A Unified Approach of Parameter Estimation

Ahmed Guellil, Tewfik Kernane

A Unified Approach of Parameter Estimation

Ahmed Guellil[1] and Tewfik Kernane[2]
[1]*Department of Probability and Statistics, Faculty of Mathematics*
University of Sciences and Technology USTHB,
BP 32 El-Alia, Algeria
[2]*Department of Mathematics, Faculty of Science*
King Khaled University, Abha Kingdom of Saudi Arabia
e-mail: guellilamed@yahoo.fr, tkernane@gmail.com

Abstract

We introduce a new distance and we use it to parameter estimation purposes. We observe how it operates and we use in its place the usual methods of estimation which we call the methods of the new approach. We realize that we obtain a discretization of the continuous case. Moreover, when it is necessary to consider truncated data nothing is changed in computations.

Key words and phrases: Parameter estimation, minimum distance estimation, family of auxiliary distributions, type-I censoring.

1 Introduction

In the traditional approach of estimation there are three following basic elements: a family of theoretical probability distributions, an empirical law and some estimation methods. We choose a method according to its properties and the problem at hand. The empirical distribution and the family of theoretical laws are datum of the problem whatever the method chosen. We propose a new viewpoint where the empirical law corresponding to a given theoretical one is perceived as being an empirical conditional distribution with the knowledge of the data. It becomes then an estimate of the conditional theoretical law knowing the observations before being an estimation for the theoretical distribution from which it emanated.

We introduce a new distance and we use it to estimate. We observe then how it operates and use in its place the usual methods of estimation which we call the methods of the new approach. We notice then that this leads to a unification of the methods of estimation since we do not make any more distinction between fixed type-I censored data and complete samples and between discrete

and continuous cases. We thus obtain a considerable lightening in the procedures of computation in estimation problems. The distinction in the traditional approach between truncated or type-I censored data and complete samples is not really justified since all samples are in fact truncated. Indeed, a sample is not truncated if it covers the totality of the support of the distribution from which it was drawn, if not it is truncated. Moreover it is natural to consider that the sample describes only the parts of the distribution which capture the data. The other parts are obtained by deduction. Also, the discretization for the continuous case obtained with the new approach is justified. Indeed, practically all usual distributions can be reconstituted exactly starting from two or three points of their graphs. We can then estimate them starting from two or three points which represent their graphs empirically. In addition to the unification of several methods of estimation we note that the estimations with the new measure have the following specific properties. It does not require that the family of candidate theoretical distributions to be made up of the same type of laws. There is always a solution which will be acceptable in general. If the ratios of the frequencies of an empirical distribution coincide with those of the theoretical one from which it emanated then, from the first we can find the second with certainty. If the ratios of the frequencies of the empirical distribution coincide with those of the theoretical one which it best fits, then the estimations obtained are optimal in the sense that one cannot improve them. We checked also on some examples, analytically and numerically, that when we make tending the ratios of the frequencies of the empirical distribution towards those of the theoretical one, then the estimates tend towards the true parameters. This last property implies convergence of the estimators. We prove the convergence of the estimators obtained with the new measure for a broad class of usual laws. Moreover, with the new measure we achieve more flexibility in computation compared to the method of maximum likelihood.

We can distinguish in this paper three different parts. The first is on the subject of a new distance, presented in section 2. We can be interested and study it as a mathematical object without necessarily referring to its applications in statistics. That is a metric which does not have none equivalent in the theory of mathematics. We noted some of its remarkable properties, this promises new prospects. The second relates to the use of this distance in problems of estimation in statistics. That gives birth to a new method of estimate, presented in section 3. The study suggested in this part is not at all exhaustive. But the results obtained are already interesting and encouraging. The third part relates to a new approach of estimation. We can look at this new approach separately; this is the discretization of the methods of the continuous case. By adopting it we widen the field of application of the usual methods of estimation. It is presented in section 4. In sections 5 and 6 we gave using examples a practical illustration of the possibilities of the new method and the new approach of estimation. In section 7 we showed what the users of statistics gain immediately in the light of our work in comparison with the traditional approach. Lastly, in section 8 we gave in short a reminder of the whole of the results obtained.

2 A New Distance Between Probability Distributions

In statistics, we use distances to measure the difference between probability distributions. Usually these distances are conceived in the same manner, the differences between distributions are almost always expressed by using variations in geometric sense between their graphs. We introduce a distance which operates differently. It is based on relativist properties of probability measures. But its interest is due especially to the fact that it is not equivalent to usual distances.

Definition 1 *Consider two probability measures P and Q defined on the same measurable space $(\Omega, \mathcal{F})$, f and g being their respective probability distributions not necessarily with respect to the same measure and E an event from this space. We say that f and g have same variations on E, if their restrictions on E define the same probability measure on E equipped with the sigma algebra trace of $\mathcal{F}$ on E.*

Example 2 *Let f be a density of a probability measure P and E an event such that $P(E) > 0$. The restriction of f on E and the conditional distribution of f with respect to E define the same probability measure on E and consequently they have the same variations on E.*

Example 3 *Let f be a probability distribution and c a positive constant. The functions f and $g = f + c$ have the same variations in the geometric sense but they do not have the same variations within the meaning of the above definition.*

Proposition 4 *Let f and g be two probability distributions defined and positives on a part E not reduced to only one element. If in any point (x, y) of $E \times E$, we have*

$$\frac{f(x)}{f(y)} = \frac{g(x)}{g(y)} \tag{1}$$

then f and g have same variations on E.

