ROBUST STATISTICS: A FUNCTIONAL APPROACH

By Anne Ruiz-Gazen

Toulouse School of Economics

Abstract For a given statistical method, good properties are usually obtained under strong hypotheses that are likely not to be verified in practice. In this context, the robustness of a method refers to its capability of keeping good properties even when the hypotheses are not verified. Statistics imply the treatment of data sets that are assumed to be the realization of random variables. The assumptions usually concern the probability distribution from which the data set is generated. One field in Statistics called "robust statistics" consists in defining some measures of robustness and proposes robust methods in the sense of these measures. After giving some basic notions of statistics, we present different well-known measures of robustness from the "robust statistics" field. In order to illustrate the ideas we consider the simple case of estimating a "mean" using the arithmetic mean, the median, the trimmed and the winzorised mean. The paper is an introductory presentation of robust statistics for readers who are not necessarily statisticians but it contains some technical details. The reader not interested in such details can read the introduction of each section and subsection where general ideas are given in simple words.

Résumé Les bonnes propriétés d'une méthode statistique donnée sont généralement obtenues sous des hypothèses fortes qui ne sont pas vérifiées en pratique. Dans ce contexte, la robustesse d'une méthode consiste en sa capacité à garder de bonnes propriétés même lorsque les hypothèses ne sont pas vérifiées. La Statistique implique le traitement de données qui sont supposées être la réalisation de variables aléatoires. Les hypothèses portent en général sur la distribution de probabilités sous-jacente aux données. Un domaine de la Statistique appelé "statistique robuste" consiste à définir des mesures de robustesse et à proposer des méthodes robustes au sens des mesures précédemment définies. Après avoir présenté des notions de base en statistique, nous donnons différentes mesures de robustesse bien connues dans le domaine. Pour illustrer les idées, nous considérons le cas simple de l'estimation d'une "moyenne" par la moyenne arithmétique, la médiane, la moyenne tronquée et la moyenne winzorisée. Le papier se veut un exposé introductif à la statistique robuste qui s'adresse à des lecteurs non nécessairement statisticiens mais il contient toutefois des détails techniques. Le lecteur qui n'est pas intéressé par ces détails pourra se contenter de lire l'introduction des différentes sections et sous-sections.

1. Introduction. Statistics is a branch of mathematics which deals with the collection and the analysis of data bases. Statistical analysis helps making decision in scientific, industrial or societal problems and is either descriptive or inferential. Descriptive statistics consist in summaries and descriptions of the collected data sample using some graphical or numerical measures. Inferential statistics use probability theory in order to deduce properties of a population from the properties of a sample drawn from the population. The data are assumed to be generated according a Data Generating Process (DGP). Assumptions are made on the probability distributions that underlie the DGP leading to the definition of some models. For example, data are assumed to be generated by a gaussian or normal

AMS 2000 subject classifications: Primary 6202, 62F35; secondary 62F10

Keywords and phrases: breakdown point, influence function, maxbias, median, trimmed mean, winsorized mean

distribution. Statistical methods are defined and studied under these assumptions which are more or less stringent according to the model. In practice these assumptions are usually not verified by the DGP and statisticians face some robustness questions: what is the behavior of a given statistical method when the model is not completely valid? How can we measure such a behavior in order to compare different methods? If the existing methods are not reliable for small assumptions violations, is it possible to propose more reliable methods?

The Data Generating Process may deviate from the assumed model in different ways. In this paper, the deviation comes from the fact that all the observed data do not follow the assumed probabilistic distribution. The model is well adapted to a majority of the data but not to all of them. The statistician is still interested in the model fitted by the majority of the observations but he has to cope with a fraction of observations that follow a different distribution than the assumed one. This small fraction of observations are usually called outliers or outlying observations. In practice such a deviation is not unusual: it is frequent that data bases contain gross errors because of measurement or copying errors for instance. In this context, a robust statistical procedure is one which is not too much influenced by potential outlying observations. What is precisely meant by "not too much influenced" will be detailed in Section 4.

The type of deviation we focus on may seem not general enough but it already leads to many statistical complications that are not easily overcome and which make the field of robust statistics challenging and complex from a theoretical and an applied point of view. Moreover, it is important to point out that the robustness of a procedure is always relative to a given model and, consequently, robustness measures and methods are to be proposed for each particular statistical model. And there exist many different models depending on the type of data and the objectives of the analysis! In the present paper, only a simple problem and model are tackled which consist in estimating the mean of a gaussian distribution with simple and intuitive estimators. Note that the whole robustness theory on this "simple" problem is already much more intricate than the material we propose below.

