TESTING VARIOUS ROBUST ESTIMATORS OF THE MOMENTS OF A DISTRIBUTION

Frédéric GALLIANO

AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France

Created on August 2, 2018
Updated on October 9, 2019

Contents

1 LITERATURE SURVEY OF THE MOST COMMON ESTIMATORS

1.1 Definitions

Let's note $[X_i, Y_i]$ $(i = 1 \dots N)$ a set of random variables X and Y drawn from the joint *Probability Distribution Function* (PDF), $g(x, y)$. We note the marginalized PDF of X, $f(x) = \int_{-\infty}^{\infty} g(x, y)\,\mathrm{d}y$. The question is how do we estimate the following moments, in the discrete finite case:

The mean: $\mu(X) \equiv \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$;

The variance: $V(X) \equiv \int_{-\infty}^{\infty} [x - \mu(X)]^2 f(x)\,\mathrm{d}x$ and the corresponding *standard deviation*, $\sigma(X) \equiv \sqrt{V(X)}$;

The skewness: $\gamma_1(X) \equiv \int_{-\infty}^{\infty} \left[\frac{x - \mu(X)}{\sigma(X)}\right]^3 f(x)\,\mathrm{d}x$;

The covariance: $V(X,Y) \equiv \iint_{-\infty}^{\infty} [x-\mu(X)]\,[y-\mu(Y)]\,g(x,y)\,\mathrm{d}x\,\mathrm{d}y$ and the corresponding *correlation coefficient*, $\rho(X,Y) \equiv \dfrac{V(X,Y)}{\sigma(X)\sigma(Y)}$.

1.2 Classical Estimators: Expectations

These moments can easily be estimated using the discrete expression of the statistical expectation. However, they are known to not be robust (*i.e.* sensitive to outliers; *cf.* Sect. 2). The expressions are:

$$\hat{\mu}(X) = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

$$\hat{\sigma}(X) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[X_i-\hat{\mu}(X)]^2} \tag{2}$$

$$\hat{\gamma}_1(X) = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left[\frac{X_i-\hat{\mu}(X)}{\hat{\sigma}(X)}\right]^3 \tag{3}$$

$$\hat{\rho}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{X_i-\hat{\mu}(X)}{\hat{\sigma}(X)}\right]\left[\frac{Y_i-\hat{\mu}(Y)}{\hat{\sigma}(Y)}\right]. \tag{4}$$

The last one (Eq. 4) is also known as the *Pearson correlation coefficient*.

1.3 Percentile-Based Estimators

Noting $F(x) \equiv p(X \leq x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x$, the *Cumulative Distribution Function* (CDF), we can define the three quartiles: $Q_1 = F^{-1}(0.25)$, $Q_2 = F^{-1}(0.5)$ and $Q_3 = F^{-1}(0.75)$, Q_2 being the *median* (*cf.* Fig. 1). In the discrete case, $F^{-1}(v)$ is simply the value of the $v \times N^{\text{th}}$ ordered element of X. We can also define the c^{th} percentile as $P_c = F^{-1}(c/100)$. Estimators based on these quantities are less-sensitive to outliers.

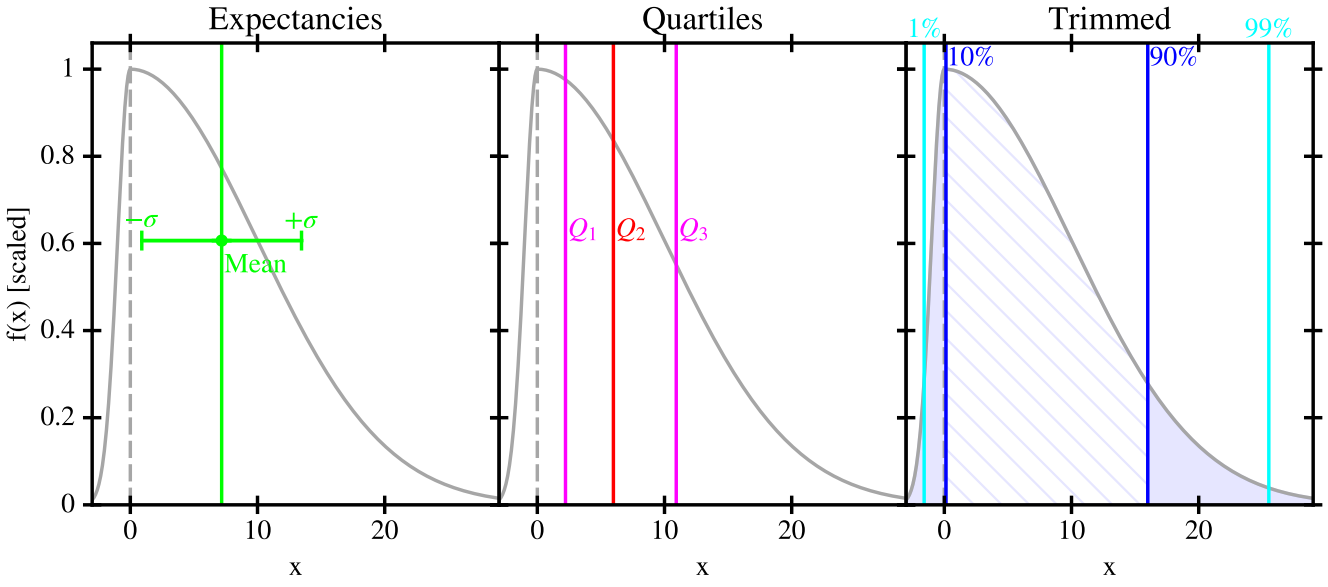

Figure 1: *Visual representation of several estimators.*

1.3.1 Estimators of the Mean

Two, easy-to-compute, well-known robust estimators of the mean use these quantities:

The median: $\mathrm{med}(X) = Q_2$, which is robust, but can be biased if the PDF is skewed;

The trimean: $\mu_3(X) = \dfrac{Q_1 + 2Q_2 + Q_3}{4}$, which accounts for the asymmetry of the PDF.

