



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

**Editors**

H. JOSEPH NEWTON  
 Department of Statistics  
 Texas A&M University  
 College Station, Texas  
 editors@stata-journal.com

NICHOLAS J. COX  
 Department of Geography  
 Durham University  
 Durham, UK  
 editors@stata-journal.com

**Associate Editors**

CHRISTOPHER F. BAUM, Boston College  
 NATHANIEL BECK, New York University  
 RINO BELLOCCO, Karolinska Institutet, Sweden, and  
 University of Milano-Bicocca, Italy  
 MAARTEN L. BUIS, University of Konstanz, Germany  
 A. COLIN CAMERON, University of California–Davis  
 MARIO A. CLEVES, University of Arkansas for  
 Medical Sciences  
 WILLIAM D. DUPONT, Vanderbilt University  
 PHILIP ENDER, University of California–Los Angeles  
 DAVID EPSTEIN, Columbia University  
 ALLAN GREGORY, Queen’s University  
 JAMES HARDIN, University of South Carolina  
 BEN JANN, University of Bern, Switzerland  
 STEPHEN JENKINS, London School of Economics and  
 Political Science  
 ULRICH KOHLER, University of Potsdam, Germany

FRAUKE KREUTER, Univ. of Maryland–College Park  
 PETER A. LACHENBRUCH, Oregon State University  
 JENS LAURITSEN, Odense University Hospital  
 STANLEY LEMESHOW, Ohio State University  
 J. SCOTT LONG, Indiana University  
 ROGER NEWSON, Imperial College, London  
 AUSTIN NICHOLS, Urban Institute, Washington DC  
 MARCELLO PAGANO, Harvard School of Public Health  
 SOPHIA RABE-HESKETH, Univ. of California–Berkeley  
 J. PATRICK ROYSTON, MRC Clinical Trials Unit,  
 London  
 PHILIP RYAN, University of Adelaide  
 MARK E. SCHAFER, Heriot-Watt Univ., Edinburgh  
 JEROEN WEESIE, Utrecht University  
 IAN WHITE, MRC Biostatistics Unit, Cambridge  
 NICHOLAS J. G. WINTER, University of Virginia  
 JEFFREY WOOLDRIDGE, Michigan State University

**Stata Press Editorial Manager**

LISA GILMORE

**Stata Press Copy Editors**

DAVID CULWELL, SHELBI SEINER, and DEIRDRE SKAGGS

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go “beyond the Stata manual” in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

<http://www.stata-journal.com>

**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
<b>Printed &amp; electronic</b>		<b>Printed &amp; electronic</b>	
1-year subscription	\$115	1-year subscription	\$145
2-year subscription	\$210	2-year subscription	\$270
3-year subscription	\$285	3-year subscription	\$375
1-year student subscription	\$ 85	1-year student subscription	\$115
1-year institutional subscription	\$345	1-year institutional subscription	\$375
2-year institutional subscription	\$625	2-year institutional subscription	\$685
3-year institutional subscription	\$875	3-year institutional subscription	\$965
<b>Electronic only</b>		<b>Electronic only</b>	
1-year subscription	\$ 85	1-year subscription	\$ 85
2-year subscription	\$155	2-year subscription	\$155
3-year subscription	\$215	3-year subscription	\$215
1-year student subscription	\$ 55	1-year student subscription	\$ 55

Back issues of the *Stata Journal* may be ordered online at

<http://www.stata.com/bookstore/sjj.html>

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

<http://www.stata-journal.com/archives.html>

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to [sj@stata.com](mailto:sj@stata.com).



Copyright © 2015 by StataCorp LP

**Copyright Statement:** The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, **STATA**, Stata Press, Mata, **mata**, and NetCourse are registered trademarks of StataCorp LP.

# Time-efficient algorithms for robust estimators of location, scale, symmetry, and tail heaviness

Wouter Gelade  
University of Namur  
Centre of Research in the Economics of Development (CRED)  
Namur, Belgium  
wouter.gelade@unamur.be

Vincenzo Verardi  
University of Namur  
Centre of Research in the Economics of Development (CRED)  
Namur, Belgium  
and Université libre de Bruxelles  
ECARES and iCite  
Brussels, Belgium  
vincenzo.verardi@unamur.be

Catherine Vermandele  
Université libre de Bruxelles  
Laboratoire de Méthodologie du Traitement des Données (LMTD)  
Brussels, Belgium  
catherine.vermandele@ulb.ac.be

**Abstract.** The analysis of the empirical distribution of univariate data often includes the computation of location, scale, skewness, and tail-heaviness measures, which are estimates of specific parameters of the underlying population distribution. Several measures are available, but they differ by Gaussian efficiency, robustness regarding outliers, and meaning in the case of asymmetric distributions. In this article, we briefly compare, for each type of parameter (location, scale, skewness, and tail heaviness), the “classical” estimator based on (centered) moments of the empirical distribution, an estimator based on specific quantiles of the distribution, and an estimator based on pairwise comparisons of the observations. This last one always performs better than the other estimators, particularly in terms of robustness, but it requires a heavy computation time of an order of  $n^2$ . Fortunately, as explained in Croux and Rousseeuw (1992, *Computational Statistics* 1: 411–428), the algorithm of Johnson and Mizoguchi (1978, *SIAM Journal of Scientific Computing* 7: 147–153) allows one to substantially reduce the computation time to an order of  $n \log n$  and, hence, allows the use of robust estimators based on pairwise comparisons, even in very large datasets. This has motivated us to program this algorithm for Stata. In this article, we describe the algorithm and the associated commands. We also illustrate the computation of these robust estimators by involving them in a normality test of Jarque–Bera form (Jarque and Bera 1980, *Economics Letters* 6: 255–259; Brys, Hubert, and Struyf, 2008, *Computational Statistics* 23: 429–442) using real data.