The problem of the sensitivity of usual statistical methods in presence of outliers has been already considered many years ago (see for instance [15], [16] and also [6] and discussion for a historic perspective). The theory discussed in the present paper is the one introduced by Huber and Hampel in the sixties and seventies of the previous century in the continuity of Tukey's work ([17]). Their work with some coauthors is described in detail in the books [11] and [8]. There exist thousands of references on robust statistics and some references will be given all along the present paper but many essential references will not be present. For example, the recent books by [12] and [9] will not be quoted anymore in what follows.

In order not to give a too superficial insight into robust statistics, some notions of statistics are needed. Some basic concepts are given in section 2 while general principles of robust statistics are detailed in section 3. Only a few formula are given and all of them are explained in simple words. The interested reader should refer to the literature for details and a more mathematically rigorous presentation.

2. Some basic statistical concepts. Statistics encompass many different "domains" for which adapted methods have been developed. Concerning the collection of data, there exist the domains of survey sampling and experimental design. Concerning data analysis, it is not easy to give an exhaustive list but there are among others, linear and generalized linear models, time series and stochastic processes, spatial statistics, survival analysis and quality control.

For each of these domains most of the statistical inferential procedures can be divided in three different types: estimation, prediction and hypotheses testing. In the following we will

only consider estimation problems. In particular, as a simple but illustrative example, we will consider the problem of estimating a "mean" parameter for some simple "parametric model".

In statistics, parametric models are to be distinguished from non-parametric models. The former models are precisely defined up to a finite number of parameters that need to be estimated from the data while the latter are more flexible but usually do not lead to a complete description of the structure of the data set. The advantage of nonparametric statistics over parametric statistics is that no precise assumption on the distribution of the data is necessary but on the other hand, parametric models describe the data completely and are easier to interpret. Robust statistics have been mostly devoted to parametric models but the theory can be applied as soon as the quantities to estimate are defined as functional of the probability distribution of the DGP. The work in [4] gives an example of a robustness study in a nonparametric context for estimating production frontiers. In other respects, it is sometimes wrongly believed that nonparametric statistics are robust simply because they do not rest on precise distribution assumptions but it is not true in general. Nonparametric procedures may be very sensitive to a few outliers present in the data as exemplified in [1] and [4].

Note that statistics can also be divided into two main streams which are the frequentist and the Bayesian approaches. In the following, only frequentist methods are considered but the interested reader can refer to [8] (p. 55) or [11] (chapter 15) for a discussion on robust Bayesian methods.

Let us now focus on the problem of estimating parameters in a parametric model. A very simple example consists in assuming that the data follow a normal distribution denoted by $\mathcal{N}$. It is well known from probability theory that such a distribution is characterized by its first two moments namely the mean μ that can be any real number and the variance σ^2 that can be any positive number. Let us assume in order to simplify the results and even if it is highly unrealistic, that the variance σ^2 is known and is equal to one. In that case, the only unknown parameter is the mean μ and we will say that μ is to be *estimated* by using the information contained in the data set. Let me give a very simple illustration of the proposed example. Imagine that you are teaching a course on a given topic, for instance linear algebra in mathematics, and that you need to evaluate the level of your students in linear algebra after the course. One can imagine that the level of one given student is a unknown parameter, the μ parameter in the example. You may have access to μ if you were able to design a magic instrument that you could plug in the brain of your student. But because there is no magic instrument, you generally decide to get an idea of the μ value of each of your students by organizing exams and use the obtained grade or better, the mean of different grades as an *estimate* of μ. In that case, the data consist in the grades of a given student in linear algebra. They can be seen as realizations of a random variable since if you organize several exams, the conditions for the student may vary. For instance, the student may have slept well the day before the exam and be in perfect shape, the questions of the exam being on the part of the course he prefers while on the other extreme, the student may have a headache and the questions may be on a part of the program he dislikes. Modeling the data distribution by a normal distribution is one possibility but assuming that the variance is known is not realistic in this example. For a normal distribution, the estimator of μ which consists in taking the average of the grades is optimal in a sense that is detailed below. Note that an estimator is a random variable while the value it takes on a particular data set is a fixed value and is called an estimate.

The quality of a parameter estimator for a parametric model is usually measured by indicators which are its bias, its precision and its efficiency. Let us denote by θ the parameter we want to estimate and by F_θ the probability distribution we assume for the data generating process. For a given data set of size n assumed to be generated independently and from the F_θ distribution, we will define estimators of θ and these estimators will be denoted by $\hat{\theta}$.