1.3.2 Estimator of the Standard-Deviation

Similarly, the "*Median Absolute Deviation*" (MAD)[1] is a robust estimator of the standard deviation:

$$\mathrm{MAD}(X) = k \times \mathrm{med}\left(|X - \mathrm{med}(X)|\right). \tag{5}$$

However, it requires the assumption of a tuning parameter k. For a normal law, $k = 1.4826$.

1.3.3 Estimators of the Skewness

Bowley (1920)'s estimator, based on the quartiles, is defined as:

$$\gamma_{\mathrm{B}}(X) = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}. \tag{6}$$

This formula can be generalized to other percentiles, with the parameter $\alpha < 0.5$:

$$\gamma_{\mathrm{B}}(X;\alpha) = \frac{F^{-1}(1-\alpha) + F^{-1}(\alpha) - 2F^{-1}(0.5)}{F^{-1}(1-\alpha) - F^{-1}(\alpha)}. \tag{7}$$

Since the value of α in the expression above is arbitrary, Groeneveld & Meeden (1984) have proposed an estimator integrating over α:

$$\gamma_{\mathrm{GM}}(X) = \frac{\displaystyle\int_0^{0.5} F^{-1}(1-\alpha) + F^{-1}(\alpha) - 2F^{-1}(0.5)\,\mathrm{d}\alpha}{\displaystyle\int_0^{0.5} F^{-1}(1-\alpha) - F^{-1}(\alpha)\,\mathrm{d}\alpha}. \tag{8}$$

1.3.4 Estimators of the Correlation Coefficient

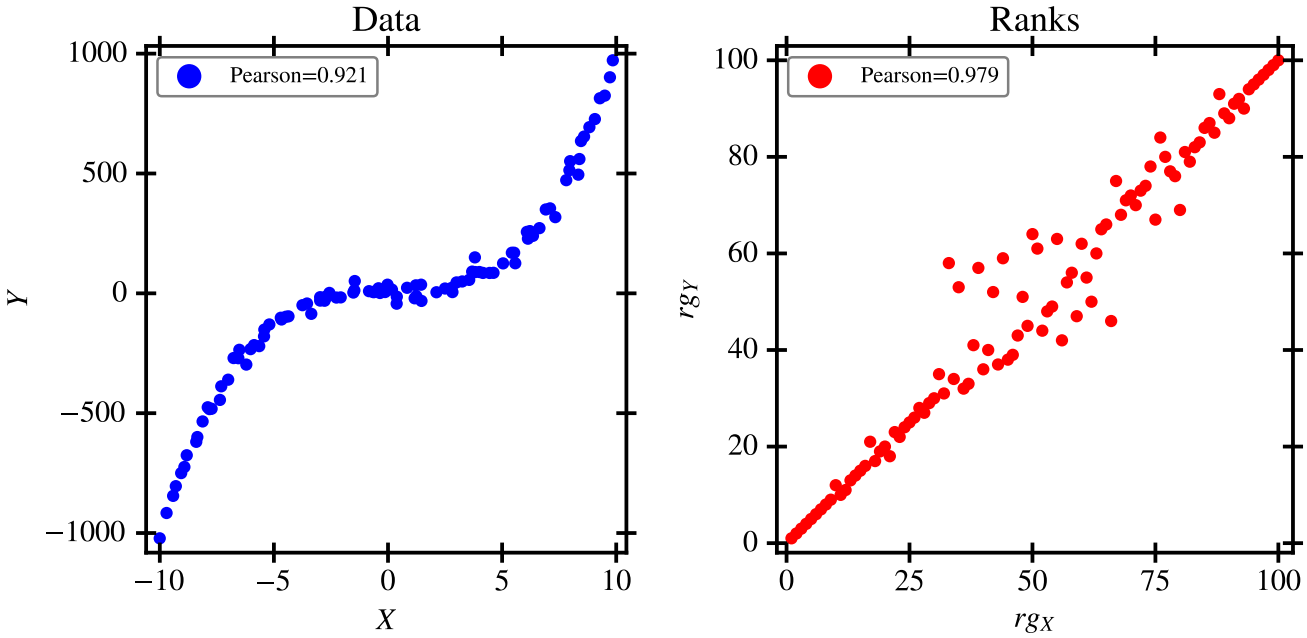

Figure 2: *Demonstration of the Spearman rank correlation coefficient.* The left panel shows a non-linear relation between X and Y and its Pearson correlation coeffient. The right panel shows the relation between the ranks of the data. The ranking linearizes the relation. The Pearson correlation coefficient of this new trend is the Spearman correlation coefficient of the data X and Y.

