Université de Montréal

Analyse statistique de la pauvreté et des inégalités

par

Mame Astou Diouf

Département de sciences économiques
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae Doctor (Ph.D.)
en sciences économiques

Mai 2008

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée:

Analyse statistique de la pauvreté et des inégalités

présentée par

Mame Astou Diouf

a été évaluée par un jury composé des personnes suivantes:

Prof. Yves Sprumont : président-rapporteur

Prof. Jean-Marie Dufour : directeur de recherche

Prof. Silvia Gonçalves : membre du jury

Prof. Russell Davidson : examinateur externe (McGill University)

Mme Guylaine Racine : représentant du doyen de la FES (École de service social)

Sommaire

Malgré les larges écart-types estimés dans plusieurs études de pauvreté et d'inégalités empiriques, la plupart des études dans ce domaine n'ont pas recours à l'inférence statistique. Deux types d'inférence sont généralement utilisés pour les mesures de pauvreté et d'inégalités: les distributions asymptotiques et le bootstrap. Bien que ces méthodes puissent ne pas être toujours fiables, aucune étude n'a encore proposé de méthode d'inférence exacte valide pour de tels problèmes. Nous proposons de telles méthodes.

Dans le premier article, nous construisons des bandes de confiance pour des fonctions de distribution en inversant des tests d'adéquation basés sur des statistiques de Kolmogorov-Smirnov (KS) standardisées et améliorées. Le test de KS, bien que populaire, ne permet pas de discriminer grandement entre les distributions qui diffèrent le plus dans les queues. Pour corriger ce problème, des statistiques de KS pondérées basées sur les principes de Wald, du multiplicateur de Lagrange et du ratio de vraisemblance ont été proposées respectivement par Anderson et Darling (1952), Eicker (1979) et Berk et Jones (1979). Toutefois, ces dernières souffrent de problèmes dus à leurs dénominateurs qui peuvent être proches de zéro. Pour y remédier, nous proposons des statistiques de KS améliorées obtenues en ajoutant un terme de régularisation au dénominateur des statistiques d'Anderson-Darling et d'Eicker. Nous en déduisons des bandes de confiance exactes pour les fonctions de distribution et montrons que, dans le cas continu, ces bandes de confiance sont indépendantes de la distribution testée sous l'hypothèse nulle et qu'elles sont conservatrices dans le cas non continu tout en bénéficiant de propriétés de monotonicité qui améliorent les bandes de confiance sans altérer leur fiabilité.

Dans les deuxième et troisièmes articles, nous proposons des intervalles de confiance exacts pour les mesures de pauvreté de Foster, Greer et Thorbecke (FGT, 1984) et les mesures d'inégalités les plus populaires, respectivement. Nous observons d'abord que ces mesures peuvent se réécrire comme des fonctions de moyennes de variables aléatoires, ces dernières étant elles-mêmes des fonctionnelles de fonctions de distribution de variables bornées et non bornées. Ensuite, nous utilisons des techniques de projection pour déduire des intervalles de confiance à distance finie pour la moyenne d'une variable aléatoire

bornée à partir de bandes de confiance de la fonction de distribution sous-jacente. Lorsque la variable aléatoire n'est pas bornée, nous proposons un principe de projection généralisé qui s'applique aux fonctions de distributions dont les queues sont bornées par des lois de Pareto. Enfin, nous appliquons ces procédures aux mesures de pauvreté FGT et aux mesures d'inégalités (les mesures d'entropie généralisée, de déviation logarithmique et d'Atkinson et les indices de Theil, de Lorenz, de Gini et de variation logarithmique).
Dans les trois articles, des études Monte Carlo sont effectuées pour analyser la performance des méthodes d'inférence et illustrer le choix du paramètre de régularisation. Elles montrent que les statistiques régularisées donnent des tests plus puissants que celles existantes, lorsqu'elles sont appliquées à des distributions qui diffèrent le plus dans les queues. De même, les bandes de confiance de fonctions de distribution et les intervalles de confiance pour la moyenne basés sur ces statistiques produisent de meilleurs résultats. Dans certains cas, les intervalles asymptotique et bootstrap ne produisent pas de résultats fiables alors que les intervalles proposées sont robustes et plus courts. Pour illustration, nous analysons dans les articles 2 et 3 les profils de pauvreté et d'inégalités des ménages ruraux au Mexique en 1998 en utilisant des données du programme PROGRESA. Les résultats montrent que les intervalles asymptotiques sont souvent trop petits pour être réalistes alors que l'intervalle bootstrap peut exploser. L'analyse montre que le profil de pauvreté des ménages Mexicains dépend grandement du type de chef de ménage: les niveaux de pauvreté et d'inégalités des ménages dont le chef est un homme ou est éduqué sont moins élevés que ceux des autres ménages. De ce fait, les mesures destinées à réduire le taux d'illettrisme et à sécuriser le revenu des ménages dont le chef est une femme pourraient aider à réduire la pauvreté et les inégalités dans le Mexique rural.