For a given distribution F_θ, the expectation of a random variable X will be denoted by $E_{F_\theta}(X)$. In the simple example we consider, $\theta = \mu$ and $F_\theta = \mathcal{N}(\mu, 1)$ which is the gaussian distribution with mean μ and variance 1. If X is a random variable that follows the distribution $\mathcal{N}(\mu, 1)$, we have $E_{\mathcal{N}(\mu,1)}(X) = \mu$ and $E_{\mathcal{N}(\mu,1)}\left[X - E_{\mathcal{N}(\mu,1)}(X)\right]^2 = \sigma^2$.

A simple estimator for θ in the example $\mathcal{N}(\theta, 1)$ is the usual mean estimator:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where the X_i denote the random variables that generate the data set. Another well-known estimator is the median estimator $\hat{\theta}_2$ which is such that the probability of exceeding the median equals the probability of being below and is equal to 0.5.

In order to be more concrete let us consider the following data set of $n = 15$ observations:

$$-9.8 \quad 0.9 \quad 1.1 \quad 1.3 \quad 1.4 \quad 1.9 \quad 2.0 \quad 2.5 \quad 2.6 \quad 2.8 \quad 2.9 \quad 3.0 \quad 3.1 \quad 50.9 \quad 99.1$$

and let us assume that these data follow a gaussian distribution with mean θ unknown and variance equal to 1 even if it is not realistic for the whole data set. An estimate of θ is equal to the arithmetic mean, which is the sum of the fifteen values divided by fifteen and is equal to 11.0. Another estimate is the median of the sample which is equal to 2.5 and corresponds to the eighth observation value since the data are sorted.

Looking at the data set (which is small), we can easily detect that the first observation and the last two observations have extreme values when compared to the rest of the data. So other estimators like the trimmed mean or the winsorized mean may also be considered. Let us define these two estimators. First, a percentage α between 0 to 50% has to be fixed. The α-trimmed mean consists in calculating the mean of only $(1 - 2\alpha)$ of the observations. The percentage 2α of observations that are discarded in the average calculous corresponds to the $(\alpha \times n)$ smallest observations and the $(\alpha \times n)$ largest ones. The value $(\alpha \times n)$ is not necessarily an integer and if it is not, the trimmed mean is usually calculated by removing the integer part of $(\alpha \times n)$ observations from each end and by weighting the two observations that have become the most extreme in the sample. The weight is one minus the remainder between the value $(\alpha \times n)$ and its integer part. Concretely, on our data set, it means that if α is chosen such that two observations are to be trimmed from each end, $(\alpha = 2/15 \simeq 13.3\%)$, the α-trimmed mean will consist in calculating the mean of the eleven observations between 1.1 and 99.1 which is equal to 2.2 . If α is fixed to 10%, the observations -9.8 and 99.1 are removed while the observations 0.9 and 50.9 have a weight of 1/2 in the calculous of the mean, thus the 10%-trimmed mean is equal to 4.2. Concerning the α-winsorized mean, it consists in replacing the $(\alpha \times n)$ leftmost (respectively rightmost) observations from each end by the observations that are the minimum and the maximum once the $2(\alpha \times n)$ extreme observations removed and then calculate the mean. On our small data set, the α-winsorized mean with $\alpha = 2/15$ is obtained by calculating the mean of:

$$1.1 \quad 1.1 \quad 1.1 \quad 1.3 \quad 1.4 \quad 1.9 \quad 2.0 \quad 2.5 \quad 2.6 \quad 2.8 \quad 2.9 \quad 3.0 \quad 3.1 \quad 3.1 \quad 3.1$$

and it is equal to 2.2.

The first characteristic usually studied for an estimator is its *bias*. It consists in looking at the difference between the estimator and the true value of the parameter. This quantity is a random variable and so the bias is in fact the expectation of this difference. So, an estimator is said to be unbiased if in average, it is right on target. Theoretically, the bias of an estimator $\hat{\theta}$ of θ is the fixed quantity:

$$E_{F_\theta}(\hat{\theta}) - \theta$$

and an unbiased estimator is an estimator with zero bias.

Note that statisticians often take an asymptotic point of view which means that they look at the properties of the estimators they consider when the sample size n goes to infinity. For instance, some biased estimators may have a bias that decreases to zero when the sample size goes to infinity and in that case, the estimator is said asymptotically unbiased.

For our simple example, the mean estimator and the median estimator are unbiased estimators of the mean of the gaussian distribution. If we try to interpret this property, it has to be stressed that an unbiased estimator does not guarantee that a given estimate is close to the true value. For the small data set for instance, it does not give any idea on how far is 2.5 or 11.0 from the true mean parameter. The deviations from the true value are averaged using the probability distribution. In some sense, it means that if we could generate many different data sets with the same generating process and average all the obtained estimates then we will be closed the true value but in practice we only have one data set and so we are usually seeking for supplemental properties.

