# IS THE REAL WORLD NORMAL?

Catherine Dehon
ECARES
Université libre de Bruxelles, Brussels, Belgium

*In recent times, we have become increasingly confronted with high dimensional data sets. Statistical methods have had to adapt themselves to more complex questions from many different scientific disciplines, notably in the social sciences. But in spite of this evolution, statistics courses still rely too often on artificial examples that contribute to the myth that the real world is quite simple. Classical statistics based on parametric models also feature in undergraduate and graduate curricula. Nevertheless apparent deviations of the model cannot always be ignored. For example, we often expect large datasets to contain a small number of unusual observations, which renders classical procedures unreliable. The theory of robust statistics deals with small deviations from the model, and can be viewed as a compromise between parametric and nonparametric analysis. Should we consider introducing these modern concepts into undergraduate statistics courses?*

## INTRODUCTION

In spite of the development of research in robust and nonparametric statistics, most professors introduce statistics using solely strict parametric models. Data is assumed to be generated according to a specified distribution function $F_\theta$ , where $\theta$ is the vector of unknown parameters. Here, we concentrate only on estimation theory. It is usual in the parametric context to develop optimal procedures to estimate the vector of unknown parameters $\theta$. However, robust and nonparametric statistics are generally only introduced in more advanced curricula, followed by a small number of mathematics students.

Parametric models and "toy" examples could be the starting point for teaching statistics, but it is essential to mention from the beginning that, in practice, parametric assumptions are often violated when using real datasets (Ronchetti, 2006). The deviations from the model can essentially be classified into two categories: misspecification of the underlying distribution, or presence of a fraction of observations from an unknown, arbitrary distribution. Nonparametric statistics deal with the first issue, which is not in the scope of this note, whereas robust statistics allow us to assume that the majority of the data is generated by a parametric model whilst a small fraction of the data, called outliers, could be generated by an arbitrary distribution. Robust statistics deals with small deviations from the assumptions of the model, and can be viewed as a compromise between strict parametric modeling and more complex nonparametric tools. The idea behind a robust methodology is the construction of measures of robustness and statistical procedures that remain valid and reasonably efficient to analyze the behavior of the majority of the data. Robust methods allow the researcher to capture the pattern that underlies the vast majority of the observations of a dataset, while controlling for the influence of outliers [see, for example, Hampel, et al. (1986), Huber & Ronchetti (2009), Maronna, et al. (2006) and Heritier, et al. (2009)].

## BIAS, EFFICIENCY BUT ALSO ROBUSTNESS

In the context of teaching, the first key issue of robust statistics is to warn students of the danger of neglecting outliers in analysis and to show that classical statistics can be completely misleading on real datasets. To introduce intuitively the impact of outliers, let us start with the simple univariate one-sample location-scale model: $x_i = \mu + \sigma \varepsilon_i$, where the $\varepsilon_i$ ($i = 1, \ldots, n$) are independent and identically distributed from a standard normal distribution, noted $F_0$ with density $f_0$. We would like to estimate the vector of parameters $\theta = (\mu, \sigma)$ where $\mu$ is the location parameter and $\sigma$, the scale parameter. To facilitate the exposition, we suppose that the scale parameter $\sigma$ is known ($\sigma = 1$). Theoretically, there exists an infinity of possible estimators $T_n$ (where n is the sample size) for the location parameter $\mu$. As a starting point, we focus on the two more popular estimators of location: the sample mean and the sample median. Statistical measures are needed to compare the properties of these two estimators, but in most statistics lectures, the focus is solely on