Proof. If E is discrete the distribution generated by the restriction of f on E is $f_E = f/\sum_{x\in E} f(x)$ on E and $f_E = 0$ otherwise. If x_0 is in E such that $g(x_0) \neq 0$ then (1) implies that for all x in E, $f(x) = g(x)f(x_0)/g(x)$. By replacing f in f_E, we find the conditional distribution generated by g on E. We obtain then the result. In the same way, we obtain the result for probability densities on $\mathbb{R}$ with respect to the Lebesgue measure on $\mathbb{R}$ when E is a subset of $\mathbb{R}$ with positive probability. ■

Definition 5 *Let f and g be two probability distributions and E an event on which they are strictly positive. If E is discrete and no reduced to only one element, we call distance in variations between f and g on E the quantity*

$$d_v(f, g)_E = \sum_{(x,y)\in E} \left| \frac{f(x)}{f(y)} - \frac{g(x)}{g(y)} \right|.$$

If E is an interval of $\mathbb{R}$ and, f and g are probability densities on $\mathbb{R}$, with respect to Lebesgue measure μ on $\mathbb{R}$, we call distance in variations between f and g on E, the quantity

$$d_v(f,g)_E = \iint\limits_{E\times E} \left| \frac{f(x)}{f(y)} - \frac{g(x)}{g(y)} \right| \mu(dx)\mu(dy).$$

Note that d_v possesses the properties of symmetry and triangle inequality. But in the identity property $d_v(f,g)_E = 0 \Longleftrightarrow f \equiv g$ on E, the equality between f and g must be understood in the sense that f and g have the same variations on E.

Let d be the distance which measures the difference in two points x and y between two functions f and g by the quantity $d(f,g)(x,y) = |f(x) - g(x)| + |f(y) - g(y)|$.

Proposition 6 *We have the following property for the distance d_v :*
$d(f,g)(x,y) = 0 \Longrightarrow d_v(f,g)(x,y) = 0$, *the converse is not always true.*

Proof. Follows directly from the definitions of d and d_v. ■

3 New Method of Estimation

3.1 Frequency Tables

Let $\mathcal{F}$ be a family of probability distributions. If it contains only one type of distribution we say that it is *homogeneous* otherwise we say that it is *heterogeneous*. A heterogeneous family can be made up of several types of discrete and absolutely continuous distributions. Let us consider f in $\mathcal{F}$ and some values $y_1, ..., y_k$ from its support. We call theoretical table of frequencies of f based on $y_1, ..., y_k$ or with support $y_1, ..., y_k$ the k couples $(y_1, f_1), (y_2, f_2), ..., (y_k, f_k)$ where $f_i = f(y_i)/\sum_{j=1}^{k} f(y_j), i = 1, 2, ..., k$. We note $\bar{f}$ the distribution defined by this table. We say that the precedent table completely characterizes the family $\mathcal{F}$ if and only if there is a bijection between $\mathcal{F}$ and $\bar{\mathcal{F}} = \left\{\bar{f},\ f \in \mathcal{F}\right\}$. In this case, theoretically, from $\bar{f}$ we can determine f. $\bar{f}$ will be a representative element of f in $\bar{\mathcal{F}}$. We call $\bar{\mathcal{F}}$ the family of auxiliary distributions based on $y_1, ..., y_k$ associated to $\mathcal{F}$. We say also that the y_i, $i = 1, 2, \ldots, k$ form a basis of observations which characterizes the family $\mathcal{F}$.

Proposition 7 *Let us consider two laws of probability f and g belonging to a family of distributions $\mathcal{F}$ and having the same support E. If F is a basis of observations which characterizes the family $\mathcal{F}$ then $d_v(f,g)_F = 0$ implies that $d_v(f,g)_E = 0$.*

Proof. If $d_v(f,g)_F = 0$ then $\overline{f} = \overline{g}$ where $\overline{f}$ and $\overline{g}$ are the auxiliary distributions of f and g respectively based on F. If in addition F constitutes a basis of observations characterizing $\mathcal{F}$ then, we deduce that $f = g$. ■

It should be noted that none of the usual distances has this property and it is a key idea to justify the use of the methods of point estimation for discrete case in the continuous one.