Indeed, another important property is how precise is our estimator. If in average the estimator is right on the target, is it very dispersed around the target which won't be good or on the contrary tightened around the target? This dispersion is usually measured by the variance:

$$E_{F_\theta}\left[\hat{\theta} - E_{F_\theta}(\hat{\theta})\right]^2$$

When considering the normal model $\mathcal{N}(\mu, 1)$, the variance of the mean estimator is $1/n$ while the variance of the median estimator is approximately equal to $\pi/(2n) \simeq 1.57/n$.

Note that a more intuitive indicator that is linked with the previous two ones is the *Mean Squared Error* (MSE). It consists in calculating the squared difference between the estimator and the true parameter so that negative and positive differences cannot vanish and then taking the expectation. The definition of the MSE is:

$$E_{F_\theta}\left[\hat{\theta} - \theta\right]^2$$

The MSE is equal to the variance plus the squared bias which means that for unbiased estimators, the MSE is confounded with the variance of the estimator. Estimators can be compared in terms of mean squared error. If for estimating a parameter θ, two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are considered, we can define the relative *efficiency* of one estimator compared to the other by calculating the ratio of their mean squared errors. The estimator $\hat{\theta}_1$ is said more efficient than $\hat{\theta}_2$ if $\mathrm{MSE}(\hat{\theta}_1) < \mathrm{MSE}(\hat{\theta}_2)$. In a given class of estimators (for example the class of unbiased estimators), an estimator which is the most efficient in that class is said to be optimal.

Let us consider again the problem of estimating the mean in the $\mathcal{N}(\theta, 1)$ model with a sample of size n. It can be proved that the simple mean is the most efficient estimator among all the possible unbiased estimators. So the mean has an appealing optimal property and if we want to compare with the median estimator, the ratio of the MSE of the mean over the MSE of the median equals approximately 64%. In other words, if we want a given precision, the size of the sample when one uses the mean is only 64% the size of the sample needed when using the median. However, one should not forget that this optimality property of the mean is no longer true if the data generating process is not gaussian. If the data come from a Laplace distribution for instance, the median is much more efficient than the mean with a relative efficiency of 200%.

The indicators presented so far are basic statistical indicators for measuring the quality of estimators. In order to measure the robustness of estimators, other indicators are to be constructed and this will be done in section 4 after giving some general principles on robust statistics in section 3.

3. Some general principles. The estimators properties that have been detailed so far are based on the assumption that the data are generated according to a given parametric model. As stated in the introduction, it is common that gross-errors or outliers are present in the data set so that not all the observations can be considered as generated by the assumed model. However we still want to estimate the parameters of the parametric model followed by the majority of the data. The theory developed by Huber and Hampel in the sixties and seventies answers this question in an elegant way by considering the following principles.

3.1. *A functional approach.* First of all, a parameter is seen as a functional (function) T of the probability distribution underlying the data at the population level while an estimator of this parameter is simply the same functional but applied to the empirical distribution[1].

If we consider the mean and the median parameters for instance, for a model F_θ parametrized by θ, we have the following functionals. For the mean, $T(F_\theta) = \int x dF_\theta(x)$ and for the median, $T(F_\theta) = F_\theta^{-1}(1/2)$ where in this case, F_θ denotes the cumulative distribution function[2]. By using definitions based on functionals for estimators, it is possible to consider other models than the F_θ model and evaluate the behavior of the estimator in some neighborhood of F_θ.

One important property which is more elegant than asymptotic unbiasedness in the functional context is that the value of the functional at the parameter distribution F_θ should be equal to θ ($T(F_\theta) = \theta$). Intuitively, it is clear that we expect the functionals lead to recover the parameter we are interested in. Otherwise there is no hope that the functional applied to the empirical distribution will estimate correctly the parameter. If this property is verified, the functional is said *Fisher consistent* and only Fisher consistent functionals are considered in the robust statistics framework. Note that in the simple example introduced before, the four estimators of the "mean" parameter we have proposed are Fisher consistent.

3.2. *Neighborhood of probability distributions.* Once the estimator is defined through a functional approach, it is possible to study the behavior of this functional for other models than the original assumed F_θ model. As explained previously, our goal is to estimate the

[1] The empirical distribution is the discrete probability that gives a probability $1/n$ to each observation value in the sample

[2] $F_\theta(x)$ is the probability that a random variable following the F_θ distribution is less or equal to x. Some assumptions are needed in order the median being defined in a unique way.

parameter θ of the F_θ model even when some outlying observations do not follow this model but a different distribution denoted by G. It means that the model we have in mind for the whole data set is a mixture of two probability distributions F_θ and G with the assumption that the proportion of outliers, is less than one half. So, the model is of the form:

$$F_{\theta,\varepsilon} = (1-\varepsilon)F_\theta + \varepsilon G \tag{3.1}$$

with $(1-\varepsilon)$ greater than one half. By letting ε vary into the interval $[0; 1/2[$, the contaminated distributions $F_{\theta,\varepsilon}$ define in some sense a "contaminated" neighborhood of F_θ[3] sometimes called the *gross-error model*.

