HAL open science

A New Approach of Point Estimation from Truncated or Grouped and Censored Data

Ahmed Guellil, Tewfik Kernane

To cite this version:

Ahmed Guellil, Tewfik Kernane. A New Approach of Point Estimation from Truncated or Grouped and Censored Data. 2008. hal-00256257v2

HAL Id: hal-00256257

https://hal.science/hal-00256257v2

Preprint submitted on 28 Dec 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Approach of Point Estimation from Truncated or Grouped and Censored Data

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 Khalid University
Abha, Kingdom of Saudi Arabia
e-mail: guellilamed@yahoo.fr, tkernane@gmail.com

Abstract

We propose a new approach for estimating the parameters of a probability distribution. It consists on combining two new methods of estimation. The first is based on the definition of a new distance measuring the difference between variations of two distributions on a finite number of points from their support and on using this measure for estimation purposes by the method of minimum distance. For the second method, given an empirical discrete distribution, we build up an auxiliary discrete theoretical distribution having the same support of the first and depending on the same parameters of the parent distribution of the data from which the empirical distribution emanated. We estimate then the parameters from the empirical distribution by the usual statistical methods. In practice, we propose to compute the two estimations, the second based on maximum likelihood principle of known theoretical properties, and the first being as a control of the effectiveness of the obtained estimation, and for which we prove the convergence in probability, so we have also a criterion on the quality of the information contained in the observations. We apply the approach to truncated or grouped and censored data situations to give the flavour on the effectiveness of the approach. We give also some interesting perspectives of the approach including model selection from truncated data, estimation of the initial trial value in the celebrate *EM* algorithm in the case of truncation and merged normal populations, a test of goodness of fit based on the new distance, quality of estimations and data.

Key words and phrases: EM algorithm, Minimum distance, Model selection from truncated data, Point estimation, Truncated data, Grouped and censored data.

1 Introduction

Point estimation is the most popular forms of statistical inference (see Lehmann and Casella [10]). We introduce in this paper a new statistical point estimation approach which found be useful in special practical situations such as truncated and grouped and censored data. The data are said to be truncated when measuring devices fail to report observations below and/or above certain readings. For example, truncated data frequently arise in the statistical analysis of astronomical observations (see Efron and Petrosian [6]) and in medical data (see Klein and Zhang [9]), and if the truncation is ignored this can cause considerable bias in the estimation. There exists in the literature many approaches of estimation from "incomplete data" such as maximum likelihood based approach of the EM algorithm (Hartley [7], Dempster et *al* [5]), or nonparametric methods such as Kaplan-Meier (Kaplan and Meier [8]) or Lynden-Bell estimators (Lynden-Bell [11]). The purpose of the present paper is to investigate another approach which consists on combining two new methods of estimation and to apply it in the fixed type I censored or grouped and censored data situations.

In the first method, we remark that in estimation problems we deal in general with three functions: a theoretical probability law $f(\cdot, \theta)$ of a random variable X, depending on a parameter θ (real or vector valued), an empirical distribution $\widehat{f}$ constructed from a sample of observations drawn from the random variable X, and an estimation $\widetilde{f}$ (from an estimation $\widetilde{\theta}$ of θ) obtained through the empirical law $\widehat{f}$. The empirical distribution $\widehat{f}$ is considered as a representative distribution of f, but in practice it is reduced to only few of its characteristics such as the mean and variance. The variational aspect of $\widehat{f}$ is often neglected while its importance. We can easily find, for instance, two distributions having the same support, mean and variance while their variations differ significantly, or conversely having the same variations but their supports and characteristic parameters are different. But two probability distributions with same support and same variations in each subset of the support are necessarily the same. We introduce then a new distance which measures the difference between variations of two distributions on a finite number of points and to use it for estimation purposes by the method of minimum distance. Since the new measure is not equivalent to classical ones it will give new insights that could not be investigated by classical distances.

In the second method, we remark that the empirical distribution arising from a sample of observations can be viewed in fact as a conditional distribution as it is built from the knowledge of the data. It will be then an estimation of the theoretical conditional distribution with respect to the observations before being an estimation for the parent distribution. This theoretical conditional distribution is represented by the auxiliary distribution introduced in this paper. To determine this distribution in discrete case, we have simply to take the conditional distribution with respect to the observed values and we proceed analogously for the continuous case. It should be noted that in discrete case it is known as the truncated distribution which is the conditional distribution given a truncation (see for example Shaw [13]) but it is presented here in a general framework. We have to deal with two discrete probability distributions having the same finite support, a theoretical distribution and its empirical representation with respect to the observations. The parameters of the

former are those of the parent distribution and the aim is to estimate them from the first instead of the parent one as commonly used. We use classical tools such as the method of moments or maximum likelihood principle. The setting that seems to us most suitable for illustrating our approach is the one of truncated or grouped and censored data. In usual practical problems, truncation can be on left or right or in either situations, and the "cut off" can be deterministic or random. In our approach, the truncation may be on any part of the range of the distribution so that the setting is more general. Also, classical approaches for truncated data are in general custom-made depending on specific problems and distributions, or subjective based methods. Instead, our approach is quite general and might be used in any situation where the underlying complete data come from a known family of distributions. We confine ourselves as a first presentation to fixed type I and grouped and censored data.

In the subsequent section, we propose a variational distance between probability distributions. In Section 3, we define a truncation of data and associated empirical and theoretical distributions and we use two different methods for estimation from truncation, a first method using minimum of the new distance introduced in this paper and a second method based on traditional tools of estimations such as the method of maximum likelihood. In Section 4, we present the new approach and we illustrate the procedure by three examples: a binomial probability law, a normal distribution and a Gamma density function. We present also a basic feature of the new approach which prove the accuracy of the method and some illustrative examples. In Section 5, we give some elements of comparison with the classical approach of estimation. In Section 6, we list some perspectives of the new approach: model selection from truncated data using the new distance, estimation of the first trial value in the celebrate *EM* algorithm for incomplete data in the case of truncation and merged normal distributions, a goodness of fit test based on the new distance, decision making about the quality of estimations and data. Finally, concluding remarks are made some pointing to other possible extensions and applications.

2 A New Distance Between Probability Distributions

As is usual, given a sample of n independent and identically distributed observations, $(x_1, ..., x_n)$, drawn from an unknown discrete random variable X falling in a discrete family of probability laws $\mathcal{P} = \{f(\cdot, \theta), \theta \in \mathbb{R}^r\}$ depending on a parameter θ (real or vector valued), i.e., $f(x, \theta) = P(X = x)$, one can summarize the sample into k couples $(y_1, \widehat{f}_1), ..., (y_k, \widehat{f}_k)$, $k \leq n$, where the y_i are the different values taken by the sample and $\widehat{f}$ is the empirical law $\widehat{f}_j = n_j/n$, where n_j represents the absolute frequency of the value y_j, $j = 1, ..., k$.

Usually, it is hoped that $\widehat{f}_j \approx f(y_j, \theta)$, in a certain probabilistic sense. But if the empirical distribution arises from truncated data, we do not hope in general having $\widehat{f}(x) \approx f(x, \theta)$, for the values x in the support of $\widehat{f}$, since the complete sample size n is usually not reported. However, we expect reasonably to have approximately

$$\frac{\widehat{f}(x)}{\widehat{f}(y)} \approx \frac{f(x, \theta)}{f(y, \theta)}, \tag{1}$$

for any points in its support, only if the sample has serious irregularities.

Introduce the following *distance of proportional variations* between $f(\cdot,\theta)$ and $\widehat{f}$

$$d_v(\widehat{f}, f(\cdot,\theta)) = \sum_{i,j\in\{1,\ldots,k\}} \left| \frac{\widehat{f_i}}{\widehat{f_j}} - \frac{f(y_i,\theta)}{f(y_j,\theta)} \right|. \tag{2}$$

It turns out that this new distance, as we will show, measures the variations between probability distributions.