The *Spearman rank correlation coefficient* is not a *percentile-based* method, but there is ranking involved. This estimator is the Pearson correlation coefficient (Eq. 4), replacing the variables by their rank, rg_X and rg_Y (*cf.* Fig. 2). It therefore does not quantify only a *linear* correlation, but a more general, monotonic correlation. It provides a more robust linear correlation coefficient estimator, as the outliers are replaced by their ranks.

[1] also known as the *Median Absolute Deviation About the Median* (MADAM).

1.4 Trimmed Moments

The expectation estimators of Eqs. (1)-(4) can become robust if we suppress the outliers. This can be done if we define a trimming threshold $\epsilon < 0.5$, and trim out the values lower than $F^{-1}(\epsilon)$ and higher than $F^{-1}(1-\epsilon)$. Numerically, it is simply the mean of the sorted sample between indices $1+\epsilon N$ and $(1-\epsilon)N$. The drawback is that the choice of the trimming fraction, ϵ, is arbitrary.

1.5 M-estimators

M-estimators are a generalization of *Maximum-Likehood Estimators* (MLE). They consist in minimizing a *loss function*, $\rho(X_i, \theta)$. The value of θ at minimum is the estimator, $\hat{\theta}$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \rho(X_i, \theta), \tag{9}$$

If ρ is not differentiable, the M-estimator is a *ρ-type estimator*. However, if it is differentiable, the M-estimator is a ψ-type estimator, and it can be simplified by solving:

$$\sum_{i=1}^{N} \psi(X_i, \theta)\bigg|_{\theta=\hat{\theta}} = \sum_{i=1}^{N} \frac{\partial \rho(X_i, \theta)}{\partial \theta}\bigg|_{\theta=\hat{\theta}} = 0. \tag{10}$$

Several loss functions can be found in the literature. One of the most used is *Tukey's biweight* (for *bisquare weight*; Mosteller & Tukey, 1977). Posing $u = (x - l_x)/(c \times s_x)$, where l_x is the location parameter, s_x is the the scale parameter, and c, a tuning parameter, it is defined as:

$$\begin{cases} \rho(u) &= \frac{1}{6}\left[1-(1-u^2)^3\right] & \text{if} \quad |u|<1 \\ &= 0 & \text{elsewhere} \end{cases} \tag{11}$$

$$\begin{cases} \psi(u) &= u(1-u^2)^2 & \text{if} \quad |u|<1 \\ &= 0 & \text{elsewhere.} \end{cases} \tag{12}$$

In astrophysics, it has been used by Beers et al. (1990) and implemented in `astropy`.

Eq. (10) can be solved to build estimators of the scale and location (Lax, 1975; Mosteller & Tukey, 1977):

$$\hat{\mu}_{\mathrm{M}}(X) = l_x + \frac{\sum_{|u_i|<1}(x_i - l_x)(1-u_i^2)^2}{\sum_{|u_i|<1}(1-u_i^2)^2} \tag{13}$$

$$\hat{\sigma}^2_{\mathrm{M}}(X) = N \times \frac{\sum_{|u_i|<1}(x_i - l_x)^2(1-u_i^2)^4}{\left[\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)\right]^2} \tag{14}$$

The location and scale parameters can be taken as $l_x = \mathrm{med}(X)$ and $s_x = \mathrm{MAD}(X)$. We can also iterate, replacing them with their M-estimates, until a given accuracy is reached. The tuning parameter is usually taken as $c = 6$ for the mean and $c = 9$ for the variance. Mosteller & Tukey (1977) even suggest a slightly better variance estimator:

$$\hat{\sigma}^2_{\mathrm{M}}(X) = N \times \frac{\sum_{|u_i|<1}(x_i - l_x)^2(1-u_i^2)^4}{\left[\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)\right]\left[-1+\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)\right]}, \tag{15}$$

which converges towards Eq. (2).

Improvement: I could not find any M-estimator of the skewness in the litterature. However, a simple *W-estimator*, weighting the expectation formula with weights $w(u) = \psi(u)/u$ ($c = 9$), provides the best skewness estimator I have tested so far:

$$\hat{\gamma}_{\mathrm{M}}(X) = \frac{\sum_{|u_i|<1} u_i^3(1-u_i^2)^2}{\sum_{|u_i|<1}(1-u_i^2)^2}. \tag{16}$$

Joana Frontera-Pons confirmed that this expression was correct.

Posing $v = (y - l_y)/(c \times s_y)$, the covariance can be estimated, with $c = 9$, as (Mosteller & Tukey, 1977; Beers et al., 1990):

$$\hat{V}_{\mathrm{M}}(X,Y) = N \times \frac{\sum_{|u_i|<1,|v_i|<1}(x_i - l_x)(1-u_i^2)^2(y_i - l_y)(1-v_i^2)^2}{\left[\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)\right]\left[\sum_{|v_i|<1}(1-v_i^2)(1-5v_i^2)\right]} \tag{17}$$

2 TESTING THE ROBUSTNESS WITH MONTE-CARLO SIMULATIONS

2.1 Univariate Tests

2.1.1 Simulated Distributions

To test the robustness of these estimators, we have drawn samples from a split-normal law (Villani & Larsson, 2006):

$$f(x) = \frac{1}{\sqrt{2\pi}(1+\tau)\lambda} \times \begin{cases} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\lambda}\right)^2\right] & \text{if } x < \mu \\ \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\tau\lambda}\right)^2\right] & \text{if } x \geq \mu. \end{cases} \tag{18}$$

We vary the asymmetry parameter of the distribution ($\tau = 0.02 - 50$), the sample size ($N = 10 - 100\,000$) and the fraction of outliers ($\phi = 0.1\,\% - 10\,\%$). The outliers are uniformly distributed between 0 and 10^4.We keep the position and scale parameters constant ($\mu = 100$ and $\lambda = 9$, respectively). A few of these distributions are displayed in Fig. 3.