Note that the objective is to estimate θ and not calculate ε or G. The proportion ε is called the contamination proportion and the outliers are the observations that follow the G distribution.

3.3. *Equivariance properties.* Generally, the estimator we are interested in should verify some equivariance properties. Let me consider again the example of estimating a mean parameter and imagine that the data correspond to Celsius temperature in degrees. From this data, you estimate the mean by taking the arithmetic mean of the values but then you decide to transform your data on the Fahrenheit scale and reestimate the mean by taking the arithmetic mean of your transformed data. What you expect about your estimates is that when you apply the transformation that transforms the Celsius scale to the Fahrenheit scale to the estimate obtained from the initial data, the estimate you obtain will be the one you calculate on your transformed data. The fact that an estimator is capable to "follow" some transformation is called an equivariance property[4]. In the example on temperature, the transformation is an affine transformation and we are interested in estimators of the mean that are affine equivariant. Note that the mean and the median are both affine equivariant estimators. Depending on the problem considered and the parameter to estimate, different equivariance properties are required.

4. Measuring robustness. In the previous section, we introduced a general but rigorous framework that allows us to define some robustness measures for estimators. The three measures we develop below are not the only existing measures but they are the best known methods and the most widely used measures especially concerning the influence function and the breakdown point.

One robustness property consists in checking whether the functional of interest is continuous on a neighborhood of F_θ. This property is called the qualitative robustness. It will not be discussed further in the present paper since it is quite complex to define in detail and not widely used in practice. Moreover it is closely related with the breakdown point which is easier to understand and will be detailed in subsection 4.2. The interested reader can refer to [11] for details and [13] for a recent discussion on that property.

In the example we consider, the median and the α-trimmed mean ($\alpha \neq 0$) are qualitatively robust while the mean and the winsorized mean are not. So, in what follows, the winsorized mean won't be considered further.

4.1. *Influence function and B-robustness.* The most famous concept of robustness in statistics is without any doubt the influence function and the measures of robustness that

[3]Many other neighborhood are defined and studied in [11].

[4]Equivariance should not be confounded with invariance which means that a quantity remains unchanged after a transformation of the data.

are derived from it. It consists in some sense in a derivative of the functional at the F_θ model and is widely used in practice. As stated in [8] (p.84), it describes the effect of an infinitesimal contamination of the model on the estimator, standardized by the mass of the contamination. It gives a quantitative measure of the behavior of an estimator in a neighborhood of the model F_θ that can be further analyzed. In general, the desired property of an influence function is its boundedness. When an estimator has a bounded influence function, it is said *B-robust*.

Before giving the functional definition of the influence function, let me introduce the concept from the sample point of view. For a given estimator, the idea is to calculate the difference between the estimate at a given sample and the estimate when an observation x is added to the sample and divide by $1/(n+1)$. In the small data example, for the mean estimate, this calculous will lead to :

$$16 \times \left[11.0 - \frac{165.7 + x}{16}\right] = 165.7 - 15x$$

which is an affine function of x. Let us remark that if $x = 11.0$ which is the mean value of the original sample, the value of the function is zero. Such a function is called a *sensitivity curve* and is denoted by SC. If we denote by $x_1, x_2, \ldots, x_n$ the n observations and $\hat{\theta}(x_1, x_2, \ldots, x_n)$ the estimate at the sample, we have:

$$\mathrm{SC}(x; x_1, x_2, \ldots, x_n, \hat{\theta}) = \frac{\hat{\theta}(x_1, x_2, \ldots, x_n, x) - \hat{\theta}(x_1, x_2, \ldots, x_n)}{1/(n+1)}.$$

For the mean it gives:

$$\mathrm{SC}(x; x_1, x_2, \ldots, x_n, \hat{\theta}) = \sum_{i=1}^{n} x_i - nx$$

which is always an affine function. Such a function can be plotted and there also exists a version where instead of adding an observation x to the sample, one observation x_i is replaced by x. But for both versions the drawback of this sensitivity curve is that it depends on the sample. For each data set there is a different sensitivity curve and not only because the sample size changes. The functional approach is a very elegant way to circumvent the problem and gives a general expression for a given functional. The idea is to consider a neighborhood of the model F_θ of the form:

$$\tilde{F}_{\theta,\varepsilon} = (1-\varepsilon)F_\theta + \varepsilon\delta_x$$

where δ_x denotes the Dirac probability measure which gives a mass one at the point x and then calculate (when it is possible) the influence function:

$$\mathrm{IF}(x; F_\theta, T) = \lim_{\varepsilon \to 0} \frac{T(\tilde{F}_{\theta,\varepsilon}) - T(F_\theta)}{\varepsilon}$$