In continuous case also, any sample $x_1, ..., x_n$ is summarized into k couples $(y_1, \widehat{f_1}), ...$ $, (y_k, \widehat{f_k})$, $k \leq n$. This can be done uniquely, by grouping for example the sample in classes where the y_i are the mid-classes (or class means) and $\widehat{f_i} = \widehat{f}(y_i)$ where $\widehat{f}$ is an empirical density estimator, or the data is presented in a grouped and censored form. The proportional variational distance d_v in this case, between the density $f(x,\theta)$ of X and its empirical law $\widehat{f}$, is thus defined as (2). One of its main powerful feature is that when using traditional distances we have to use the sample size n through the expression of $\widehat{f_i} = n_i/(nh_n)$, where h_n is the size of class intervals; but sometimes, as for truncated data situations where measuring devices fail to report even the number of sample points in certain ranges, then the real size n is not known, but a truncated sample size n_t is instead used. Using the ratios $\widehat{f_i}/\widehat{f_j}$ will clear up the effect of the truncated sample size which can lead to considerable bias in the estimation.

Note that d_v possesses the properties of symmetry and triangle inequality. But in the identity property $d_v(f,g)(x,y) = 0 \Longleftrightarrow f \equiv g$, the equality between f and g must be understood in the sense that f and g have the same variations on the points x and y. It should be stressed that this new measure is not equivalent to classical ones and should then give new insights and information about other characteristics and features of probability distributions.

From now on f shall represent a theoretical probability law in both discrete or continuous cases and $\widehat{f}$ shall represent the corresponding empirical law in both cases. Denote by $\Omega = \{x \in \mathbb{R}, f(x,\theta) > 0\}$ the set of *atoms* of f or *support*. Let $\mathcal{F}$ be the $\sigma-$algebra generated by sets $A = B \cap \omega$ where the ω are the Borel sets of $\mathbb{R}$ and $B \subset \Omega$. For all $A \in \mathcal{F}$, we have $P(A) = \int_A f(x,\theta)\mu(dx)$,where μ is the Lebesgue measure on $\mathbb{R}$. In discrete case, we have $P(A) = \sum_{x\in A} f(x,\theta)$.

For all $i \geq 1$, we set $\Omega_i = \Omega$, $\mathcal{F}_i = \mathcal{F}$ and $P_i = P$. Let $\Omega^n = \Omega_1 \times ... \times \Omega_n$, $\mathcal{F}^{(n)} = \mathcal{F}_1 \otimes ... \otimes \mathcal{F}_n$ and $P^{(n)} = P_1 \otimes ... \otimes P_n$. The probability space $\left(\Omega^n, \mathcal{F}^{(n)}, P^{(n)}\right)$ represents the space of samples of size n from the random variable X. We omit the subscript n in $\left(\Omega^n, \mathcal{F}^{(n)}, P^{(n)}\right)$ for notational convenience and shall denote the sample space as $(\Omega, \mathcal{F}, P)$.

2.1 A Notion of Variation between probability distributions

We will discuss now the measure theoretic aspect of the new distance introduced above. Let P and Q two probability measures defined on the same measurable space $(\Omega, \mathcal{F})$, f and g their respective probability densities, not necessarily with respect to the same measure and E an event of this space. We say that f and g have the same variation on E, if the respective restrictions of f and g on E, define the same probability measure on E endowed with the sigma algebra traces of $\mathcal{F}$ on E.

Definition 1 *Let f and g two probability distributions positive and defined on a part E not reduced to only one element. If in any point (x, y) of $E \times E$, we have:*

$$\frac{f(x)}{f(y)} = \frac{g(x)}{g(y)} \tag{3}$$

then we say that f and g have same variations 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.*

Definition 3 *Let f and g two probability distributions and E an event on which they are strictly positive. If E is discrete and not reduced to only one element, and one of the distributions f and g being discrete and the other may not be discrete, 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:

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

Let be given a classical distance d between two functions f and g which associates for points x and y from the intersection of their domain of definitions, the quantity $d\,(f, g)\,(x, y) = |f(x) - g(x)| + |f(y) - g(y)|\,.$

Proposition 4 *We have the following properties for the distance d_v :*
1. *$d(f, g)(x, y) = 0 \Longrightarrow d_v(f, g)(x, y) = 0$, the converse is not always true.*
2. *Let $\widehat{f}$ be a kernel density estimation. Then $\lim_{n\to\infty} d_v(\widehat{f}, f) = 0$ in probability.*
3. *Let f and g be two functions defined on $\mathbb{R}$ and $E \subset \mathbb{R}$ satisfying:*

$$\forall\,(x, y) \in E \times E,\ d_v(f, g)(x, y) = 0.$$

If

$$\int_{\mathbb{R}} f\ d\mu = \int_{\mathbb{R}} g\ d\mu = 1,$$

where μ is the Lebesgue measure on $\mathbb{R}$, then

$$\mu\left(\overline{E}\right) = 0 \implies f = g \quad \mu - \textit{almost surely on } \mathbb{R}.$$

Proof. 1. Follows directly from the definitions of d and d_v.
2. Follows from the fact $\lim_{n\to\infty} d(\widehat{f}, f) = 0$ in probability (see Parzen [12]), then $\lim_{n\to\infty} d_v(\widehat{f}, f) = 0$ in the same probabilistic notion of convergence.
3. Fix $y_0 \in E$, we have $f(x)/f(y_0) = g(x)/g(y_0)$ for all $x \in E$. This implies that

$$\int_E f(x)dx = 1 \iff \int_E f(y_0)\frac{g(x)}{g(y_0)}dx = \frac{f(y_0)}{g(y_0)}\int_E g(x)dx = 1.$$

We deduce that $f(y_0) = g(y_0)$, and the result follows. ■

3 Truncated Data

The truncated data specification, or generally *incomplete data,* implies the existence of two sample spaces $\mathcal{X}_o$ and $\mathcal{X}_t$, such that the complete sample space is given by $\Omega = \mathcal{X}_o \cup \mathcal{X}_t$. The observed data $\mathbf{x}_o = (x_1, ..., x_{n_t})$, where n_t is the truncated sample size, are a realization from $\mathcal{X}_o$ and the unobserved data $\mathbf{z} = \left(x_1^*, ..., x_{n-n_t}^*\right)$, where n is the complete unknown sample size, are from $\mathcal{X}_t$. The complete data $\mathbf{x} = \mathbf{x}_o \cup \mathbf{z}$ is known only through the observed data $\mathbf{x}_o$ (see Dempster, Laird and Rubin [5] for further explanations about incomplete data specification).

Consider a sample of observations $x_1, ..., x_n$ drawn from a theoretical probability law $f(\cdot, \theta)$, depending on a parameter $\theta \in \mathbb{R}^r$. As usual, the data are summarized, in discrete or continuous cases (as shown in Section 2), into k couples $(y_1, \widehat{f}_1), ..., (y_k, \widehat{f}_k)$, $k \leq n$, and let $\triangle = \{u_1, ..., u_m\}$ a part from the set $\{y_1, ..., y_k\}$, $m \leq k$, which we will call *truncation.* The observed data is summarized by a truncation $\triangle_o = \{u_1, ..., u_m\}$ and an empirical estimation $\widehat{f}_o$ and assume that the unobserved data is also summarized by a set $\triangle_t = \{u_1^*, ..., u_p^*\}$ and $\widehat{f}_t$.

The structure of the new distance d_v allows the following decomposition property:

$$d_v(\widehat{f}, f(\cdot, \theta)) = d_v(\widehat{f}_o, f(\cdot, \theta)) + d_v(\widehat{f}_t, f(\cdot, \theta)) + \sum_{\substack{u_i \in \triangle_o \\ u_j^* \in \triangle_t}} \left| \frac{\widehat{f}_o(u_i)}{\widehat{f}_t(u_j^*)} - \frac{f(u_i, \theta)}{f(u_j^*, \theta)} \right| + \sum_{\substack{u_i \in \triangle_o \\ u_j^* \in \triangle_t}} \left| \frac{\widehat{f}_t(u_j^*)}{\widehat{f}_o(u_i)} - \frac{f(u_j^*, \theta)}{f(u_i, \theta)} \right|. \tag{4}$$

The following proposition is typical for the new distance and is useful for using the minimum of distance d_v.