3.2 Estimation

Let us consider k couples $(y_1, f_1), ..., (y_k, f_k)$ of a table of empirical frequencies obtained after grouping the observations of a probability law belonging to a family of distributions $\mathcal{F}$, with $f_1 + f_2 + ... + f_k = 1$. It will be said that it empirically characterizes the family $\mathcal{F}$ if the theoretical frequency table based on the y_i, $i = 1, 2, ..., k$ characterizes it too. In the sequel our starting point will be always, in the continuous as in the discrete cases, a table of empirical frequencies, based on k values $y_1, ..., y_k$, constituting a basis of observations which completely characterizes the studied family. We suppose that it is a datum of the problem and thus one does not discuss the way of obtaining it, in particular in the continuous case. We can use for example procedures to select the optimal number of bins for a regular histogram (see for example Birgé and Rozenholc [2]). When we use the maximum likelihood procedure, theoretically nothing prohibits to estimate n parameters from a table of empirical frequencies, based on k values where k is lower or equal to n. But in practice we encounter sometimes difficulties which we do not expect. In certain cases we note that the results obtained are completely aberrant. We quote from the literature some paradoxes attached to the use of the maximum likelihood procedure in these cases ([3]). When we use tables of empirical frequencies whose basis characterizes the family of theoretical probability distributions which contains the distribution which we seek we avoid in advance these difficulties**.** We will indicate by $\hat{f}$ the discrete empirical distribution represented by this table. We notice that it is completely given if the ratios $f_i/f_j = \hat{f}(y_i)/\hat{f}(y_j)$ $i, j = 1, 2, ..., k$ are known and if $\hat{f}$ arises from a sample of a given theoretical distribution f, then from the law of large numbers $\hat{f}(y_i)/\hat{f}(y_j)$ tends to $f(y_i)/f(y_j)$ when the sample size tends to infinity. This result remains valid even when the support S represents a fixed type-I censored sample. When grouping in classes if one withdraws several classes and their frequencies, the frequencies of the remaining classes keep this property. Whether the sample considered is truncated or not and that the distribution from which it belongs is discrete or absolutely continuous, we can measure the difference in variations between $\hat{f}$ and a theoretical distribution f in $y_1, ..., y_k$ by

$$d_v(\widehat{f}, f)(y_1, ..., y_k) = \sum_{i,j \in \{1,...,k\}} \left| \frac{\widehat{f}_i}{\widehat{f}_j} - \frac{f(y_i)}{f(y_j)} \right|.$$

Since $\widehat{f}$ converges in probability towards f then $d_v(\widehat{f}, f)$ converges in probability towards 0.

Let us consider two probability distribution f and g which does not belong necessarily to the same type of laws and not equal to zero in $y_1, ..., y_k$ If $d_v(\widehat{f}, f)(y_1, ..., y_k) < d_v(\widehat{f}, g)(y_1, ..., y_k)$, we say that $\widehat{f}$ is more close to f than to g, in the sense of dv. We thus define a new method of estimation.

Example 8 *We simulated* 10000 *samples of size* 100 *from a binomial distribution* $\mathcal{B}(8, 0.1)$ *and* 10000 *others from a* $\mathcal{B}(15, 0.15)$. *For each sample obtained we kept only the observations belonging to* $\{0, 1, 2, 3\}$ *with their frequencies. Then, starting from the empirical distribution thus defined we tried to identify the law simulated among the two binomial distributions considered. The correct distribution is selected for* $98,8\%$ *of cases when we used samples from the former and for* $99,43\%$ *of cases when from the latter.*

Example 9 *We simulated* 10000 *samples of size* 1000 *from* $\mathcal{W}(1.2, 1.5)$ *and we omitted the observations below the threshold* 1.25. *Each truncated sample was summarized into* 11 *classes. We selected between* $\mathcal{W}(1.2, 1.5)$ *and the Gamma distribution* $G(2, 0.5)$ *using the metric* d_v. *The distance* d_v *has selected the correct distribution, that is* $\mathcal{W}(1.2, 1.5)$, 98.16%.

Let us consider in a problem of estimation, a family of the theoretical laws $\mathcal{F}$ and an empirical distribution $\widehat{f}$ with support $y_1, ..., y_k$ which constitutes a basis of observations characterizing $\mathcal{F}$. If it exists f belonging to $\mathcal{F}$ such as $d_v(\widehat{f}, f)(y_1, ..., y_k) = 0$, we say that f is an exact solution.

Proposition 10 *The exact solution, when it exists, is optimal in the sense that we cannot improve it.*

Proof. Indeed, in this case there is in $\mathcal{F}$ a distribution whose table of frequencies coincides exactly with that of $\hat{f}$, it is unique and it is f. ■

Criterion 11 (of quality) *Let* $\hat{f}$ *be an empirical distribution and* f *the theoretical one which best fits when we estimates by a given method. If* $d_v(\hat{f}, f) = 0$ *then according to the preceding proposition the estimate obtained is optimal in the sense that it is the best possible improvement of the estimation.*

We have there a quality criterion when it holds, not only it supplants all the usual criteria but more since it gives a total and definitive guarantee of the optimality of the estimates. One will further show with examples that in some cases we can very easily find estimates possessing this property. We will also show by using examples that, when one makes tending $d_v(\hat{f}, f)$ towards 0 the differences between the estimates and the estimated values tend towards 0 and at end one obtains their exact values. The latter property which remains to be proved in the general case implies immediately convergence of estimates. For the moment there is already the following result.

3.3 Convergence in Probability of the Minimum Distance Estimator

Let $X_1, ..., X_n$ a sample with $X_i \sim f(x, \theta)$, $\theta = (\theta_1, ..., \theta_s)^t \in \Theta \subseteq \mathbb{R}^s$, with

$$f(x, \theta) = K(x) \times \exp\left\{\sum_{k=1}^{s} \theta_k T_k(x) + A(\theta)\right\}, \tag{2}$$

$x \in \mathcal{X} \subseteq \mathbb{R}$, where $\mathcal{X}$ is a Borel set of $\mathbb{R}$ such that $\mathcal{X} = \{x : f(x,\theta) > 0\}$ for all $\theta \in \Theta$.

