded.

distribution, apart from two clear outliers. As can be seen from Table 4, the JB test applied on the full data rejects the null hypothesis at any significance level, a result which is due to the two outlying values. Excluding those outliers again leads to the opposite conclusion. The robust normality tests detect in both situations the normal behavior which is present in the large majority of the observations.

Of course we do not know whether the so-called outliers in these two data sets are contaminated observations, or whether the true distribution has a long left tail. We therefore recommend to apply both the robust and the classical test. If they contradict each other, the researcher is stimulated to study the nature of these ‘outliers’ in detail.

6 Conclusions

In this paper we have discussed the Jarque-Bera test of normality which is not able to detect normality in the presence of outlying values. Therefore, three robust normality tests have been proposed, which are all based on the medcouple, a robust measure of skewness. The test MC3 which combines the medcouple with a left (LMC) and a right (RMC) measure of tail weight is seen to give the best overall result.

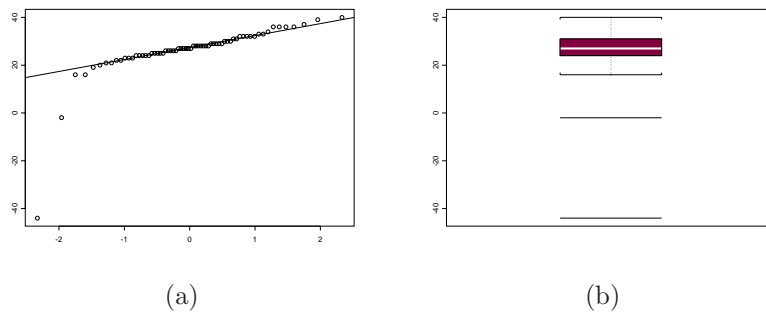


Figure 2: The speed of light data: (a) normal QQ-plot; (b) boxplot.

In our future research, we will investigate the power of these tests at smaller sample sizes and we will generalize our approach to test whether data are sampled from another distribution than the normal (where again, outlier detection rules become complicated). We then only need to compute the asymptotic mean and covariance matrix as in Table 1. This can be done quickly using a Mathematica program which is available from our website (www.agoras.ua.ac.be). Note that the Jarque-Bera test can only be generalized for distributions where the third and fourth moment exist, an assumption which often doesn't hold.

These generalized robust tests will then be used to test distributional assumptions of robust procedures. If, for example, a robust covariance matrix is computed, the underlying assumption is multivariate normality for the majority of the data. A robust test can then be applied to the Mahalanobis distances from the robust fit. The same idea applies to check normality of the residuals from a robust regression procedure.

References

- [1] Bera A., Jarque C. (1981). *Efficient tests for normality, heteroskedasticity and serial independence of regression residuals: Monte Carlo evidence*. *Economics Letter* **7**, 313–318.
- [2] Brys G., Hubert M., Struyf A. (2003a). *A comparison of some new measures of skewness*. In: *Developments in Robust Statistics, ICORS 2001*, R. Dutter, P. Filzmoser, U. Gather and P.J. Rousseeuw (eds.), Springer-Verlag, Heidelberg, 98–113.
- [3] Brys G., Hubert M., Struyf A. (2003b). *A robust measure of skewness*. *Journal of Computational and Graphical Statistics*, to appear.
- [4] Brys G., Hubert M., Struyf A. (2004). *Robust measures of tail weight*. Submitted, available at <http://www.agoras.ua.ac.be>.

- [5] Cresswell M.J. (1990). *Gendar effects in GCSE, some initial analyses*. Research Report, Associated Examining Board, Guilford, **517**.
- [6] Hoaglin D.C., Mosteller F., Tukey J.W. (1985). *Exploring data tables, trends and shapes*. John Wiley and Sons, New York, 1985.
- [7] Moors J.J.A., Wagemakers R.T.A., Coenen V.M.J., Heuts R.M.J., Janssens M.J.B.T. (1996). *Characterizing systems of distributions by quantile measures*. Statistica Neerlandica, **50**, 417–430.
- [8] Schmid F., Tiede M. (2002). *Simple tests for peakedness, fat tails and leptokurtosis based on quantiles*. Computational Statistics and Data Analysis **43**, 1–12.
- [9] Stigler S.M. (1977). *Do robust estimators work with real data*. The Annals of Statistics, **5**, 1055–1098.

Address: G. Brys, Faculty of Applied Economics, Universiteit Antwerpen, Prins-straat 13, B-2000, Antwerpen, Belgium

M. Hubert, Department of Mathematics, Katholieke Universiteit Leuven, W. de Croylaan 54, B-3001 Leuven, Belgium

A. Struyf, Postdoctoral Fellow of the Fund for Scientific Research - Flanders (Belgium), Department of Mathematics and Computer Science, Universiteit Antwerpen, Middelheimlaan 1, B-2020 Antwerpen, Belgium

E-mail: guy.brys@ua.ac.be, mia.hubert@wis.kuleuven.ac.be, anja.struyf@ua.ac.be