Improved Nonparametric Inference for the Mean of a Bounded Random Variable with Application to Poverty Measures

Mame Astou DIOUF[1] Jean Marie DUFOUR[1,2]
ma.diouf@umontreal.ca jean.marie.dufour@umontreal.ca

[1]CIREQ, Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal Québec, H3C 3J7, Canada

[2]CIRANO

May, 2005

PRELIMINARY

Abstract

We provide new finite sample nonparametric inference methods for the mean of a bounded random variable. For this purpose, we prove that the impossibility theorem of Bahadur and Savage (1956) does not apply in this case. Next, we observe that confidence intervals for the mean of a bounded random variable can actually be derived by projection from confidence intervals for the adequate distribution function and investigate finite sample nonparametric methods based on improved Kolmogorov Smirnov statistics and likelihood ratio improvement.

Further, we apply all studied inference methods on the Foster, Greer and Thorbecke (FGT, 1984) poverty measures. We show that FGT poverty measures are actually expectations of some bounded random variables, namely a mixing between a continuous bounded random variable and a mass at the poverty line. So, all inference methods for the mean of bounded random variable apply to this case. We study the relative performance of such methods.

Monte Carlo simulations demonstrate the necessity of using finite sample nonparametric approaches. The asymptotic and bootstrap inference methods appear not reliable in finite sample. On the contrary, the finite sample nonparametric inference methods we propose are robust to the framework and the sample size we use. Confidence intervals we get have a very good coverage probability (always close to 100%) and a good precision. In addition, we provide explicit expressions which make them very easy to compute.

Keywords : nonparametric inference ; mean of bounded variables ; Wald, Rao and likelihood ratio principles of improvement ; poverty measures.

1 Introduction

Since a couple of decades, there is a growing interest for poverty and inequality studies. People are more and more conscious of the importance of controlling the evolution of poverty and inequalities but studies in this area remain descriptive in the majority of cases and neglect statistical inference. More recently, some researches have investigated the performance of two types of inferences for poverty and inequality measures : inference based on asymptotic distributions and bootstrap methods (see Beran (1988) [4], Kakwani (1993) [31], Dardonni and Forcina (1999) [16], Biewen (2002)[7] Davidson and Duclos (2000) [17], Cowell and Flachaire (2002)[15], Davidson and Flachaire (2004)[18]). Most of them recommend the use of bootstrap inference rather than that of asymptotic one. They found that asymptotic approximations do not perform well for both poverty and inequality measures. In the opposite, standard bootstrap inference is proved to perform better than asymptotic inference but its results remains unsatisfactory, especially for inequality measures.

No study has investigated the applicability and the performance of finite sample non parametric inference methods for poverty and inequality measures. This paper does that for poverty measures, in particular for the Foster, Greer and Thorbecke (FGT, 1984) [23]. For this purpose, we show that the FGT poverty measures are actually expectations of a mixing between a continuous bounded random variable and a mass at the poverty line. Given that this mixing is a bounded random variable, we observe that one can use any finite sample nonparametric inference method for the mean of a bounded random variable to perform inference on the FGT poverty measures. However, following Bahadur and Savage (1956)[3], one can think that this problem has no solution. In fact, Bahadur and Savage (1956) proved that non-parametric inference cannot be performed for the mean of a random variable when observations are independant and identically distributed from an unknown distribution function with finite mean [see Dufour (2003)[20] for more details]. But given that the support of the distribution function we study is bounded, this impossibility theorem does not apply in this case. Nonparametric inference methods can be investigated.

Such methods have been provided in the literature by Anderson (1969)[1], Hora and Hora (1990)[28] and Fishman (91)[25]. However, comparing the performance of these methods to that of asymptotic and bootstrap inference for the mean of a random variable with Beta distribution, Sutton and Young (1997)[39] show that the first methods are strongly reliable but yield confidence intervals of larger width than the latest ones. In this paper, we propose improved finite sample nonparametric inference methods for the mean of a bounded random variable.

The contribution of this paper is in two directions.

First, we provide solutions to the problem of performing finite sample nonparametric inference for the mean of a bounded random variable. The key idea of our approach is to observe that confidence intervals for the mean of a bounded random variable can actually be derived by projection from confidence intervals for the adequate distribution function (see Dufour (1990)[19]). Moreover, any confidence interval for a distribution function with a bounded support can be used for that purpose. We investigate finite sample nonparametric methods based on various versions of the Kolmogorov Smirnov statistic (standard one and standardized and studentized ones). We derive improved methods using modified standardized and studentized Kolmogorov Smirnov statistics and likelihood based statistics that dominates Kolmogorov Smirnov statistics (according to Berk and Jones (1979) [6]).

Second, we apply all inference methods on the Foster, Greer and Thorbecke (FGT, 1984)[23] poverty measures using the definition of these measures as expectations of some bounded random variables. The inference methods we derive provide new inference methods for poverty measures. Monte Carlo simulations compare the performance of the confidence intervals based on such methods to that of the previously used inference approaches. The performance is evaluated with respect to two main characteristics : the accuracy and the precision represented respectively by the level of confidence and the width of the confidence interval. The results demonstrate the necessity of using finite sample nonparametric approaches. They confirm that asymptotic inference has a very bad performance in finite sample and gives a very low level of confidence. Bootstrap inference also fails to give accurate results in finite sample when the distribution presents a high probability of zero values. Moreover, even when the probability of zero values is low, the bootstrap confidence interval has a very bad precision in small sample. On the contrary, the finite sample nonparametric inference methods we propose are totally robust to the the sample size and the form of the distribution being studied. Confidence intervals we get have a very good coverage probability which is always close to 100% and a good precision. Lastly, we provide explicit expressions which make the confidence bands very easy to compute.

The remaining of this paper is organized as follows. Section 2 presents the inference methods for the mean of a bounded random variable which are available in the literature. Asymptotic methods (asymptotic and bootstrap inference) and exact exact methods (Anderson (1969)[1], Hora and Hora (1990)[28] and Fishman (91)[25] inference methods) are exposed.

The third section presents the projection principle we stated before and interprets the Anderson's confidence interval as an illustration of this principle. It also generalizes the Anderson's confidence interval and presents some improvment we perform on this method.

Sections 4 and 5 present improved nonparametric confidence intervals for the mean of a bounded random variable. These improvements are done in three directions : according to the Wald, the Rao and the likelihood ratio principles. We improve the confidence interval for distribution function using these principles and apply the projection principle to get improved confidence intervals for the mean of a bounded random variable.

Section 6 applies the inference methods we discuss on the Foster, Greer and

Thorbecke (FGT, 1984) poverty measures. We present briefly the FGT poverty measures and show how the previous inference methods can be applied to them. We present Monte Carlo evidence of the good performance of the finite sample nonparametric inference methods we provide.

Section 7 concludes.

2 Confidence Intervals for the Mean of a Bounded Random Variable : a Review

There exists a variety of methods that perform inference for the mean of a bounded random variable. These methods are devided in two groups : asymptotic and exact methods.

The asymptotic methods are essentially asymptotic and bootstrap inference methods. They are certainly the most widely used inference methods but given the asymptotic arguments they rely, they suffer from limits in finite sample.

In parallel, exact non parametric procedures also exist but are not well known. One can cite among them Anderson (1969)[1], Hora and Hora (1990)[28] and Fishman (91)[25].

The relative performance of these asymptotic and exact procedures has been studied by Sutton and Young (1997)[39]. They compute the level of confidence and the width (which represent respectively the accuracy and the precision) of the confidence intervals these methods yields for the mean of a Beta distribution with different values of Beta parameters. Their results show that asymptotic and standard boostrap procedures are not reliable in small samples but appear to be more precise than nonparametric inference methods. In fact, their level of confidence is lower than the theoritical one but the width of their intervals is better (less) than the width of the exact method intervals. Moreover, the precision of the asymptotic method intervals is alterated when the random variable have a high probability of being null. On the contrary, the accuracy of exact methods is very good (almost always 100%) but is obtained at the cost of a worse precision than the asymptotic methods.

In a very close area, other studies provide one-sided nonparametric inference methods for the mean of particular random variables. Breth (1976)[10] uses the Kolmogorov confidence interval for distribution function to derive nonparametric confidence intervals for a mean using censored distribution. Breth, Maritz and Williams (1978)[11] provide a distribution-free lower confidence limit for the mean of a nonnegative random variable using the Kolmogorov upper bound for distribution function and Kaplan (1987)[32] does so using the Markov inequality and a martingale argument.

Note that nonparametric confidence bands for the mean of a bounded random variable lead to one-sided confidence interval when the random variable has only one bound.

2.1 The asymptotic methods

2.1.1 Asymptotic inference

This inference method is well known and widely used. It relies on convergence theorems that state the asymptotic normality of most of the statistics in the literature.

Let X be a random variable with a distribution function $F(x)$. To perform asymptotic inference for the expected value of X, $E(X)$, one can use the usual t-statistic

$$W = \frac{\widehat{E(X)} - E(X)}{\widehat{V[\widehat{E(X)}]}}$$

The empirical mean $\overline{X}$ is often used as an estimation of $E(X)$.

Given that W is asymptotically normally distributed, one can build a $(1-\alpha)$ level confidence interval for E(X) by

$$\widehat{E(X)} - z_{(1-\frac{\alpha}{2})} * \left[\widehat{V[\widehat{E(X)}]}\right]^{1/2} \leq E(X) \leq \widehat{E(X)} + z_{(1-\frac{\alpha}{2})} * \left[\widehat{V[\widehat{E(X)}]}\right]^{1/2} \quad (1)$$

where $z_{(1-\frac{\alpha}{2})}$ is the $(1-\frac{\alpha}{2})$ percentile of the standard normal distribution.

The asymptotic inference has been often criticized with respect to its poor finite sample performance due to the low speed of convergence of the real distributions of statistics. Confidence intervals built on this basis are not reliable at all. They yield a level of confidence far smaller than the targeted one (see Sutton and Young (1997)[39]).

2.1.2 Bootstrap inference

As asymptotic inference, bootstrap inference is well known. A variety of bootstrap methods have been implemented and used to peform adequate inference in accordance with the properties of the studied data. The percentile-t bootstrap is one of the simpler bootstrap procedures.

Let X and W defined as above. The percentile-t bootstrap confidence interval is similar to the asymptotic one where the asymptotic normal critical values are replaced by those from the bootstrapped distribution. That is :

$$\widehat{E(X)} - D^W_{(1-\frac{\alpha}{2})} * \left[\widehat{V[\widehat{E(X)}]}\right]^{1/2} \leq E(X) \leq \widehat{E(X)} - D^W_{(\frac{\alpha}{2})} * \left[\widehat{V[\widehat{E(X)}]}\right]^{1/2} \quad (2)$$

where $D^W_{(\frac{\alpha}{2})}$ and $D^W_{(1-\frac{\alpha}{2})}$ are respectively the $\frac{\alpha}{2}$ and $(1-\frac{\alpha}{2})$ percentiles of the bootstrapped distribution of W.

This method has been often proved to yield better finite sample performance than the asymptotic one but remain unsatisfactory in multiple cases (see Sutton and Young (1997)[39]). In fact, in presence of distributions with heavy tails, multiple outliers values or a high probability of having null values, the percentile-t booststrap fails to give accurate results. One must investigate the origin of the bootstrap's problem and use the adequate bootstrap method to get some acceptable to results. Moreover, bootstrap inference involves a resampling procedure which is very computationnally expensive.All this leads against the use of the bootstrap based inference methods.

2.2 The exact methods

2.2.1 The Hora and Hora (1990) confidence interval

Hora and Hora (1990)[28] provide a finite sample $(1-\alpha)$ percent confidence interval for the mean μ of a bounded random variable. The confidence interval they derive is symmetric with respect to the sample mean.

Theorem 1 [Hora and Hora (90)] *Let X be a random variable with an unknown continuous cumulative distribution function $F(x)$ with range $[a,b]$ (for numbers $a<b$) such that $F(a)=0$ and $F(b)=1$. Suppose that we have a sample of n draws from $F(x)$. Then a $(1-\alpha)$* percent confidence interval for the mean of X is *defined by$\overline{X_n} \pm \Delta_\alpha$ where $\overline{X_n}$ is the empirical mean and Δ_α is the Kolmogorov -Smirnov percentile value.*

In the opposite of the asymptotic methods, the Hora and Hora confidence interval totally distribution-free which performance doesn't rely at all on sample sizes. It is easy to compute and depend only on the confidence level, the sample mean and the Kolmogorov-smirnov percentile value. Sutton and Young (1997) show that this confidence interval yield a very high level of coveragebut its precision is far worse than the precision of asymptotic methods.

2.2.2 The Fishman (1991) confidence interval

Fishman (1991)[25] derives a finite sample $(1-\alpha)$ percent confidence interval for the mean μ of a bounded random variable. Its confidence interval is based on the Hoeffding's inequality (1963) :

$$\begin{aligned} \Pr[\overline{X}_n - \mu &\geq \varepsilon] \leq e^{nf(\varepsilon,\mu)} \\ \textit{where } f(\varepsilon,\mu) &= (\varepsilon+\mu)\ln\frac{\mu}{\varepsilon+\mu} + (1-\varepsilon-\mu)\ln\frac{1-\mu}{1-\varepsilon-\mu} \text{ for } 0<\varepsilon<1-\mu \end{aligned}$$

Theorem 2 [Fishman (91)] *Let $X_1,...,X_n$ denote iid random variables with $E(X_i)=\mu$ and $Pr[0 \leq X_i \leq 1]=1$. Then*

$$\Pr[\mu_1(\overline{X}_n,n,\alpha) < \mu < \mu_2(\overline{X}_n,n,\alpha)] \geq 1-\alpha \tag{3}$$

with

$$\begin{aligned} \mu_1(\overline{X}_n,n,\alpha) &= \{t, 0<t\leq \overline{X}_n \leq 1 \textit{ and } e^{nf(\overline{X}_n - t,t)} = \alpha/2\} \textit{ for } \overline{X}_n > 0 \\ &= 0 \textit{ for } \overline{X}_n = 0 \end{aligned}$$

$$
\begin{aligned}
\mu_2(\overline{X}_n, n, \alpha) &= \{t, 0 \le \overline{X}_n \le t < 1 \ \textit{and} \ e^{nf(t-\overline{X}_n, 1-t)} = \alpha/2\} \ \textit{for} \ \overline{X}_n < 1 \\
&= 1 \ \textit{for} \ \overline{X}_n = 1
\end{aligned}
$$

The Fishman's confidence interval applies to bounded random variable of range $[0, 1]$ but can be generalized for any support $[a, b]$.