The family (2) is a large family of distributions, one finds there, for example, the family of the normal laws, and the family of the laws of Poisson. We assume that the support $\mathcal{X}$ does not depend on θ. Denote by $\widetilde{\theta}_n$ the estimator by the minimum of metric d_v between the empirical and theoretical distributions $\widehat{f}_n$ (based on a sample of size n) and $f(\cdot,\theta)$, that is

$$\widetilde{\theta}_n = \arg\min_{\theta} d_v(f(\cdot,\theta), \widehat{f}_n).$$

This estimator falls into the class of M-estimators. Using well known theorems on the convergence of M-estimators (see for example Amemiya [1]) we will prove that $\widetilde{\theta}_n$ converges in probability to the true parameter.

Proposition 12 *Let $X_1, ..., X_n$ be a sample from the family of distributions (2). If the set of natural parameters Θ is convex and the true parameter θ is an interior point of Θ, then the estimator $\widetilde{\theta}_n$ by the minimum of the distance of variations d_v converges in probability to the true parameter θ, i.e.,*

$$\widetilde{\theta}_n \xrightarrow{P} \theta.$$

Proof. Since we search for a minimum of the criterion function d_v, it suffices to show, under the assumptions of the family (2) and the convexity of the set Θ, that $d_v(\theta, \underline{x})$ seen as a function of θ is a convex function (see Amemiya [1]). Hence, this reduces the problem to the convexity of

$$\delta_{ij}(\theta) = \left| \frac{f(y_i,\theta)}{f(y_j,\theta)} - \frac{\widehat{f}(y_i)}{\widehat{f}(y_j)} \right|.$$

For $\lambda, \mu \in \mathbb{R}$ with $\lambda + \mu = 1$, and $\theta^{(1)}, \theta^{(2)} \in \Theta$, we have

$$\delta_{ij}(\lambda\theta^{(1)} + \mu\theta^{(2)}) = \left| C_{ij} \exp\left\{ \sum_{k=1}^{s} \left[\lambda\theta_k^{(1)} + \mu\theta_k^{(2)} \right] (T_k(y_i) - T_k(y_j)) \right\} - A_{ij} \right| \tag{3}$$

where $C_{ij} = K(y_i)/K(y_j)$ and assume that $C_{ij} > 0$ and $A_{ij} = \widehat{f}(y_i)/\widehat{f}(y_j)$.
we have from the convexity of the exponential function that

$$\begin{aligned} \exp\left\{ \sum_{k=1}^{s} \left[\lambda\theta_k^{(1)} + \mu\theta_k^{(2)} \right] (T_k(y_i) - T_k(y_j)) \right\} \quad \leq \quad & \lambda \exp\left\{ \sum_{k=1}^{s} \theta_k^{(1)} (T_k(y_i) - T_k(y_j)) \right\} \\ & + \mu \exp\left\{ \sum_{k=1}^{s} \theta_k^{(2)} (T_k(y_i) - T_k(y_j)) \right\}, \end{aligned}$$

then

$$C_{ij} \exp\left\{ \sum_{k=1}^{s} \left[\lambda\theta_k^{(1)} + \mu\theta_k^{(2)} \right] (T_k(y_i) - T_k(y_j)) \right\} - A_{ij} \leq$$

$$\lambda C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(1)} \left(T_k(y_i) - T_k(y_j)\right)\right\} + \mu C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(2)} \left(T_k(y_i) - T_k(y_j)\right)\right\}$$

$$- (\lambda + \mu) A_{ij} \leq \lambda \left[C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(1)} \left(T_k(y_i) - T_k(y_j)\right)\right\} - A_{ij}\right] +$$

$$\mu \left[C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(2)} \left(T_k(y_i) - T_k(y_j)\right)\right\} - A_{ij}\right].$$

Introducing the absolute value we get

$$\delta_{ij}(\lambda\theta^{(1)}+\mu\theta^{(2)}) = \left|C_{ij} \exp\left\{\sum_{k=1}^{s} \left[\lambda\theta_k^{(1)} + \mu\theta_k^{(2)}\right] \left(T_k(y_i) - T_k(y_j)\right)\right\} - (\lambda + \mu) A_{ij}\right|$$

$$\leq \lambda \left|C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(1)} \left(T_k(y_i) - T_k(y_j)\right)\right\} - A_{ij}\right|$$

$$+\mu \left|C_{ij} \exp\left\{\sum_{k=1}^{s} \theta_k^{(2)} \left(T_k(y_i) - T_k(y_j)\right)\right\} - A_{ij}\right| = \lambda\delta_{ij}(\theta^{(1)}) + \mu\delta_{ij}(\theta^{(2)}).$$

Hence $\delta_{ij}(\theta)$ is a convex function of θ, which implies the convexity of $d_v(\theta, \underline{x})$ seen as a function of θ and then the convergence in probability of the minimum of distance d_v estimator. ■