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org 

the fact that these two estimators are unbiased, consistent, and that the sample mean is optimal in the sense of efficiency under the normality assumption, the asymptotic efficiency of the sample median being only of 64%. At this point, it is essential to go beyond simply measuring the quality of an estimator using the concepts of biasness and efficiency, and to introduce information about its robustness. The influence function is a measure of robustness that could be introduced in an undergraduate course (Hampel, 1971). The influence function is a local measure of robustness, that reflects the effect of a single outlier. The simplest way to introduce this measure is to start from the point of view of a sample. Let $x_1, \ldots, x_n$ be a set of observations generated by the assumed model $F_\theta$, $T_{1,n}(x_1, \ldots, x_n)$, the sample mean and $T_{2,n}(x_1, \ldots, x_n)$, the sample median. The empirical influence function (EIF) of an estimator $T_n$ is given by the estimate $T_n(x_1, \ldots, x_n, x)$ when an arbitrary observation x ($-\infty \leq x \leq \infty$) is added to the sample. We generated 100 observations from a standard normal distribution and computed the EIF for the sample mean and the sample median (Figure 1, left panel). We see that the EIF of the sample mean is unbounded, meaning that one observation is sufficient to break down the estimator. On the other hand, the sample median is robust in the sense that its EIF is bounded. The median is said to be B-robust.

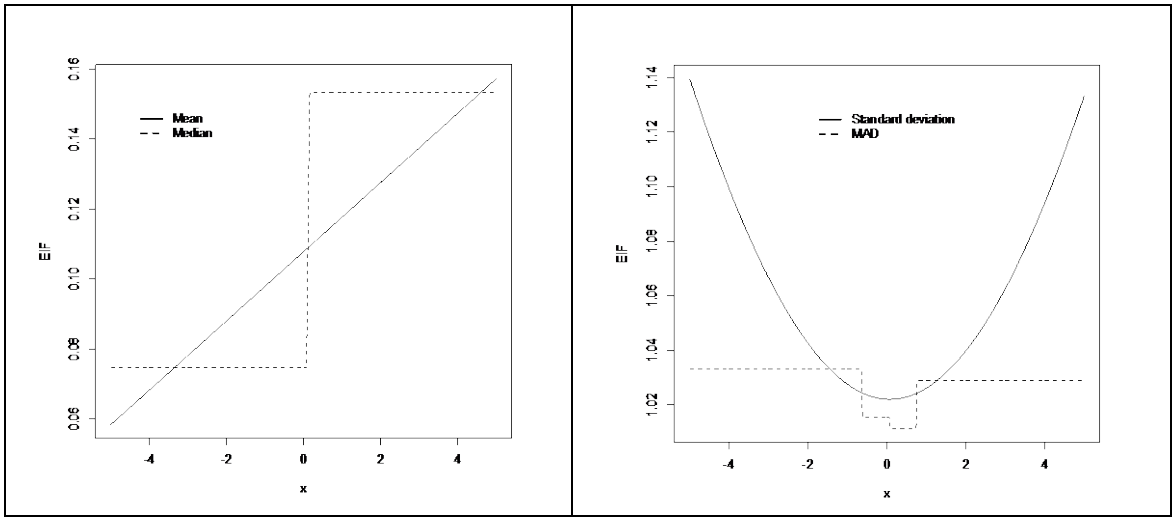


Figure 1: Empirical influence functions of the sample mean and the sample median (left panel). Empirical influence functions of the variance and the MAD (right panel)

We can also measure the effect of a single outlier on estimator $T_n$ , using a scaled version of the EIF called the sensitivity curve:

$$SC(x_1,\ldots,x_n,x,T_n)=\frac{T_n(x_1,\ldots,x_n,x)-T_n(x_1,\ldots,x_n)}{1/(n+1)}.$$