This confidence interval is one of the most general nonparametric confidence intervals for a bounded random variable which exist in the literature. Sutton and Young (1997) show that the Fishman's inference method yields confidence intervals of a high confidence level than the Hora and Hora (1990) method and better precision. However its width is higher than the width of the asymptotic methods. Moreover, the confidence bands are not explicitly defined. They are computed as a zero of a function. As a consequence, the accuracy of this inference method relies in a large part on the accuracy of the iterative procedure that will be used to derive the bounds μ_1 and μ_2. Second, this confidence interval depends on the empirical mean $\overline{X}_n$ which is very sensitive to outliers. These remarks reduce the reliability of the Fishman's confidence interval.

2.2.3 The Anderson (1969) confidence interval

Anderson (1969) [1] derives a finite sample nonparametric confidence interval for the mean of a continuous bounded random variable. Its confidence interval involves the significance points of the Kolmogorov goodness-of-fit test.

Theorem 3 [Anderson (1969)] *Let X be a random variable with an unknown continuous cumulative distribution function $F(x)$ with range $[a, b]$ (for numbers $a < b$) such that $F(a) = 0$ and $F(b) = 1$. Suppose that we have a sample of n draws from $F(x)$ and let $x_{(1)} < ... < x_{(n)}$ be the ordered observations. We state $x_{(0)} = a$ and $x_{(n+1)} = b$. Let $F_n(x)$ be the empirical cumulative distribution function of the sample.*

Let β and γ define a $(1-\alpha)$ confidence interval for $F(x)$ i.e. such that

$$
P\ [F_n(x) - \beta \ \le F(x) \le F_n(x) + \gamma, \quad \forall x] = 1 - \alpha
$$

The following inequalities hold simultaneously with probability $1-\ \alpha$.

$$
\begin{aligned}
E[X] &\le \frac{1}{n}\left[(r+1)\ x_{(r+1)} + \sum_{j=r+2}^{n} x_{(j)}\right] + \beta\ \left[b - x_{(r+1)}\right] \qquad \textit{for}\ r = [n\beta] \\
E[X] &\ge \frac{1}{n}\left[\sum_{j=1}^{n-s-1} x_{(j)} + (s+1)\ x_{(n-s)}\right] - \gamma\ \left[x_{(n-s)} - a\right] \qquad \textit{for}\ s = [n\gamma]
\end{aligned} \tag{4}
$$

The Anderson's (1969) inference method has some properties similar to the Hora and Hora (1990) and the Fishman's (1991) ones. In fact, it provides a non parametric confidence interval and is robust to sample size. Moreover, Sutton and Young (1997) show that the Anderson's inference method provides an interval with a level of confidence of 100% but a larger width than the asymptotic and bootstrap one. However, the width of this confidence interval is better than the width of the previous exact methods and its confidence interval is explicitly defined and is very simple to compute. As a last remark, note that the Anderson's confidence interval apply only for bounded continuous random variables. This restricts its applicability compared to the others exact methods which hold for all bounded variables.

Anderson (1969) generalizes his confidence interval to the mean of any strictly increasing function of X over the interval [a,b].

Corollary 1 *Under the hypotheses of theorem 3, if g(x) is a strictly increasing function over the interval [a,b] then the following inequalities hold simultaneously with probability* $(1-\alpha)$

$$E[g(X)] \leq \frac{1}{n}\left[(r+1)\ g[x_{(r+1)}] + \sum_{j=r+2}^{n} g[x_{(j)}]\right] + \beta\ \left(g(b) - g[x_{(r+1)}]\right) \qquad for\ r = [n\beta]$$

$$E[g(X)] \geq \frac{1}{n}\left[\sum_{j=1}^{n-s-1} g[x_{(j)}] + (s+1)\ g[x_{(n-s)}]\right] - \gamma\ \left(g[x_{(n-s)}] - g(a)\right) \qquad for\ s = |n\gamma| \tag{5}$$

3 Generalization of the Anderson's Confidence Interval

Anderson (1969) provides a very inference method for the mean of a bounded continuous random variable. Excepted its restriction to continuous cases, this inference method is the best nonparametric inference method for the mean of bounded random variables. In this section we investigate the ways to generalize it and further, we discuss how to improve it.

3.1 A projection principle

Proposition 1 [A projection principle] Let *X be a bounded random variable with a continuous distribution function* $F(x)$ *with range* $[a,b]$ for numbers $a<b$. *Suppose that a n-sample of X is available. If a* $(1-\alpha)$ *confidence interval for the distribution function* $F(x)$ *is defined by*

$$F_n^L(x)\ \leq F(x) \leq F_n^U(x)\ \forall x$$

Then the following inequalities hold with probability $(1-\alpha)$.

$$b - \int_a^b F_n^U(x)\ dx \leq E[X] \leq b - \int_a^b F_n^L(x)\ dx$$

Proof *Propoition 1 follows directly from the definition of the mean of X :*

$$E[X] = \int_a^b x \, dF(x) = b - \int_a^b F(x) \, dx$$

The implications of proposition 1 are very helpful for the problem we are interested with.

First, it appears that non parametric confidence bands for the mean can be derived from non parametric confidence bands for the distribution function F(x) by a simple projection argument. Moreover, any confidence interval for a distribution function with bounded support can be used to this end. Anderson(1969) chooses the Kolmogorov's confidence interval for the distribution function of X for that. As an illustration, we recover the Anderson's confidence interval with this method in appendix 2. We use other confidence intervals which provides of better performance for the distribution function.

Another implication of proposition 1 is the properties of the confidence interval for the mean it induces. In fact, since the projection conserves the properties of the confidence interval for the distribution function, the confidence interval for the mean will exhibit the same advantages of the initial confidence interval but will suffer from the same limits too. As an application, the confidence interval of Anderson has the same propoerties than the Kolmogorov's one. That is first, for F continuous, the bounds it defines are independent of the distribution F that is tested under the null. This makes the Anderson's confidence interval very easy to compute. Second, the confidence interval is too conservative which leads to high level of confidence but large width. Last, the Kolmogorov confidence interval is uniform with respect to the empirical distribution function $F_n(x)$.This implies a bad adjustment for the tails of the distribution which can be improved to derive better inference for the mean. We use these limits to generalize the Anderson' confidence interval and improve it.

As a last remark, let's draw the attention of the reader to the fact that this projection principle can be stated in a more general case. In fact, the same argument apply to general functions of the distribution function F(x). This has been stated and used in various areas for several studies as Dufour (1990), Dufour and Neifar (1994) [21], Dufour and Taamouti (1999) [22] etc.

The corollary 1 of Anderson (1969) is an application of this general principle. As a consequence, the new inference methods for the mean we provide in the remaining of the paper can be applied to any other functional of distribution functions. This includes all the centered and non centered moments of X and several other possibilities.

3.2 Generalisation to non continuous case

The last subsection highlights some identified limits of the Anderson's confidence interval. The first one of these is its restriction to continuous cases. The following proposition uses a property of the Kolmogorov's confidence interval for distribution function to generalize the Anderson's confidence interval to any bounded random variable.

Definition : *A confidence interval is conservative at the level $(1-\alpha)$ if its level of confidence is greater or equal to the nominal level $(1-\alpha)$ for all distributions and greater to $(1-\alpha)$ for at least one distribution.*

Proposition 2 *The exact confidence interval of Anderson (1969) is conservative for the mean of any bounded random variable.*

Proof *This proposition is a direct application of the conservative property of the Kolmogorov confidence interval for non continuous cases as stated by Kolmogorov (1941) and proven by Noether (1963), Conover (1972) and Gleser(1985).*

A much simpler proof is provided as follows. Define D_n the Kolmogorov statistic for any continuous distribution function F

$$D_n = \max_x \mid \widehat{F}_n(x) - F(x) \mid$$

Let G(x) be another distribution function. Then,

$$D_n = \sup_x \mid F_n(x) - F(x) \mid = \sup_{v \in [0,1]} \left| \frac{1}{n} \sum_{k=1}^{n} I_{(F[x_k] \leq v)} - v \right| \geq \sup_{v \in V} \left| \frac{1}{n} \sum_{k=1}^{n} I_{(G[x_k] \leq v)} - v \right|$$

for a $V \subset [0,1]$.

In other words, the $(1-\alpha)$ level Anderson's confidence interval for the mean of continuous bounded random variable defines a confidence interval with level greater equal to $(1-\alpha)$ for the mean of any bounded random variable.

3.3 Improving the Anderson's confidence interval for bounded non continuous variables

The performance of confidence intervals are commonly measured by its level of confidence (accuracy) and its width (precision). Given that the Anderson's confidence interval provides a confidence interval of higher level for non continuous cases one can expect that its width will be higher than what we would want. This corresponds to the second limit of the Anderson's confidence interval we highlighted in the beginning of this section.

A good idea to avoid this outcome is to compute a more adequate confidence interval for the distribution function we get. For this purpose, we must compute by simulation the adequate values of β and γ according to the general form of the distribution function we deal with. This can be done easily and provides a more precise confidence interval for the mean.

We illustrate this method for a special case we are interested in. The interest of this special case will appear at the end of this paper with the application we perform for poverty measures.

Corollary 2 *Let Y be a continuous random variable with distribution F and X be the transformed random variable $X = (\frac{z-Y}{z})^{\alpha}\ I_{[0\leq Y\leq z]}$. The adequate Kolmogorov statistic for goodness of fit test is of the form*

$$D_n = \max_{v\in[p,1]} \left| \frac{1}{n}\sum_{k=1}^{n} I_{[G[x_k]\leq v]} - v \right| \tag{6}$$

where $p = Proba(Y > z)$

Proof : *See appendix 3.*

Note that the distribution function $G(x)$ is such that $G(x)$ is continuous on $(a, b]$ and $G(a) = p > 0$ and $G(b) = 1$. As a consequence $G_X(X)$ is a mixing :

$$G_X(X) \begin{cases} = p & \text{with probability p} \\ \sim U_{(p,1]} & \text{with probability 1-p} \end{cases}$$

The adequate distribution of D_n can be easily computed numerically in this case using the following procedure :

- generate n observations from a mixing of p with probability p and an uniform law $U_{(p,1]}$ with probability 1-p
- Compute the Kolmogorov statistic for this sample
- repeat N times

The adjusted values of β and γ are simply the adequate quantiles of the distribution of D_n.

Note that the same procedure applies to continuous case where the only difference is that the values of the distribution function are drawn from an uniform law ($F_Y(Y) \sim U_{[0,1]}$).

In practice, the true value of p is generally unknown but can be estimated on data. The following proposition says how this estimation affects the performance of the computed confidence interval.

Proposition 3 [Monotonicity] : Let *$G^1(x)$ and $G^2(x)$ be two distribution functions with range [a,b]. Suppose that $G^1(x)$ and $G^2(x)$ are non continuous at a finite number of points belonging respectively to sets V^1 and V^2 with $V^1 \subseteq V^2$. If,*

$$\forall\ x_l \in V^1 \cap V^2,\ G^1(x_l) \leq G^2(x_l)$$

then the confidence interval for $G^1(x)$ is conservative for the distribution function $G^2(x)$.

A special case of this proposition is given by the following corollary.

Corollary 3 : *Let $G^1(x)$ and $G^2(x)$ be two distribution functions with range [a,b]. If $G^1(x)$ and $G^2(x)$ are continuous on (a,b] such that $G^2(a) = p_2 > p_1 = G^1(a) => 0$ and $G^1(b) = 1 = G^2(b)$*

Then the confidence interval for $G^1(x)$ is conservative for the distribution function $G^2(x)$.

The following corollary follows as a consequence of the previous proposition and corollary.

Corollary 4 : *The confidence interval for the mean of the random variable with distribution function $G^1(x)$ is also conservative for the mean of the random variable which distribution function is $G^2(x)$.*

4 Inference Based on Non Uniform Confidence Bands

The previous section provides the mean to avoid the first two limits of the Anderson's confidence interval we gave before. We present another improvement of this confidence interval based on non uniform versions of the Kolmogorov statistic.

Intuitively, the choice of such statistics can be easily motivated. In fact, in addition to its conservative nature, the Kolmogorov confidence interval has been often criticized because of it is uniform with respect to the empirical distribution function. This is another source of imprecision for the projected confidence interval for the mean.

4.1 The Anderson Darling and Eicker Statistics

In the literature, weighted versions of the Kolmogorov statistic have been proposed to provide better inference for the tail of distributions.

One of these statistics is the standardized Kolmogorov statistic provided by Anderson and Darling (1952) [2]. It is defined as follows :

$$AD_n = \sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F^{1/2}(x)[1 - F(x)]^{1/2}} \mid$$

Another statistic is the studentized Kolmogorov statistic provided by by Eicker (1979)[24] :

$$E_n = \sup_x V_n \text{ with } V_n = \begin{cases} \sqrt{n} \mid \frac{F_n(x) - F(x)}{F_n^{1/2}(x)[1 - F_n(x)]^{1/2}} \mid & \text{if } F_n(x) \notin \{0, 1\} \\ 0 \text{ if } F_n(x) \in \{0, 1\} \end{cases}$$

As the standard Kolmogorov statistic, these statistics provide confidence intervals for continuous distribution functions.

Proposition 4 [The AD confidence interval for distribution function] *Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function $F(x)$. Let c_{AD} such that $Pr[E_n \leq c_{AD}] = 1-\alpha$. Then the following inequalities define a $(1-\alpha)$ confidence interval for F(x)*

$$\begin{aligned} F_n^L(x) &\leq F(x) \leq F_n^U(x) \quad (7) \\ \text{where } F_n^L(x) &= \frac{2F_n(x) + \frac{c_{AD}^2}{n} - \sqrt{\Delta}}{2(1+\frac{c_{AD}^2}{n})} \text{ and } F_n^U(x) = \frac{2F_n(x) + \frac{c_{AD}^2}{n} + \sqrt{\Delta}}{2(1+\frac{c_{AD}^2}{n})} \\ \text{with } \Delta &= \left[2F_n(x) + \frac{c_{AD}^2}{n}\right]^2 - 4F_n^2(x)\left[1 + \frac{c_{AD}^2}{n}\right] \end{aligned}$$

Proof c_{AD} *is defined by*

$$P[\sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F^{1/2}(x)[1-F(x)]^{1/2}} \mid \leq c_{AD}] = 1 - \alpha$$

It follows that with probability $1-\alpha$

$$(1 + \frac{c_{AD}^2}{n})F^2(x) - \left[2F_n(x) + \frac{c_{AD}^2}{n}\right] F(x) + F_n^2(x) \leq 0 \qquad \forall\ x$$

This is the case if and only if the inequalities in (7) are satisfied.