Mots clés : inférence exacte ; Kolmogorov-Smirnov ; Anderson-Darling ; Eicker ; pauvreté ; inégalité ; moyenne ; régularisation; distribution de Pareto.
JEL codes: C01, C12, C14, O11.

Summary

Despite the growing interest in poverty and inequality studies and the large standard errors found in many empirical studies, most of the work in this area neglects statistical inference. Two types of inference procedures for poverty and inequality measures have been considered: asymptotic distributions and bootstrapping. These methods can be quite unreliable, even with fairly large samples, but no study has proposed provably valid exact inference procedures for such problems. We propose such ones.

In the first paper, we build nonparametric confidence bands for distribution functions by inverting goodness-of-fit tests based on improved standardized Kolmogorov-Smirnov statistics (KS, henceforth). Despite its popularity, the KS test does not allow to discriminate a lot between distributions that differ mostly through their tails. To correct this drawback, weighted KS statistics based on the three common principles in econometrics (the Wald, Lagrange multiplier, and likelihood-ratio principles) are proposed respectively by Anderson and Darling (1952), Eicker (1979), and Berk and Jones (1979). However, they also suffer from drawbacks because standard errors can be very close to zero. To correct these, we propose improved weighted KS statistics obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics and derive from them exact nonparametric confidence bands for distribution functions. We show that in the continuous case, these confidence bands are independent of the distribution assumed under the null hypothesis and are conservative for noncontinuous distributions. In the noncontinuous case, we derive monotonicity properties to narrow the confidence bands without altering their reliability.

In the second and third papers, we develop such inference methods for the Foster, Greer and Thorbecke (FGT, 1984) poverty measures (paper 2) and the most popular inequality measures (paper 3): the generalized entropy measures, the Theil index, the Lorenz curve, the Gini index, the Atkinson measures, the mean logarithmic deviation, and the logarithmic variation. We first observe that these poverty and inequality indicators can be interpreted as functions of the expectations of random variables which are themselves functional of distribution functions, where the involved variables can be either bounded

or unbounded. Using projection techniques, we then derive finite-sample nonparametric confidence intervals for the mean of a bounded random variable from confidence bands for the distribution of the underlying variable. When the random variable is unbounded, we propose a generalized projection principle for distribution functions which tails are bounded by a Pareto distribution. Then, we apply these procedures to the FGT poverty measures and to inequality measures.

Monte Carlo simulations are performed in the three papers to study the relative performance of the inference methods and illustrate how to choose the regularization parameter. The results show that the regularized statistics yield more powerful goodness-of-fit tests than the existing ones when applied to distributions with more discrepancy in the tails. Likewise, the CBs for distribution functions and the confidence intervals based on these regularized statistics have a better performance. The simulations show that asymptotic and bootstrap confidence intervals for the mean can fail to provide reliable inference, while the proposed methods are robust and yield shorter confidence intervals. As an illustration, we analyze the profile of poverty and inequality of Mexico in 1998 using households' survey data (papers 2 and 3). The results show that the widths of the asymptotic confidence intervals are often too small to be realistic while those of the bootstrap can be 10 times larger than the widths delivered by exact methods. The study shows that the poverty profile of Mexican households depends greatly on the type of households' head: poverty levels and inequality among households with a male head or an educated head are much smaller than those among other households. Hence, policies aimed at reducing illiteracy and at securing the income of households with a female head could help reduce poverty and inequality in rural Mexico.

Keywords : nonparametric inference; Kolmogorov-Smirnov; Anderson-Darling; Eicker; empirical distribution; mean; poverty; inequality; regularization; Paretian heavy tail.
JEL codes: C01, C12, C14, O11.

Table des matières

Liste des tableaux

Liste des graphiques

A mon père: Gana Diouf. Je sais que tu apprécies cette thèse à sa juste valeur.

A ma mère: Yaye Awa N'Diongue. Sans toi, rien de tout cela n'aurait été possible.