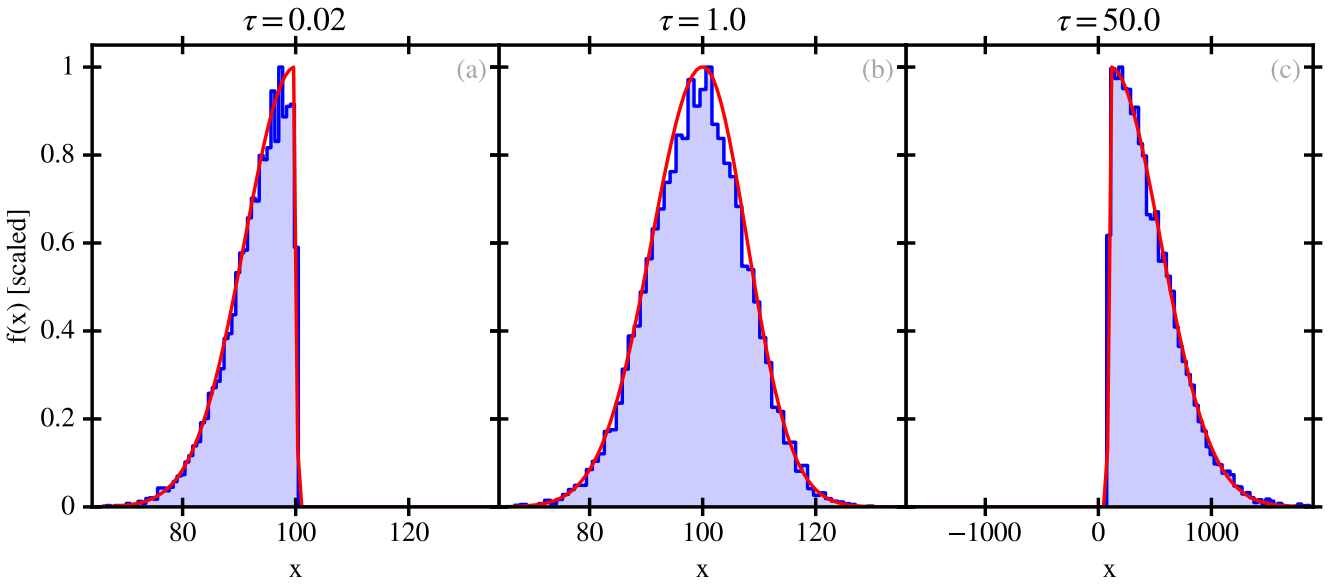

Figure 3: *Select simulated distributions.* We show three extreme values of the skewness, parameterized by the τ parameter of the split-normal distribution. The theoretical disribution is the red line. The blue histogram shows the drawn sample ($N = 100\,000$).

2.1.2 Accuracy of the Different Estimators

We have applied the different estimators of Sect. 1 to the distributions of Sect. 2.1.1. We have studied the relative error by comparing these estimators to the true value of the moments of the distribution ($|M_{\text{est}}/M_{\text{true}} - 1|$). Fig. 4 shows these errors for the different estimators as a function of the asymmetry parameter, τ, in the case a large sample ($N = 100\,000$). Fig. 5 shows the same errors as a function of the sample size, medianing over τ.

Looking at Figs. 4-5, we can conclude the following points regarding the accuracy of the estimators.

1. When the fraction of outliers is low (0.1 %; top panels), the expectation-based estimators (Sect. 1.2) are obviously among the best. However, as soon as the fraction of outliers reaches 1 % and higher, they become unreliable and are among the worst.

2. The median-based estimators (Sect. 1.3) for the mean are consistently good. The MAD is among the best for the standard-deviation, however it is not the case for any distribution (*cf.* Sect. 2.1.3). However, the corresponding skewness estimators are not very good.

3. The trimmed estimators are logically good when the fraction of outliers is lower than their trimming threshold. And they logically become comparable to their expectation counterparts when they do not trim enough. One could therefore conclude that the higher the trimming fraction, the better. It works for the mean. However, these estimators are more biased for the standard-deviation and the skewness, when ϵ is higher. This is because, if we cut too much flesh out of the distribution, we bias its scale and shape estimators. Thus, unless the fraction of outliers can be guessed, these estimators are problematic.