Proposition 5 [The Eicker confidence interval for distribution function] *Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function $F(x)$. Let c_E such that $Pr[E_n \leq c_E] = 1-\alpha$. Then, the following inequalities define a $(1-\alpha)$ confidence interval for F(x)*

$$F_n(x) - \frac{c_E}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} \leq F(x) \leq F_n(x) + \frac{c_E}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} \quad (8)$$

Proof c_E *is defined by*

$$P[\sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2}} \mid \leq c_E] = 1 - \alpha$$

It follows that with probability $1-\alpha$

$$-c_E \leq \frac{\sqrt{n}[F_n(x) - F(x)]}{F_n^{1/2}(x)[1-F_n(x)]^{1/2}} \leq c_E \qquad \forall\ x$$

This leads to the inequalities *(8).*

Figure 2 (see appendix 1) shows that the Anderson-Darling and Eicker statistics yield confidence intervals for distribution function that provide a better adjustment for the tail of distribution than the standard Kolmogorov statistic. In fact the Eicker statistic provides confidence regions that contract to zero width in the tail. The improvement of the Anderson Darling statistic is less clear. Its width converges to $(c_{AD}^2/n)/(1+\frac{c_{AD}^2}{n})$ in the tails of the distribution which is to compare to $\gamma+\beta$, the uniform width of the standard Kolmogorov confidence interval (in the Anderson's theorem, 1969). We simulate the critical values of c_{AD}, γ and β for $n = 100$, 200, 500 and 1000. The results show that the width of the uniform confidence interval is better than those of the AD confidence interval in the tails of the distribution for $n = 100$ but the AD confidence interval become better in the tails for $n \geq 200$.

Applying the principle of projection we outlined for the Anderson's confidence interval, we derive a $(1-\alpha)$ confidence interval for the mean of X based on the AD and Eicker confidence intervals for $F(x)$.

Theorem 4 [Confidence interval for the mean of a bounded continuous random variable based on the Anderson Darling statistic]

Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function F(x) with range [a,b]. Define $x_{(0)} = a$ and $x_{(n+1)} = b$ Let $F_n^L(x)$ and $F_n^U(x)$ represent the AD confidence interval for F(x). Then a $1-\alpha$ confidence interval for E(X) is

$$\begin{aligned} E(X) &\leq \sum_{k=1}^{n}[F_n^L(x_{(k)}) - F_n^L(x_{(k-1)})] * x_{(k)} + [1 - F_n^L(x_{(n)})] * b \qquad (9) \\ E(X) &\geq F_n^U(x_{(0)}) * a + \sum_{k=1}^{n}[F_n^U(x_{(k)}) - F_n^U(x_{(k-1)})] * x_{(k)} \end{aligned}$$

Proof : *The proof of this theorem is a direct application of the principle of projection. Proposition 4 states that*

$$F_n^L(x_{(k)}) \leq F(x) \leq F_n^U(x_{(k)}) \quad \forall k = 1, ..., n \qquad (10)$$

Given the definition of $F_n^L(x)$ and $F_n^U(x)$, one can see that $F_n^L(x_{(k)}) \geq 0\ \forall k$ and $F_n^U(x_{(k)}) \leq 1\ \forall k$. As a consequence, the confidence bands of Owen are effective for all $x_{(i)}$. The distribution which satisfies the first inequality in (10) with the largest mean is the distribution with jumps $F_n^L(x_{(1)})$ at $x_{(1)}$; $F_n^L(x_{(k)}) - F_n^L(x_{(k-1)})$ at $x_{(k)}$ for k=2,...,n and $1 - F_n^L(x_{(n)})$ at $x_{(n+1)} = b$. Idem, The distribution which satisfies the second inequality in (10) with the smallest mean is the distribution with jumps $F_n^U(x_{(0)})$ at $x_{(0)} = a$; $F_n^U(x_{(k)}) - F_n^U(x_{(k-1)})$ at $x_{(k)}$ and $1 - F_n^U(x_{(n)}) = 0$ at $x_{(n+1)} = b$. This yields the inequalities (9).

Theorem 5 [Confidence interval for the mean of a bounded continuous random variable based on the Eicker statistic]

Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function F(x) with range [a,b]. Define $x_{(0)} = a$ and $x_{(n+1)} = b$. Let c_E such that $Pr[E_n \le c_E] = 1 - \alpha$. The following inequalities define a more than $(1 - \alpha)$ confidence interval for the the mean of X.

$$E(x) \le \left[\frac{k_E^I + 1}{n} - \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k_E^I + 1}{n}\right)^{1/2}\left(1 - \frac{k_E^I + 1}{n}\right)^{1/2}\right\}\right] * x_{(k_E^I+1)}$$
$$+ \sum_{k=k_E^I+2}^{n}\left[\frac{1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1 - \frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1 - \frac{k-1}{n}\right)^{1/2}\right\}\right] * x_{(k)}$$

$$E(x) \ge \sum_{k=1}^{k_E^S}\left[\frac{1}{n} + \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1 - \frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1 - \frac{k-1}{n}\right)^{1/2}\right\}\right] * x_{(k)}$$
$$+ \left[1 - \frac{k_E^S}{n} - \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k_E^S}{n}\right)^{1/2}\left(1 - \frac{k_E^S}{n}\right)^{1/2} + \varepsilon\right\}\right] * x_{(k_E^S+1)}$$

where

$$k_E^I = [\frac{nc_E^2}{(n + c_E^2)}] \qquad \text{and} \qquad k_E^S = [\frac{n^2}{(n + c_E^2)}]$$

and $[k]$ represents the integer part of k.

Proof : *See appendix 4.*

For the same reason as previously mentioned for the former inference methods, the following corollary holds.

Corollary 5 *The non uniform confidence intervals for distribution function and the corresponding ones for the mean of a continuous bounded random variable are conservative for the non continuous case.*

Moreover, for the same reason than the Anderson's confidence interval, the Eicker and Anderson Darling inference methods can be adjusted to yield inference of better performance for bounded non continuous random variables.

4.2 The Regularized Anderson Darling and Eicker Statistics

The previous subsection two non uniform confidence intervals for distribution functions and the corresponding confidence intervals for the mean based on some statistics provided by the literature. These inference methods are proved to improve the previous one based on the non weighted Kolmogorov statistic.

We investigate a better improvement of the non weighted Kolmogorov statistic inference method. In fact, we study the power of the goodness of fit tests using the Anderson-Darling and Eicker statistics. The first part of Table1 (see appendix 1) prove these tests suffer from a problem of poor power of test against some specific alternatives which can be avoided. We improve the power of these tests by modifying the weight of the previous statistics.

These new statistics can be viewed as some regularized versions of the Anderson-Darling and Eicker statistics. They are obtained by shifting the denominator of the statistics by an additionnal term $\varepsilon(x)$:

$$\begin{aligned} AD_n^R &= \sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F^{1/2}(x)[1-F(x)]^{1/2} + \varepsilon(x)} \mid \qquad (11) \\ E_n^R &= \sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \varepsilon(x)} \mid \end{aligned}$$

Table1 (appendix 1) shows that this regularization have a major impact on the power of the goodness-of-fit tests. We provide below the confidence intervals for distributions functions and for the mean using such statistics. As a consequence of the improvement in the power of tests, theses confidence intervals will perform better than the previous ones.

However, we saw that the width of the Anderson-Darling confidence interval converges to a non zero constant at the tails of the distributions leading to worse inference than the Eicker confidence interval. We expect the Regularized AD statistic to perform less than the regularized Eicker one for a given function $\varepsilon(x)$. We prefer to escape this discussion here. Moreover, the Regularized Anderson confidence interval for distribution function and for the mean are expressed relatively to the roots of an equation of degree 4. Relatively to the poor gain in performance these confidence intervals yield compared to those of the Regularized Eicker inference method, the generated extra complexity can appears too expensive. For these reasons, we choose not to provide these results here.

Proposition 6 [The Regularized Eicker confidence interval for distribution function] *Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function $F(x)$. Let c_{RE} such that $Pr[E_n^R \leq c_{RE}] = 1-\alpha$. Then the following inequalities define a $(1-\alpha)$ confidence interval for F(x)*

$$F_n(x) - \frac{c_{RE}}{\sqrt{n}} \left\{ F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \varepsilon(x) \right\} \leq F(x) \leq F_n(x) + \frac{c_{RE}}{\sqrt{n}} \left\{ F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \varepsilon(x) \right\} \qquad (12)$$

Proof : *c_{RE} is defined by*

$$P[\sup_x \sqrt{n} \mid \frac{F_n(x) - F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \varepsilon(x)} \mid \leq c_{RE}] = 1-\alpha$$

It follows that with probability $1-\alpha$

$$-c_{RE} \leq \frac{\sqrt{n}[F_n(x) - F(x)]}{F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \varepsilon(x)} \leq c_{RE} \qquad \forall\ x$$

This leads to the inequalities *(12).*

As expected, this confidence interval generally provides a better performance than the existing ones. Figure 2 (see appendix 1) shows that this confidence interval is of better width than the uniform one for the tails of the sample. It also shows that this confidence interval performs much better in the center of the distribution compared to the Eicker's confidence interval. However its width converges to $\varepsilon(x)$ in the tail of the distribution whereas those of the Eiker confidence interval converge to zero. For this reason, this later is narrower in the very last values of X. However, in general, the regularized Eicker statistic yield a better inference than the Eicker statistic.

In this section, we use a constant $\varepsilon(x)$ and choose it to maximize the power of the goodness of fit tests the statistics induce. However, it is clear that one can investigate a function $\varepsilon(x)$ that would give narrower confidence interval for the distribution function $F(x)$. Intuitively, such function would give lower values in the center of the distribution and larger ones in the tails.

We apply the principle of projection we outlined for the Anderson's confidence interval to derive a $(1-\alpha)$ confidence interval for the mean of X based on this later confidence interval for $F(x)$. Non constant $\varepsilon(x)$ would give quite similar formulas.

Theorem 6 [Improved Nonparametric confidence interval for the mean of a bounded continuous random variable]
Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function F(x) with range [a,b]. Define $x_{(0)} = a$ and $x_{(n+1)} = b$. Let c_{RE} such that $Pr[E_n^R \leq c_{RE}] = 1-\alpha$. The following inequalities define a more than $(1-\alpha)$ confidence interval for the the mean of X.

$$E(x) \leq \left[\frac{k^I_{RE}+1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k^I_{RE}+1}{n}\right)^{1/2}\left(1-\frac{k^I_{RE}+1}{n}\right)^{1/2} + \varepsilon\right\}\right] * x_{(k^I_{RE}+1)} + \frac{c_{RE}\varepsilon}{\sqrt{n}} * b$$
$$+ \sum_{k=k^I_{RE}+2}^{n}\left[\frac{1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}\right] * x_{(k)}$$

$$E(x) \geq \frac{c_{RE}\varepsilon}{\sqrt{n}} * a + \left[1 - \frac{k^S_{RE}}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k^S_{RE}}{n}\right)^{1/2}\left(1-\frac{k^S_{RE}}{n}\right)^{1/2} + \varepsilon\right\}\right] * x_{(k^S_{RE}+1)}$$
$$+ \sum_{k=1}^{k^S_{RE}}\left[\frac{1}{n} + \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}\right] * x_{(k)}$$

where

$$
\begin{aligned}
k_{RE}^{I} &= \left[\frac{2c_{RE}\varepsilon n^{3/2}+c_{RE}^{2}n+\sqrt{\Delta^{I}}}{2(c_{RE}^{2}+n)}\right] \text{ and } \Delta^{I}=c_{RE}^{3}n^{2}[c_{RE}-4c_{RE}\varepsilon^{2}+4\varepsilon n^{1/2}] \\
k_{RE}^{S} &= \left[\frac{c_{RE}^{2}n+2n^{2}-2c_{RE}\varepsilon n^{3/2}-\sqrt{\Delta^{S}}}{2(c_{RE}^{2}+n)}\right] \text{ and } \Delta^{S}=c_{RE}^{3}n[c_{RE}n-4c_{RE}n\varepsilon^{2}+4\varepsilon n^{3/2}]
\end{aligned}
$$

and $[k]$ represents the integer part of k.

Proof : *See appendix 5.*

For the same reason as previously mentioned, the following corollary holds.

Corollary 5 *The improved confidence intervals for distribution function and for the mean of a continuous bounded random variable are conservative for the non continuous case.*

Moreover, for the same reason than the Anderson's confidence interval, the improved Anderson's confidence interval can also be adjusted to yield inference of better performance for the mean of non continuous bounded random variables.

4.3 Choice of epsilon

We introduce in this section two new statistics for goodness of fit tests. As explained above, the addition of the term ε in the denominator of the statistics E_n and AD_n improve the power of the goodnees of fit tests these tests provide. A natural question arises : how to choose ε ? We restrict ourself to deal first with constant ε. Given that, it seems logical to choose ε such that to maximize the power of the goodness of fit test. One can guess that not only one ε may be adequate. It is reasonble to expect one "optimal" value by tested hypothesis and perhaps by sample size too. So how to discriminate among all these values ?

Our objective is to improve the power of the goodness of fit tests performed with standardized or studentized Kolmogorov statistics but in the same time to distort the original distributions the less necessary. In other words, we look for the lowest ε that increase sufficiently the power of the performed tests.