# 1 Introduction

When analyzing univariate data, one must estimate location, scale, skewness (SK), and tail-heaviness (kurtosis) parameters of the underlying distribution. Together, these measures effectively characterize the distribution. Several estimators for these parameters are available, but they do not all share the same properties; they differ in Gaussian efficiency, robustness regarding outliers, smoothness of the influence function (IF), and meaning in the case of asymmetric distributions.

In this article, we systematically compare, for each type of parameter (location, scale, SK, and kurtosis), three estimators of different natures. The first one, generally considered the “classical” estimator, is based on the first, second, third, or fourth (centered) moment of the empirical distribution; the second one is defined on the basis of specific quantiles of the distribution; and the third one is based on pairwise comparisons of the observations. We compare the measures by breakdown point (that is, maximal outlier contamination each can withstand), Gaussian efficiency (that is, relative asymptotic variances [ASVs]), and smoothness of the IF (that is, relative sensitivity of the estimator to changing a fraction of points in the sample).

The estimators of the third category perform very nicely, in terms of both efficiency and robustness. This contrasts with the other estimators. The classical estimators of the first category are highly efficient but not robust to outliers. The quantile-based estimators of the second category have the opposite property: they are very robust but not efficient. The pairwise-based estimators of the third category are typically as robust as the quantile-based ones but more efficient (though not always as efficient as the classical estimators). In this sense, these pairwise-based estimators combine the best of two worlds.

However, because these estimators are based on pairwise comparisons, the heaviness of their computation may make them unfeasible in practice. To overcome this excessive computational complexity, we follow the idea developed in Croux and Rousseeuw (1992), which consists in applying the very efficient deterministic algorithm of Johnson and Mizoguchi (1978) for reducing the computation time from an order of  $n^2$  to an order of  $n \log n$ . Commands are programmed to make the estimators of the third type available for applied researchers.

The article is structured as follows. In section 2, we introduce various estimators for the location, scale, SK, and tail heaviness of the distribution from which the data have been generated, and we compare their (asymptotic) Gaussian efficiency and robustness properties. In section 3, we show how these estimators can be used to test the normality of the distribution, following the idea of the Jarque and Bera (1980) statistical test even when outliers are present. In sections 4 and 5, we illustrate the use of the algorithm and the associated commands. We provide an example in section 6, and we conclude in section 7.

Many measures of location, scale, SK, and tail heaviness (kurtosis) have been studied in the statistical literature. In this section, we compare the (asymptotic) Gaussian efficiency and robustness performance of three types of estimators: 1) the “classical” estimators based on (centered) moment of the distribution; 2) the estimators built from specific quantiles of the distribution; and 3) the estimators based on pairwise comparisons or combinations of the observations.

## 2.1 Definitions

We compare the estimators in terms of breakdown value, Gaussian efficiency, and IFs.

The asymptotic breakdown value is the maximal outlier contamination that an estimator can withstand before breaking down (that is, before leading to arbitrary values).

Gaussian efficiency is related to the ASV of the estimator under a Gaussian distribution. The lower the asymptotic variance, the more efficient the estimator.

The IF at a point  $x$  measures the effect of a perturbation of the distribution on the estimator by adding a small probability mass at point  $x$ . We are mostly interested in whether the IF is bounded and smooth. When it is unbounded, the effect of an outlier on the estimator can be arbitrarily large. This implies that the estimator is not robust to outliers. When the IF is smooth, a small change in a data point has only a small effect on the estimator. Thus the IF smoothness tends to improve efficiency.

Note that the IF can also be used to obtain asymptotic confidence intervals (see Hampel et al. [1986, 85 and 226]). Alternatively, jackknifing can be used to obtain confidence intervals.

## 2.2 Location estimators

There is a consensus in applied statistics that the sample mean ( $\bar{x} = 1/n \sum_{i=1}^n x_i$ ) and the sample median  $\{Q_{0.5} = F_n^{-1}(0.5)\}$  are two complementary location estimators: the mean is very efficient with Gaussian data but fragile to outliers (and meaningless with highly asymmetric data), while the median is very robust (and meaningful for asymmetries) but rather inefficient. Both are extensively used in practice.

Less well known is the midpoint estimator using pairwise comparisons introduced by Hodges and Lehmann (1963). The Hodges–Lehmann (HL) estimator is defined by  $\text{HL} = \text{med} \{(x_i + x_j)/2; i < j\}$ .

In terms of robustness, the HL, like the median, outperforms the mean. Figure 1 shows the IFs under the standard Gaussian distribution,  $\Phi$ . The IF of the HL and the median, unlike the mean, are bounded. Thus they both have positive asymptotic breakdown values (see table 1). The median, with a breakdown value of 50%, is more robust to outliers than HL.

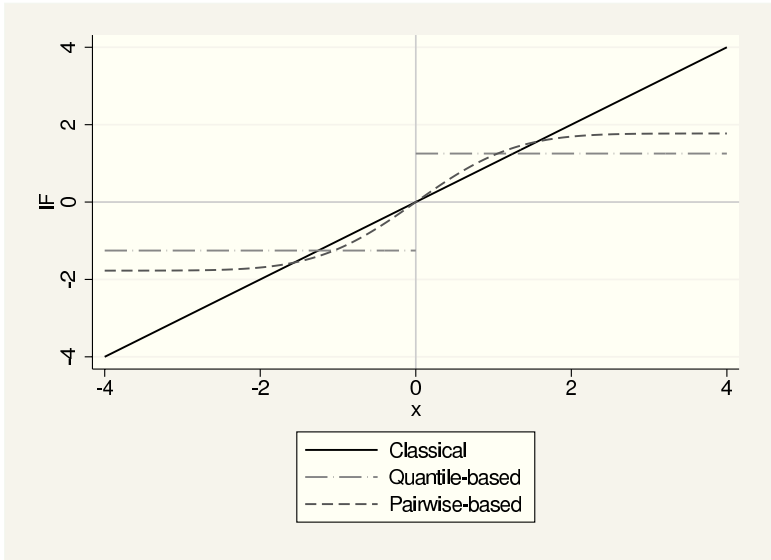


Figure 1. IFs of the location estimators under the standard Gaussian distribution. The IF of HL is bounded and appears as a smooth version of that of the median.