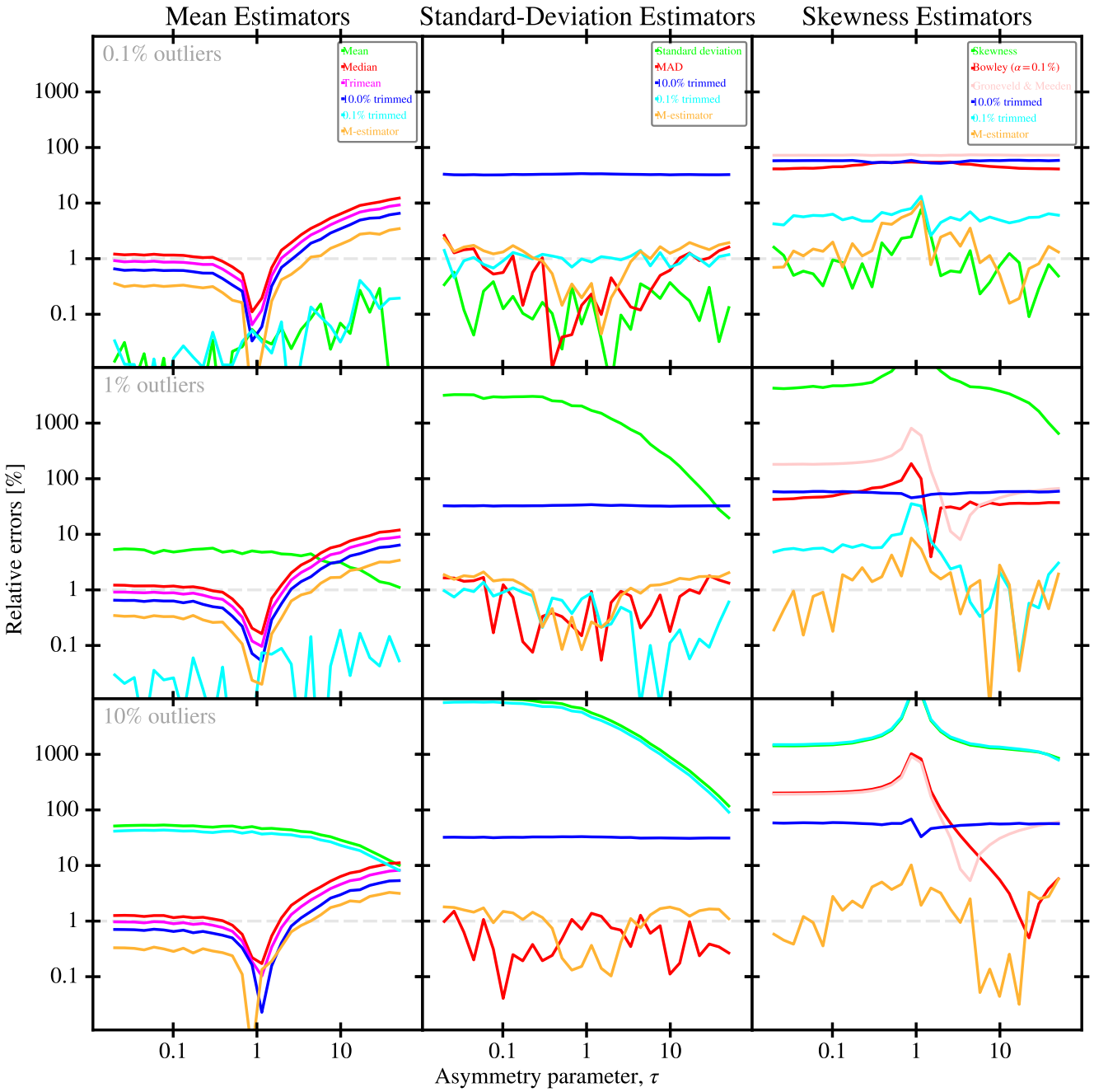

Figure 4: *Robustness of select estimators.* Each panel shows the relative error on the moment as a function of the asymmetry parameter, τ. The columns show different moments, and the rows vary the fraction of outliers. Each estimator is color coded. These curves correspond to the $N = 100\,000$ samples.