Table 1 (see appendix 1) relates the results we get for specified hypotheses and $n = 500$. We choose $n = 500$ because the simulations we perform for our application are for $n = 100$, $200, 500$ and 1000. We perfom goodness of fit tests for the Singh Maddala distribution which we use for our distribution but also for the Normal distribution which is the most widely used by studies. For the case we choose, one can see first that the presence of ε highly increases considerably the power of the tests : $\varepsilon = 0$ is not a good choice. Second, the power of test is very low for small ε but becomes acceptable with moderately larger values. We see that $\varepsilon \doteq 0.15$ is suffisant to yield an acceptable power of test. More greater values of ε yield small increases of the power that are not significant. We will use this value in the remaining of the paper.

Note that one can investigate the optimal ε for each sample size and distribution to improve the performance of the inference method we propose. We perform tests for different n but we don't provide the results here. However, as we expected, the choice of ε would in general depend on n. Moreover, the impact of the distribution function will be less important as the sample size increase. For example, for n=100 the additionnal increase in the power of test is more significant as we increase the value of ε. The optimal value of ε would be $\varepsilon \doteq 0.5$ for the normal case and $\varepsilon \doteq 0.3$ for the Singh Maddala case.

5 Likelihood Based Nonparametric Confidence Interval for the Mean of a Bounded Random Variable

The projection principle we stated earlier says that any confidence interval for the distribution function of a bounded random variable can be used to perform inference for the mean of this later. Using this, one can investigate other confidence intervals that dominate the Kolmogorov one to try to improve the results we already get.

Berk and Jones (1979)[6] propose two statistics based on a likelihood criterion instead of the moment one to standardize the Kolmogorov statistic. They show that these statistics are more efficient, in the sense of Bahadur efficiency, than any weighted Kolmogorov statistic. Owen (1995)[36] uses one of these statistics to derive non parametric confidence bands for continuous distribution functions.

Theorem 7 [Owen (1995)] *Let $x_1, ...x_n$ be an iid sample from a continuous distribution function F(x). Define the likelihood criterion*

$$\begin{aligned} R_n &= \sup_{-\infty \leq x \leq +\infty} K[F_n(x), F(x)] \\ \textit{where } K(\widehat{p}, p) &= \widehat{p} \log(\frac{\widehat{p}}{p}) + (1-\widehat{p}) \log(\frac{1-\widehat{p}}{1-p}) \end{aligned}$$

Let λ_n define a $1-\alpha$ confidence interval for R_n i.e

$$\Pr[R_n > \lambda_n] = 1 - \alpha$$

Then a $1-\alpha$ confidence interval for the distribution function is

$$L(x) \leq F(x) \leq H(x) \tag{13}$$

with $L(x) = \min\{p \ , \ K[F_n(x), p] \leq \lambda_n\}$ and $H(x) = \max\{p \ , \ K[F_n(x), p] \leq \lambda_n\}$

The intuition of this confidence interval is not so hard. In fact, remember that $nF_n(x)$ is binomial with parameters n and F(x). Thus, $-nK(\widehat{p}, p)$ is the log-likelihood ratio for the probability parameter p based on a binomial observation of $n\widehat{p}$ successes in n trials. It follows that Owen's confidence bands are computed by performing a likelihood ratio test on the distribution of $F_n(x)$. Only candidates $F(x)$ with sufficiently large likelihood at each x belong to the confidence interval.

Figure 2 shows that the resulting interval is narower in the tails and wider in the center than the uniform Kolmogorov confidence interval. Moreover, it also performs better than the regularized Eicker confidence interval. However, for the same reason as that stated before its width is higher than that of the Eicker confidence interval at the very end of the sample.

Owen (1995) gives an approximation for λ_n for n=1,...,1000 and $\alpha = 0.05$ and 0.01.

$$\begin{aligned}\lambda_n^{0.95} &= \frac{1}{n}[3.0123 + 0.4835\log(n) - 0.00957\log^2(n) - 0.001488\log^3(n)] \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } n = 2, ..., 100 \\ &= \frac{1}{n}[3.0806 + 0.4894\log(n) - 0.02086\log^2(n)] \quad \text{for } n = 101, ..., 1000\end{aligned}$$

and

$$\begin{aligned}\lambda_n^{0.99} &= \frac{1}{n}[-4.626 - 0.541\log(n) + 0.0242\log^2(n)] \qquad \text{for } n = 2, ..., 100 \\ &= \frac{1}{n}[-4.71 - 0.512\log(n) + 0.0219\log^2(n)] \qquad \text{for } n = 101, ..., 1000\end{aligned}$$

Using this non parametric likelihood confidence interval, we derive a confidence interval for the mean of a bounded continuous random variable by projection.

Theorem 8 [Likelihood Based Nonparametric confidence interval for the mean of a bounded continuous random variable]
Let $x_{(1)}, ...x_{(n)}$ be an ordered iid sample drawn from a continuous distribution function F(x) with range [a,b]. Define $x_{(0)} = a$ and $x_{(n+1)} = b$ Let L(x) and H(x) represent the Owen's confidence interval for F(x). Then a $1-\alpha$ confidence interval for E(X) is

$$\begin{aligned}E(X) &\leq L(x_{(1)}) * x_{(1)} + \sum_{k=2}^{n}[L(x_{(k)}) - L(x_{(k-1)})] * x_{(k)} + [1 - L(x_{(n)})] * x_{(n+1)} \\ E(X) &\geq H(x_{(0)}) * x_{(0)} + \sum_{k=1}^{n}[H(x_{(k)}) - H(x_{(k-1)})] * x_{(k)} \qquad (14)\end{aligned}$$

Proof : *The proof of this theorem relies on the same trick than that of the non uniform confidence intervals we have derived. The theorem of Owen (95) gives that*

$$L(x_{(k)}) \leq F(x_{(k)}) \leq H(x_{(k)}) \quad \forall k = 1, ..., n$$

Given the definition of L(x) and H(x), one can see that for $x_{(0)}$, $L(x_{(0)}) = 0$ and $H(x_{(0)}) = 1 - e^{-\lambda_n}$ and for $x_{(n)}$: $L(x_{(n)}) = e^{-\lambda_n}$ and $H(x_{(n)}) = 1$. As a consequence, the confidence bands of Owen are effective for all $x_{(i)}$. The distribution which satisfies the first inequality in (13) with the largest mean is the distribution with jumps $L(x_{(1)})$ at $x_{(1)}$; $L(x_{(k)}) - L(x_{(k-1)})$ at $x_{(k)}$ for k=2,...,n and $1 - L(x_{(n)})$ at $x_{(n+1)} = b$. Idem, The distribution which satisfies the second inequality in (13) with the smallest mean is the distribution with jumps $H(x_{(0)})$ at $x_{(0)} = a$; $H(x_{(k)}) - H(x_{(k-1)})$ at $x_{(k)}$ and $1 - H(x_{(n)}) = 0$ at $x_{(n+1)=b}$. This yields the inequalities (14).

For the same reason as the previously mentioned for the former inference methods, the following corollary holds.

Corollary 5 *The likelihood based confidence intervals for distribution function and for the mean of a continuous bounded random variable are conservative for the non continuous case.*

Moreover, for the same reason than the Anderson's confidence interval and the improved Anderson's confidence interval, the likelihood based inference method can also be adjusted to yield inference of better performance for non continuous bounded random variables.

6 Application to FGT Poverty Measures

Since the last ten years, many studies have been performed on asymptotic and bootstrap inferences for poverty and inequality measures.

Studies on asymptotic inference generally use the same theoritical basis to derive the asymptotic law of estimations. They suppose that a ramdomly drawn sample of n observations from the same distribution is available. Then, they apply the central limit theorem or the Slutsky theorem to derive the asymptotic law and estimate confidence intervals of the statistics they consider. As a consequence, the derived law is generally normal with a variance-covariance structure that can be consistently estimated. This procedure has been applied to a variety of poverty and inequality measures. One can cite among the major contributions Bishop, Formby and Zheng (1997) [8], Rongve (1997) [37], Kakwani (1993) [31], Dardonni and Forcina (1999) [16], Davidson and Duclos (2000) [17] and Zheng (2001) [40].

Many studies have also been performed on bootstrap inference for poverty and inequality measures. Among the major contributions, one can cite Knight (1989)[33], Beran (1988) [4], Burr (1994) [12], Hall (1992) [27], Mills and Zandvakili (1997)[34], Biewen (2002)[7], Cowell and Flachaire (2002)[15].

Most of these studies find evidence that (1) statistical inference is essential even when large samples are available and (2) the use of bootstrap inference is recommended rather than that of asymptotic one.

More recently Davidson and Flachaire (2004)[18] study the finite sample performance of asymptotic and bootstrap inference for the Theil inequality indice and the FGT poverty measures. They show that standard bootstrap inference gives better performance than asymptotic one for poverty and inequality measures. The authors explain this by the low convergence of the law of the statistics to the asymptotic one. However, their results show that both methods have bad performance for inequality measures but the performance of the standard bootstrap inference is acceptable for poverty measures.

6.1 The FGT Poverty Measures

Since Sen (1976)[38], many poverty measures have been exhibited going from the well known headcount ratio to the so used Foster, Greer and Thorbecke (FGT, 1984)[23] poverty measures. These measures have been established to fit some axioms stated as a minimum regularity conditions that must verify a good measure of poverty. Among all these measures the FGT poverty measures are, for sure, the most popular ones due to the large variety of desirable properties they fit, in particular their decomposability. We provide inference methods for these measures.

Let,

- Y be the revenue of people with a continuous distribution function $F_Y(y)$ on a positive support ; y_i represents the revenue of individual i
- z be the poverty line
- n be the sample size

The FGT poverty measures are defined for $\alpha > 0$ by :

$$P_\alpha(Y,z) = \int_0^z (\frac{z-y}{z})^\alpha \; dF(y)$$

Let X be the random variable

$$X = (\frac{z-Y}{z})^\alpha \; \mathrm{I}_{[0 \leq Y \leq z]}$$

Then, one can see that

$$P_\alpha(Y,z) = \int_0^{+\infty} (\frac{z-y}{z})^\alpha \; I_{[y \leq z]} \; dF(y) = E[X]$$

X is a bounded random variable with range $[0,1]$ and cumulative distribution function

$$G_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - F_Y[z(1-x^{1/\alpha})] & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Note also that X has a probability mass of $G_X(0) = 1 - F_Y(z)$ at point 0.

As a consequence, the FGT poverty measures are the expected value of a bounded random variable and can be computed as such. All procedures that perform inference for a bounded random variable apply to FGT measures.

We are interested in testing the null hypothese

$$H_0 : P_\alpha = P_\alpha^0 \text{ against the alternative } H_1 : P_\alpha \neq P_\alpha^0$$

The adequate asymptotic statistic for this test is the t-statistic

$$W = \frac{\widehat{P_\alpha} - P_\alpha}{\widehat{V[\widehat{P_\alpha}]}}$$

If an iid sample y_i, $i = 1, ..., n$ of Y is available, unbiased estimators of P_α and its variance are given by

$$\widehat{P}_\alpha = \frac{1}{n}\sum_{i=1}^{n}(\frac{z - y_i}{z})^\alpha \ I\ [y_i \leq z] \text{ and } \widehat{V\ (\widehat{P}_\alpha)} = \frac{1}{n}[\widehat{P}_{2\alpha} - \widehat{P}_\alpha^2]$$

The asymptotic normality of $\widehat{P_\alpha}$ has been proved by Kakwani (1993). Then, the asymptotic and bootstrap inference methods we described above apply directly. The corresponding asymptotic and bootstrap confidence intervals are respectively

$$\widehat{P_\alpha} - z_{(1-\frac{\delta}{2})} * \left[\widehat{V[\widehat{P_\alpha}]}\right]^{1/2} \leq P_\alpha \leq \widehat{P_\alpha} + z_{(1-\frac{\delta}{2})} * \left[\widehat{V[\widehat{P_\alpha}]}\right]^{1/2}$$

$$\widehat{P_\alpha} - D^W_{(1-\frac{\delta}{2})} * \left[\widehat{V[\widehat{P_\alpha}]}\right]^{1/2} \leq P_\alpha \leq \widehat{P_\alpha} + D^W_{(\frac{\delta}{2})} * \left[\widehat{V[\widehat{P_\alpha}]}\right]^{1/2}$$

Using the definition of P_α in term of X, one can also apply the finite nonparametric inference methods we have discussed above (Fishman (91), Anderson (1969), improved Anderson inference methods, likelihood based inference method) to poverty measures. We use the conservative property of the confidence intervals to apply them to poverty measures. Moreover,we can use the adjustment we discussed for the special discrete variable X the poverty measures involve. Here, the parameter p in the corollary 2 corresponds to the probability of being rich.

6.2 Monte Carlo Simulations

We investigate the performance of the above methods with Monte Carlo evidence. We use a modified version of the set up of Davidson and Flachaire (2004)[18] that is based itself on Brachman, Stich and Trede (1996)[9]. We simulate the revenue of the population from a mixing law :

$$Y = \begin{cases} z \text{ with probability } 1 - P_0 \\ SM(a, b, c) \text{ with probability } P_0 \end{cases}$$

where $SM(a, b, c)$ represents a Singh Maddala distribution with parameters a, b and c for the cumulative distribution function spécification

$$F(y) = 1 - \frac{1}{[1 + ay^b]^c}$$

Following Davidson and Flachaire (2004), we set $a = 100$, $b = 2.8$ and $c = 1.7$. $\alpha = 2$ is used for the poverty measure P_α. Davidson and Flachaire (2004) set the poverty line z to half the median of the distribution SM(a,b,c). In order to keep the same logic, we use the same poverty line. For the specifications they choose, the true value for P_2 is $P_2^0 = 0.013017$. Our corresponding true value is $P_2^0 = P_0 * 0.013017$.

We choose $P_0 = 0.1$ and perfom $N = 10,000$ Monte Carlo replications for sample sizes $n = 100, 200, 500, 1000$.

The choice of $P_0 = 0.1$ is done to illustrate a case where bootstrap inference fails even for $n = 1000$.

For a SM(a,b,c) with the values we choose, the probability of being poor is about 0.11. Then, in our set up, the true value of p (used in the adjusted case) is $p = (1 - P_0) + 0.89 * P_0$. We use two values in our simulations. The first one corresponds to the case where the true value is known and the second one to the case where the proportion of rich for the Singh Maddala distribution is over estimated and set to 0.95.

Let's discuss some points concerning the framework we use for our simulations.