Proposition 5 *Let be given a truncated data $\triangle_o$ with corresponding empirical estimation $\widehat{f}_o$. Then $\lim_{n_t \to \infty} d_v(\widehat{f}_o, f) = 0$ in probability.*

Proof. We have from Proposition 1 that $\lim_{n \to \infty} d_v(\widehat{f}, f) = 0$ in probability. Then, from the decomposition property (4) we obtain $\lim_{n \to \infty} d_v(\widehat{f}_o, f) = \lim_{n_t \to \infty} d_v(\widehat{f}_o, f) = 0$ in probability. ■

3.1 An Auxiliary Distribution

Define the empirical distribution $\widetilde{f}$ corresponding to a given truncation $\triangle$ by:

$$\widetilde{f}(x) = \begin{cases} \widetilde{f}_i & \text{if} \quad x = u_i, \quad i = 1, ..., m, \\ 0 & \text{otherwise,} \end{cases}$$

where the $\widetilde{f}_i$ satisfy the following set of proportional allocation equations $\widetilde{f}_i / \widetilde{f}_j = \widehat{f}_i / \widehat{f}_j$, for $i, j = 1, ..., m$ and $\widetilde{f}_1 + ... + \widetilde{f}_m = 1$.

Define the following auxiliary distribution from $f(\cdot, \theta)$, which is akin to the proportional allocation procedure for missing values (see Hartley [7]).

$$h(x, \theta) = \begin{cases} \dfrac{f(x, \theta)}{f(u_1, \theta) + f(u_2, \theta) + ... + f(u_m, \theta)} & \text{if} \quad x = u_i, \quad i = 1, ..., m, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Remark 6 *If the truncation is random, that is, there exists a random variable T such that we observe, for example, the random variable X only if $X > T$ or $X < T$, then the probability law used in (5) is replaced by the conditional law of X with respect to $\{X > T\}$ or $\{X < T\}$ respectively.*

The auxiliary distribution h was found be useful for estimation problems in truncated data. Indeed, it is well known in classical estimation from truncated data (see Hartley [7]) that missing values could be recovered by "proportional allocation" procedures, then the auxiliary distribution h, which is already based on proportional allocation, will be an intuitive and natural tool for estimation purposes from truncated data. The function h is a theoretical probability distribution depending on the same parameters of those of f. It has also the same *support* as that of $\widetilde{f}$.

Definition 7 *We call $\widetilde{f}$ and $h(\cdot, \theta)$ the empirical and theoretical distributions of a given truncation $\triangle = \{u_1, ..., u_m\}$ from a sample of observations $(x_1, ..., x_n)$.*

4 The Approach of Estimation

We will use mainly two methods of estimation. The first method is a minimum distance estimation using the metric d_v between the empirical and theoretical distributions $\widehat{f}$ and $f(\cdot, \theta)$. The second is similar to traditional ones such as the method of substitution or maximum likelihood principle, by considering $\widetilde{f}$ as an empirical estimation of $h(\cdot, \theta)$. The first is based on variational difference between distributions and the second in the sense of an euclidean difference and hence they treat different aspects of the sample of observations. If for a given data they give different estimations, we cannot suspect the approaches but we can say that the data do not restore in a coherent way all aspects of the probability distribution from which it emanated. If on the other hand they give significantly the same estimations we can assert that the estimation is credible since through different aspects it has given the same distribution. That is the distribution which fits the best the empirical distribution. Practically, we propose to calculate the estimations by the two methods and take the second one since based on maximum likelihood principle of good known theoretical properties. We use then the first as a tool of decision on whether the estimation is credible or not. The estimation will then be considered as credible in cases where the two methods give approximately the same estimation.

4.1 Convergence in Probability of the Minimum Distance Estimator

Let $X_1, X_2, ..., X_n$ a sample with $X_i \sim f(x, \theta)$, $\theta = (\theta_1, ..., \theta_s)^t \in \Theta \subseteq \mathbb{R}^s$, with

$$f(x, \theta) = K(x) \times \exp\left\{\sum_{k=1}^{s} \theta_k T_k(x) + A(\theta)\right\}, \tag{6}$$

$x \in \mathcal{X} \subseteq \mathbb{R}$, 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 (6) is very rich, 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 8 *Let $X_1, X_2, ..., X_n$ be a sample from the family of distributions (6). If the set of natural parameters Θ is covex 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 (6) 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

$$\delta_{ij}(\lambda\theta^{(1)} + \mu\theta^{(2)}) = \left| C_{ij} \exp\left\{ \sum_{k=1}^{s} \left[\lambda\theta_k^{(1)} + \mu\theta_k^{(2)}\right] (T_k(y_i) - T_k(y_j)) \right\} - A_{ij} \right| \tag{7}$$

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

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

then

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

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

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

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

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. ■

4.2 A Maximum Likelihood Principle with the Auxiliary Distribution

We firstly begin in a general situation, that of the one-parameter exponential family, to show how to use the procedure explained below in the case of the new method. Consider the one-parameter exponential family with density

$$f(x,\theta)=K(x)\times\exp[\theta T(x)-A(\theta)], \tag{8}$$

where θ is the parameter, T a statistic, $K(x)$ a function of x and A is a function of the parameter θ. Let us use the maximum likelihood principle. Consider a sample of observations $x_1,...,x_n$ from which we derive the support $\triangle=\{y_1,...,y_k\}$. We then construct the auxiliary distribution from the support $\triangle$, expressed in the following form

$$h(x,\theta)=\frac{K(x)\times\exp[\theta T(x)-A(\theta)]}{\sum_{i=1}^{k}K(y_i)\times\exp[\theta T(y_i)-A(\theta)]}. \tag{9}$$

We have to maximize the likelihood function given in our case by

$$\mathbf{L}_h\left(y,\theta\right)=\prod_{i=1}^{k}h(y_i,\theta). \tag{10}$$

Without loss of generality, we assume that the class intervals are the same. Then, we have

$$\log\mathbf{L}_h\left(y,\theta\right)=\sum_{i=1}^{k}\log h(y_i,\theta)=\sum_{i=1}^{k}\frac{n_i}{n}\log\left[\frac{K(y_i)\times\exp[\theta T(y_i)-A(\theta)]}{\sum_{i=1}^{k}K(y_i)\times\exp[\theta T(y_i)-A(\theta)]}\right], \tag{11}$$

taking the derivative and solving the score equation on θ we obtain an estimator of the parameter θ satisfying the relation

$$\sum_{i=1}^{k} \frac{n_i}{n} \frac{\sum_{i=1}^{k} T(y_i) \times f(y_i, \theta)}{\sum_{i=1}^{k} f(y_i, \theta)} = \sum_{i=1}^{k} \frac{n_i}{n} T(y_i). \tag{12}$$

The later result may be obtained directly by the method of moments, but we have presented the maximum likelihood method since it is widely used in statistical inference.

In order to test the performance of the proposed approach, we use synthetic data sets which were generated by simulation from three examples of probability law: binomial law, normal density and a Gamma distribution. The examples were selected from various simulation studies from different family of probability distributions and the two methods have shown their effectiveness and never deviate significantly from the true parameter. The reason for using synthetic data sets is that the true parameters for the synthetic datasets are known and the accuracy of results obtained by using the two new methods can be compared.

4.3 Examples

Binomial distribution. We generated a synthetic data set of size 500 from a binomial law $\mathcal{B}(n, p)$ with $n = 10$ and $p = 0.3$, and denote by $f(y; p) = C_n^y p^y (1-p)^{n-y}$ its probability mass function. The data are summarized in the following table.