4 New Approach of Estimation

4.1 Foundation

Let us consider in a problem of estimation the family of theoretical distributions $\mathcal{F}$ and an element f belonging to $\mathcal{F}$. We have in an obvious way, $d_v(\widehat{f}, f)(y_1, ..., y_k) = d_v(\widehat{f}, \bar{f})(y_1, ..., y_k)$ where $\bar{f}$ is the representative of f in $\bar{\mathcal{F}}$, $\bar{\mathcal{F}}$ being the family of auxiliary distributions based on $y_1, ..., y_k$, associated to $\mathcal{F}$. $\bar{f}$ is a discrete probability distribution with same support as $\hat{f}$ and depend on the same parameters of f. If the theoretical table of frequencies based on $y_1, ..., y_k$ characterizes completely the family $\mathcal{F}$ then the determination of f is equivalent to the determination of $\bar{f}$. When $\mathcal{F}$ is homogeneous, for determining $\bar{f}$, instead of dv we can also make use of the usual methods (method of moments, method of maximum likelihood, Bayesian Methods, ... etc.). Then they will be called the methods of the new approach. When proceeding in this way, all occurs as if one replaces the family of the theoretical distributions $\mathcal{F}$ by the corresponding family $\bar{\mathcal{F}}$. We note also what follows:

1. In discrete case, if the usual methods of estimation are used it is as if one estimates in a traditional way starting from truncated samples. This supposes that it is considered that any sample which does not completely cover the support of the distribution from which it is resulting is truncated in a deterministic way, the truncation being the parts which do not appear in the observations.

2. In continuous case, often in practice one associates with the sample of observations an optimal discrete distribution in a certain way and one uses it to estimate. Then when replacing dv by the usual methods we obtain a discretization of the continuous case.

3. In discrete case $\bar{f}$ represents the conditional distribution of f knowing the observations $y_1, ..., y_k$. In the continuous case $\bar{f}$ is calculated in a similar manner. It seems that there also it has the same interpretation except that this type of calculation does not exist in the theory of probability.

For reason of coherence only with what has just been said in 1, 2 and 3, we propose to view the empirical distribution as being the conditional empirical distribution knowing the observations, since it is calculated knowing the observations, even if that is not obvious in the continuous case. One then conceives it more easily as being an estimate of $\bar{f}$ before being for f.

5 Analytical computation

In this part we will organize a discussion around some very simple examples to try to reveal the specificity of the new approach and its contribution compared to the traditional one. Let us consider a table of frequencies based on two observations x and y with their respective frequencies n_1 and n_2. Starting from such table, with the new method one can estimate only one parameter. Such table characterizes practically all the families of usual laws when one has to estimate only one parameter. We can obtain such a table when the sample considered is not truncated but of small size or is truncated and grouped in two classes only. In the light of the new distance we will see in the examples which follow that, according to whether one estimates only one parameter or two simultaneously, even if the sample is not of small size, it will be henceforth preferable to group it in two or three classes only because one can gain in the precision of the estimations. Indeed, the two or three points obtained have more weight to represent the theoretical points of the distribution which they describe empirically and the method of estimation with d_v practically always gives in this case an optimal solution in the most general meaning.

5.1 Estimation of the parameter of the exponential distribution

Assume we want to estimate, from the preceding table, the probability density f_λ given by $f_\lambda\ (x) = \lambda e^{-\lambda x}$ if $x > 0$ and $f_\lambda(x) = 0$ otherwise, $\lambda > 0$, and F denotes its cdf.

a. Suppose it is a summary of a not truncated sample. Then the estimators of λ by the methods of maximum likelihood of the classical approach $\hat{\lambda}$ and the new one $\hat{\lambda}_N$ are respectively: $\hat{\lambda} = (n_1 + n_2) / (n_1 x + n_2 y)$ and $\hat{\lambda}_N = (\log(n_1) - \log(n_2)) \diagup (y - x)$ .As we can see, in general $\hat{\lambda}$ is different from $\hat{\lambda}_N$. When we compute $\tilde{\lambda}$, the estimation obtained using d_v, we find that it is equal to $\hat{\lambda}_N$. $\tilde{\lambda}$ is here optimal in the general sense. If

$$\frac{n_1}{n_2} = \frac{f(x)}{f(y)} + \varepsilon$$

then

$$\hat{\lambda}_N(\varepsilon) = \lambda + \varepsilon k$$

k being a constant. Then $\hat{\lambda}_N(\varepsilon)$ tends towards λ when ε tends towards 0. We can check that the difference between λ and $\hat{\lambda}(\varepsilon)$ does not tend towards 0 when ε tends towards 0. If the sample size tends towards infinity then, from the law of large numbers, the differences between the ratios of the empirical relative frequencies and those theoretical which their correspond tend towards 0 and consequently $\hat{\lambda}_N$ tends to λ. But one can have these variations close to 0 same for samples of finite sizes. It is noticed that the first solution here is always acceptable but the second not. The second is not acceptable only if there are anomalies in the sample of observations and then one is warned. We are not able to detect the sample deficiency from the first. The second is not acceptable when $x < y$ and $n_2 < n_1$ or conversely, but it is not what one expects, since the exponential law being decreasing, $x < y$ we must have $n_1 > n_2$. Now if in a problem the preceding exact solution is not acceptable and we have to propose an estimate of λ, that is always possible with the new method. Put

$$\alpha(\lambda) = \left|\frac{f(x)}{f(y)} - \frac{n_1}{n_2}\right| + \left|\frac{f(y)}{f(x)} - \frac{n_2}{n_1}\right| \text{ and } E = \{\alpha(\lambda), \lambda > 0\}$$