We previously drew the attention of the reader to the fact that we choose our framework in order to illustrate a case where bootstrap fails. In fact, the performance of the other inference methods (other than bootstrap) are totally robust to the origin of the data. In particular, the accuracy and the precision of the nonparametric approaches remain whatever the framework. So does the improvement we realize relatively to the benchmark case on the precision of confidence intervals with the methods we present in the rest of the paper. Moreover, in the regular case, say the set up of Davidson and Flachaire (2004), bootstrap inference also fails when n is less than 200. This is a problem when we think that poverty studies often need to investigate the profile of populations' subgroupwhere sample sizes are most of the time very small.

Another important point is that it is not necessary to deal with a so extreme case that we choose to make bootstrap fails. In fact, the reason why bootstrap fails to give good results is that the proportion of 0 is very high in the sample of X. As a consequence, the results also apply when individuals whom revenue are set to z with probability $1 - P_0$ get revenue drawn from a distribution fonction with range $[z, +\infty)$ whereas others have revenue drawn from any continuous distribution allowing a positive probability of being poor (in order to have some interesting data). The value of P_0 can also be increased. For example with our framework, when P_0 is set to 0.2, the accuracy and the precision of the bootstrap confidence interval are bad even for n greater than 500 but smaller than 1000.

6.3 Simulations results

The results we obtain with our simulations are reported in table 2 (see appendix 1).

Table 2 shows the results we get. We see that as proven by the literature, asymptotic confidence interval is not reliable at all. Asymptotic method gives the most precise confidence interval but fails to provide an acceptable level of coverage even at $n = 1000$. This suggests that the degree of precision of the confidence interval it yields is not desirable because its cost is not acceptable.

Bootstrap inference gives better accuracy than asymptotic one but yields a level of confidence which remains far from the nominal level for sample sizes less than 1000. Moreover, its precision is very bad with a very large width of the confidence intervals.

In the opposite, finite sample nonparametric methods are strongly reliable. They provide a confidence level almost equal to 100% for all n and even for the adjusted cases. The Fishman confidence interval seems to experience problems of accuracy for n=100 but its level of confidence remains nevertheless acceptable in this case (90%).

Concerning the precision of the confidence intervals, the width of the Fishman's confidence interval is better than the continuous (conservative) Anderson's one. However, when we apply the adjustment for the discrete case, this later becomes better for all sample sizes and in particular for small samples. The width of the Anderson's confidence interval is devided by almost 5 when the true probability of being poor is used. The over estimation of p gives slightly better precision than the true value at the cost of a negligeable lower level of coverage.

The improved Anderson's confidence interval gives better results. The width of the Anderson's confidence interval is divided at least by 3 in the continuous conservative case. In addition, the discrete adjustment is once more precise. We go from a confidence interval with a width almost 20 times superior to the asymptotic one for the continuous Anderson's confidence interval to an adjusted improved confidence interval which width is about 2 times superior to the asymptotic width.This improvement is cost free, the level of coverage of the improved confidence intervals remaining close to 100 %.

The precision of the likelihood based confidence interval is also very good. The width of the confidence intervals are better than the width of the Kolmogorov based one in particular in the continuous (conservative) case. In the discrete case, the improvment is more visible for $n > 100$. However, the first improved inference method we provide seems to act better for adjusted discrete cases. Finally, for moderate sample sizes (about 1000) the improved inference methods we provide seem to perform equivalently.

To resume, we have provided highly reliable and precise confidence intervals for the mean of a bounded random variable in general and for FGT poverty measures in particular. These simulation results prove that it is essential to have finite sample nonparametric inference methods. The bad accuracy of the asymptotic and bootstrap inference would motivate people to use these proposed methods rather than the former. Moreover the improved inference methods we provide are very simple to compute with explicit solutions that doesn't depend on iterative procedures. And what is most important, they are robust to the distribution and the size of the samples. Moreover, the high level of coverage of our confidence intervals prove their reliability which is not the case with asymptotic inference.

7 Conclusion

Several outcomings can be issued from this paper.

First, we confirm that asymptotic inference provides the best possible inference precision within all the inference methods but is not reliable at all in finite sample. This is completely in accordance with the literature.

Second, we prove that standard bootstrap inference fails totally, both in accuracy and in precision, when it is performed on a variable with a high probability of being null.

These two first conclusions gives evidence for the advantage of using nonparametric inferences which are known to be totally insensitive to sample size or variables' specifications.

Third, we improve the standardized and studentized Kolmogorov statistics to get more powerful goodness of fit tests and non uniform confidence intervals for distribution functions that better fit the tails of distributions.

Fourth, we provide finite sample nonparametric methods to perform inference for the mean of a bounded random variable. We emphasize a simple projection principle which provides such inference methods using the confidence interval for the corresponding distribution function. Two confidence intervals for distribution function are used : one based on an improved studentizes kolmogorov statistic and another based on a likelihood criterion. The approaches we propose have been proven to provide very good performances.

To end, we show how all these inference methods apply to poverty measures. The performance of each method has been investigated and we emphasize the additional profit of taking into account the particular discrete specification of the random variable we use to compute the FGT poverty measure.

Monte Carlo simulations demonstrate the necessity of using finite sample nonparametric approaches. The asymptotic and bootstrap inference methods are not reliable in finite sample. On the contrary, the finite sample nonparametric inference we propose are robust to the framework and the sample size we use. Confidence intervals we get have a very good coverage probability that is always close to 100% and a good precision too. In addition, we provide explicit expressions which make them very easy to compute.

References

[1] T. W. ANDERSON. Confidence limits for the value of an arbitrary bounded random variable with a continuous distribution function. *Bulletin of The International and Statistical Institute* **43**, 249–251 (1969).

[2] T. W. ANDERSON AND D. A. DARLING. Asymptotic theory of certain goodness-of-fit criteria based on stochastic process. *Annals of Mathematical Statistics* **23**, 193–212 (1952).

[3] R. R. BAHADUR AND L. R. SAVAGE. The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics* **27**(4), 1115–1122 (1956).

[4] R. BERAN. Prepivoting test statistics : a bootstrap view of asymptotic refinements. *Journal of American Statistical Association* **83**(403), 687–697 (1988).

[5] R. H. BERK AND D. H. JONES. Relatively optimal combinations of test statistics. *Scan J Statist.* (1977).

[6] R. H. BERK AND D. H. JONES. Goodnees-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* (1979).

[7] M. BIEWEN. Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics* **108**, 317–342 (2002).

[8] J. A. BISHOP, J. P. FORMBY, AND B. ZHENG. Statistical inference and the sen index of poverty. *International Economis Review* **38**(2) (1997).

[9] K. BRACHMAN, A. STICH, AND M. TREDE. Evaluating parametric income distribution models. *Allgemeines Statistisches Archiv* **80**, 285–298 (1996).

[10] M. BRETH. Nonparametric confidence intervals for a mean using censored distribution. *Journal of Royal Statistical Society serie B* **35**, 251–254 (1976). Notes, Comments and Queries.

[11] M. BRETH, J. S. MARITZ, AND E. J. WILLIAMS. On distribution-free lower confidence limits for the mean of a nonnegative random variable. *Biometrika* **65**, 529–534 (1978).

[12] D. BURR. A comparison of certain bootstrap confidence intervals in the cox model. *Journal of American Statistical Association* **89**, 1290–1296 (1994).

[13] R. C. H. CHENG AND T. C. ILES. One-sided confidence bands for cumulative distribution function. *The Annals of Statistics* **30**(2), 155–159 (1988).

[14] W. J. CONOVER. A kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association* **67**, 591–596 (1972).

[15] F. A. COWELL AND E. FLACHAIRE. Sensitivity of inequality measures to extreme values. working paper, STICERD - Distributional Analysis Research Programme Papers 60, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE (2002).

[16] V. DARDANONI AND A. FORCINA. Inference for lorenz curve orderings. *Econometrics Journal* **2**, 49–75 (1999).

[17] R. DAVIDSON AND J.-Y. DUCLOS. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* **68**, 1435–1464 (2000).

[18] R. DAVIDSON AND E. FLACHAIRE. Asymptotic and bootstrap inference for inequality and poverty measure. working paper, EUREQUA, GREQAM, Departement of Economics Mc Gill University (2004).

[19] J. M. DUFOUR. Exact tests and confidence sets in linear regressions with autocorrelated errors. *Econometrica* **58**(2) (1990).

[20] J. M. DUFOUR. Identification, weak instruments and statistical inference in econometrics. presidential address to the canadian economics association. *Canadian Journal of Economics* **36**(4) (2003).

[21] J. M. DUFOUR AND M. NEIFAR. Méthodes d'inférence exactes pour un modèle de régression avec erreurs ar(2) gaussiennes. *L'Actualité économique* (1994).

[22] J. M. DUFOUR AND M. TAAMOUTI. Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica* (1999).

[23] J. E.FOSTER, J. GREER, AND E. THORBECKE. A class of decomposable poverty measures. *Econometrica* **52**, 761–765 (1984).

[24] F. EICKER. The asymptotic distribution of the suprema of the standardized empirical process. *The Annals of Statistics* **7**(1) (1979).

[25] G. S. FISHMAN. Confidence intervals for the mean in the bounded case. *Statistics and Probability Letters* **12**, 223–227 (1991).

[26] L. J. GLESER. Exact power of goodness-of-fit tests of komogorov type for discontinuous distributions. *Journal of the American Statistical Association* **80**(392) (1985).

[27] P. HALL. The bootstrap and edgeworth expansion. *Springer Series in statistics. New York : Springer Verlag* (1992).

[28] J. A. HORA AND S. C. HORA. Nonparametric bounds for errors in dollar unit sampling based on the cumulative distribution function. *Proceedings of the Decision Sciences Institute* pp. 115–117 (1990).

[29] A. HOROWITZ. The bootstrap. *Handbook of Econometrics* **5** (2000). J. J. HECKMAN and E. E. LEAMER (eds) Elsevier Science.

[30] D. JAESCHKE. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics* **7**(1) (1979).

[31] N. KAKWANI. Statistical inference in the measurement of poverty. *Review of Economics and Statistics* **75**, 632–639 (1993).

[32] H. M. KAPLAN. A method of one-sided nonparametric inference for the mean of a nonnegative population. *The American Statistician* **41**, 157–158 (1987).

[33] K. KNIGHT. On the bootstrap of the sample mean in the infinite variance case. *Annals of Statistics* **17**, 1168–1175 (1989).

[34] J. MILLS AND S. ZANDVAKILI. Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* **12**, 103–150 (1997).

[35] G. E. NOETHER. Note on kolmogorov statistic in the discrete case. *Metrika* **7**, 115–116 (1963).

[36] A. B. OWEN. Nonparametric likelihood confidence bands for a distribution function. *American statistical Association* **90**, 516–521 (1995).

[37] I. RONGVE. Statistical inference for poverty indices with fixed poverty lines. *Applied Economics* **29** (1997).

[38] A. SEN. Poverty : an ordinal approach to measurement. *Econometrica* **44**, 219–231 (1976).

[39] V. K. SUTTON AND D. M.YOUNG. A comparison of distribution-free confidence interval methods for estimating the mean of a random variable with bounded support. Baylor University, Department of Information Systems (1997).

[40] B. ZHENG. Statistical inference for poverty indices with relative poverty lines. *Journal of Econometrics* **101**, 337–356 (2001).

APPENDIX 1 : Tables and Figures

TABLE 1 : POWER AND LEVEL OF THE GOODNESS OF FIT TESTS FOR DIFFERENT ε

Table 1.0 : Critical points of the Statistics' Distributions

$n = 500$ $\alpha = 0.05$ $N = 1,000,000$

Epsilon	0.005	0.07	0.1	0.15	0.2	0.3	0.4
Dn_unif	0.0604	0.0603	0.0604	0.0604	0.0604	0.0604	0.0604
Eicker	4.8010	4.8004	4.7936	4.8004	4.7979	4.7965	4.7997
Reg Eicker	32.6779	2.9203	2.5837	2.2902	2.09E+00	1.79E+00	1.58E+00
A-D	3.3925	2.4711	2.6233	2.7256	2.0863	2.0586	8.5900
Reg A-D	4.2185	2.6525	2.4800	2.2515	2.06E+00	1.78E+00	1.57E+00

Epsilon	0.5	0.75	1	10	100	1000	1000000
Dn_unif	0.0603	0.0603	0.0604	0.0604	0.0604	0.0604	0.0604
Eicker	4.8019	4.7996	4.7982	4.7965	4.7901	4.7936	4.7965
Reg Eicker	1.4080	1.1114	0.9215	0.1290	1.34E-02	1.35E-03	1.35E-06
A-D	1.9852	2.4415	1.8377	2.6262	2.0124	2.8392	1.5794
Reg A-D	1.4014	1.1094	0.9204	0.1290	1.34E-02	1.35E-03	1.35E-06

Table 1.1 : $n = 500$ $N = 10,000$ $H0: \ X \sim N(0,1)$ vs $H1: X \sim N(0.1, 1.1)$

	Epsilon	0.005	0.07	0.1	0.15	0.2	0.3	0.4
Dn_unif	Puiss	63.87	64.07	64.07	64.23	64.07	63.41	63.58
	Niveau	5.1	5.0	5.1	5.2	4.8	4.8	5.0
Eicker	Puiss	5.08	5.04	4.91	4.79	5.04	4.90	5.07
	Niveau	5.1	5.0	4.8	5.0	5.0	5.1	5.1
Reg Eicker	Puiss	0.57	52.32	62.63	66.45	67.05	66.48	67.09
	Niveau	4.9	5.0	5.1	5.1	5.0	5.0	4.8
A-D	Puiss	28.57	27.75	28.18	27.95	28.15	28.53	28.49
	Niveau	4.8	5.2	5.2	5.2	5.1	5.1	4.8
Reg A-D	Puiss	67.66	84.02	81.82	80.29	78.40	74.95	74.10
	Niveau	4.7	5.1	5.0	5.0	5.0	4.8	5.0