Table 1.

y_i	0	1	2	3	4	5	6	7
n_i	15	71	108	134	97	47	23	5

Our aim is to estimate the parameter p, with the knowledge of $n = 10$, from different truncation of data.

For illustrating the two methods, consider the truncation $\triangle = \{2, 3, 4, 5\}$ with truncated sample size $n_t = 386$. We have then a truncation proportion of $Q = 100(n - n_t)/n = 22,8$ % in data. For the first method, we have to search the value of the parameter p which minimizes the distance d_v, that is:

$$\min_p d_v(\widehat{f}, f) = \min_p \sum_{\substack{i,j \in \triangle \\ i \neq j}} \left| \frac{f(y_i; p)}{f(y_j; p)} - \frac{n_i}{n_j} \right|,$$

Using computer algebra package, we obtain the result $\widetilde{p}_1 = 0.299$.

For the second method, the empirical distribution $\widetilde{f}$ given the truncation $\Delta = \{2, 3, 4, 5\}$ is given by $\widetilde{f}(2) = 108/386$, $\widetilde{f}(3) = 134/386$, $\widetilde{f}(4) = 97/386$, $\widetilde{f}(5) = 47/386$ and $\widetilde{f}(x) = 0$ if $x \notin \Delta$.

The auxiliary distribution $h(\cdot, p)$ is given by:

$$h(x, p) = \begin{cases} \dfrac{f(x, p)}{f(2, p) + f(3, p) + f(4, p) + f(5, p)} & \text{if } x = u_i, \quad u_i \in \{2, 3, 4, 5\} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

By the method of substitution, the estimation of p is obtained by solving the equation:

$$\sum_{u_i \in \{2,3,4,5\}} u_i \times h(u_i, p) = \sum_{u_i \in \{2,3,4,5\}} u_i \times \widetilde{f}(u_i) \tag{14}$$

Using a computer algebra package we obtain the result $\widetilde{p}_2 = 0.3$.

In the following table we present the estimations $\widetilde{p}_1$ from the first method using minimum distance approach using the distance d_v, and $\widetilde{p}_2$ from the auxiliary distribution, of the parameter p, for known n, according to the truncation $\triangle = \{u_1, ..., u_m\}$ considered.

Table 2. The estimations $\widetilde{p}_1$ and $\widetilde{p}_2$ by the new approach of the parameter p of the binomial probability law $\mathcal{B}(n,p)$ with $p = 0.3$ and known $n = 10$.

n°	$\triangle$	Truncated sample size n_t	Proportion of truncation Q (%)	$\widetilde{p}_1$	$\widetilde{p}_2$
1	$\{0,1,2,3,4,5,6,7\}$	500	0	0.305	0.298
2	$\{0,1,2,3,4,5\}$	472	5.6	0.295	0.293
3	$\{1,2,3,4,5\}$	457	8.6	0.288	0.292
4	$\{0,1,2,3,4\}$	425	15	0.295	0.293
5	$\{1,2,3,4\}$	410	18	0.287	0.292
6	$\{0,2,3,4,5\}$	401	19.8	0.295	0.298
7	$\{2,3,4,5\}$	386	22.8	0.299	0.3
8	$\{0,1,3,4,5\}$	364	27.2	0.295	0.289
9	$\{0,2,3,4\}$	354	29.2	0.295	0.301
10	$\{1,3,4,5\}$	349	30.2	0.287	0.287
11	$\{2,3,4\}$	339	32.2	0.305	0.305
12	$\{0,3,4,5\}$	293	41.4	0.295	0.293
13	$\{2,4,5,6,7\}$	280	44	0.308	0.307
14	$\{0,1,2,5,6,7\}$	269	46.2	0.298	0.299
15	$\{0,1,4,5,6,7\}$	258	48.4	0.3013	0.295
16	$\{0,4,5,6,7\}$	187	62.6	0.3071	0.302
17	$\{0,5,6,7\}$	90	82	0.3014	0.301
18	$\{0,5\}$	62	87.6	0.2937	0.294

As previously said, the two estimations by the new approach, $\widetilde{p}_1$ and $\widetilde{p}_2$, are accurate in all cases and close to each other. Furthermore, the truncation proportion has no effect on the quality of estimations. The two estimations are also not sensitive to small cell probabilities as for truncations including the value $y_8 = 7$. It should be noted that the classical estimation by maximum likelihood without truncation is $\widehat{p} = 0.297$, and considering our approach we obtained the estimations $\widetilde{p}_1 = 0.3053$ for the first method and $\widetilde{p}_2 = 0.2978$ for the second.

Normal distribution. Consider a sample of size 400 drawn from a normal population with mean $m = 0$ and standard deviation $\sigma = 1$. Consider the data falling in 11 fixed class intervals as shown in the following table, with mid-classes u_i and absolute frequencies n_i

Table 3.

y_i	-2.581	-2.06	-1.533	-1.009	-0.485	0.039	0.563	1.086	1.610	2.134	2.658
n_i	5	8	23	48	71	89	72	43	25	10	6

The number of bins can be selected from an optimal procedure developed by Birgé and Rozenholc [2]. Let the following table where we estimate simultaneously m and σ by the minimum distance procedure with dv. We denote the estimations by $\tilde{m}_1$ and $\tilde{\sigma}_1$. In each line of the table the estimates are made starting from the table of frequencies based on the observations indicated in the first column. The truncated sample size is denoted by n_t. We have then a truncation proportion of $Q = 100(n - n_t)/n$ in data.

Table 4.

S	n_t	$Q\%$	$\widetilde{m}_1$	$\tilde{\sigma}_1$
$\{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}\}$	400	0	0.083	1.130
$\{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9\}$	384	4	0.003	1.092
$\{y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9\}$	379	5.25	0.054	0.977
$\{y_3, y_4, y_5, y_6, y_7, y_8, y_9\}$	371	7.25	0.052	0.993
$\{y_4, y_5, y_6, y_7, y_8, y_9\}$	348	13	0.043	1.017
$\{y_5, y_6, y_7, y_8, y_9\}$	300	25	0.052	1.012
$\{y_3, y_4, y_5, y_6\}$	231	42.25	0.303	1.104
$\{y_6, y_7, y_8, y_9\}$	229	42.75	-0.225	1.140
$\{y_6, y_7, y_8\}$	204	49	-0.065	1.052
$\{y_3, y_5, y_7\}$	166	58.5	0.052	0.993
$\{y_2, y_3, y_4, y_5\}$	150	62.5	-0.137	0.904
$\{y_3, y_4, y_5\}$	142	64.5	-0.151	0.893

Remark 9 *In practice, the bins are in fact chosen after obtaining the truncated sample so the results should be more efficient, but this does not affect the preceding results obtained after grouping the whole sample and truncate from the bins since the aim is to give some feel about the accuracy of the estimations. Also we can avoid grouping the observations by considering empirical frequencies obtained from kernel density estimations.*

4.3.1 Gamma probability density

Consider a sample of size 800 drawn from a Gamma distribution $G(a, b)$ with density given by

$$f(x \mid a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} \exp\left(-\frac{x}{b}\right), \quad x \geq 0, \tag{15}$$

and parameters $a = 7$ and $b = 3$. Consider the data falling in 16 fixed class intervals as shown in the following table, with mid-classes u_i and absolute frequencies n_i :

Table 5.

u_i	5.89	8.72	11.56	14.39	17.23	20.06	22.89	25.73	28.56	31.39
n_i	11	40	60	108	118	104	100	74	63	53

34.23	37.06	39.89	42.73	45.56	48.39
27	21	11	5	3	2

In the following table we show the estimations $\widetilde{b}_1$ from the minimum of distance d_v and $\widetilde{b}_2$ by the second method for the parameter b, with known $a = 10$, according to the truncation $\triangle$ considered.