Table 1. A comparison of the three location estimators’ performance with respect to Gaussian efficiency, asymptotic breakdown value, and boundedness of the IF. HL is less robust but more efficient than its robust alternative, the median.

	Type	ASV( $\cdot$ , $\Phi$ )	Asymptotic breakdown val.	Bounded IF?
$\bar{x}$	Classical	1	0%	No
$Q_{0.5}$	Quantile based	$\pi/2$	50%	Yes
HL	Pairwise based	$\pi/3$	29%	Yes

The mean is the most efficient location estimator. Between HL and the median, HL is more efficient, with an ASV under the standard Gaussian distribution of  $\pi/3$  against  $\pi/2$  for the median (see table 1). This is also illustrated by the IFs, where the IF of the HL appears as a smooth version of that of the median.

Although the HL, because it is based on pairwise comparisons, has nice properties, it seems to have a high computational complexity and to be unusable in big samples. However, a simple algorithm proposed by Johnson and Mizoguchi (1978) allows one to substantially reduce this computational complexity from  $O(n^2)$  to  $O(n \log n)$ . We program this estimator in Stata and describe the associated command in section 5. However, we do not believe that it brings enormous advantages for the median and therefore do not strongly advise using it systematically instead of the median.

## 2.3 Scale estimators

To estimate the scale parameter of the underlying distribution, we let the classical estimator be the standard deviation:  $s = \{1/n \sum_{i=1}^n (x_i - \bar{x})^2\}^{1/2}$ . It is the most efficient estimator of the scale parameter,  $\sigma$ , when using Gaussian data. But, like the mean, it completely lacks robustness: its IF is unbounded (see figure 2); thus it has an asymptotic breakdown value of 0% (for example, see Rousseeuw and Croux [1993, 1275]).

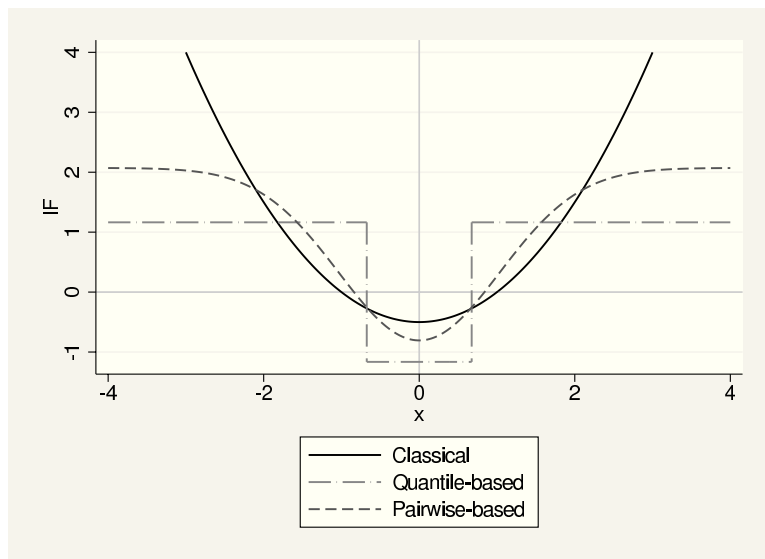


Figure 2. IFs of the scale estimators under the standard Gaussian distribution. The IF of  $Qn$  is bounded and appears as a smooth version of that of IQR.

There are, however, several robust alternatives to estimating the standard deviation. First, there are two alternatives based on quantiles. A commonly used one is the interquartile range (IQR),  $IQR = d \times (Q_{0.75} - Q_{0.25})$ , where setting  $d = 0.7413$  ensures consistency for  $\sigma$  at Gaussian distributions. A second one is the median absolute deviation (MAD),  $MAD = b \times \text{med}_i |x_i - \text{med}_j x_j|$ , where  $b = 1.4826$  makes it consistent for  $\sigma$  at Gaussian distributions. The MAD is very robust, but it aims at symmetric distributions only; it essentially finds the symmetric interval (around the median) that contains 50%



of the data, which does not seem to be a natural approach at asymmetric distributions. The IQR does not have this problem, because the quartiles need not be equally far away from the median. Because the MAD does have this problem of assumed symmetry, we focus mainly on the IQR.

A very interesting but relatively unknown scale estimator is the  $Qn$  statistic of Rousseeuw and Croux (1993),  $Qn = d \times (|x_i - x_j|; i < j)_{(k)}$ , where  $d = 2.2219$  ensures consistency for Gaussian distributions, and  $k = \binom{h}{2} \cong \binom{n}{2}/4$  with  $h = (n/2) + 1$ . In other words, the statistic  $Qn$  corresponds approximately to the 25th percentile of the  $\binom{n}{2}$  distances  $(|x_i - x_j|, i < j)$ .

The  $Qn$  estimator generally outperforms the other robust estimators in both efficiency and robustness. Its Gaussian efficiency, at 83%, is surprisingly high and substantially higher than that of IQR and MAD (see table 2).

Table 2. A comparison of three scale estimators' performance with Gaussian efficiency, asymptotic breakdown value, and boundedness of the IF.  $Qn$  is more robust and more efficient than its robust alternative, IQR.

	Type	ASV( $\cdot, \Phi$ )	Asymptotic breakdown val.	Bounded IF?
$s$	Classical	0.5	0%	No
IQR	Quantile based	1.3605	25%	Yes
$Qn$	Pairwise based	0.6077	50%	Yes

Also, in terms of robustness,  $Qn$  outperforms the other estimators. It has an asymptotic breakdown value of 50%, which is higher than that of IQR. Finally,  $Qn$  is also applicable to asymmetric distributions, and its influence is smooth, unlike that of IQR (see figure 2).