For the sample mean, it is straightforward to derive the sensitivity curve:

$$SC(x_1,\ldots,x_n,x,T_{1,n})=\sum_{i=1}^{n}x_i-nx$$

which is, as the EIF, an affine and unbounded function of x. The derivation of the SC for the median is also fairly straightforward. It would thus be feasible to introduce this concept of robustness in introductory statistics lectures. The drawback at sample level is that the values of the empirical influence functions and sensitivity curves depend on the sample. To circumvent this drawback, we need to go a step further and introduce the functional representation T of the estimator $T_n$ (see Ruiz-Gazen, 2012). The need for this new mathematical concept, not often used outside the robust world, is arguably one of the main obstacles to introducing robust statistics at undergraduate level. The statistical functional T of an estimator of location $T_n$ in the one-sample location-scale model maps any statistical distribution $F_\theta$ to a real number $T(F_\theta)$. For the empirical distribution function corresponding to the sample $F_n$, we require that $T(F_n) = T_n$. The statistical function associated with the sample mean is $T(F) = E_F[X]$ (only defined for distributions that have

a finite first moment) and for the sample median $T(F) = F^{-1}(1/2)$. Moreover, these two statistical functionals for location parameter are Fisher consistent at the model distribution $F_\theta$ with $\theta = (\mu,\sigma)$ meaning that $T(F_\theta) = \mu$. This concept is similar to that of asymptotic unbiasedness, and verifies that the statistical functional recovers the interested parameter. The introduction of a statistical functional is an elegant way to consider models in the neighborhood of $F_\theta$. As mentioned in the introduction, in robust statistics we consider that the majority of the data comes from a specified distribution $F_\theta$, whilst a small fraction $\varepsilon$ of the data comes from an arbitrary unknown distribution G, leading to the contaminated model: $F_{\theta,\varepsilon} = (1-\varepsilon) F + \varepsilon G$, with $0 < \varepsilon < 1$ the level of contamination, and $G \in \Im$ the family of all possible distribution functions. A local contamination can be represented by the Dirac probability measure which gives a mass of one at the point x: $G = \Delta x$. The influence function (IF) is then defined by the influence of an infinitesimal level of contamination at x:

$$IF(x,T,F_\theta) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon\Delta_x) - T(F)}{\varepsilon}.$$

The influence function can be viewed as the population limit of the sensitivity curve introduced at the sample level. At the standard model $F_0$ with $\theta = (0,1)$, the influence functions of the mean and the median are given by: $IF(x,T_1, F_0) = x$ and $IF(x,T_2, F_0) = sign(x)/2f_0(0)$. An estimator associated with a bounded influence function is said to be B-robust. Another way to verify B-robustness is to check if the supremum of the absolute value of the IF (called the gross-error-sensitivity, GES) is finite or not. In the robust literature, it has been shown that the median has the smallest possible value of GES provided that $2f_0(0) \neq 0$ (Hampel et al., 1986).

It seems rather surprising that though the median is introduced as a standard statistical robust tool in the one-sample location-scale model where $\mu$ and $\sigma$ are the two unknown parameters, its natural companion for estimating the scale parameter, the median absolute deviation (MAD), is almost systematically ignored:

$$MAD = cmedian_{1 \le i \le n} \left| x_i - median_j x_j \right|,$$

where c is chosen such that the MAD functional is Fischer-consistent at the assumed model. Indeed, the interquartile range is often the sole example given of a robust estimator for the scale parameter. Does this difficulty stem from the specification of the constant c, or is it due to the low efficiency of the MAD under the normality (37%)? The previous robustness concepts (EIF, IF, GES) can be applied in the same manner to compare the variance and the MAD (see Figure 1, right panel) leading to the same conclusion: the MAD estimator is B-robust whereas the standard deviation estimator is a non-robust estimator.

## M-ESTIMATOR INSTEAD MLE

Having introduced measures of robustness, it now becomes interesting to discuss methods for building robust estimators. The extension of the class of maximum likelihood estimators (MLE) to the class of M-estimators (Huber, 1964) is one way to do this. The MLE is a particular case of the M-estimator, thus it would make more sense to start by introducing the class of M-estimators, followed by the MLE and its efficiency properties. For reasons of simplicity of exposition, we only consider here the location parameter $\mu$ to illustrate and investigate estimation problems. As the ML estimator, the M-estimator of location $T_n$ is defined by an optimization problem:

$$\min_t \sum_{i=1}^{n} \rho\left(\frac{x_i - t}{\hat{\sigma}}\right)$$

where $\rho(u)$ is an even function, non-decreasing for positive u, and $\hat{\sigma}$ is a robust, consistent estimator of scale such as the MAD (note that the inclusion of scale is required to guarantee equivariance properties). When the $\rho$-function is convex, the minimization problem is equivalent to the solution of the estimating equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - t}{\hat{\sigma}}\right) = 0,$$

where $\psi(u) = \frac{\partial}{\partial u}\rho(u)$. However, if ρ is bounded, the estimating equation can have many solutions (local minima or maxima), which increases computational difficulties. If the ψ-function is the score function

$$\psi_{score}(u) = \frac{-f_0'(u)}{f_0(u)},$$

we recover the MLE, of which the asymptotic variance, under regularity conditions, reaches the lower bound of the Cramer-Rao inequality. In the gaussian model, the score function is given by ψ(u) = u leading to the sample mean which is efficient but not robust. Note that the ψ-function is unbounded in this situation. With the Cauchy distribution, the score function is given by ψ(u) = sign(u) leading to the sample median, which is efficient and robust. But in the majority of situations, we must find a trade-off between efficiency and robustness. A very appealing property of the class of M-estimators is that the form of the influence function solely depends on the ψ-function: a bounded ψ-function yields a bounded IF. Indeed the influence function of a location M-functional T at the standard distribution $F_0$ is given by

$$IF(x, T, F_0) = \frac{\psi(x)}{E_{F_0}[\psi'(X)]}.$$

A classical choice for ρ is the Huber (1964) ρ-function defined as

$$\rho_c^H(u) = \begin{cases} u^2 if\ |u| \le c \\ 2c|u| - c^2 if\ |u| > c \end{cases}.$$

The positive tuning constant c reflects the trade-off between efficiency and robustness, and can be seen as an intermediate case between $\rho(u) = u^2$ and $\rho(u) = |u|$. The corresponding location estimator is the median for c tending to zero, and the mean for c tending to infinity. As shown in Figure 2, the ρ-function is convex and the ψ-function is bounded, leading to a B-robust estimator of location (IF is bounded). Another popular family of ρ-functions is the Biweight family proposed by Tukey:

$$\rho_c^B(u) = \min(1, 1 - (1 - (\frac{u}{c})^2)^3)$$

where c is a positive tuning constant. In this family, the ρ-function is bounded, leading to a bounded but also redescending ψ-function (see Figure 2, right panel). Thus the influence of very large outliers is not only bounded but actually vanishes at ±∞. However the drawback is that $\psi_c^B$ is not monotone, potentially leading to multiple solutions.