E is a part of $\mathbb{R}$ which is bounded below by 0. It admits then a lower bound say α_0. If α_0 is in E then there is $\lambda_0 > 0$ such that $\alpha(\lambda_0) = \alpha_0$. In this case the estimation of λ is λ_0. If α_0 is not in E then, whatever the strictly positive integer n, there exists $\lambda > 0$ such that $|\alpha(\lambda) - \alpha_0| < 1/n$. Put $A_n = \{\lambda > 0 / \ |\alpha(\lambda) - \alpha_0| < 1/n\}$. A_n is a decreasing sequence and then there exists A_0 such that $\lim_{n\to\infty} A_n = A_0$. In this case, each value λ from A_0 can be considered as an estimation of λ with the new approach.

b. Assume now that the table given is that of a fixed type-I censored data. For example in a not truncated grouped data one kept only the centers of two classes and their corresponding frequencies. With the new approach the table is enough and the solution is exactly the same as previously. But in this case the preceding estimate of the traditional approach is not valid here. One must use the methods of truncated data. One then needs the part of the support of f represented here by x and y. To be able to carry out calculations let us suppose that this table is the summary of the observations falling into the interval $[0, c]$ with $c > 0$. That is a right truncated sample. We consider the observed likelihood

$$L_{obs} = \left(\frac{f(x)}{F(c)}\right)^{n_1} \left(\frac{f(y)}{F(c)}\right)^{n_2}.$$

We have to consider that n_T observations are greater than c and have been discarded, but n_T is unknown. In order to compute the complete likelihood we have to determine the conditional distribution of n_T given that the observations

follows an exponential distribution to be able to implement the EM algorithm which require the computation of the conditional expectation of the complete log-likelihood function. It is then not possible to have an analytic solution and a recursive procedure is used to achieve a numerical solution. In general it is not always easy to use the method of maximum likelihood as let it believe the examples on the usual laws. Although Maximum likelihood estimators have good statistical properties in large samples, they often cannot be reduced to simple formulas, so estimates must be calculated using numerical methods.

5.2 Estimation of the parameters of a normal distribution

Let us consider a normal law $N(m, \sigma)$.

5.2.1 Estimation of the average

Solving the following equation in m :

$$\frac{n_1}{n_2} - \frac{\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(y-m)^2}{2\sigma^2}\right)} = 0$$

we obtain

$$\tilde{m} = \frac{1}{-\frac{x}{\sigma^2} + \frac{y}{\sigma^2}} \left(-\ln \frac{n_1}{n_2} - \frac{1}{2}\frac{x^2}{\sigma^2} + \frac{1}{2}\frac{y^2}{\sigma^2} \right)$$

It should be noted that $\tilde{m}$ is function of σ. When solving precedent equation after replacing (n_1/n_2) by $(f(x)/f(y)) + \varepsilon$, we obtain:

$$\tilde{m}(\varepsilon) = \frac{1}{-\frac{x}{\sigma^2} + \frac{y}{\sigma^2}} \left(-\ln \left(\frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{e^{-\frac{(y-m)^2}{2\sigma^2}}} + \varepsilon \right) - \frac{1}{2}\frac{x^2}{\sigma^2} + \frac{1}{2}\frac{y^2}{\sigma^2} \right)$$

where $\lim_{\varepsilon \to 0} \tilde{m}(\varepsilon) = m$.

5.2.2 Estimation of the Variance

Solving the following equation in σ,

$$\ln \frac{n_1}{n_2} = -\frac{(x-m)^2}{2\sigma^2} + \frac{(y-m)^2}{2\sigma^2}$$

we have:

1. If $\frac{n_1}{n_2} = 1$ and $-2mx + 2my + x^2 - y^2 = 0$,any value σ belonging to $\mathbb{R}$ is solution.
2. If $\frac{n_1}{n_2} = 1$ and $-2mx + 2my + x^2 - y^2 \neq 0$, there is no solution.
3. If $\frac{n_1}{n_2} \neq 1$,one obtains:

$$\tilde{\sigma} = \left| \frac{1}{2 \ln \frac{n_1}{n_2}} \sqrt{2} \sqrt{2mx \ln \frac{n_1}{n_2} - 2my \ln \frac{n_1}{n_2} - x^2 \ln \frac{n_1}{n_2} + y^2 \ln \frac{n_1}{n_2}} \right|$$

If $\frac{n_1}{n_2} = \frac{f(x)}{f(y)} + \varepsilon$, one obtains $\lim_{\varepsilon \to 0} \tilde{\sigma}(\varepsilon) = \sigma$.