	Epsilon	0.5	0.75	1	10	100	1000	1000000
Dn_unif	Puiss	63.57	64.46	63.71	63.73	64.15	63.41	63.85
	Niveau	5.0	5.6	4.8	5.3	4.8	4.9	4.7
Eicker	Puiss	4.36	4.96	4.47	4.84	5.08	4.87	4.85
	Niveau	5.1	5.1	5.0	5.0	4.9	5.0	4.9
Reg Eicker	Puiss	66.84	66.72	65.59	63.97	64.14	63.41	63.85
	Niveau	5.3	5.7	5.0	5.3	4.8	4.9	4.7
A-D	Puiss	28.51	28.26	28.75	28.37	29.36	28.61	28.68
	Niveau	5.5	5.1	5.1	4.7	5.4	5.2	5.1
Reg A-D	Puiss	72.58	71.09	68.95	64.35	64.16	63.41	63.85
	Niveau	5.0	5.6	5.0	5.3	4.8	4.9	4.7

Figure 1.1 : graph of table 1.1

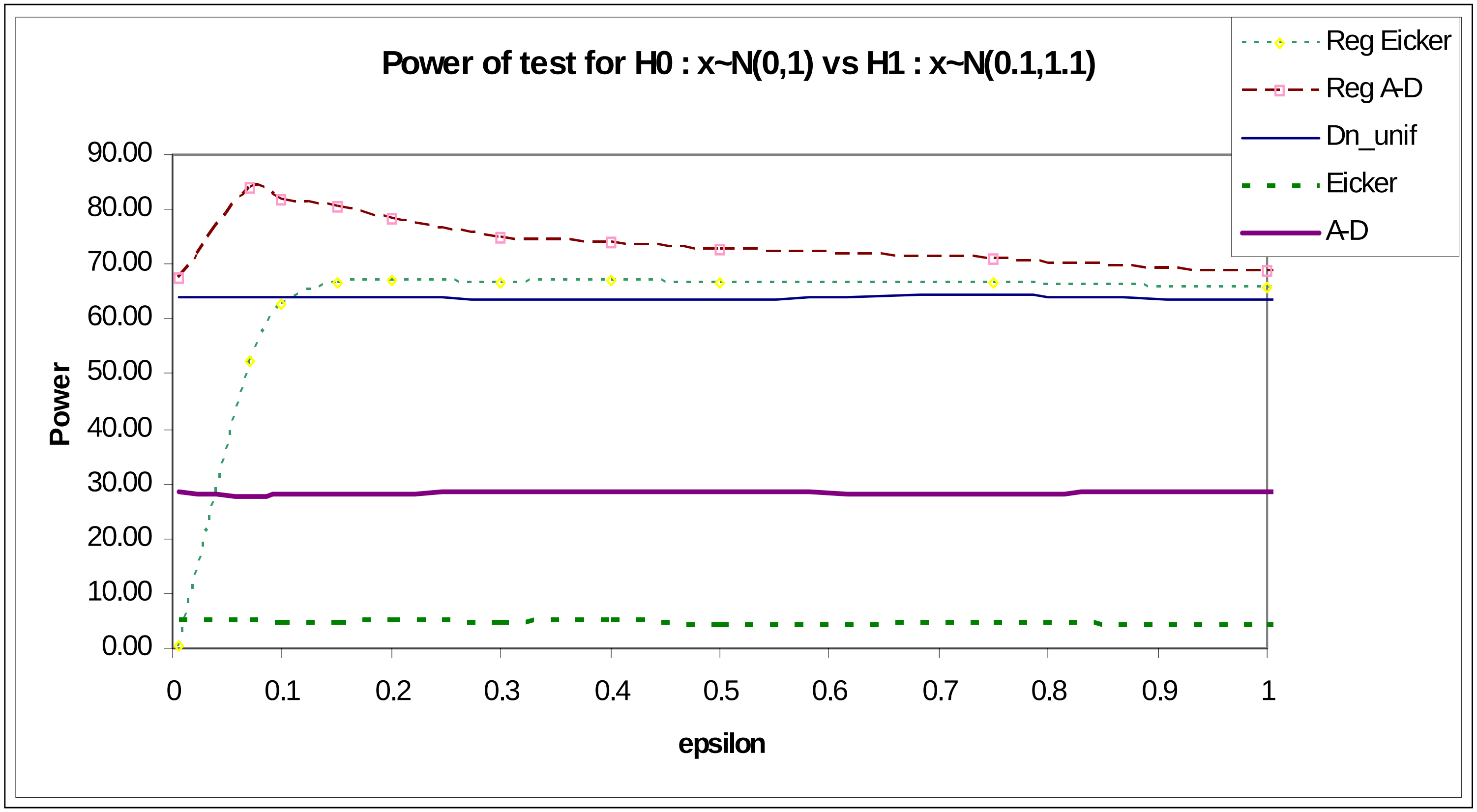


Table 1.2 : $n = 500$ $N = 10,000$ $H0: \ X \sim SM[100, 2.8, 1.7]$ vs $H1: X \sim SM[100, 2.89, 1.7]$

	Epsilon	0.005	0.07	0.1	0.15	0.2	0.3	0.4
Dn_unif	Puiss	62.51	62.27	61.82	62.23	61.94	62.10	61.73
	Niveau	4.7	5.3	5.1	5.4	5.1	4.9	4.8
Eicker	Puiss	17.67	18.27	17.71	17.70	18.20	18.46	17.39
	Niveau	5.0	4.8	4.9	5.2	5.0	5.3	4.6
Reg Eicker	Puiss	8.89	51.58	57.08	60.51	61.57	62.40	62.04
	Niveau	4.9	5.2	5.1	5.4	5.3	4.7	4.9
A-D	Puiss	3.90	4.06	3.68	4.02	3.75	3.85	3.82
	Niveau	4.9	5.6	4.9	5.0	5.1	4.8	5.2
Reg A-D	Puiss	14.29	58.22	58.44	59.69	60.14	61.27	61.27
	Niveau	4.8	5.6	5.1	5.2	5.3	4.6	5.0

	Epsilon	0.5	0.75	1	10	100	1000	1000000
Dn_unif	Puiss	63.50	62.75	61.63	62.71	63.44	61.71	62.01
	Niveau	4.9	5.0	5.1	4.7	5.4	4.9	4.7
Eicker	Puiss	17.40	17.14	16.99	17.70	18.00	18.21	17.46
	Niveau	5.0	5.0	4.8	4.9	4.9	5.0	4.9
Reg Eicker	Puiss	62.31	63.21	61.96	62.76	63.43	61.71	62.01
	Niveau	5.0	5.2	4.9	4.7	5.4	4.9	4.7
A-D	Puiss	3.60	3.94	3.85	3.67	3.78	3.98	3.95
	Niveau	5.3	4.8	5.5	5.2	4.9	4.9	4.7
Reg A-D	Puiss	62.98	62.38	61.54	62.74	63.43	61.71	62.01
	Niveau	4.9	5.2	4.9	4.7	5.4	4.9	4.7

Figure 1.2 : graph of table 1.2

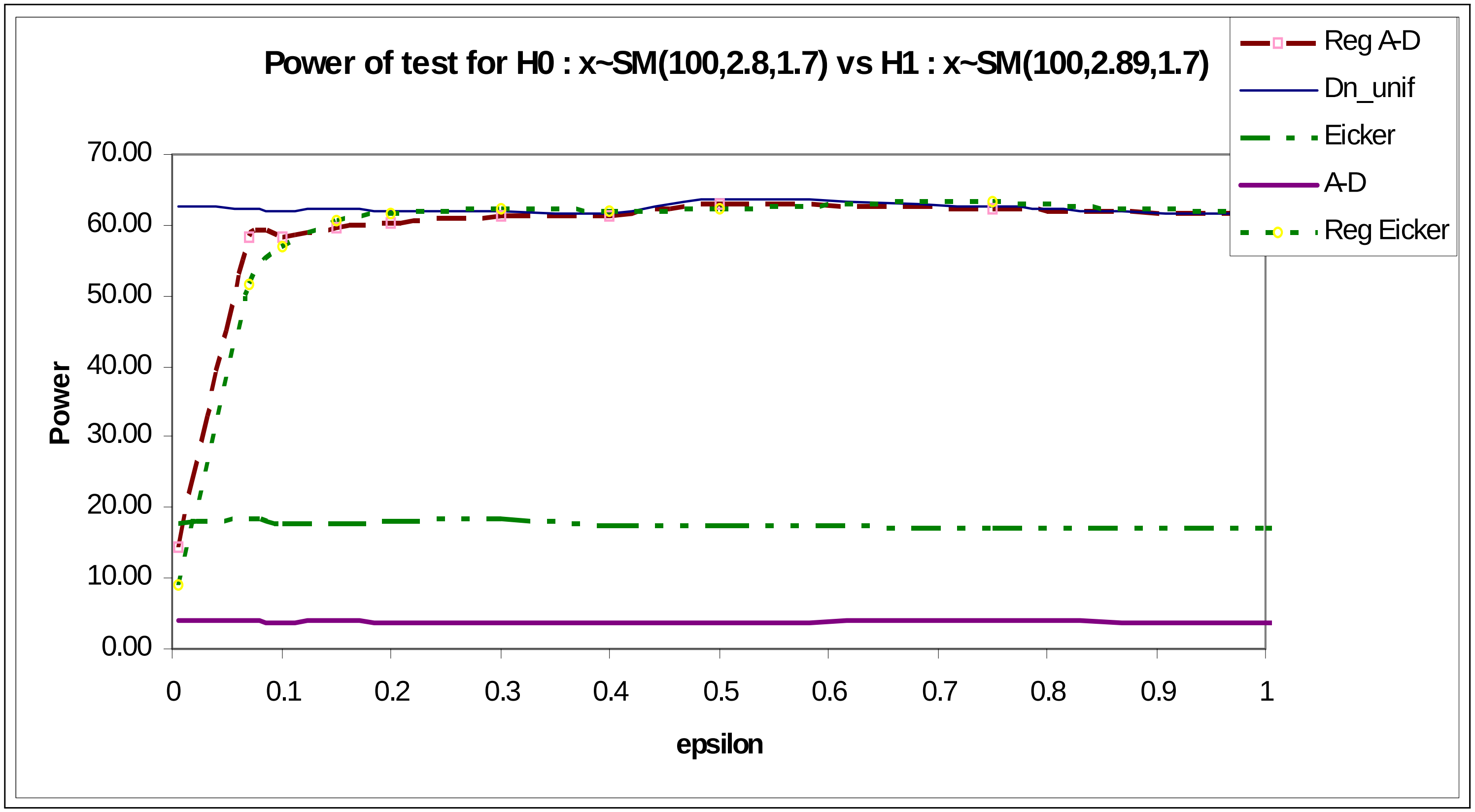


TABLE 2 : RESULTS OF MONTE CARLO SIMULATIONS

Table 2.0 : Critical points of the Statistics' Distributions

n	100	200	500	1000
Dn_unif	0.13402	0.09518	0.06041	0.04274
Dn_Pr=0.89	0.02900	0.01900	0.01100	0.00727
Dn_Pr=0.95	0.02500	0.01500	0.00700	0.00700
Reg Eicker	2.69002	2.40370	2.29023	2.26624
Reg E_Pr=0.89	0.83825	1.03709	1.03375	0.98140
Reg E_Pr=0.95	0.77982	0.73149	0.74536	0.74611
Lamdan	0.04900	0.02500	0.01050	0.00540
Lamdan_Pr=0.89	4.74E-05	3.00E-03	1.38E-03	1.93E-04
Lamdan_Pr=0.95	1.94E-03	0.00E+00	1.17E-03	8.81E-04

Table 2.1 : Level of Confidence Intervals

				Hora and Hora uniforme		
Nb pop	Asymp	Bootstrap	Fishman	continu	Pr=0.89	Pr=0.95
100	45.65%	38.91%	89.34%	100%	100%	100%
200	61.95%	62.70%	94.32%	100%	100%	100%
500	77.67%	92.07%	99.43%	100%	100%	100%
1000	85.95%	96.34%	100%	100%	100%	100%

	Dn (uniforme)			Reg Eicker (non uniforme)			Likelihood based IC		
Nb pop	continu	Pr=0.89	Pr=0.95	continu	Pr=0.89	Pr=0.95	continu	Pr=0.89	Pr=0.95
100	100%	99.91%	99.77%	100%	100%	100%	99.96%	98.68%	98.84%
200	100%	99.94%	99.78%	100%	100%	100%	100%	98.93%	98.93%
500	100%	100%	99.50%	100%	100%	100%	100%	99.86%	99.43%
1000	100%	100%	99.99%	100%	100%	100%	100%	100%	100%

Table 2.2 : Confidence Intervals

		IC asymptotique				IC bootstrap				ICFishman (91)			
Nb pop	Nb pauv	IC		Etendue	Etendue max	IC		Etendue	Etendue max	IC		Etendue	Etendue max
100	2	-8.78E-04	0.003	0.004	0.038	-1.57E-04	11.474	11.474	47212.320	0	0.041	0.041	0.077
200	3	-6.50E-04	0.003	0.004	0.023	-6.16E-06	21.938	21.938	91596.732	0	0.023	0.023	0.046
500	6	-1.98E-04	0.003	0.003	0.011	2.49E-04	133.444	133.444	1251176.522	0	0.011	0.011	0.020
1000	12	1.51E-04	0.002	0.002	0.006	4.62E-04	0.063	0.062	294.772	1.00E-04	0.007	0.007	0.012