A mes soeurs chéries: Diago, Fatoumata Binetou, Awa et la toute choyée Oumy Khairy. Vous me manquez tellement.

Je vous adore tous.

Remerciements

Je remercie du fond du coeur mon directeur de thèse Jean-Marie Dufour pour son encadrement, sa patience et son soutien à la fois moral et financier tout au cours de ces longues années.

Je remercie le personnel administratif du département de sciences économiques de l'Université de Montréal et du CIREQ pour toute l'aide apportée durant la préparation de ce doctorat. Je remercie le département de sciences économiques de l'Université de Montréal, le CIREQ, la Chaire de Recherche du Canada en économètrie (Jean-Marie Dufour) et la Chaire William Dow en économie politique (Université McGill) pour le soutien financier alloué.

Je remercie chaleureusement Nour Meddahi pour m'avoir aguillé durant les premières années si importantes de ce processus. Je remercie également tout le corps professoral du doctorat de sciences économiques de l'Université de Montréal pour la formation apportée et leur aide à divers moments.

Merci également à toutes les personnes qui ont de près ou de loin aidé à la réalisation de ce projet si cher. Merci à Peter Lanjouw pour m'avoir fourni gracieusement les données utilisés pour l'application empirique (voir page 110). Merci à Albert Touna Mama pour son aide très apprécié. Merci à un collègue du Fonds Monétaire International qui se reconnaitra, pour ses encouragements aux premières aurores de chaque journée. Un grand merci à une personne spéciale, une soeur de coeur: Kadiata Kane, pour avoir été là aux moments cruciaux.

J'envoie mes sincères remerciements à mes amis et aux membres de ma famille, petite et grande, pour leur soutien, leur aide et pour avoir cru en moi.

Introduction Générale

Durant les dernières décennies, il y eut un intérêt croissant pour les études de pauvreté et d'inégalités. Toutefois, en dépit des larges écart-types trouvés dans les études empiriques, la plupart des analyses dans ce domaine sont restées descriptives, ne procédant pas à une inférence statistique rigoureuse. Deux types de procédures inférentielles ont été proposés: les distributions asymptotiques et le bootstrap; voir Beran (1988), Kakwani (1993), Rongve (1997), Mills et Zandvakili (1997), Dardanoni et Forcina (1999), Biewen (2002), Davidson et Duclos (2000), Zheng (2001) et Davidson et Flachaire (2007). La plupart de ces études recommandent l'utilisation du bootstrap au lieu des approximations asymptotiques parce que ce dernier peut ne pas être fiable quand il est appliqué à des échantillons de taille petite voire modérée. Ces études reconnaissent cependant également les limites du bootstrap standard, en particulier que la procédure peine souvent à performer en présence de distributions avec des queues épaisses ou des masses de probabilité comme c'est le cas dans les études de pauvreté et d'inégalités. Dans ce cadre, des procédures spécifiques doivent être implémentées pour améliorer les résultats du bootstrap mais le choix de la procédure adéquate requiert de connaître la nature du problème à l'origine de l'échec du bootstrap standard, ce qui n'est pas trivial quand la distribution étudiée est inconnue. Des études montrent que les méthodes d'inférence asymptotique et bootstrap ne produisent pas de résultats satisfaisants lorsque appliquées aux mesures d'inégalités. Entre autres, Davidson et Flachaire (2007) montrent que les distributions asymptotiques donnent une pauvre approximation des véritables distributions des statistiques quand la taille de l'échantillon est petite ou moyenne. Ils montrent de plus que le bootstrap i.i.d. donne des tests de pauvre niveau lorsque appliqué à l'indice d'inégalité de Theil avec une distribution de revenue Singh-Maddala. En dépit de toutes ces problèmes, aucune étude n'a, à notre connaissance, proposé de méthode d'inférence nonparamétrique à distance finie valide pour les mesures de pauvreté et d'inégalités. Dans cette thèse, nous nous intéressons à ce problème.

Dans le premier article, nous étudions plusieurs bandes de confiance basées sur des

fonctions de distribution. La première est basée sur le test de KS (KS, ci-dessous) qui est l'un des tests non paramétriques d'adéquation de lois le plus populaire. Ce dernier est fondé sur la statistique de KS qui est le supremum sur toutes les observations de la différence entre la fonction de distribution supposée sous l'hypothèse nulle et la fonction de distribution empirique de l'échantillon. Le test rejette la fonction de distribution testée si elle est trop loin de celle empirique, le seuil de rejet étant défini par le point critique de la statistique. Le test doit sa popularité à l'une de ses propriétés très pratiques: la distribution de la statistique de KS est indépendante de la fonction de distribution supposée sous l'hypothèse nulle lorsque celle-ci est continue et par conséquent, les points critiques de la statistique ne dépendent pas de la distribution testée et le même ensemble de points critiques peut être utilisé pour tester toutes les distributions continues. En inversant ce test, il est possible de construire une bande de confiance pour les fonctions de distribution qui bénéficient des mêmes propriétés que le test de KS.