5.3 Remarks

1. As shown in the examples above, if there is a table of frequencies based on two observations and one estimates only one parameter, then with dv one easily obtains optimal estimates in the most general sense of the term. It is not always easy when the table is based on k observations $y_1, ..., y_k$ with $k \geq 3$. If the table is thus formed and that we cannot determine a total exact solution one proposes to take the various couples of possible observations in $\{y_1, ..., y_k\}$ and to determine the exact solution each time when it is possible and approached otherwise. Each estimation is weighted by the sum of the frequencies of the elements of the couple and we calculate their mean. For example in the case of the first example if there are exact solutions for the various couples we take $\tilde{\lambda} = \left(1 / \sum_{i,j=1, i \neq j}^{k} (n_i + n_j) \right) \sum_{i,j=1, i \neq j}^{k} (n_i + n_j) \frac{(\ln(n_i) - \ln(n_j))}{y_j - y_i}$. We notice that here for each couple the estimation converges towards the true value when the differences between the ratios of the empirical relative frequencies and corresponding theoretical ones tend towards 0, then it is the same for the latter.

2. In the first example we have obtained the same solution with dv and the method of maximum likelihood of the new approach. It is not an isolated case. We noted in various examples considered in this document, when we estimate only one parameter, they always give concordant results.

6 Numerical Example

Even in the discrete case the two approaches are different since, contrary to the traditional one, with the new we do not distinguish truncated samples from those not truncated. In traditional approach of truncated samples all parts of the support of the estimated distribution which are supposed to be observed are used in calculations through the conditional theoretical distribution. With the new one we use only the observations. Now, if we consider the samples which do not cover all the support of the distribution from which they emanated are truncated, the truncations being the parts which do not appear in the observations and we apply the traditional approach, we fall in the new one. For this reason we do not insist on the discrete case, we give only examples concerning the continuous case. It is not easy to present a comparative study of the numerical results of the two approaches, since to the same estimate of the new it corresponds two estimates of the traditional according to whether it is considered that the sample is truncated or not. In addition, in the traditional approach when the sample is truncated the nature of truncation is used in calculations. Then the frequency

table, without indication of the parts observed, is not enough. It is necessary at each time to indicate the intervals represented by the observations in the table. For all these reasons we present the estimates of the two approaches only when that makes better to underline the specificity of the new one. For example, we simulated synthetic data of size 400 from the standard normal distribution and we grouped them into 11 classes represented by the observations $y_1, ..., y_{11}$ and their frequencies. We obtain $y_3 = -1.5331$, $y_6 = 0.0386$ and $y_8 = 1.0863$ with their respective absolutes frequencies $n_3 = 23$, $n_6 = 89$ and $n_8 = 43$. In the table presented hereafter, in the part before the line of n_8 we consider the two observations y_3 and y_6. The distance dv in these two points between the empirical distribution and the standard normal distribution is null as one takes $n_3 = 27500$ and $n_6 = 89000$. We fix then $n_6 = 89000$ and give ascending values for n_3, more and more near to 27500 as indicated in the table and we estimate m when σ is known and σ when m is known. At each time we estimate them with the method of minimal distance with dv, the method of moments of the new approach and the method of maximum likelihood of the classical approach. We note estimates obtained with dv and with maximum likelihood of the new approach respectively by $\widetilde{m}$ and $\hat{m}_{Mnew}$ for average and $\tilde{\sigma}$ and $\hat{\sigma}_{Mnew}$ for the standard deviation and we note $\hat{m}_{CLH}$ and $\hat{\sigma}_{CLH}$ those obtained with the classical maximum likelihood procedure for truncated samples. For this last, the observed part is assumed to be $[-1.7951, -1.2712[\cup [-0.22335, 0.30055[$.

$y_3 = -\mathbf{1.5331}$, $y_6 = \mathbf{0.038690}$, $y_8 = \mathbf{1.0863}$, $n_6 = \mathbf{89000}$					
$\mathbf{n_3}$	**23000**	**24000**	**26000**	**27000**	**27500**
$\tilde{m}$	0.113 69	0.08661	0.03568	0.01167	−0.000001
$\hat{m}_{Mnew}$	0.11369	0.08661	0.03568	0.01167	−0.000001
$\hat{m}_{CLH}$	**0.110 75**	**0.08444**	**0.03478**	**0.01128**	**0.000155**
$\tilde{\sigma}$	0.931 64	0.946 64	0.976 94	0.992 28	1.0
$\hat{\sigma}_{Mnew}$	0.931 64	0.94664	0.976 94	0.99228	1.0
$\hat{\sigma}_{CLH}$	**0.92171**	**0.93701**	**0.967796**	**0.98335**	**0.991165**
$\mathbf{n_8}$	**43000**	**44444**	**47273**	**48214**	**49371**
$\tilde{m}$	−0.02224	−0.017549	−0.00785	−0.00762	0.000002
$\hat{m}_{Mnew}$	0.03676 3	0.05190 7	0.088443	0.102 94	0.0000005
$\tilde{\sigma}$	0.917 67	0.935 46	0.97180	0.987 16	1.0
$\hat{\sigma}_{Mnew}$	1.068 9	1.108 0	1. 196 8	1.242	1