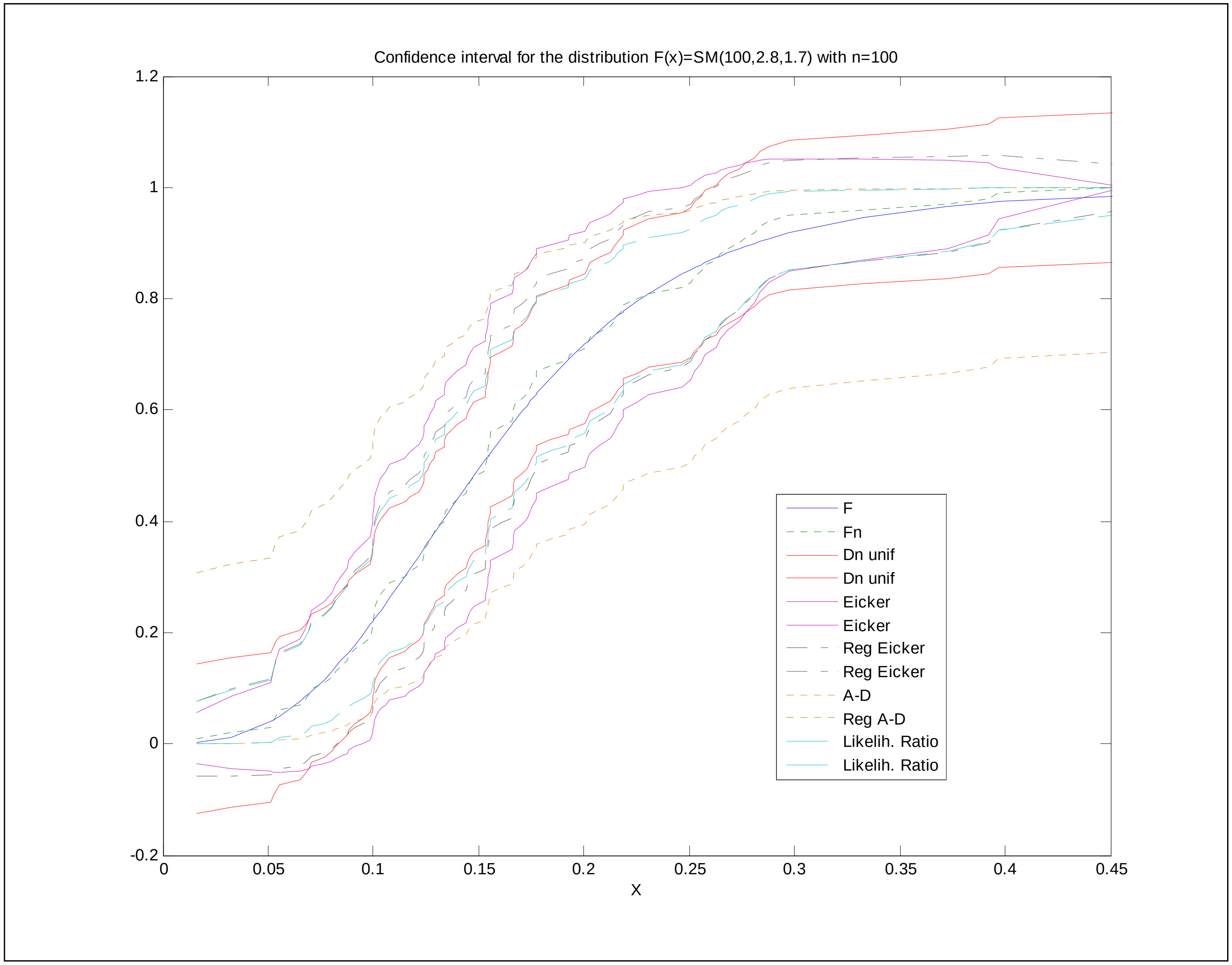
		IC Hora and Hora continu uniforme				IC Hora and Horauniforme discret Pr=0.89				IC Hora and Horauniforme discret Pr=0.95			
Nb pop	Nb pauv	IC		Etendue	Etendue max	IC		Etendue	Etendue max	IC		Etendue	Etendue max
100	2	-1.33E-01	0.135	0.268	0.268	-2.78E-02	0.030	0.058	0.058	-2.38E-02	0.026	0.050	0.050
200	3	-9.40E-02	0.096	0.190	0.190	-1.78E-02	0.020	0.038	0.038	-1.38E-02	0.016	0.030	0.030
500	6	-5.91E-02	0.062	0.121	0.121	-9.68E-03	0.012	0.022	0.022	-5.68E-03	0.008	0.014	0.014
1000	12	-4.14E-02	0.044	0.085	0.085	-5.94E-03	0.009	0.015	0.015	-5.67E-03	0.008	0.014	0.014

		IC Dn (uniforme) continu				IC Dn (uniforme) discret Pr=0.89				IC Dn (uniforme) discret Pr=0.95			
Nb pop	Nb pauv	IC		Etendue	Etendue max	IC		Etendue	Etendue max	IC		Etendue	Etendue max
100	2	0	0.135	0.135	0.148	8.77E-06	0.030	0.030	0.043	2.11E-05	0.026	0.026	0.038
200	3	0	0.096	0.096	0.105	1.25E-05	0.020	0.020	0.029	3.11E-05	0.016	0.016	0.025
500	6	0	0.062	0.062	0.067	3.04E-05	0.012	0.012	0.017	1.34E-04	0.008	0.008	0.012
1000	12	0	0.044	0.044	0.048	7.62E-05	0.009	0.009	0.011	8.45E-05	0.008	0.008	0.011

		IC Reg Eicker (non uniforme) continu				IC Reg Eicker (non uniforme) discret Pr=0.89				IC Reg Eicker (non uniforme) discret Pr=0.95			
Nb pop	Nb pauv	IC		Etendue	Etendue max	IC		Etendue	Etendue max	IC		Etendue	Etendue max
100	2	0	0.045	0.045	0.081	2.18E-06	0.015	0.015	0.035	3.11E-06	0.014	0.014	0.034
200	3	0	0.029	0.029	0.051	9.99E-07	0.013	0.013	0.028	1.10E-05	0.010	0.010	0.023
500	6	0	0.019	0.019	0.029	6.58E-06	0.009	0.009	0.017	3.66E-05	0.007	0.007	0.013
1000	12	2.6756E-10	0.014	0.014	0.020	4.02E-05	0.007	0.007	0.010	1.03E-04	0.005	0.005	0.009

		Likelihood based IC continu				Likelihood based IC discret (Pr=0.89)				Likelihood based IC discret (Pr=0.95)			
Nb pop	Nb pauv	IC		Etendue	Etendue max	IC		Etendue	Etendue max	IC		Etendue	Etendue max
100	2	1.95E-05	0.076	0.076	0.091	9.57E-05	0.033	0.033	0.042	8.88E-05	0.035	0.035	0.044
200	3	3.59E-05	0.040	0.040	0.048	1.53E-04	0.018	0.018	0.023	1.54E-04	0.018	0.018	0.022
500	6	9.32E-05	0.018	0.018	0.022	2.00E-04	0.012	0.011	0.015	2.66E-04	0.009	0.009	0.012
1000	12	1.72E-04	0.010	0.010	0.013	2.85E-04	0.007	0.007	0.009	3.37E-04	0.006	0.006	0.008

FIGURE 2 : $F(x)$ is a $SM(a,b,c)$ with $a = 100,\ b = 2.8,\ \ c = 1.7$



APPENDIX 2 : Proof of theorem 3 with the projection principle

By definition,

$$E[X] = \int_a^b x \; dF(x) = b - \int_a^b F(x) \; dx$$

So, integrating the confidence interval for F(x) provide a confidence interval for E[X], that is,

$$b - \int_a^b \max\{F_n(x) + \gamma, 1\} dx \leq E[X] \leq b - \int_a^b \max\{F_n(x) - \beta, 0\} dx \qquad (15)$$

Let r and s be defined as in the theorem 3.

- The left hand-side of (*15*) yields :

$$\begin{aligned}
E[X] &\geq b - [\int_a^{x_{(n-s)}} F_n(x) + \gamma \; dx + [b - x_{(n-s)}] * 1] \\
&\geq x_{(n-s)} - \gamma[x_{(n-s)} - a] - \frac{1}{n}\sum_{k=1}^{n-s}(k-1)[x_{(k)} - x_{(k-1)}] \\
&\geq x_{(n-s)} - \gamma[x_{(n-s)} - a] - \frac{1}{n}\sum_{k=1}^{n-s}(k-1)x_{(k)} + \frac{1}{n}\sum_{k=0}^{n-s-1} k x_{(k)} \\
&\geq x_{(n-s)} - \gamma[x_{(n-s)} - a] + \frac{1}{n}\sum_{k=1}^{n-s-1} x_{(k)} + \frac{1}{n}x_{(n-s)} - \frac{1}{n}(n-s)x_{(n-s)} \\
&\geq \frac{1}{n}[\sum_{k=1}^{n-s-1} x_{(k)} + (n+1-(n-s))x_{(n-s)}] - \gamma[x_{(n-s)} - a] \\
&\geq \frac{1}{n}[\sum_{k=1}^{n-s-1} x_{(k)} + (s+1)x_{(n-s)}] - \gamma[x_{(n-s)} - a]
\end{aligned}$$

This last inequality is indeed the left hand-side inequality in theorem 3.

- The right hand-side of (*15*) yields :

$$
\begin{aligned}
E[X] &\leq b - 0 * [x_{(r+1)} - a] - \int_{x_{(r+1)}}^{b} F_n(x) - \beta \; dx \\
&\leq b + \beta[b - x_{(r+1)}] - \frac{1}{n} \sum_{k=r+2}^{n+1} (k-1)[x_{(k)} - x_{(k-1)} \\
&\leq b + \beta[b - x_{(r+1)]} - \frac{1}{n} \sum_{k=r+2}^{n+1} (k-1)x_{(k)} + \frac{1}{n} \sum_{k=r+1}^{n} kx_{(k)} \\
&\leq b + \beta[b - x_{(r+1)}] - \frac{1}{n} \sum_{k=r+2}^{n} x_{(k)} + \frac{b}{n} - \frac{(r+1)x_{(r+1)}}{n} - \frac{n+1}{n} b \\
&\leq \frac{1}{n}[(r+1)x_{(r+1)} + \sum_{k=r+2}^{n} x_{(k)}] + \beta[b - x_{(r+1)}]
\end{aligned}
$$

Again, this last inequality is similar to the right hand-side one of theorem 3.

APPENDIX 3 : Proof of Corollary 3

Remind from the last section that $X = (\frac{z-Y}{z})^{\alpha}\ \mathrm{I}_{[0\leq Y\leq z]}$ is a random variable with a probability mass of $G_X(0) = 1 - F_Y(z)$ at point 0 and distribution function

$$G_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - F_Y[z(1 - x^{1/\alpha})] & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Actually, X follows a mixing of a probability mass at 0 and a continuous distribution function. Its distribution can be rewritten

$$G_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - F_Y(z) \ \equiv p & \text{if } x = 0 \\ p + \int_0^x h(u)\ du & \text{if x} > 0 \end{cases}$$

with some adequate density h[1]. It follows that $G_X(X)$ is also a mixing :

$$\begin{aligned} G_X(X) &= p & \text{with probability p} \\ &\sim U_{(p,1]} & \text{with probability 1-p} \end{aligned}$$

The Kolmogrov statistic corresponding to this special case is

$$\begin{aligned} D_n &= \max_{0\leq x\leq 1} \mid G_n(x) - G(x) \mid = \max\{|\widehat{p} - p|, \max_{0<x\leq 1} \mid G_n(x) - G(x) \mid\} \\ &= \max\{|\widehat{p} - p|, \max_{0<x\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} I_{[x_k\leq x]} - G(x) \mid\} \\ &= \max\{|\widehat{p} - p|, \max_{0<x\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} I_{[G(x_k)\leq G(x)]} - G(x) \mid\} \\ &= \max\{|\widehat{p} - p|, \max_{p<v\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} I_{[G(x_k)\leq v]} - v \mid\} \\ &= \max_{p\leq v\leq 1} \left| \frac{1}{n}\sum_{k=1}^{n} I_{[G[x_k]\leq v)]} - v \right| \end{aligned}$$

[1] Actually, h is the density of the random variable $T = (\frac{z-Z}{z})^{\alpha}$ where Z is a truncation of Y on [0,z]

APPENDIX 4 : Proof of theorem 5

Suppose that we are under the hypotheses of theorem 5. Then, given c_E such that $Pr[E_n \leq c_{RE}] = 1-\alpha$ proposition 5 gives us a $1-\alpha$ confidence interval for F(x) :

$$F_n(x) - \frac{c_E}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} \leq F(x) \leq F_n(x) + \frac{c_E}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} \tag{16}$$

- The left hand-side of (*16*) yields for $k = 0, ..., n+1$:

$$\begin{aligned} F_n(x_{(k)}) - \frac{c_E}{\sqrt{n}} F_n^{1/2}(x_{(k)})[1-F_n(x_{(k)})]^{1/2} &\leq F(x_{(k)}) \\ or\ \frac{k}{n} - \frac{c_E}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} &\leq F(x_{(k)}) \end{aligned} \tag{17}$$

  Given that $F(x_{(k)}) \geq 0\ \forall k$, this bound is effective only for k such that $g(k) \geq 0$ where $g(k)$ represents the lower bound defined above i.e.

$$\begin{aligned} &\frac{k}{n} - \frac{c_E}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \geq 0 \Leftrightarrow k \geq c_E\sqrt{n}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \\ &\Leftrightarrow k^2 \geq c_E^2 n\frac{k}{n}[1-\frac{k}{n}] \Leftrightarrow (n+c_E^2)k^2 - nc_E^2 k \geq 0 \\ &\Rightarrow (n+c_E^2)k - nc_E^2 \geq 0 \Rightarrow k \geq \frac{nc_E^2}{(n+c_E^2)} = k_E^{*I} \end{aligned}$$

  Then, $g(k) \geq 0$ for $k \geq k_E^{*I}$ i.e $k = k_E^I + 1, ..., n$ with $k_E^I = [k_E^{*I}]$, the integer part of k_E^{*I}

The distribution which satisfies (*17*) with the largest mean is the distribution with jumps of

1. $g(k_E^I + 1) = \frac{k_E^I+1}{n} - \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k_E^I+1}{n}\right)^{1/2}\left(1-\frac{k_E^I+1}{n}\right)^{1/2}\right\}$ at $x_{(k_E^I+1)}$
2. $g(k) - g(k-1) = \frac{1}{n} - \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}$ at $x_{(k)}$ for $k = k_E^I + 2, ..., n$
3. $1 - g(n) = 1 - \{1 - \frac{c_E}{\sqrt{n}}[\left(\frac{n}{n}\right)^{1/2}\left(1-\frac{n}{n}\right)^{1/2}]\} = 0$ at b

This yields

$$E(x) \leq \left[\frac{k_E^I+1}{n} - \frac{c_E}{\sqrt{n}}\left\{\left(\frac{k_E^I+1}{n}\right)^{1/2}\left(1-\frac{k_E^I+1}{n}\right)^{1/2}\right\}\right] * x_{(k_E^I+1)}$$

$$+\sum_{k=k_E^I+2}^{n}\left[\frac{1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}\right] * x_{(k)}$$

- The right hand-side of (*16*) yields for $k = 0, ..., n+1$:

$$\begin{aligned} F(x_{(k)}) &\leq F_n(x_{(k)}) + \frac{c_E}{\sqrt{n}} F_n^{1/2}(x_{(k)})[1 - F_n(x_{(k)})]^{1/2} \\ or\ F(x_{(k)}) &\leq \frac{k}{n} + \frac{c_E}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1 - \frac{k}{n}]^{1/2} \end{aligned} \tag{18}$$

Given that $F(x_{(k)}) \leq 1\ \forall k$, this bound is effective only for k such that $h(k) \leq 1$ where $h(k)$ represents the upper bound defined above i.e.

$$\frac{k}{n} + \frac{c_E}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1 - \frac{k}{n}]^{1/2} \leq 1 \Leftrightarrow \frac{c_E}{n^{3/2}}[k]^{1/2}[n-k]^{1/2} \leq [1 - \frac{k}{n}]$$
$$\Rightarrow c_E^2 k[n-k] \leq n(n^2 - 2nk + k^2)$$
$$\Rightarrow (n + c_E^2)k^2 - n(2n + c_E^2)k + n^3 \geq 0$$

Let

$$\begin{aligned} \Delta_E^S &= n^2(2n + c_E^2)^2 - 4n^3(n + c_E^2) = n^2(4n^2 + 4nc_E^2 + c_E^4) - 4n^4 + 4n^3 c_E^2 \\ &= n^2 c_E^4 \end{aligned}$$

The roots of the equations $(n + c_E^2)k^2 - n(2n + c_E^2)k + n^3 = 0$ are

$$k_{E1}^{*S} = \frac{n^2}{(n + c_E^2)} \text{ and } k_{E2}^{*S} = \frac{2n^2 + 2nc_E^2}{2(n + c_E^2)} = n$$

and $h(k) \leq 1$ for $k \leq k_{E1}^{*S}$ i.e $k = 0, ..., k_E^S$ with $k_E^S = [k_{E1}^{*S}]$, the integer part of k_{E1}^{*S} (Note that $h(k_E^S) < 1$).