Despite its very nice statistical performance,  $Qn$  may seem difficult to use in practice because of its high computational complexity. Indeed, according to its definition, we must determine an order statistic of  $\binom{n}{2}$  pairwise differences. However, the algorithm proposed in Johnson and Mizoguchi (1978) can be used for the midpoint estimate. We programmed the  $Qn$  estimator with this efficient algorithm and called the command `sqn`. We provide a detailed description of this command in section 5.

## 2.4 SK estimators

The most used SK estimator is the Fisher estimator:  $\gamma_1 = 1/n \sum_{i=1}^n \{(x_i - \bar{x})/s\}^3$ . Because this SK measure relies on the mean and the standard deviation, it is not surprising that its resistance to outliers is null. More precisely, its asymptotic breakdown value is equal to 0%, and its IF is unbounded (see figure 3 and, for example, Groeneveld [1991]). Alternative estimators of SK, such as  $(\bar{x} - \text{mode})/s$  and  $(\bar{x} - Q_{0.5})/s$  proposed by Pearson (1916), still rely on the standard deviation and are thus just as fragile with respect to outliers as the classical SK estimator.

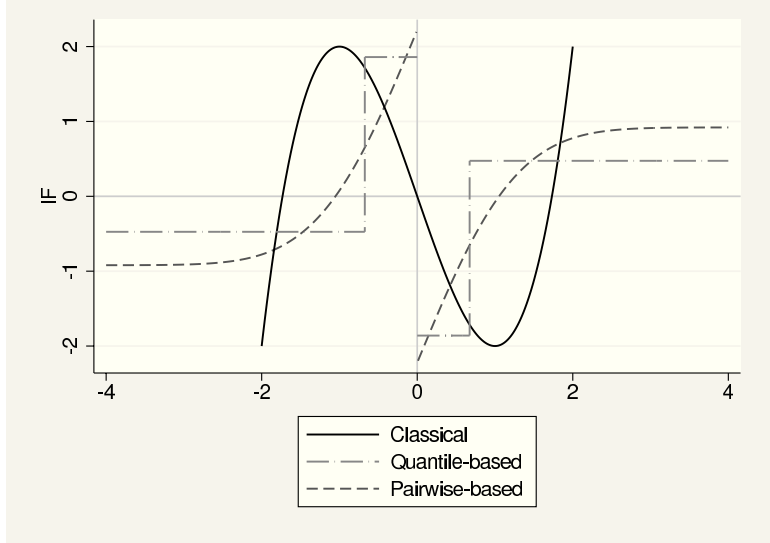


Figure 3. IFs of the SK estimators under the standard Gaussian distribution. The IF of the MC is bounded and appears as a smooth version of that of  $SK_{0.25}$ .

Fortunately, there again are several robust alternatives. First, Hinkley (1975) proposed the following quantile-based estimator,

$$SK_p = \frac{(Q_{1-p} - Q_{0.5}) - (Q_{0.5} - Q_p)}{Q_{1-p} - Q_p} = \frac{Q_p + Q_{1-p} - 2Q_{0.5}}{Q_{1-p} - Q_p}$$

where  $0 < p < 0.5$ . This is a generalization of Yule and Kendall's (1968) (YK) SK estimator, which can be obtained by setting  $p$  equal to 0.25:  $SK_{YK} = SK_{0.25}$ .

An alternative robust SK operator called “medcouple” (MC) was proposed by Brys, Hubert, and Struyf (2004). It is based on pairwise comparisons and replaces the quantiles  $Q_p$  and  $Q_{1-p}$  in  $SK_p$  with actual data points. More precisely,  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  denotes the  $n$ -order statistics associated with the sample; then

$$MC = \text{med}_{x_{(i)} \leq Q_{0.5} \leq x_{(j)}} h(x_{(i)}, x_{(j)})$$

where, for all  $x_{(i)} \neq x_{(j)}$ , the kernel function  $h$  is given by

$$h(x_{(i)}, x_{(j)}) = \frac{(x_{(j)} - Q_{0.5}) - (Q_{0.5} - x_{(i)})}{x_{(j)} - x_{(i)}}$$

For the special case  $x_{(i)} = x_{(j)} = Q_{0.5}$ , we define the kernel as follows:  $m_1 < \dots < m_k$  denotes the indices of the order statistics that are tied to the median  $Q_{0.5}$  (that is,  $x_{(m_l)} = Q_{0.5}$  for all  $l = 1, \dots, k$ ). Then

$$h(x_{(m_i)}, x_{(m_j)}) = \begin{cases} -1 & \text{if } i + j < k + 1 \\ 0 & \text{if } i + j = k + 1 \\ +1 & \text{if } i + j > k + 1 \end{cases}$$

Because of the denominator,  $h(x_{(i)}, x_{(j)})$ , and hence MC, is always between  $-1$  and  $+1$  (like  $SK_p$ ). The MC is zero for symmetric distributions, while it is positive and negative for right- and left-tailed distributions, respectively.

Overall, the MC outperforms the YK  $SK$  estimator. In terms of robustness, they are comparable: both have an asymptotic breakdown value of 25%<sup>1</sup> (see table 3), and both have bounded IFs. However, the IF of the MC is smoother than that of  $SK_p$ .

Table 3. A comparison of three  $SK$  estimators' performance with Gaussian efficiency, asymptotic breakdown value, and boundedness of the IF. The MC is as robust as yet more efficient than its alternative,  $SK_{0.25}$ .

	Type	ASV( $\cdot, \Phi$ )	Asymptotic breakdown val.	Bounded IF?
Fisher	Classical	6	0%	No
$SK_{0.25}$	Quantile based	1.8421	25%	Yes
MC	Pairwise based	1.25	25%	Yes

The big difference between the MC and  $SK_{YK}$  lies in efficiency. The Gaussian efficiency of the former is substantially higher than that of  $SK_{YK}$ . The Gaussian efficiency of these robust estimators is even better than that of the classical Fisher estimator of  $SK$ .

As for  $Qn$ , at first sight, the computational complexity of MC is of the order of  $O(n^2)$ . As for HL and  $Qn$ , the Johnson and Mizoguchi (1978) algorithm can be used to compute MC with a complexity of  $O(n \log n)$ . We programmed this  $SK$  estimator in Stata, and we describe the commands in section 5.