Table 6. The estimations $\widetilde{b}_1$ and $\widetilde{b}_2$ by the new approach of the parameter b of the Gamma probability distribution $G(a, b)$ with $b = 3$ and known $a = 7$.

n°	$\triangle$	n_t	Q (%)	$\widetilde{b}_1$	$\widetilde{b}_2$
1	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9$ $u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}\}$	800	0	3.018	3.054
2	$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9,$ $u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}\}$	787	1.625	2.980	3.065
3	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}\}$	779	2.625	3.012	3.068
4	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$	731	8.625	2.895	3.059
5	$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$	720	10	3.063	3.075
6	$\{u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$	680	15	3.157	3.119
7	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\}$	678	15.25	2.864	3.002
8	$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9\}$	667	16.625	2.978	3.018
9	$\{u_3, u_4, u_5, u_6, u_7, u_8, u_9\}$	627	21.625	3.086	3.062
10	$\{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$	615	23.125	2.859	2.960
11	$\{u_2, u_3, u_4, u_5, u_6, u_7, u_8\}$	604	24.5	2.908	2.977
12	$\{u_4, u_5, u_6, u_7, u_8, u_9\}$	567	29.125	3.046	3.016
13	$\{u_2, u_3, u_4, u_5, u_6, u_7\}$	530	33.75	2.908	2.978
14	$\{u_2, u_3, u_4, u_5, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}\}$	443	44.625	3.018	3.080
15	$\{u_1, u_2, u_3, u_4, u_5, u_6\}$	441	44.875	2.775	2.894
16	$\{u_1, u_2, u_3, u_4, u_8, u_9, u_{10}, u_{11}, u_{15}\}$	439	45.125	2.969	3.048
17	$\{u_1, u_2, u_3, u_4, u_5, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}\}$	406	50.75	3.018	3.031
18	$\{u_1, u_2, u_3, u_4, u_5\}$	337	57.875	2.788	2.931
19	$\{u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}\}$	256	67.625	2.990	3.212
20	$\{u_{10}, u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{16}\}$	122	84.75	2.894	2.822

The estimations from the two methods are also accurate in this case of gamma distribution for the parameter b. Here also the truncation proportion does not affect the quality of estimations. When we consider the complete data, the classical estimation is $\widehat{b} = 3.04$ and the two new estimations are $\widetilde{b}_1 = 3.018$ and $\widetilde{b}_2 = 3.054$.

As it was noticed in the examples above, the two methods lead to approximately the same estimation results. Nevertheless, if the two estimations are significantly different, it seems related to the quality of the selected data. An important feature of this new approach is that the quality of estimations is uninfluenced by the truncation proportion. The following section will give further insights of the new approach.

4.4 A Basic Feature of the New Approach

The preceding results have shown the effectiveness of the new approach and worked well in simulation experiments. Furthermore, the proposition below will give an insight of a major feature of the new approach by considering the one parameter exponential family. We will prove that for all truncation considered formed by more than two points, from a sample of observations; if the ratios of the relative frequencies of the u_i are equal to the theoretical ones, then we may obtain the true value of the parameter. We may conjecture that when considering an arbitrary law of probability depending on r parameters, such that we have a truncation composed by $r+1$ points

having exact empirical ratios of the relative frequencies then we obtain the true values of the r parameters.

Proposition 10 *Consider a probability distribution f from the one-parameter exponential family with density*

$$f(x,\theta) = K(x) \times \exp[\theta T(x) - A(\theta)], \tag{16}$$

where $\theta \in \mathbb{R}$ is the parameter, T a statistic, $K(x)$ a function of x and A is a function of the parameter θ. Assume that we wish to estimate the parameter θ. If we consider a truncation having two points x and y with empirical frequencies f_1 and f_2 satisfying $f_1/f_2 = f(x,\theta)/f(y,\theta)$, then, using the approach considered here, we obtain the true value of θ.

Proof. 1. If we consider the minimum of distance d_v the result is immediate.
2. Consider now the second method to estimate m. Consider two values x and y from the exponential family with density given by (16), with $\widetilde{\theta}$ being the estimation by the new approach, and assume that their empirical frequencies f_1 and f_2 are such that

$$\frac{f_1}{f_2} = \frac{f(x,\widetilde{\theta})}{f(y,\widetilde{\theta})}.$$

We obtain

$$\overline{u} = xf_1 + yf_2 = \frac{xK(x)\exp\left(\widetilde{\theta}T(x)\right) + yK(y)\exp\left(\widetilde{\theta}T(y)\right)}{K(x)\exp\left(\widetilde{\theta}T(x)\right) + K(y)\exp\left(\widetilde{\theta}T(y)\right)}.$$

Then, we solve on θ the following equation:

$$\left(x - \frac{xK(x)\exp\left(\widetilde{\theta}T(x)\right) + yK(y)\exp\left(\widetilde{\theta}T(y)\right)}{K(x)\exp\left(\widetilde{\theta}T(x)\right) + K(y)\exp\left(\widetilde{\theta}T(y)\right)}\right) K(x)\exp\left(\theta T(x)\right)$$

$$+\left(y - \frac{xK(x)\exp\left(\widetilde{\theta}T(x)\right) + yK(y)\exp\left(\widetilde{\theta}T(y)\right)}{K(x)\exp\left(\widetilde{\theta}T(x)\right) + K(y)\exp\left(\widetilde{\theta}T(y)\right)}\right) K(y)\exp\left(\theta T(y)\right) = 0,$$

after straightforward algebra we obtain

$$(x-y)\exp\left(\widetilde{\theta}T(y) + \theta T(x)\right) + (y-x)\exp\left(\widetilde{\theta}T(x) + \theta T(y)\right) = 0,$$

yielding the true value $\widetilde{\theta} = \theta$. The proof is complete. ■

Remark 11 *Note that the frequencies f_1 and f_2 need not be exact, that is f_1 may be different from $f(x,\theta)$ and also f_2, but we require only that their ratio is equal to the theoretical one $f(x,\theta)/f(y,\theta)$.*

Examples

Binomial distribution. Consider again the binomial distribution $\mathcal{B}(n,p)$ with $n=10$ and $p=0.3$ and assume n is known and we wish to estimate p. Assume we have the following truncation with only two points $\triangle=\{0,1\}$. The exact ratio of their probability distribution is given by $f(0,p)/f(1,p)=7/30$, which is a rational value that will simplify the example. Choose the absolute frequencies of the two values considered as being $n_1=7$ and $n_2=30$ for the values $u_1=0$ and $u_2=1$ respectively, in order for having $f_1/f_2=f(x,p)/f(y,p)=7/30$. Using the first approach, that of the minimum of distance d_v, we have to solve

$$\min_p d_v(\widehat{f},f)=\min_p\left[\left|\frac{C_{10}^0(1-p)^{10}}{C_{10}^1p(1-p)^9}-\frac{7}{30}\right|+\left|\frac{C_{10}^1p(1-p)^9}{C_{10}^0(1-p)^{10}}-\frac{30}{7}\right|\right],$$

and we get the true value $\widetilde{p}_1=0.3$.

Using the second method we have to solve the following equation on p

$$\frac{0\times C_{10}^0(1-p)^{10}+1\times C_{10}^1p(1-p)^9}{C_{10}^0(1-p)^{10}+C_{10}^1p(1-p)^9}=\frac{30}{37},$$

and we obtain also the exact result $\widetilde{p}_2=0.3$.

Gamma distribution. Consider the Gamma probability distribution $G(a,b)$ with $a=10$ and $b=5$. Assume that a is known and we wish to estimate b. Consider the truncation $\triangle=\{u_3,u_8\}$ with $u_3=30.13$ and $u_8=60.02$. We have the following value of the ratio $f(u_3,b)/f(u_8,b)\approx 0.799$ (the result is an approximate result since for probability density functions it is difficult to get an exact rational value but we will show that the estimations are very close to the true value). Consider the absolute frequencies $n_3=79.93$ (or 80) and $n_8=100$ for the values $u_3=30.13$ and $u_8=60.02$ respectively. We have then $n_3/n_8\approx f(u_3,b)/f(u_8,b)$. Using the minimum of distance d_v, we have to solve

$$\min_b d_v(\widehat{f},f)=\min_b\left[\left|(79.93/100)-((30.13/60.02)^9\times\exp(-(1/b)\times(30.13-60.02)))\right|\right.$$

$$\left.+(100/79.93)-((60.02/30.13)^9\times\exp(-(1/b)\times(60.02-30.13)))\right],$$

and we get the result $\widetilde{b}_1\approx 5$.