For a given functional T and a model F_θ, IF is a function of x. For the example where θ is the mean of the distribution F_θ and for $\hat{\theta}$ the mean estimator, we have $T(F) = \int t\, dF(t)$ and a simple calculous leads to $\mathrm{IF}(x; F_\theta, T) = \theta + x$ which is also an affine function of x. In general because the estimators we consider are affine equivariant, there is no loss of generality in considering that $\theta = 0$ (if not, we can always make an affine transformation such that it becomes true). For the mean, we have that $\mathrm{IF}(x) = x$. The interpretation of this influence function is that if we contaminate the distribution by a contamination in x, if the

x is large it will have a large influence on the mean estimator because the influence function is not bounded in x. An important criterion for an estimator is to have a bounded influence function and this property is called B-robustness. The mean estimator in our example is not B-robust.

The influence function of the median and the α-trimmed mean have been calculated in the literature (see for instance, [18]) but their expressions are not detailed further in the present paper. Rather a plot of the influence functions is given in Figure 1 for the mean (solid line), the α-trimmed mean with $\alpha = 2/15$ (dotted line) and the median (long dashed line). As can be seen from the plot, except for the mean, the influence functions of the estimators are bounded and so the trimmed mean and the median are B-robust estimators.

Other concepts have been derived from the influence function such as the gross-error sensitivity which, if the influence function is bounded, gives its maximum absolute value. The question of the smoothness of the influence function is also sometimes considered. None of the two "robust" estimators we have defined previously (median and trimmed mean) have a smooth influence function but it is possible to define other estimators (the well-known M-estimators) through their influence function. It means that one can design the influence function he dreams of (bounded, smooth,...) and obtain an estimator with the designed influence function.

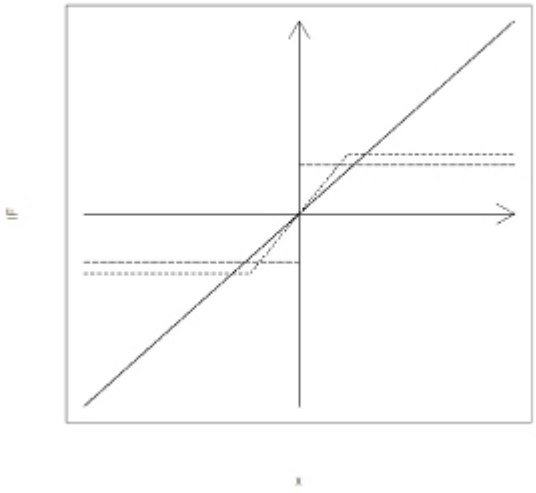

FIGURE 1. *Influence functions plot: solid line for the mean, long dashed line for the median, dotted line for the α-trimmed mean with $\alpha = 2/15$.*

4.2. *Bias curve and breakdown point.* Another widely used measure of robustness in the field of robust statistics is the breakdown point whose sample-based version, proposed by [5] is easy to understand. For a "mean" parameter, the breakdown point is the minimal fraction of observations one has to replace before the estimator tends to infinity. Let us consider the example of the mean estimator. If only one observation in the data set is replaced by an extremely large value, the mean will explode or "breakdown" which means that the breakdown point of the mean is $1/n$ where n is the sampling size of the data set.

At the limit, when the size of the sample goes to infinity, the value of the breakdown point of the mean is zero which is the worst possible case. On the contrary, for the median, the breakdown is 50% which is the best possible among the affine equivariant estimators of a "mean" parameter[5]. So, one can replace up to half of the observations minus one without causing an "explosion" of the median estimator. The breakdown point of the α-trimmed estimator is α and so choosing α is equivalent to choosing the breakdown point of the estimator. The sample-based version of the breakdown point has the property that it does not usually depend on the data set but only on the sample size and in that case, its limit when n goes to infinity is usually considered.

Let us consider an estimator of the "mean" parameter with a breakdown point $\varepsilon^* > 0$. This measure does not give us any idea concerning the behavior of the estimator if the percentage of contamination ε is between 0 and ε^*. Because the breakdown point is larger than ε, we know that the estimator is not going to infinity but we would like more information. The maximum deviation curve defined in [2] addresses the problem. It consists, for each ε in $[0; \varepsilon^*[$, in calculating the maximum absolute deviation between the estimate at the given data set and the estimate calculated when some of the observations have been replaced arbitrarily. By taking the maximum over the possible values of the good observations, it is possible to define a quantity that does not depend on the data set at hand as in [2]. Another possibility is to define the notion at the functional level.