## 2.5 Tail-heaviness estimators

Tail heaviness is traditionally measured using kurtosis:  $\gamma_2 = 1/n \sum_{i=1}^n \{(x_i - \mu_n)/\sigma_n\}^4$ . The parameter  $\gamma_2$  is equal to three for distributions with tails similar to the normal, greater than three for leptokurtic distributions (that is, those with heavier tails than the normal), and smaller than three for platokurtic distributions (that is, those with lighter tails than the normal). However, this parameter also measures the “peakedness” of a

1. The asymptotic breakdown value of  $SK_p$  is 100% and is thus bigger than 25% when setting  $p$  higher than 0.25 (the value used in  $SK_{YK}$ ). Doing this, however, also reduces efficiency.

distribution, and one disadvantage of using kurtosis is the difficulty of grasping what kurtosis really estimates. Another disadvantage is that its interpretation, and thus its use, is restricted to symmetric distributions. Moreover, as usual for estimators relying on the mean and the standard deviation, the kurtosis coefficient is very sensitive to outliers in the data (0% asymptotic breakdown value and unbounded IF<sup>2</sup>; see figure 4 and Ruppert [1987]).

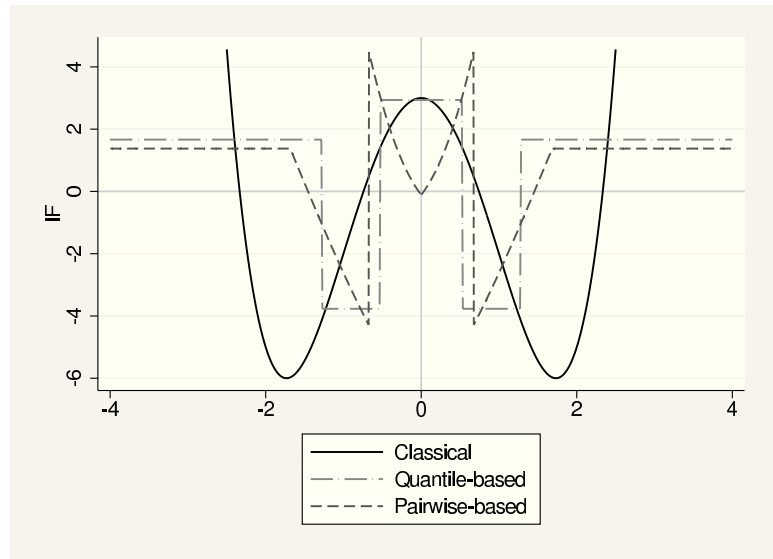


Figure 4. IFs of the tail-heaviness estimators under the standard Gaussian distribution

To overcome these problems, Brys, Hubert, and Struyf (2006) proposed two measures of left- and right-tail weight for univariate continuous distributions. These measures can be applied to symmetric as well as asymmetric distributions that do not need to have finite moments. Moreover, they unambiguously measure tail heaviness (not peakedness), and they are robust against outlying values.

---

2. The form of this IF shows that contamination at the center has far less influence than that in the extreme tails. This suggests that  $\gamma_2$  is primarily a measure of tail behavior and only to a lesser extent of peakedness.

More precisely, Brys, Hubert, and Struyf (2006) defined left- and right-tail measures as measures of SK that are applied to the half of the probability mass on the left and on the right, respectively, of the median  $Q_{0.5}$ . They use the two previously mentioned robust estimators,  $SK_p$  ( $0 < p < 0.5$ ) and MC.

By applying the MC to each side of the distribution, we get the left medcouple (LMC) and the right medcouple (RMC):  $LMC = -MC(x < Q_{0.5})$  and  $RMC = MC(x > Q_{0.5})$ . Similarly, using the  $SK_p$  SK estimator, we get the left quantile weight (LQW) and the right quantile weight (RQW):  $LQW_p = -SK_{p/2}(x < Q_{0.5})$  and  $RQW_p = SK_{p/2}(x > Q_{0.5})$ . Thus  $LQW_{0.25}$ , for instance, considers the quantiles  $Q_{0.125}$  and  $Q_{0.375}$  around the center of the left side of the distribution ( $Q_{0.25}$ ). For these estimators, a higher value of these estimators means a heavier tail. For comparison, note that the tail weights of the normal distribution are 0.2.

The performance of these robust measures of tail heaviness is strictly connected to the performance of their underlying estimators (MC and  $SK_p$ ) and so follows the same pattern as before. The LMC and RMC have a higher Gaussian efficiency than  $LQW_{0.25}$  and  $RQW_{0.25}$  (they are also more efficient than the classical kurtosis for Gaussian data). In terms of robustness, all of their IFs are bounded (see figure 4 and Brys, Hubert, and Struyf [2006, 740–741]), and the asymptotic breakdown value is the same for LMC and RMC and  $LQW_p$  and  $RQW_p$  when  $p = 0.25$  (see table 4). For the latter estimators, increasing  $p$  increases robustness but decreases efficiency; decreasing  $p$  does the opposite. This presents another advantage of the LMC and RMC; they do not require one to (somewhat arbitrarily) fix a value for  $p$ .

Table 4. A comparison of the tail-heaviness estimators’ performance with respect to Gaussian efficiency, asymptotic breakdown value, and boundedness of the IF

	Type	ASV( $\cdot, \Phi$ )	Asymptotic breakdown val.	Bounded IF?
$\gamma_2$	Classical	24	0%	No
$LQW_{0.25}$ ( $RQW_{0.25}$ )	Quantile based	3.71	12.5%	Yes
LMC (RMC)	Pairwise based	2.62	12.5%	Yes

Note that the efficient algorithm of Johnson and Mizoguchi (1978) allows one to again substantially reduce the computational complexity needed to compute LMC and RMC. These tail-weight measures have been implemented as separate options in the MC code (see section 5).