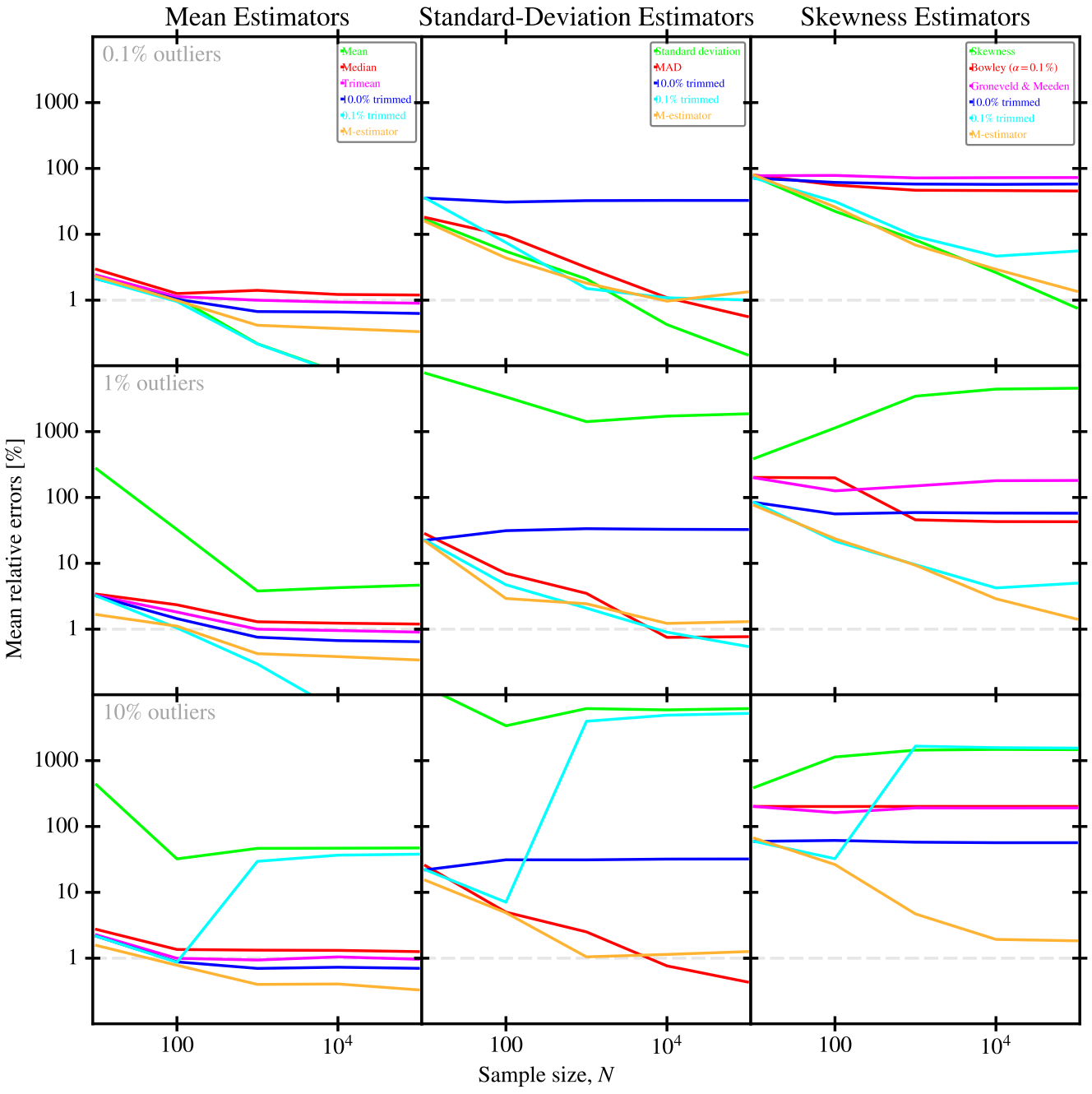

Figure 5: *Robustness of select estimators* as a function of the sample size. These curves have been averaged over τ.

4. M-estimators are the most consistent and are always very good (percent-level accuracy).

2.1.3 Additional Comparison Between the MAD and the M-Estimator of Scale

Since the MAD seems to work as well as the M-estimator for the simulation of Sect. 2.1.1, we have performed another test using a Student's t law ($\mu = 0$, $\sigma = 1$), varying the number of degrees of freedom (ν; left panel of Fig. 6):

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\lambda}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\lambda}\right)^2\right)^{-\frac{\nu+1}{2}}. \tag{19}$$

The corresponding standard-deviation is $\sigma = \lambda\sqrt{\nu/(\nu-2)}$. We have generated 100 samples of size 100 000, for each ν.

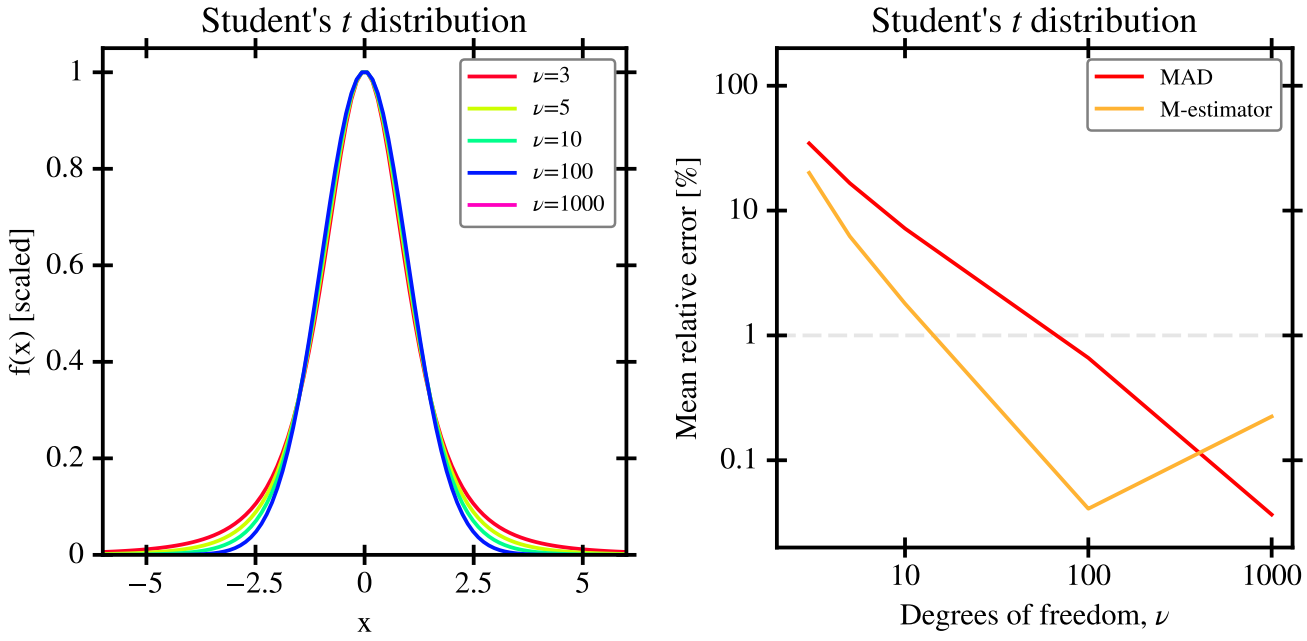

Figure 6: *Tests with Student's t distribution.* The left panel shows the distributions for different degrees of freedom, ν. The right panel compares the accuracies of the MAD and the M-estimator, as a function of ν.