Malgré le fait que le test de KS est pratique, il souffre d'un inconvénient majeur: il discrimine faiblement entre les distributions qui diffèrent principalement au niveau de leurs queues, ce qui altère les performances du test et des bandes de confiance. En particulier, la bande de confiance de KS a souvent été critiquée en raison de son caractère uniforme. Sa largeur est constante pour toutes les observations; par conséquent, ses bornes ne convergent pas vers 0 et 1 dans les queues de distributions, contrairement aux fonctions de distribution qu'elles bornent. Pour corriger cette contreperformance, nous utilisons des statistiques pondérées de KS basées sur les trois principes fondamentaux en économétrie: les principes de Wald, du multiplicateur de Lagrange et du ratio de vraisemblance. Ces statistiques ont été proposées par Anderson et Darling (1952), Eicker (1979) et Berk et Jones (1979), respectivement.

Les statistiques d'Anderson-Darling et d'Eicker sont des statistiques de KS standardisées pour lesquelles la différence entre la distribution théorique et celle empirique est divisée par une sorte d'écart-type. Ces statistiques permettent une meilleure discrimination entre les distributions qui diffèrent principalement au niveau de leurs extrémités. En utilisant ces statistiques, nous proposons des bandes de confiance exactes dont la largeur diminue

au fur et à mesure que les observations s'éloignent du centre de la distribution. Toutefois, les statistiques d'Anderson-Darling et d'Eicker ont leurs propres inconvénients. Les poids au niveau des dénominateurs de ces statistiques deviennent très proches de zéro pour les observations dans les queues, ce qui induit un comportement erratique des statistiques. Pour y remédier, nous proposons des statistiques obtenues par l'ajout d'un terme de régularisation au dénominateur des statistiques d'Anderson-Darling et d'Eicker. Ces statistiques conservent les avantages des statistiques de KS pondérées, mais ne souffrent pas d'instabilité. En inversant les statistiques régularisées, nous proposons des bandes de confiance exactes améliorées pour les fonctions de distribution.

La statistique de Berk-Jones est le supremum, sur toutes les observations, du ratio de log-vraisemblance entre les fonctions de distribution empirique et théorique utilisée comme distance entre ces deux fonctions. Il a été prouvé que cette statistique domine toutes les statistiques pondérées de KS, au sens de Bahadur et constitue donc une bonne référence de comparaison pour nos méthodes d'inférence. Cette statistique a été utilisée par Owen (1995) pour construire une bande de confiance non paramétrique pour les fonctions de distribution continues.

Nous montrons que dans le cas continu, les distributions des statistiques basées sur des fonctions de distribution empiriques sont indépendantes de la fonction de distribution testée sous l'hypothèse nulle, ainsi que leurs points critiques. Par conséquent, les bandes de confiance qu'elles permettent de construire dépendent de la distribution testée uniquement par l'échantillon. Ces bandes sont construites avec le même ensemble de points critiques pour toutes les fonctions de distribution continues, ce qui les rend facile à calculer. Pour les fonctions de distribution discontinues, nous dérivons des propriétés de monotonicité qui exploitent l'emboîtement des ensembles images de différentes distributions pour réduire la largeur des intervalles de confiance sans toutefois en altérer la fiabilité.

Dans les deuxième et troisième articles de cette thèse, nous proposons des intervalles de confiance pour la moyenne d'une variable aléatoire que nous appliquons aux mesures de pauvreté de Foster, Greer et Thorbecke (1984, ci-dessous FGT) et aux mesures

d'inégalités les plus populaires: la mesure d'entropie généralisée—qui inclut l'indice de Theil, la courbe de Lorenz, l'indice de Gini, la mesure d'Atkinson, la mesure de déviation logarithmique et l'indice de variation logarithmique. Pour ce faire, nous observons que les mesures de pauvreté peuvent se réécrire comme la moyenne d'une variable aléatoire bornée—un mélange entre une variable aléatoire continue et une masse de probabilité au seuil de pauvreté—et proposons que les méthodes d'inférence nonparamétrique exactes pour la moyenne d'une variable aléatoire bornée leur soient appliquées (article 2).