### 3 Normality test based on SK and tail-heaviness estimators

As stated above, these descriptive statistics can be used to characterize the underlying distribution. In particular, they can be used to test for normality. For example, the

Jarque and Bera (1980) test relies on the classical SK and kurtosis coefficients to test for normality. More precisely, under the normality assumption ( $\gamma_1 = 0$  and  $\gamma_2 = 3$ ), we can write

$$\sqrt{n} \begin{pmatrix} \gamma_1 \\ \gamma_2 - 3 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 6 & 0 \\ 0 & 24 \end{pmatrix} \right\}$$

which leads to the Jarque–Bera test statistic,

$$T = n \left\{ \frac{\gamma_1^2}{6} + \frac{(\gamma_2 - 3)^2}{24} \right\} \approx \chi_2^2$$

The Jarque–Bera test is a very popular and interesting test for normality; it outperforms such tests as the Kolmogorov–Smirnov test, the Cramér–von Mises test, and the Durbin test for many alternative distributions (Shapiro, Wilk, and Chen 1968). Unfortunately, despite its good power properties and computational simplicity, the Jarque–Bera test is highly sensitive to outliers because it is constructed from the moment-based SK and kurtosis measures.

Brys, Hubert, and Struyf (2008) proposed robust alternatives to the Jarque–Bera test. These authors explain that the Jarque–Bera test can be seen as a special case of the following general testing procedure: Let  $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \dots, \widehat{\theta}_k)'$  be a vector of estimators of  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$  (a vector of characteristic parameters of the underlying distribution) such that, under the null hypothesis of normality,

$$\sqrt{n} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

Then, the general test rejects the null hypothesis of normality at level  $\alpha$  if

$$T = n (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \boldsymbol{\Omega}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) > \chi_{k;1-\alpha}^2$$

where  $\chi_{k;1-\alpha}^2$  is the  $(1 - \alpha)$ -quantile of the  $\chi^2$  distribution with  $k$  degrees of freedom. Brys, Hubert, and Struyf (2008) then propose to use the robust SK estimator MC or the tail-heaviness estimators LMC and RMC.

Three tests have been studied. The first one uses the SK estimator MC; for this,  $k = 1$ ,  $\widehat{\boldsymbol{\theta}} = \text{MC}$ , and  $\boldsymbol{\Omega} = 1.25$ . The second one uses the left and right tail-heaviness estimators, LMC and RMC; in this case,  $k = 2$ ,  $\widehat{\boldsymbol{\theta}} = (\text{LMC}, \text{RMC})'$ ,  $\boldsymbol{\theta} = (0.199, 0.199)'$ , and

$$\boldsymbol{\Omega} = \begin{pmatrix} 2.62 & -0.0123 \\ -0.0123 & 2.62 \end{pmatrix}$$

The third test combines MC, LMC, and RMC; here  $k = 3$ ,  $\widehat{\boldsymbol{\theta}} = (\text{MC}, \text{LMC}, \text{RMC})'$ ,  $\boldsymbol{\theta} = (0, 0.199, 0.199)'$ , and

$$\boldsymbol{\Omega} = \begin{pmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{pmatrix}$$

This last test seems to have the best overall performance.

Efficiently implementing the different estimators using the pairwise combinations described above involves a simple algorithm by Johnson and Mizoguchi (1978). Given a number  $k$  and an  $n \times q$  matrix  $\mathbf{M}$  with sorted (nonincreasing) rows, this algorithm finds the  $k$ th maximal element of  $\mathbf{M}$  in time  $O(n \log n)$ . In this section, we illustrate this algorithm.

The algorithm repeatedly guesses a new candidate for the  $k$ th maximum of  $\mathbf{M}$ . After each guess, it discards some of the elements of  $\mathbf{M}$  (because they cannot be the  $k$ th maximum). In this way, it systematically reduces this set of candidates until it guesses the  $k$ th maximum of  $\mathbf{M}$ . The key to the efficiency of the algorithm is that at every guess, it (efficiently) discards many elements of  $\mathbf{M}$ . By doing this, the algorithm needs few attempts before finding the  $k$ th maximum.

We present the algorithm in more detail below. It is part Stata code and part *pseudocode* (in *slanted type*).

```

matrix excludeleft = J(n,1,1)          // length-n vector with value 1 everywhere  1
matrix excluderight = J(n,1,q)         // length-n vector with value q everywhere  2
while kth maximum is not found {       3
    scalar m = new guess for kth maximum using nonexcluded elements of M         4
    scalar nr bigger = number of elements a in M with a > m                       5
    scalar nr smaller = number of elements a in M with a < m                     6
    if nr bigger >= k                    // kth maximum is bigger than m           7
        excluderight[i] = position smallest element bigger than m in row i, for all i  8
    else if nr smaller >= (n*q)-k        // kth maximum is smaller than m         9
        excludeleft[i] = position biggest element smaller than m in row i, for all i 10
    else                                // kth maximum equals m                   11
        m is the kth maximum element of M                                       12
}
```

In lines one and two, the algorithm initializes the data structure, which maintains the elements that have already been discarded. At any time, for any row  $i$ , all elements to the left of position `excludeleft[i]` are discarded. For example, if `excludeleft[1]` is four, the first three elements in row one are discarded. Similarly, `excluderight[i]` contains the position in row  $i$  to the right of which all elements have been discarded.

The loop on line three continues until the  $k$ th maximum has been found. It first makes a new guess  $m$  for the  $k$ th maximum (line four) and then calculates the number of elements in  $\mathbf{M}$  that are greater and smaller than  $m$  (lines five and six). This can be done efficiently because each row is sorted. Then, if there are more than  $k$  elements that are greater than  $m$  (line seven), the  $k$ th maximum must be greater than  $m$ . We can thus discard all elements smaller than or equal to  $m$  (line eight). Similarly, we can discard all elements greater than or equal to  $m$  if there are more than  $(n \times q) - k$  elements smaller than  $m$ . Finally, if the  $k$ th maximum is neither among the elements strictly greater than nor among those strictly smaller than  $m$  (line 11), then  $m$  must be the  $k$ th maximum, and we have found the  $k$ th maximum.