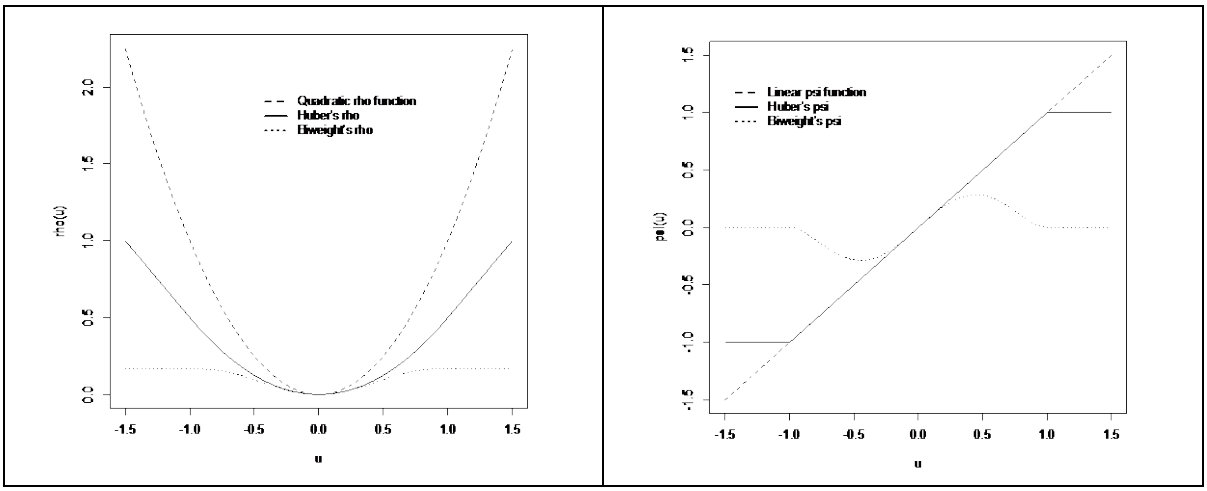


Figure 2: Quadratic, Huber and Biweight loss functions (left panel)
and the associated ψ-functions (right panel)

In conclusion, in order to achieve an adequate compromise between robustness and efficiency, we would like to find a smooth, bounded and eventually redescending ψ-function, which remains close to the score function at the center (see Croux & Dehon, 2012).

CORRELATION MEASURES

Measures of robustness, such as the influence function plotted below, are general tools of the robust methodology and can be applied within different models to classical, robust or nonparametric estimators. In practice, one of the most commonly used statistical measures is the correlation coefficient which measures the association between two quantitative variables. To facilitate the exposition, consider the bivariate normal distribution $F_\rho$ with population coefficient ρ. Let $\{(x_i, y_i), 1 \leq i \leq n\}$ be a bivariate sample of size n. The classical Pearson estimator of correlation, depending on sample means, is well known for being non-robust, which can be formally shown by means of the unboundedness of its influence function. The influence function, which measures the influence of an infinitesimal amount of contamination at a given value on the functional, can also be used to compare the three well-know nonparametric correlation estimators: the Kendall, the Spearman and the Quadrant correlation. On Figure 3 (as the expressions of the IF depends on ρ, we set $\rho = 0.5$), the interpretation is easy: the IF of the Pearson correlation is unbounded, proving its lack of robustness, the IF of the Spearman and Kendall correlation are bounded and smooth, and the IF of the Quadrant correlation is bounded but not smooth, indicating that small changes in the data may have a relatively large (but bounded) impact on the correlation.