From the second method, we compute $\overline{u}=46.7438$ and solve on b the following equation

$$(30.13-46.7438)\times 30.13^9\times\exp(-30.13/b)$$
$$+(60.02-46.7438)\times 60.02^9\times\exp(-60.02/b)=0.$$

The result is $\widetilde{b}_2\approx 5$.

Now assume that the parameters a and b are unknown and show how to jointly estimate them using the new approach. Since now there are two unknown parameters, we need to have three points from the support, so consider $u_1=34.7702$, $u_2=57.5008$ and $u_3=74.5487$ with their corresponding absolute frequencies $n_1=102, n_2=100$ and $n_3=34$. We have to find a and b which minimize the distance d_v that is $\min_{a,b} d_v(\widehat{f},f)$. The result is $\widetilde{a}\approx 10.0454$ and $\widetilde{b}\approx 4.9739$.

5 Elements of Comparison with the Classical Approach

Our aim here is not to give a detailed comparison study which needs to be investigated thoroughly, but only some elements of appreciation. A major feature which characterizes this new approach from the others is that when we have exact ratios of frequencies we obtain the true parameter and when their difference from the theoretical ratios decrease the quality of estimation increase even if we are using only a part from the sample of observations. This is not the case for classical approaches. In classical approaches, quality considerations are only viewed through mean properties of estimators or their asymptotic behaviour. By combining the two proposed methods we have in fact a point criterion. Another characteristics is that the proportion of truncation has any effect on the quality of estimations. The first method uses a well known method of minimum distance but with a new one which has an important advantage of being symmetric, the property of which many traditional distances do not have. However, the estimations are obtained in this case implicitly so it is difficult to find explicit expressions and study their properties to compare them with classical ones. Using the new distance we hope having fast convergent estimators since we expect that the influence of the errors in the frequencies will be slight in the new approach as we are using ratios of frequencies. Consider now the second method of the new approach. We use classical procedures of estimation such as the maximum likelihood principle using the auxiliary distribution. We may obtain the estimators and study their properties as commonly used and then preserves the advantages of classical methods. In classical approach, given a sample, the estimation of certain parameters such as the mean and variance do not change according to the family of parent distributions. The latter information is not used and this disadvantages the approach. However, in the new approach the estimations of the mean and variance change according to the distribution from which the data emanated.

The following two examples show the effectiveness of using the auxiliary distribution.

Example. Consider the following frequency table:

Table 7.

x_i	2	3	$Total$
n_i	n_1	n_2	n
$\widehat{f}(x_i) = f_i$	$f_1 = (n_1/n)$	$f_2 = (n_2/n)$	1

Any sample of observations that satisfies the preceding frequency table may belong from one of the following distributions:

$$g_1(x) = \begin{cases} \frac{x}{6} & if\ x \in \{1,2,3\}, \\ 0 & otherwise, \end{cases} \quad \text{or} \quad g_2(x) = \begin{cases} \frac{x-1}{6} & if\ x \in \{2,3,4\}, \\ 0 & otherwise. \end{cases}$$

The decision for determining which of the two distributions is more appropriate for table 7, depends intuitively on the values n_1 and n_2 (or f_1 and f_2). However, if we use the classical maximum likelihood, we obtain that the samples of observations were generated from distribution h_1 whatever the values of n_1 and n_2, that is:

$$\left(\frac{1}{6}\right)^{n_1} \times \left(\frac{2}{6}\right)^{n_2} < \left(\frac{2}{6}\right)^{n_1} \times \left(\frac{3}{6}\right)^{n_2}.$$

We will show by using the new approach that the decision is more relevant. Determine first the auxiliary distributions, h_1 and h_2, based on the truncation $\triangle = \{2, 3\}$, for g_1 and g_2 respectively. We obtain

$$h_1(x) = \begin{cases} 2/5 & if \quad x = 2, \\ 3/5 & if \quad x = 3, \\ 0 & otherwise, \end{cases} \quad \text{and} \quad h_2(x) = \begin{cases} 1/3 & if \quad x = 2, \\ 2/3 & if \quad x = 3, \\ 0 & otherwise. \end{cases}$$

By using the maximum likelihood for h_1 and h_2, we have to decide according to the quantities $(2/5)^{n_1} \times (3/5)^{n_2}$ and $(1/3)^{n_1} \times (2/3)^{n_2}$. Solving the following inequality

$$\left(\frac{2}{5}\right)^{n_1} \times \left(\frac{3}{5}\right)^{n_2} \leq \left(\frac{1}{3}\right)^{n_1} \times \left(\frac{2}{3}\right)^{n_2},$$

which is equivalent to $(6/5)^{\alpha} (9/10)^{1-\alpha} \leq 1$, where $\alpha = n_1/n_2$, we obtain $0 < \alpha \leq -\log(9/10)/\log(4/3) = x_0 \approx 0.36624$. If $0 < \alpha < x_0$, the data were generated from g_2 and if $x_0 < \alpha < 1$, the data were generated from g_1. We cannot make any decision about the case $\alpha = x_0$.

Example. Consider a binomial distribution with parameters $n = 4$ and p is unknown, from which we consider some samples of observations of size 15 given in table 8 by their absolute frequencies and chosen in order for having $\overline{x} = 8/15$.

Table 8.

	Values						
samples	0	1	2	3	4	$\widehat{p}$	$\widetilde{p}$
1	7	8	0	0	0	0.133	0.222
2	9	5	0	1	0	0.133	0.184
3	9	4	2	0	0	0.133	0.139
4	10	3	1	1	0	0.133	0.134
5	10	4	0	0	1	0.133	0.216
6	12	0	2	0	1	0.133	0.196
7	13	0	0	0	2	0.133	0.385

It is clear that the information given by the samples are not the same, nevertheless the classical estimation method gives us the same estimation $\widehat{p} = 8/(15 \times 4) \approx 0.133$. If we use the second method of the new approach, we have to solve the following equation for each sample:

$$0 \times h(0, p) + 1 \times h(1, p) + 2 \times h(2, p) + 3 \times h(3, p) + 4 \times h(4, p) = \overline{x},$$

where $h(x, p)$ is the corresponding auxiliary distribution. The estimations given by the new method differ from sample to another as shown in the latest column of table 8, which is natural since each sample provides a different information about the parent distribution. We can also use the minimum of distance d_v and we get also the same conclusion.

6 Perspectives for the New Approach

6.1 Model Selection From Truncated Data

The fact that the distance d_v is a metric allow to propose various applications of this new measure. We can use it for model selection amongst different probability families. We choose two or more possible candidate parametric families of distributions, and for each alternative family, estimate the parameters to select a specific candidate. Determine the distance between the specific candidate and the empirical distribution using the new metric d_v. Finally, select the family which yields the minimum distance. In view of the new approach this can also be done in case of truncated data as opposed to classical approaches (see for example Cox [3], [4]), Taylor and Jakeman [16]) for model selection which can be used, from the best of our knowledge, only for complete data.

To investigate this perspective thoroughly, samples of various sizes from known distributions should be simulated, and the method for model selection applied, we can score the selection as correct or not after repeating the process a large number of times, the probability of correct selection could be estimated according to a given sample size.

We can also use the new distance in cases where classical goodness of fit tests cannot reject two candidate families. We can choose the one which yields the minimum of distance d_v.

In the following examples, we shall select, in the first, between binomial distributions from truncated data. In the second example, we select between a Weibull and a Gamma distributions from right truncated data.