This algorithm requires time  $O(n \log n)$  (Johnson and Mizoguchi 1978). Note, however, that one should not explicitly calculate the entire matrix  $\mathbf{M}$ , which would be of

complexity  $O(n^2)$ . Indeed the algorithm needs to inspect only some elements, and only those elements should be calculated. This substantially reduces the running time of the algorithm. Table 5 illustrates the time saved when using this algorithm with a comparison of the running time<sup>3</sup> of the different estimators based on pairwise comparisons when using this algorithm and when using standard algorithms.

Table 5. Running time (in seconds) of the implemented estimators (HL,  $Qn$ , and MC) using the efficient Johnson and Mizoguchi (1978) algorithm (left) and using standard algorithms (right). When  $n$  is greater than 10,000, the running time of the standard algorithm is not reported, because it took too long to compute.

$n$	HL efficient	HL naïve	$Qn$ efficient	$Qn$ naïve	MC efficient	MC naïve
500	0.2	0.1	0.2	0.1	0.1	0.1
1,000	0.4	0.5	0.4	0.6	0.3	0.5
2,000	0.9	2.9	0.9	2.6	0.7	2.0
5,000	2.0	23.0	3.0	22.0	2.0	15.0
10,000	6.0	113.0	7.0	117.0	5.0	72.0
50,000	46.0	/	44.0	/	33.0	/
100,000	105.0	/	108.0	/	74.0	/

## 5 Commands

We programmed the following commands in Stata to estimate the described statistics using the efficient algorithm of Johnson and Mizoguchi (1978).

- For the HL statistic of Hodges and Lehmann (1963), the command is  
`mhl varname [if] [in]`
- For the  $Qn$  statistic of Rousseeuw and Croux (1993), the command is  
`sqn varname [if] [in]`
- For the MC, the command is  
`medcouple varname [if] [in] [, lmc rmc nomc]`
- For the robust test of normality, we created the command  
`robjb varname [if] [in] [, level(#) {skewness|kurtosis} right]`

3. We used Stata/SE 12 and a computer with a 2.66 GHz Intel dual-core processor and 2GB of RAM.



### 5.1 Options for `medcouple`

`lmc` specifies calculating the MC only for observations smaller than the median. This is an indicator of the heaviness of the left tail.

`rmc` specifies calculating the MC only for observations larger than the median. This is an indicator of the heaviness of the right tail.

`nomc` specifies not to calculate the global MC. This is useful when one is interested in only the tail heaviness.

## 5.2 Options for `robjb`

`level(#)` specifies the confidence level for inference. The default is `level(0.95)`.

`skewness` specifies that a test exclusively based on SK be run.

`kurtosis` specifies that a test exclusively based on the heaviness of the tails be run.

`right` specifies that a test exclusively based on the heaviness of the right tail be run.

## 6 Example

To illustrate the usefulness of the estimators, we will analyze the body weight of 64 different animal species. The dataset we use is available online.<sup>4</sup> These data have been made available by Rice University, University of Houston–Clear Lake, and Tufts University.

We first calculate the classical, quantiles-based, and pairwise-based estimates of location, scale, SK, and tail heaviness (see table 6). We can easily calculate the classical and quantiles-based estimates by using the formulas provided in the theoretical section and using the standard `summarize` and `centile` Stata commands (see the do-file pertaining to the example for further details). To compute the pairwise-based estimates, we use the commands as follows:

```
mhl body
sqn body
medcouple body, lmc rmc
```

---

4. See [http://onlinestatbook.com/stat\\_sim/transformations/body\\_weight.html](http://onlinestatbook.com/stat_sim/transformations/body_weight.html).

Table 6. Classical, quantiles-based, and pairwise-based estimates of location, scale, SK, and tail heaviness

		Classical		Quantiles based		Pairwise based
Location	$\bar{x}$ :	3,111,355.5	$Q_{0.5}$ :	3,500.0	HL:	94,307.5
Scale	$s$ :	13,033,900.0	IQR:	166,221.28	$Qn$ :	6,665.438
SK	$\gamma_1$ :	5.461	$SK_{0.25}$ :	0.976	MC:	0.985
Tails	$\gamma_2$ :	32.770	LQW:	−0.052	LMC:	−0.090
			RQW:	0.883	RMC:	0.915

If we look at only the classical estimators, we conclude that the average animal weight is very high but with a huge dispersion. The asymmetry is large and positive, and the tails are very heavy. When we look at the equivalent robust statistics, we see that the median weight is much lower than the mean weight. The robust dispersion is also much smaller than that suggested by classical estimators, and the right SK is extreme. As for tail heaviness, the right tail is extremely heavy, while the left one is slightly lighter than the left tail of the normal (recall that the normal has tail weights of 0.2 and that higher values indicate heavier tails). The difference between classical and robust estimators indicates that outliers are present in the dataset.

For this problem, we can transform the data to reduce the excessive importance of very big animals (such as dinosaurs). Given that weights are strictly positive, we consider a logarithmic transformation and redo the above descriptive statistics analysis (see figure 5 and table 7) as follows:

```
generate lbody = ln(body)
mhl lbody
sqn lbody
medcouple lbody, lmc rmc
```

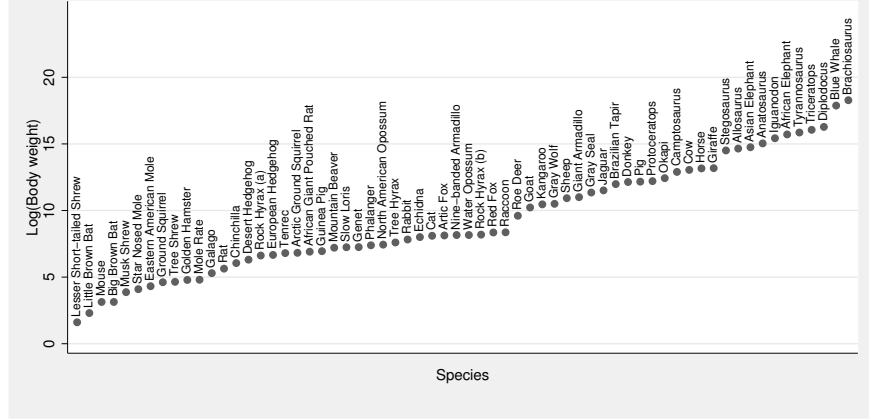


Figure 5. Logarithm of the body weights, in increasing order, of 64 animal species