Let us consider the gross-error model defined by (3.1) and define, for each percentage ε of contamination in $[0; 1]$, the maximum bias function of an estimator associated with a functional T by:

$$B(\varepsilon; T, F_\theta) = \sup_{F_{\theta,\varepsilon}} |T(F_{\theta,\varepsilon}) - T(F_\theta)|.$$

This expression is a function of ε and the breakdown point is derived as:

$$\varepsilon^*(T, F) = \inf\{\varepsilon > 0 | B(\varepsilon; T, F_\theta) = \infty\}.$$

The *maxbias curve* is obtained by plotting the maximum bias as a function of ε as in Figure 2.

For our simple "mean" example, it is not possible to draw the maxbias curve of the mean since its breakdown point is zero. The maxbias curves of the median is plotted on Figure 2. For the trimmed mean, the maxbias curve is given in [3] and is very close to the maxbias curve of the median but the difference is that it becomes infinite when $\varepsilon \geq \alpha$ while it becomes infinite when $\varepsilon \geq 1/2$ for the median.

5. Strategies for the construction of robust estimators. In the previous section, several measures of robustness have been derived. But the question on how to choose a robust estimator among the many different possibilities has not been discussed yet. The main philosophy in robust statistics is to take into account not only the question of efficiency (which is usually the only concern of usual parametric statistics) but also the question of robustness of a procedure. The problem is that a high efficiency requirement is often in conflict with a high robustness requirement which means that some compromise has to be found by the statistician. Two strategies, we describe briefly and without any mathematical

[5]In fact, [14] have proved that the maximal breakdown point of an affine equivariant estimator of a "mean" parameter is the ratio of the integer part of $(n + 1)/2$ and n which tends to $1/2$ when n tends to infinity.

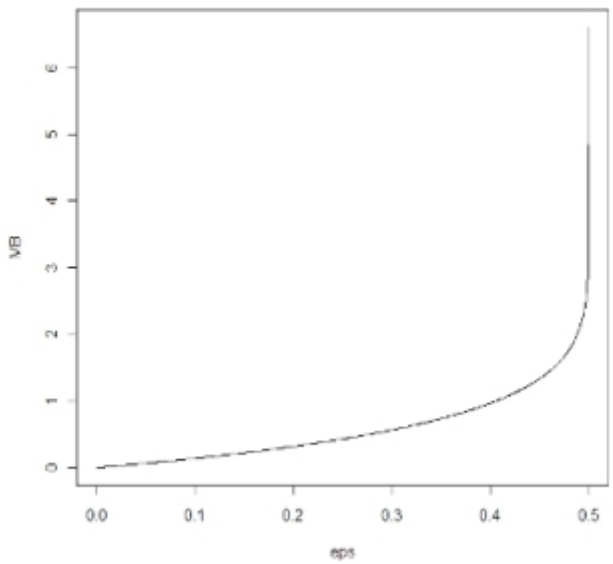

FIGURE 2. *Maxbias curve of the median*

formalism in the following subsections, have been developed in the pioneer and fundamental work by Huber and Hampel.

5.1. *Minimax strategies.* This strategy has been defined by [10] and the idea is to optimize the "worst" that can happen over a neighborhood of the target model. The neighborhood is the gross-error model and the "worst" has to be understood in the sense of a particular measure. For estimating a "mean" parameter for instance, the measure can be the asymptotic bias or the asymptotic variance. So, the idea is to *mini*mize the *max*imum of the asymptotic bias or the asymptotic variance over all possible distribution which belongs to the gross-error neighborhood of the target model (the model we assume for the majority of the data). It means in particular that a least favorable distribution has to be found in the defined neighborhood. For the asymptotic bias, under some assumptions on the model, the minimax estimator is the median. For the asymptotic variance, the theory is deeper and the estimators it leads to will not be detailed further[6]. The theory underlying this strategy is very elegant but as stated by Huber himself, "it does not carry beyond problems possessing a high degree of symmetry".

5.2. *The infinitesimal approach.* This infinitesimal approach is the one developed in [8] and is also called the approach based on the influence function. In general, the idea is to take into account the qualitative robustness, the influence function and the breakdown point of estimators. In order to define some optimal robust estimator, one possibility is to look for the most efficient estimator but among a class of robust estimators. Indeed, [7] proposed to minimize the asymptotic variance subject to an upper bound on the gross-error sensitivity which is the maximum absolute value of the influence function of the estimator. Results have been obtained for different domains of statistics including the case of estimating a "mean" parameter and such estimators are said optimal B-robust estimators. Contrarily to the previous strategy, the infinitesimal strategy allows only infinitesimal devi-

[6]In [11] p.97, three asymptotically minimax estimators are given for the normal model.

ations from the model but because of that, it works for more general families of distributions.