Selection from Binomial distributions. We simulated 10000 samples of size 100 from a Binomial distribution $\mathcal{B}(8, 0.1)$ and each time we retained only the observations belonging from $\{0, 1, 2, 3\}$ with their frequencies. Then we tried to identify the law simulated starting from the corresponding table of frequencies. We used the distance d_v to select between the original distribution of each simulated sample and the distribution $\mathcal{B}(10, 0.15)$ and we score the selection as correct if the distance between the empirical distribution and the original one is less than with the alternative one $\mathcal{B}(15, 0.15)$. The correct distribution was selected $98,8\%$. Conversely, we simulated 10000 samples of size 100 from a Binomial distribution $\mathcal{B}(10, 0.15)$ and we select with $\mathcal{B}(8, 0.1)$, the correct distribution was selected $99,43\%$.

Selection between Weibull and Gamma distributions. We simulated 10000 samples of size 1000 from the weibull distribution $\mathcal{W}(1.2, 1.5)$ and we truncated them on right by considering only observations above the cut-off 1.25. Each truncated sample was summarized into 11 classes. We selected between $\mathcal{W}(1.2, 1.5)$ and the Gamma distribution $G\,(2, 0.5)$. The distance d_v has selected the correct distribution, that is $\mathcal{W}(1.2, 1.5)$, 98.16%.

We can also find, before selecting between distribution, the best fit from the family of gamma distributions $G\,(a, b)$ of the truncated data from a given probability density say $\mathcal{W}(1.2, 1.5)$. We have then to solve an optimization problem of finding the minimum of a function of two variables, $\min_{a,b} d_v(\widehat{f}, f)$ where $\widehat{f}$ is the empirical distribution and $f \equiv G\,(a, b)$, using well known methods such as Lavenberg-Marquardt using a computer algebra package. Also it should be better to choose the number of

bins for each truncated sample by an optimal procedure, for example that of Birgé and Rozenholc [2].

6.2 Estimation of the initial trial value in *EM* Algorithm

The initial starting value is of great importance in convergence behaviour of algorithms such as *EM* Algorithm. Usually, as for the latter, the initial trial value is guessed. Surprisingly, we will show that our procedure gives an estimation of the starting value instead of having to guess. The approach will be illustrated by the following classical example which was the basis of the *EM* algorithm.

Example of Hartley (1958) revisited. Hartley [7] used an algorithmic procedure to estimate the parameter of a Poisson distribution from data on the pollution of a sort of seeds by the presence of noxious weed seeds quoted from Snedecor [15] and truncated them by missing the frequencies of the values 0 and 1 as shown in the following table 9 (Table 1 in Hartley [7])

Table 9.

Values	missing	0	1							
	observed			2	3	4	5	6	7	9
frequencies n_i				26	16	18	9	3	5	1

Hartley [7] has guessed the frequencies of the missing values 0 and 1 by taking $n_0 = 4$ and $n_1 = 14$, and after 4 steps of his algorithmic procedure, which has been the basis of the well known EM algorithm for incomplete data (Dempster, Laird and Rubin [5]), has reached the estimation $\widehat{\lambda} = 3.026$ (see table 1 p.177 Hartley [7]). Using the second method, we get the estimation $\widetilde{\lambda}_2 = 3.1149$. And by proportional allocation procedure we can see that the frequencies we get are $n_0 = 4.29$ and $n_1 = 13.38$ which are close to the guessed values. Using the distance d_v we obtain the estimation $\widetilde{\lambda}_1 = 3.8447$, and by removing the last value which has a small frequency $n_7 = 1$, we obtain a better result $\widetilde{\lambda}_1 = 3.4441$, which are also appreciable as starting values since in practice the true parameter is unknown.

Initial trial value for mixture Normal Populations. We shall present an application of the previous method used for truncated data in the situation where we have a mixture population of two normal distributions. In classical methods, we use the merged distribution $f = \alpha f_1 + (1-\alpha) f_2$ and we estimate the parameters α, m_1 and m_2 using for example the EM algorithm which is based on maximizing the complete likelihood of the merged distribution by an algorithmic procedure from a guessed initial trial value. However, the problem of occurrence of several local maxima is well-known for the setting of EM algorithm. Also, Seidel, Mosler and Alker [14] pointed out that the likelihood-ratio test in mixture models depends on the choice of the initial trial value for the EM algorithm. If the initial trial value is close to the true value it is clear that the algorithm will converge in few steps to the true local maximum. We will show that using the new approach we get an accurate estimated initial trial value.

Assume we have a merged sample from two samples of observations of sizes n_1 and n_2 from two normal distributions $f_1 = N(m_1, \sigma_1)$ and $f_2 = N(m_2, \sigma_2)$, with

$m_1 \neq m_2$. By assuming that σ_1 and σ_2 are known, our aim is to estimate the means m_1 and m_2, and also the merging proportion α of each population.

We will use a method based on truncations. The main idea being to split the range of the merged sample into three suitably chosen parts. A central part where the observations are highly merged, a left and right truncated parts where the observations become mainly from one of the distributions considered. If for example $m_1 < m_2$, then to estimate m_1 we have to use the chosen right truncated part (left truncation $\triangle$).

The procedure is summarized as follows:

1. We compute the sample mean m_g of the merged observations.

2. For determining the location of the two means m_1 and m_2, we compute the empirical standard deviation S_l of the observations less than m_g, and S_r for those that are greater. Assume that $S_l < S_r$, in this case if $\sigma_1 < \sigma_2$ then we deduce that m_1 is situated on the left of m_g. Otherwise, it will be assumed to be on its right. We follow the same idea for the case $S_l > S_r$. If $\sigma_1 = \sigma_2$ we pass directly to the third step.

3. Assume that m_1 is on the left. It is well known that for a normal distribution $N(m, \sigma)$ we have $P(]m - \sigma, m + \sigma[) \simeq 0.68$. We hope that on the left of $\sup_l = m_g - \sigma_2$ the number of observations generated from $N(m_2, \sigma_2)$ is negligible, and on the right of $\min_r = m_g + \sigma_1$ the number of observations generated from $N(m_1, \sigma_1)$ is also negligible. Hence, to estimate m_1, we consider only the part of observations situated on the left of $m_g - \sigma_2$, and to estimate m_2 we consider the part situated on the right of $m_g + \sigma_1$.

The following example will provide some feel for the accuracy of the procedure.

Example. We consider the case where $\sigma_1 = \sigma_2$. consider two samples of observations generated from $N(m_1, \sigma_1)$ and $N(m_2, \sigma_2)$, where $m_1 = 1.3$ and $m_2 = 2.4$, with known $\sigma_1 = \sigma_2 = 1$ and sizes $n_1 = 300$ and $n_2 = 200$. We combine them to obtain a merged sample of size $n = 500$. We have chosen the distributions in such a way that the histogram (Fig.1) of the merged sample does not show directly the existence of a mixture of two distributions. When the histogram of the merged population is bimodal the situation is more easier, since when taking a suitably left (or right) part we get more accurate estimation from the situation that this part will have a negligible number of observations from the second distribution.