Table 7. Classical, quantiles-based, and pairwise-based estimates of location, scale, SK, and tail heaviness for transformed data

	Classical	Quantiles based	Pairwise based
Location	$\bar{x}$ : 9.313	$Q_{0.5}$ : 8.161	HL: 9.289
Scale	$s$ : 4.135	IQR: 4.207	$Qn$ : 4.281
SK	$\gamma_1$ : 0.304	$SK_{0.25}$ : 0.465	MC: 0.386
Tails	$\gamma_2$ : 2.192	LQW: 0.499	LMC: 0.515
		RQW: 0.241	RMC: 0.241

When we do this transformation, we see that the differences between classical and robust estimators become much smaller. The mean is only slightly larger than the median, the dispersion estimate is very similar for all the methods, and the SK estimate suggests moderate positive SK. As for tail heaviness, the classical estimator is close to 3, which is the value of the kurtosis for the normal and therefore suggests standard tails. Nevertheless, the robust estimator for the latter indicates a heavy left tail. Indeed, the left-tail weights of about 0.5 are substantially above 0.2 (the tail weight of normal tails). This last point is very important.

Let's imagine that we want to test for the normality of the  $\log(\text{body})$  variable. The classical and the robust estimators give different results. The standard Jarque–Bera statistic is 2.726, which is much smaller than the critical value of  $\chi^2_{2;0.95} = 5.99$ . This implies that the standard Jarque–Bera test would not reject the null hypothesis of normality. On the other hand, we can also calculate the robust test statistic involving MC, LMC, and RMC as follows:

`robjb lbody`

This robust test statistic is equal to 9.266, which is larger than the critical value of  $\chi^2_{3,0.95} = 7.815$ . This indicates the rejection of the null hypothesis of normality. This means that even though the logarithmic transformation substantially reduced the effect of atypical individuals, outliers still bias the classical estimations. In particular, we believe that the heaviness of the left tail is not satisfactorily identified by the classical kurtosis coefficient, and this affects the result of the normality test.

## 7 Conclusion

Different statistics are available to estimate the location, the scale, the SK, and the tail heaviness of a distribution. Some of these estimators are based on pairwise comparisons of the observations; these estimators, apparently much heavier and more complex to compute, perform better, specifically in terms of robustness and efficiency. The algorithm of Johnson and Mizoguchi (1978) allows for a substantial reduction in the computation time of these robust estimators from an order  $n^2$  to an order of  $n \log n$ . This makes it possible to determine these estimators even in very large datasets. To make these estimators available for applied researchers, we have programmed them in Stata, following the efficient algorithm of Johnson and Mizoguchi (1978). We have also programmed a robust version of the Jarque–Bera (1980) test of normality, for which the test statistic involves some of these estimators of SK and tail heaviness.

## 8 Acknowledgment

We gratefully acknowledge the financial support of the National Fund for Scientific Research.

## 9 References

- Brys, G., M. Hubert, and A. Struyf. 2004. A robust measure of skewness. *Journal of Computational and Graphical Statistics* 13: 996–1017.
- . 2006. Robust measures of tail weight. *Computational Statistics and Data Analysis* 50: 733–759.
- . 2008. Goodness-of-fit tests based on a robust measure of skewness. *Computational Statistics* 23: 429–442.
- Croux, C., and P. J. Rousseeuw. 1992. Time-efficient algorithms for two highly robust estimators of scale. In *Computational Statistics*, ed. Y. Dodge and J. Whittaker, vol. 1, 411–428. Heidelberg: Physica-Verlag.
- Groeneveld, R. A. 1991. An influence function approach to describing the skewness of a distribution. *American Statistician* 45: 97–102.
- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

- Hinkley, D. V. 1975. On power transformations to symmetry. *Biometrika* 62: 101–111.
- Hodges, J. L., Jr., and E. L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34: 598–611.
- Jarque, C. M., and A. K. Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6: 255–259.
- Johnson, D. B., and T. Mizoguchi. 1978. Selecting the  $K$ th element in  $X + Y$  and  $X_1 + X_2 + \cdots + X_m$ . *SIAM Journal on Scientific Computing* 7: 147–153.
- Pearson, K. 1916. Mathematical contributions to the theory of evolution. XIX: Second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London, Series A* 216: 429–457.
- Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88: 1273–1283.
- Ruppert, D. 1987. What is kurtosis?: An influence function approach. *American Statistician* 41: 1–5.
- Shapiro, S. S., M. B. Wilk, and H. J. Chen. 1968. A comparative study of various tests for normality. *Journal of the American Statistical Association* 63: 1343–1372.
- Yule, G. V., and M. G. Kendall. 1968. *An Introduction to the Theory of Statistics*. 14th ed. London: Griffin.

### About the authors

Wouter Gelade is a doctoral researcher of the Fonds National de La Recherche Scientifique at the CRED of the University of Namur, Belgium, and holds a PhD in computer science. His current research interests are development economics, microfinance, and applied econometrics.

Vincenzo Verardi is an associate researcher of the Fonds National de La Recherche Scientifique and a professor of economics and econometrics at the University of Namur and at the Université libre de Bruxelles, Belgium. His research interests are applied econometrics, development economics, political economics, and public finance.

Catherine Vermandele teaches statistics at the Université libre de Bruxelles, Belgium, and is responsible for the LMTD. She is particularly interested in nonparametric statistics, robust statistical methods, and sampling theory.