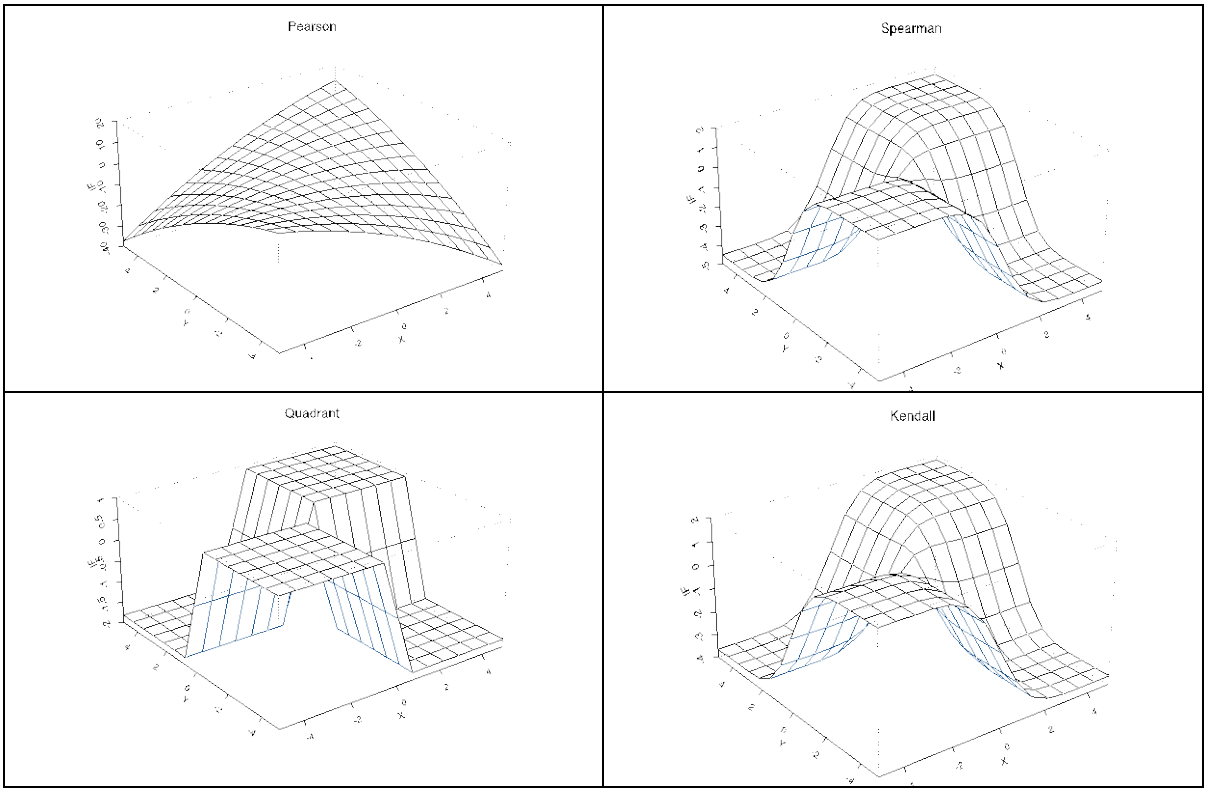


Figure 3: IF of the Pearson, the Kendall, the Spearman and the Quadrant functional at the bivariate normal distribution with ($\rho = 0.5$)

For teaching purposes, the intuition behind the influence function is easy to explain, and the graphical forms associated with it are easy to interpret. But even in this very simple setup, the mathematical derivation of the expression of the IF for the Pearson, the Kendall, the Spearman and the Quadrant correlation is not straightforward. Figure 3 shows the IF for the transformed Kendall, Spearman and Quadrant correlation since the classically used expressions leads to associate

functionals that are not Fisher-consistent. Thus even though the concept of influence function can be introduced intuitively, it would be unthinkable to introduce the mathematical derivation (Croux & Dehon, 2010) of the influence function for the functional associated with bivariate correlation measures in undergraduate statistical courses.

## CONCLUSION

Being able to combine both rigorous mathematical developments with the introduction of robust statistics at undergraduate level is not an easy task. The mathematical treatment can be adapted to the level of the course but basic ideas, tools and intuition on robust statistics should be introduced in all undergraduate statistical courses, especially for non-mathematicians who are likely to apply later on the statistical methods they have learned to real data that may then very well contain outliers.

## REFERENCES

Croux, C., & Dehon, C. (2012). Robust estimation of location and scale. In A. H. El-Shaarawi and W. Piegorsch (Eds.), *Encyclopedia of Environmetrics, 2nd edition* (pp. 802-807). Chichester, UK: John Wiley & Sons Ltd.

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, *19*(4), 497-515.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York: Wiley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics, 42*, 1887-1896.

Heritier, S., Cantoni, E., Copt, S. & Vitoria-Feser, M. P. (2009). *Robust methods in biostatistics.* United Kingdom: Wiley Series in Probability and Statistics.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Huber P. J. & Ronchetti E. M. (2009). *Robust Statistics*. New York: Wiley, 2nd edition.

Maronna, R. A., Martin, R. D & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York: Wiley.

Ronchetti, E. (2006). The historical development of robust statistics. In B. Chance & A. Rossman (Eds.), *Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil*. Voorburg, the Netherlands: ISI.

Ruiz-Gazen, A. (2012). Robust statistics: a functional approach. *Annals of Institut de Statistiques de l'Université de Paris, 56*(2-3), 49-64.