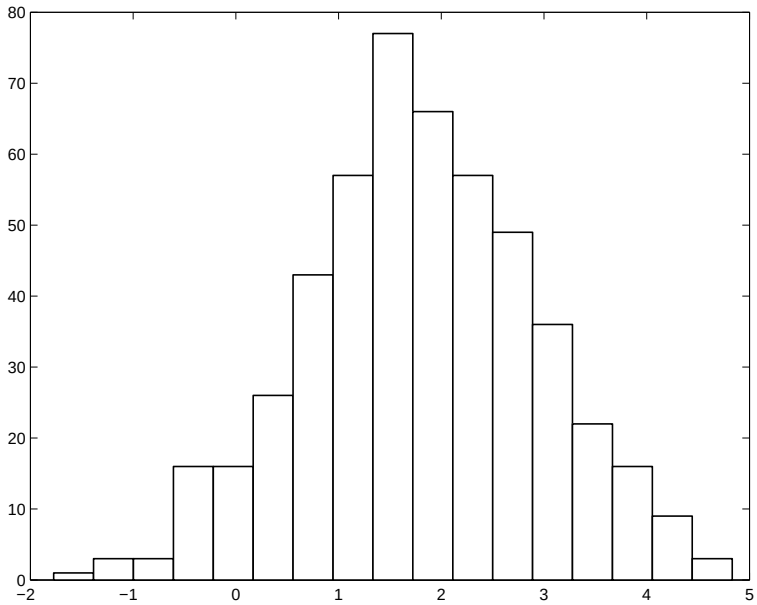

Fig 1. Merged histogram of two normal distributions $N(1.3, 1)$ and $N(2.4, 1)$.

It should be stressed that the histogram is one modal and does not show at first glance any mixture situation. Following the steps of the procedure we begin by calculating the mean of the resulting merged sample and we obtain $m_g = 1.8046$. Since the standard deviations are assumed to be equal then we compute directly $\sup_l = m_g - \sigma_2 = 0.8046$. By grouping the observations on the left of $\sup_l$ (which constitute the chosen right truncated part) in 7 classes we obtain the following table:

Table 10.

u_i	-1.5589	-1.1294	-0.6998	-0.2703	-0.1593	0.5888
n_i	1	3	6	17	24	41

Using the distance d_v we obtain for all the truncation $\widetilde{m}_1^{(dv)} = 1.244$ and by deleting u_1 we get the value $\widetilde{m}_1^{(dv)} = 1.2516$.

The sample mean of the observations on the left of $\sup_l$ is given by $\overline{u}_l = 0.1483$. Using the second method we have to solve on m the following formula

$$\frac{u_1 \times \exp\left[\frac{-(u_1-m)^2}{2\sigma^2}\right] + u_2 \times \exp\left[\frac{-(u_2-m)^2}{2\sigma^2}\right] + ... + u_k \times \exp\left[\frac{-(u_k-m)^2}{2\sigma^2}\right]}{\exp\left[\frac{-(u_1-m)^2}{2\sigma^2}\right] + \exp\left[\frac{-(u_2-m)^2}{2\sigma^2}\right] + ... + \exp\left[\frac{-(u_k-m)^2}{2\sigma^2}\right]} = \overline{u}_l. \quad (17)$$

we obtain the estimation $\widetilde{m}_1 = 1.2646$. By deleting the first value u_1 which has a weak frequency $n_1 = 1$, that is using the truncation $\triangle = \{u_2, u_3, u_4, u_5, u_6\}$, (we compute again $\overline{u}_l = 0.1734$) we obtain a better estimation $\widetilde{m}_1 = 1.3011$, which is very close to the true value $m_1 = 1.3$.

To estimate m_2, we consider the part situated on the right of $\min_r = m_g + \sigma_1 = 2.8046$. Grouping the observations on the right of $\inf_d$ (which constitute the chosen right part) in 7 classes we obtain the following table:

Table 11.

u_i	2.979	3.316	3.653	3.990	4.326	4.663
n_i	38	25	15	9	7	3

Using the distance d_v for all the truncation we get $\widetilde{m}_2^{(dv)} = 2.397$. The sample mean of the observations on the right part is given by $\overline{u}_d = 3.523$. Using formula (17) with $\overline{u}_d$, we obtain the result $\widetilde{m}_2 = 2.245$. Deleting the extreme values u_1 and u_6 we obtain $\widetilde{m}_2 = 2.412$.

The mixture proportion α can easily be estimated using the formula $\alpha \times \widetilde{m}_1 + (1-\alpha) \times \widetilde{m}_2 = m_g$.

Considering the estimations obtained, which are close to the true values of m_1 and m_2, it is clear that the EM algorithm will converge fastly to the unique solutions.

6.3 Test of Goodness of Fit Based on the New Distance

We can obtain empirical quantile estimations of d_v using Montecarlo or Bootstrapping technics, and use them in a test of goodness of fit for a specified probability distribution. We simulate N samples of the same size from the specified probability distribution and calculate the distances $d_v^{(1)}, ..., d_v^{(N)}$. We can then estimate the asymptotic distribution of d_v by

$$F_{d_v}(d) = \frac{\#d_v^{(i)} < d}{N}. \tag{18}$$

Consequently, for a sample of the same size we compute $d_v^{(obs)}$ and we reject the hypothesis that it belongs from the specified distribution if $F_{d_v}(d_v^{(obs)}) > (1-\alpha)$ for a given level of significance α.

The values $d_v^{(1)}, ..., d_v^{(N)}$ may be obtained from the empirical distribution function F_n of the sample.

6.4 Quality of Data

The fact that the new measure d_v is not equivalent to classical ones means that it treats other aspects not investigated by the latter. This may open new perspectives such as making decision about the accuracy of an estimation in cases where the classical and new estimations are close to each others. In cases where the classical estimation and the new one using d_v are significantly different then we can say that the sample of observations considered does not restore coherently all necessary information about the parent distribution from which it emanated.

7 Concluding Remarks

In the foregoing study, we have presented a new statistical point estimation method which found be useful in truncated and grouped and censored data situations. A new distance between probability distributions was introduced. It measures the difference between the variations of two given probability distributions. We introduced an auxiliary distribution based on a truncation, from a chosen family of probability distributions. This new distribution will have the same parameters to estimate as the parent one. We use then statistical methods to estimate the parameters of the random variable under study using the empirical and new auxiliary distribution in the region that captures the data, from which we determine the corresponding parent

distribution. The later is the estimation by the new method. Using the new distance introduced we also estimate by the minimum distance approach and use the resulting estimation as a control on the accuracy of estimation obtained by the former method. We have obtained a result which states that if we have to estimate the parameter of a probability distribution from the one parameter exponential family, then it suffices to have two points with exact ratio of frequencies, that is equal to the theoretical one expressed by the ratio of the value of the probability distribution on these two points, to obtain the true value of the parameter. We have conjectured that if we have in general r parameters, then it suffices to have $r+1$ points with exact ratios of their frequencies to obtain the r true parameters exactly. The later result need to be proved rigorously in a general setting for other distributions than the class considered. A large comparative study between the classical and new methods should also be investigated. We presented some perspectives of the new approach such as model selection from truncated data using the new distance, estimation of the first trial value in the celebrate *EM* algorithm in the case of truncation and for mixture of two normal populations, a test of goodness of fit based on the new distance, decision making about the quality of estimations and data.

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] Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium,* Vol. 1, 105-123. Berkeley: University of California Press.

[4] Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B n° 24, *pp* 406–424.

[5] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc.* B n° 1, *pp* 1-38.

[6] Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data, *J. Am. Stat. Assoc.* Vol. 94. No. 447. pp. 824-834.

[7] Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data, *Biometrics*, June , pp 174-194.

[8] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* Vol. 53. pp. 457-481.

[9] Klein, J. P. and Zhang, M. J. (1996). Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia, *Lifetime Data: Models in Reliability and Survival Analysis, N.P. Jewel, 175-185.*

[10] Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation.* Springer, New York.

[11] Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. R. Astr. Soc.* Vol.155. pp. 95-118.

[12] Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.,* 1065-1076.

[13] Shaw, D. (1988). On-Site samples regression problems of nonnegative integers, truncation, and endogenous stratification. *Journal of Econometrics,* 37, pp. 211-223.

[14] Seidel, W., Mosler, K., and Alker, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Ann. Ist. Stat. Math.,* 52, 481-487.

[15] Snedecor, G. W. (1956). *Statistical Methods.* 5th ed., The Iowa State College Press, Ames, Iowa.

[16] Taylor, J. A. and Jakeman, A. J. (1985). Identification of a distributional model. *Commun. Statist.- Simula. Computa.,* 14(2), 497-508.