One interesting point in the case of estimating a "mean" in a simple model like the gaussian one is that both the minimax and the infinitesimal strategies described above lead to the same results.

As stated previously, many different classes of robust estimators exist in the literature even in the simple case of estimating a "mean" parameter. Recently, in [19], some proposals have been made in order to improve the trimmed and winsorized mean estimators. All these estimators have their drawbacks and advantages and the choice depend on the criteria most valued by the statistician which in general depend on the problem at hand.

6. Conclusion. The aim of robust statistics is not only optimality but also stability of procedures. By developing stability measures in a general framework, it gives the possibility to derive the robustness properties of statistical procedures. There exist many families of robust estimators in the literature even for the simple case of estimating a "mean" parameter. The estimators we have been considering in the present paper are all examples of so-called L-estimators which means that they are linear combinations of order statistics and are easy to compute. But the classes of M- and R-estimators are also very important classes of estimators and the reader should refer for instance to [11] for details about such estimators.

As the reader may have noticed, and even though this presentation is focused on a extremely small part of the existing methodology, robust statistics is a very rich and complex field. The theory that has been developed in the pioneer work by Huber and Hampel is elegant and insightful but difficult to generalize to more complex domains of statistics. Moreover, as stated for instance in the recent paper by [16] on "the Changing History of Robustness", robust methods are not widely used by applied statisticians. My own opinion about the situation is that even when there exist softwares for robust procedures, most of the time, there are some parameters not easy to tune, like the parameter α for the trimmed or the winsorized mean. However, statisticians working in the area of robust statistics are still very active and continue spreading robust methodology in all the possible domains of statistics together with new algorithms and new softwares.

References.

[1] Cantoni, E. and Ronchetti, E. (2001). Resistant Selection of the Smoothing Parameter for Smoothing Splines. *Statistics and Computing*, **11** 141–146.
[2] Croux, C. (1996). Maximum Deviation Curves for Location Estimators. *Statistics* **28** 285–305.
[3] Croux, C., and Haesbroeck, G. (2002). Maxbias Curves of Location Estimators based on Subranges. *J. Nonparam. Statist.* **14** 295–306.
[4] Daouia, A. and Ruiz-Gazen, A. (2005) Robust Nonparametric Frontier Estimators: Qualitative Robustness and Influence Function. *Statistica Sinica* **16** 1233–1253.
[5] Donoho, D. L. and Huber, P. J. (1983). The Notion of Breakdown Point. In Peter J. Bickel, Kjell A. Doksum, and J. L. Hodges, editors, *A Festschrift for Erich L. Lehmann*, Belmont, CA: Wadsworth.
[6] Fréchet, M. (2006). Sur une limitation très générale de la dispersion de la médiane. *J. Soc. Française Statist.* **147** 2 5–15.
[7] Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* Ph.D thesis. University of California, Berkeley.
[8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Probability and Statistics, Wiley, New York.
[9] Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M.-P. (2009) *Robust methods in biostatistics.* Wiley Series in Probability and Statistics, Wiley, New York.
[10] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[11] Huber, P. J. and Ronchetti, E.(2009). Robust statistics, Wiley series in probability and mathematical statistics, Wiley, New York.
[12] Maronna, R. A. and Martin, R. D. and Yohai, V. J.(2006). *Robust statistics: theory and methods*, Wiley series in probability and statistics, Wiley, New York.
[13] Mizera, I. Qualitative robustness and weak continuity: the extreme unction? In J. Antoch, M. Hukov and P.K. Sen (eds.), *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurecková* 169–181. Institute of Mathematical Statistics, Beachwood.
[14] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.
[15] Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *J. Am. Statist. assoc.* **68** 872–879.
[16] Stigler, S. M. (2010). The Changing History of Robustness. *Amer. Statistican* **64** 4 277–281.
[17] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Olkin, I. (ed.), *Contributions to Probability and Statistics*, 448-485. Stanford Univ. Press, Stanford.
[18] Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego.
[19] Wu, M. and Zuo, Y. (2005). Trimmed and Winsorized means based on a scaled deviation. *J. Statist. Planning and Inference* **139** (2) 350–65.

Address of the author
Toulouse School of Economics
21, allée de Brienne
31000 Toulouse
France
E-mail: anne.ruiz-gazen@tse-fr.eu