The accuracies are displayed on the right panel of Fig. 6. This figure shows that the M-estimators are, this time, systematically better. It also shows that both estimators converge toward a similar accuracy at large ν, *i.e.* when the distribution tends to a gaussian. Thus, the MAD was working well on Figs. 4-5 because the distribution had the kurtosis of a normal law. This is because it was tuned for such a kurtosis with its parameter k (Eq. 5). In contrast, a Student's t has a larger kurtosis. The M-estimator accounts for it, but not the MAD.

2.2 Bivariate Tests

2.2.1 Simulated Distributions

We have simulated correlated random variables X and Y, drawn from a bivariate normal distribution, varying $\rho = -0.9$ to 0.9, varying the sample size and the fraction of outliers, the same way as in Sect. 2.1.1. We show some distributions in Fig. 7.

2.2.2 Accuracies of the Different Estimators

Fig. 8 compares the different correlation estimators as a function of ρ, and for different outlier fractions. Overall, the same conclusions as in Sect. 2.1.2 hold. The Spearman rank correlation coefficient is consistently good, but the M-estimator is consistently the best one.

2.3 Comparison to the `astropy.stats` Implementation

The `astropy.stats` module contains several functions implementing Eqs. (13)-(14) and Eq. (17). However, it has the following issues.

1. They do not iterate, *i.e.* they keep the initial estimates of the location and scale parameters (the median and the MAD).

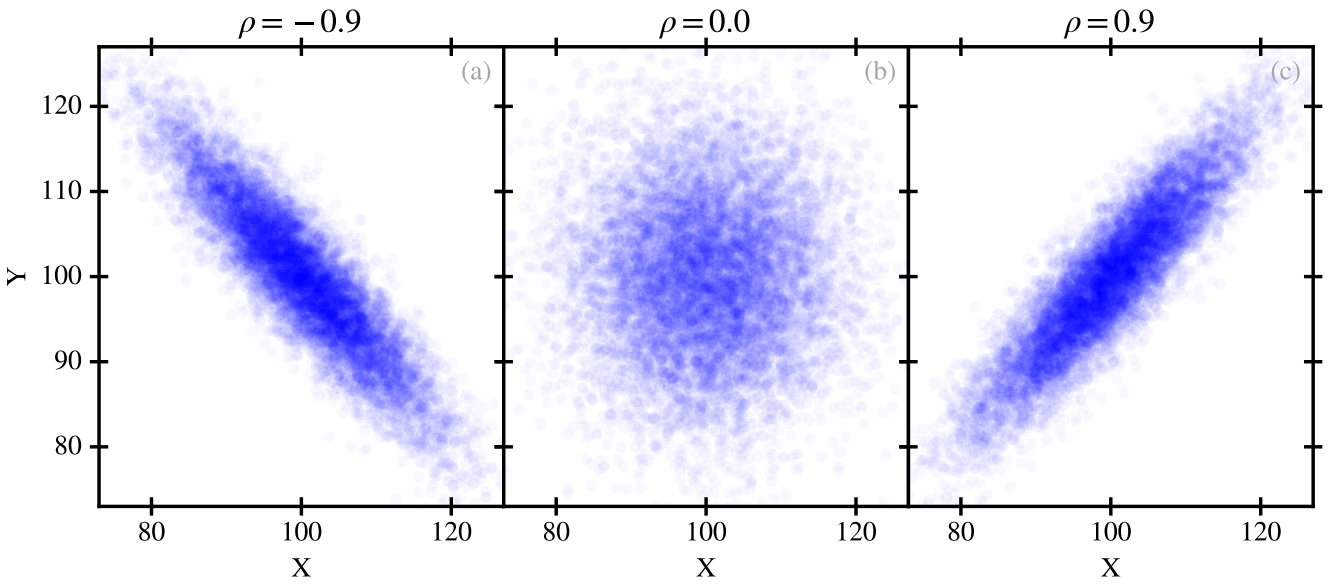

Figure 7: *Select simulated bivariate distributions.* We show three extreme values of the correlation coefficient, ρ, of these bivariate normal distributions. These plots correspond to the $N = 100\,000$ samples.