The distribution which satisfies (*18*) with the smallest mean is the distribution with jumps of

1. $h(0) = 0$ at a
2. $h(k) - h(k-1) = \frac{1}{n} + \frac{c_E}{\sqrt{n}} \left\{ \left(\frac{k}{n}\right)^{1/2} \left(1 - \frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2} \left(1 - \frac{k-1}{n}\right)^{1/2} \right\}$ at $x_{(k)}$ for $k = 1, ..., k_E^S$
3. $1 - h(k_E^S) = 1 - \frac{k_E^S}{n} - \frac{c_E}{\sqrt{n}} \left\{ \left(\frac{k_E^S}{n}\right)^{1/2} \left(1 - \frac{k_E^S}{n}\right)^{1/2} \right\}$ at $x_{(k_E^S+1)}$

This yields

$$E(x) \geq \sum_{k=1}^{k_E^S} \left[\frac{1}{n} + \frac{c_E}{\sqrt{n}} \left\{ \left(\frac{k}{n}\right)^{1/2} \left(1 - \frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2} \left(1 - \frac{k-1}{n}\right)^{1/2} \right\} \right] * x_{(k)}$$
$$+ \left[1 - \frac{k_E^S}{n} - \frac{c_E}{\sqrt{n}} \left\{ \left(\frac{k_E^S}{n}\right)^{1/2} \left(1 - \frac{k_E^S}{n}\right)^{1/2} + \varepsilon \right\} \right] * x_{(k_E^S+1)}$$

APPENDIX 5 : Proof of theorem 6

Suppose that we are under the hypotheses of theorem 6. Then, given c_{RE} such that $Pr[E_n^R \leq c_{RE}] = 1-\alpha$ proposition 6 gives us a $1-\alpha$ confidence interval for F(x) :

$$F_n(x) - \frac{c_{RE}}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} - \frac{c_{RE}}{\sqrt{n}}\varepsilon \leq F(x) \leq F_n(x) + \frac{c_{RE}}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} + \frac{c_{RE}}{\sqrt{n}}\varepsilon \tag{19}$$

- The left hand-side of (*19*) yields for $k = 0, ..., n+1$:

$$\begin{aligned} F_n(x_{(k)}) - \frac{c_{RE}}{\sqrt{n}} F_n^{1/2}(x_{(k)})[1-F_n(x_{(k)})]^{1/2} - \frac{c_{RE}\varepsilon}{\sqrt{n}} &\leq F(x_{(k)}) \\ or\ \frac{k}{n} - \frac{c_{RE}}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} - \frac{c_{RE}\varepsilon}{\sqrt{n}} &\leq F(x_{(k)}) \end{aligned} \tag{20}$$

Given that $F(x_{(k)}) \geq 0\ \forall k$, this bound is effective only for k such that $g(k) \geq 0$ where $g(k)$ represents the lower bound defined above i.e.

$$\begin{aligned} k\sqrt{n} - c_{RE}k^{1/2}[n-k]^{1/2} - c_{RE}\varepsilon n \geq 0 &\Leftrightarrow (k\sqrt{n} - c_{RE}\varepsilon n)^2 \geq c_{RE}^2 k(n-k) \\ &\Leftrightarrow k^2 n - 2kc_{RE}\varepsilon n^{3/2} + c_{RE}^2\varepsilon^2 n^2 \geq c_{RE}^2 kn - c_{RE}^2 k^2 \\ &\Leftrightarrow k^2(n+c_{RE}^2) - k(2c_{RE}\varepsilon n^{3/2} + c_{RE}^2 n) + c_{RE}^2\varepsilon^2 n^2 \geq 0 \end{aligned}$$

Let

$$\begin{aligned} \Delta^I &= (2c_{RE}\varepsilon n^{3/2} + c_{RE}^2 n)^2 - 4c_{RE}^2\varepsilon^2 n^2(n+c_{RE}^2) \\ &= 4c_{RE}^2\varepsilon^2 n^3 + 4c_{RE}^3\varepsilon n^{5/2} + c_{RE}^4 n^2 - 4c_{RE}^2\varepsilon^2 n^3 - 4c_{RE}^4\varepsilon^2 n^2 \\ &= 4c_{RE}^3\varepsilon n^{5/2} + c_{RE}^4 n^2 - 4c_{RE}^4\varepsilon^2 n^2 = c_{RE}^3 n^2[c_{RE} - 4c_{RE}\varepsilon^2 + 4\varepsilon n^{1/2}] \end{aligned}$$

Then, the roots of the equations $k^2(n+c_{RE}^2) - k(2c_{RE}\varepsilon n^{3/2} + c_{RE}^2 n) + c_{RE}^2\varepsilon^2 n^2 = 0$ are

$$k_{RE1}^{*I} = \frac{2c_{RE}\varepsilon n^{3/2} + c_{RE}^2 n - \sqrt{\Delta^I}}{2(c_{RE}^2+n)} \text{ and } k_{RE2}^{*I} = \frac{2c_{RE}\varepsilon n^{3/2} + c_{RE}^2 n + \sqrt{\Delta^I}}{2(c_{RE}^2+n)}$$

and $g(k) \geq 0$ for $k \geq k_{RE2}^{*I}$ i.e $k = k_{RE}^I + 1, ..., n$ with $k_{RE}^I = [k_{RE2}^{*I}]$, the integer part of k_{RE2}^{*I}

The distribution which satisfies (*20*) with the largest mean is the distribution with jumps of

1. $g(k_{RE}^I+1) = \frac{k_{RE}^I+1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k_{RE}^I+1}{n}\right)^{1/2}\left(1-\frac{k_{RE}^I+1}{n}\right)^{1/2} + \varepsilon\right\}$ at $x_{(k_{RE}^I+1)}$
2. $g(k) - g(k-1) = \frac{1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}$ at $x_{(k)}$ for $k = k_{RE}^I + 2, ..., n$
3. $1 - g(n) = 1 - \{1 - \frac{c_{RE}}{\sqrt{n}}[\left(\frac{n}{n}\right)^{1/2}\left(1-\frac{n}{n}\right)^{1/2} + \varepsilon]\} = \frac{c_{RE}\varepsilon}{\sqrt{n}}$ at b

This yields

$$E(x) \le \left[\frac{k^I_{RE}+1}{n} - \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k^I_{RE}+1}{n}\right)^{1/2}\left(1-\frac{k^I_{RE}+1}{n}\right)^{1/2}+\varepsilon\right\}\right]*x_{(k^I_{RE}+1)}+\frac{c_{RE}\varepsilon}{\sqrt{n}}*b$$
$$+\sum_{k=k^I_{RE}+2}^{n}\left[\frac{1}{n}-\frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2}-\left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}\right]*x_{(k)}$$

- The right hand-side of (*19*) yields for $k = 0, ..., n+1$:

$$\begin{aligned} F(x_{(k)}) &\le F_n(x_{(k)}) + \frac{c_{RE}}{\sqrt{n}}F_n^{1/2}(x_{(k)})[1-F_n(x_{(k)})]^{1/2} + \frac{c_{RE}\varepsilon}{\sqrt{n}} \\ \textit{or } F(x_{(k)}) &\le \frac{k}{n} + \frac{c_{RE}}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} + \frac{c_{RE}\varepsilon}{\sqrt{n}} \qquad (21)\end{aligned}$$

Given that $F(x_{(k)}) \le 1\ \forall k$, this bound is effective only for k such that $h(k) \le 1$ where $h(k)$ represents the upper bound defined above i.e.

$$k\sqrt{n} + c_{RE}k^{1/2}[n-k]^{1/2} + c_{RE}\varepsilon n \le n^{3/2} \Leftrightarrow n(k + c_{RE}\varepsilon\sqrt{n} - n)^2 \ge c^2_{RE}k(n-k)$$
$$\Leftrightarrow k^2 n + c^2_{RE}\varepsilon^2 n^2 + n^3 + 2kc_{RE}\varepsilon n^{3/2} - 2kn^2 - 2c_{RE}\varepsilon n^{3/2} \ge c^2_{RE}kn - c^2_{RE}k^2$$
$$\Leftrightarrow k^2(n + c^2_{RE}) + kn(2c_{RE}\varepsilon\sqrt{n} - 2n - c^2_{RE}) + n^2(c^2_{RE}\varepsilon^2 - 2c_{RE}\varepsilon\sqrt{n} + n) \ge 0$$

Let

$$\begin{aligned}\Delta^S &= n^2(2c_{RE}\varepsilon\sqrt{n} - 2n - c^2_{RE})^2 - 4n^2(c^2_{RE}\varepsilon^2 - 2c_{RE}\varepsilon\sqrt{n} + n)(n + c^2_{RE}) \\ &= 4c^2_{RE}\varepsilon^2 n^3 + 4n^4 + c^4_{RE}n^2 - 8c_{RE}\varepsilon n^{7/2} - 4c^3_{RE}\varepsilon n^{5/2} + 4c^2_{RE}n^3 \\ &\quad -4c^2_{RE}\varepsilon^2 n^3 - 4c^4_{RE}\varepsilon^2 n^2 + 8c_{RE}\varepsilon n^{7/2} + 8c^3_{RE}\varepsilon n^{5/2} - 4n^4 - 4c^2_{RE}n^3 \\ &= c^4_{RE}n^2 - 4c^4_{RE}\varepsilon^2 n^2 + 4c^3_{RE}\varepsilon n^{5/2} = c^3_{RE}n[c_{RE}n - 4c_{RE}n\varepsilon^2 + 4\varepsilon n^{3/2}]\end{aligned}$$

Then, the roots of the equations $k^2(n + c^2_{RE}) + kn(2c_{RE}\varepsilon\sqrt{n} - 2n - c^2_{RE}) + n^2(c^2_{RE}\varepsilon^2 - 2c_{RE}\varepsilon\sqrt{n} + n) = 0$ are

$$k^{*S}_{RE1} = \frac{c^2_{RE}n + 2n^2 - 2c_{RE}\varepsilon n^{3/2} - \sqrt{\Delta^S}}{2(c^2_{RE}+n)} \text{ and } k^{*S}_{RE2} = \frac{c^2_{RE}n + 2n^2 - 2c_{RE}\varepsilon n^{3/2} + \sqrt{\Delta^S}}{2(c^2_{RE}+n)}$$

and $h(k) \le 1$ for $k \le k^{*S}_{RE1}$ i.e $k = 0, ..., k^S_{RE}$ with $k^S_{RE} = [k^{*S}_{RE1}]$, the integer part of k^{*S}_{RE1} (Note that $h(k^S_E) < 1$).

The distribution which satisfies (*21*) with the smallest mean is the distribution with jumps of

1. $h(0) = \frac{c_{RE}\varepsilon}{\sqrt{n}}$ at a
2. $h(k) - h(k-1) = \frac{1}{n} + \frac{c_{RE}}{\sqrt{n}}\left\{\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right\}$ at $x_{(k)}$ for $k = 1, ..., k^S_{RE} - 1$

3. $1 - h(k_{RE}^S) = 1 - \frac{k_{RE}^S}{n} - \frac{c_{RE}}{\sqrt{n}} \left\{ \left(\frac{k_{RE}^S}{n}\right)^{1/2} \left(1 - \frac{k_{RE}^S}{n}\right)^{1/2} + \varepsilon \right\}$ at $x_{(k_{RE}^S)}$

This yields

$$E(x) \geq \frac{c_{RE}\varepsilon}{\sqrt{n}} * a + \left[1 - \frac{k_{RE}^S}{n} - \frac{c_{RE}}{\sqrt{n}} \left\{ \left(\frac{k_{RE}^S}{n}\right)^{1/2} \left(1 - \frac{k_{RE}^S}{n}\right)^{1/2} + \varepsilon \right\}\right] * x_{(k_{RE}^S+1)}$$

$$+ \sum_{k=1}^{k_{RE}^S} \left[\frac{1}{n} + \frac{c_{RE}}{\sqrt{n}} \left\{ \left(\frac{k}{n}\right)^{1/2} \left(1 - \frac{k}{n}\right)^{1/2} - \left(\frac{k-1}{n}\right)^{1/2} \left(1 - \frac{k-1}{n}\right)^{1/2} \right\}\right] * x_{(k)}$$