In the part after the line of n_8 we estimate simultaneously m and σ by the method of the minimal distance with dv and the method of moments of the new approach starting from the observations y_3, y_6 and y_8 by fixing the frequency of $n_8 = 89000$ and while taking for n_3 and n_6, the frequencies indicated. Then we observe what occurs when we make tending the differences between the ratios of the empirical frequencies and the corresponding theoretical frequencies towards 0. It is noticed that in the various examples considered, when we estimate only one parameter, the various methods of the new approach agree completely. But it is not the case when one estimates simultaneously two parameters. In the table above, when we estimate simultaneously m and σ with the method of the

moments of the new approach or the method of minimal distance with dv, when the ratios of the empirical frequencies coincide exactly with the corresponding theoretical ones we obtain their exact values. But with the method of moments, as we can see, the difference between the estimated parameters and their true values does not decrease necessarily when the difference between these ratios decreases as with the method of the minimal distance with dv. It seems that this property is specific to the estimation with dv. Here, in the various estimates with dv, at each time, the distance within the meaning of dv between the empirical distribution considered and the one to which it leads is null. Consequently the estimates with dv in that table are optimal in the most general meaning.

7 Comparison of the two approaches

A more thorough study is needed to compare the two approaches of estimation than only one section. But, by putting ourselves in the viewpoint of users of statistics, we can try to characterize what is achieved with the new approach at various levels.

7.1 Procedures

We place at disposal of statisticians all the usual methods of estimation and a new one. The remarkable fact with the new approach is that it occurs as if all is discrete except the need for grouping observations into classes in the continuous case. moreover, when it is necessary to consider fixed type-I censoring nothing change in computations. With this unification of several methods of estimation we obtain a considerable lightening of procedures compared to the traditional approach.

7.2 Computations

With the new approach, since all is discrete, there is no more the usual difficulties related to the integral calculus. With the method of maximum likelihood of the traditional approach or the new one, sometimes we encounter great difficulties when one must estimate several parameters simultaneously. But with the method of the minimal distance with d_v one can always easily propose an acceptable solution.

7.3 Credibility of estimates.

The statistician can now estimate with various methods, those of the traditional approach and of the new. If he obtains two different appreciable results it must decide for one of them. Usually we do not decide in this way since in the traditional approach we do not have criteria which give guarantees on a given specific evaluation. We have only criteria which give guarantees on average or asymptotically or by confidence interval. In this spirit, to make admitting

the new approach we should prove that it makes possible to obtain estimations better relatively to these criteria compared to those usually obtained. If one places itself in this spirit then, it is useless to continue because, for example, one cannot find better than the empirical average to estimate the average of the normal law. Of course nothing prevents us from also looking at the usual criteria in the new approach but there are new elements. One can henceforth in certain cases, without determining the estimator, affirming with certainty that the point estimation obtained with the new method is better than that obtained with maximum likelihood procedure. In other cases one can give estimators and without studying their properties one can affirm that one cannot improve them. Indeed, when the distance, within the meaning of d_v, between a given empirical distribution and the theoretical one which best fits is null, the estimate obtained is optimal in the general sense. It is noticed that when the distance within the meaning of d_v between a given empirical distribution and the one we obtain by the method of the minimal distance with d_v is not null, the solution obtained is regarded as optimal only within the meaning of the d_v. In this case perhaps it is optimal in the most general sense what must then be specified. This question remains to be studied.

8 Conclusion

We introduced a new distance and we proposed an new approach of the estimation.

1. The New distance.

We introduced a new distance and we used it in parameter estimation where we noticed what follows.

a. One can estimate even when the family of candidate theoretical distributions is not homogeneous and there is always a solution which will be acceptable in general.

b. Given a discrete empirical distribution associated to a sample belonging to a theoretical one,

- If the ratios of frequencies of the first coincide with those of the second we found exactly the latter.

- If the ratios of the frequencies of the first coincide with those of the theoretical one which best fits, then the estimations obtained are optimal in the sense that one cannot improve them.

- We showed on some examples that if we make tending the ratios of the frequencies of the first towards the corresponding theoretical ones of the second, then the estimations tend towards the true parameters. This implies immediately the convergence of the estimators. We showed the convergence in probability of the estimator for a broad class of usual laws.

c. We introduced a quality criterion, when it holds, it is stronger than of checking all the usual criteria together and we showed on some examples that in certain cases we can determine easily estimations which check it.

In addition we note a certain flexibility in calculations with *dv* compared to the method of the maximum likelihood.

2. The New approach.

We proposed an new approach of parameter estimation. When it is applied it works as if all is discrete except the need for grouping the observations in bins in continuous case. Since all is discrete there is no more the usual difficulties related to integral calculus. moreover, when it is necessary to consider fixed type-I censoring nothing is changed in computations. This unification of several methods of estimation leads to a lightening of the procedures compared to the traditional approach.

References

[1] Amemiya, T. (1985). *Advanced Econometrics.* Cambridge: Harvard University Press.

[2] Birgé, L. and Rozenholc,Y. (2006) How many bins should be put in a regular histogram. *ESAIM: Probability and Statistics*, Vol. 10, p. 24-45.

[3] Joshi V.M (1989). A counter-example against the likelihood principle: JRSS B,51, 215-216.