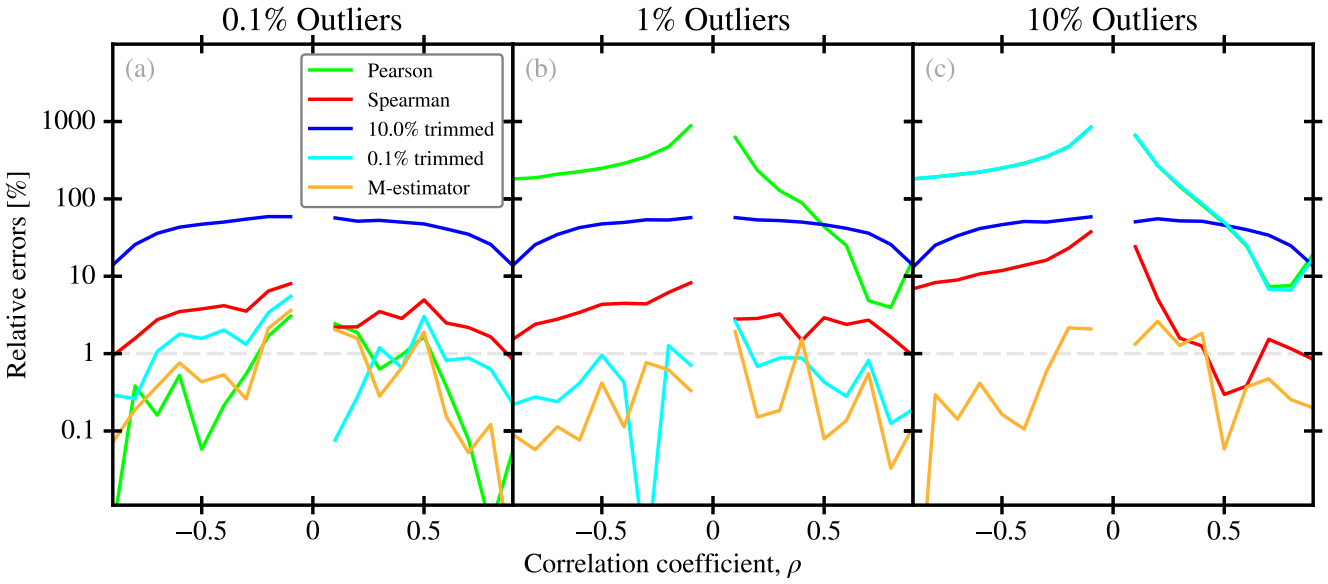

Figure 8: *Accuracy of correlation estimators*, as a function of the true value or ρ. These plots correspond to the $N = 100\,000$ samples.

2. They use the MAD with $k = 1$.

We have tried to improve the `astropy.stats` method, by iterating over the location and scale parameters, as recommended by Mosteller & Tukey (1977). We basically replace the initial values (median and MAD) by their new M-estimators and cycle until a precision of 10^{-5} is reached. The iteration process significantly improves the estimates (Fig. 9).

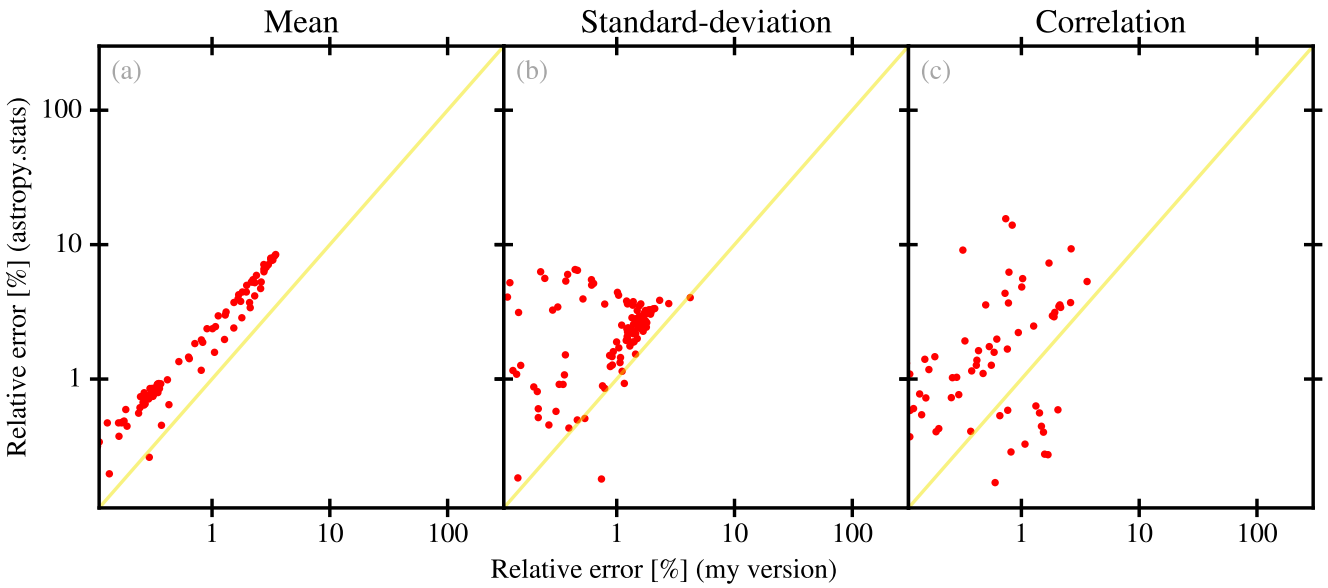

Figure 9: *Comparison to* `astropy.stats` for the $N = 100\,000$ samples.

References

Beers, T. C., Flynn, K., & Gebhardt, K. 1990, AJ, 100, 32

Bowley, A. L. 1920, Elements of statistics, Vol. 2 (PS King)

Groeneveld, R. A. & Meeden, G. 1984, Journal of the Royal Statistical Society. Series D (The Statistician), 33, 391

Lax, D. 1975, An Interim Report of a Monte Carlo Study of Robust Estimators of Width (Department of Statistics, Princeton University)

Mosteller, F. & Tukey, J. W. 1977, Data analysis and regression. A second course in statistics (Addison-Wesley Series in Behavioral Science: Quantitative Methods, Reading, Mass.: Addison-Wesley, 1977)

Villani, M. & Larsson, R. 2006, Communications in Statistics-Theory and Methods, 35, 1123