À première vue, ce problème paraît ne pas avoir de solution. En effet, d'après Bahadur et Savage (1956), il est impossible d'établir une inférence nonparamétrique pour la moyenne d'une variable aléatoire sur la base d'observations indépendantes et identiquement distribuées provenant d'une distribution inconnue dont la moyenne est finie (voir Dufour (2003) pour de plus amples détails). Toutefois dans notre cas, la nature bornée de la variable aléatoire étudiée donne une restriction suffisante pour permettre d'effectuer une inférence nonparamétrique. De tels intervalles de confiance pour la moyenne d'une variable aléatoire bornée sont proposés par Anderson (1969), Hora et Hora (1990) et Fishman (1991). Sutton et Young (1997) comparent les performances de ces méthodes à celles des méthodes bootstrap et asymptotique à l'aide de lois Beta. Ils montrent que les intervalles asymptotique et bootstrap ont une mauvaise probabilité de couverture en échantillon fini alors que les méthodes exactes sont très fiables mais produisent des intervalles plus larges que les premiers.

Nous observons que les mesures FGT sont des moyennes de variables aléatoires bornées qui sont elles-mêmes des fonctionnelles de fonctions de distribution et utilisons des techniques de projection pour déduire des intervalles de confiance à distance finie pour la moyenne d'une variable aléatoire bornée à partir de bandes de confiance de la fonction de distribution sous-jacente. Enfin, nous appliquons ces intervalles de confiance aux mesures de pauvreté FGT.

De façon similaire, nous montrons que la plupart des mesures d'inégalités peuvent se réécrire comme une fonction de moyennes de deux variables aléatoires dont l'une ou les deux peuvent ne pas être bornées (article 3). Dans ce cas, nous proposons une

généralisation du principe de projection utilisé pour les variables bornées aux variables non bornées, sous l'hypothèse que les queues de distribution étudiées sont bornées par des distributions de Pareto (voir Davidson et Flachaire, 2007 pour l'utilisation d'hypothèses similaires dans des procédures bootstrap). D'abord, nous observons que la moyenne d'une variable peut s'interpréter comme la moyenne pondérée d'une variable bornée et d'une variable non bornée, cette dernière étant la moyenne de la queue de distribution. En utilisant les techniques de projection utilisées dans le deuxième article, nous développons des intervalles de confiance pour la moyenne de la partie bornée de la variable aléatoire. Ensuite, nous établissons des bornes inférieure et supérieure pour la partie non bornée de la variable en utilisant l'hypothèse que les queues de distribution de ladite variable sont bornées par des lois de Pareto et appliquons les inégalités de Bonferroni pour calculer le niveau de l'intervalle de confiance ainsi construit. Enfin, nous appliquons ces méthodes d'inférence pour calculer des intervalles de confiance pour les mesures d'inégalités à partir de bandes de confiances des distributions sous-jacentes.

Tous ces intervalles de confiance bénéficient des mêmes propriétés pratiques que les bandes de confiance dont ils sont issus: pivotalité, conservation, monotonicité, etc.

Dans les trois articles, des études Monte Carlo sont effectuées pour analyser la performance des méthodes d'inférence et illustrer le choix du paramètre de régularisation. Elles montrent que les statistiques régularisées donnent des tests plus puissants que celles existantes, lorsqu'elles sont appliquées à des distributions qui diffèrent le plus dans les queues. De même, les bandes de confiance de fonctions de distribution et les intervalles de confiance pour la moyenne basés sur ces statistiques produisent de meilleurs résultats. Dans certains cas, les intervalles asymptotique et bootstrap ne produisent pas de résultats fiables alors qu'en revanche les intervalles exacts sont robustes à la distribution sous-jacente et à la taille de l'échantillon. Les intervalles de confiance proposés offrent une probabilité de couverture généralement plus grande que le niveau nominal tout en restant informatifs. Pour illustration, nous analysons dans les articles 2 et 3 les profils de pauvreté et d'inégalités des ménages ruraux au Mexique en 1998 en utilisant des données

du programme PROGRESA.[1] Les résultats montrent que les intervalles asymptotiques sont souvent trop petits pour être réalistes alors que l'intervalle bootstrap peuvent exploser, donnant des intervalles de largeur 10 fois supérieure à celles des méthodes exactes. L'étude montre qu'en moyenne, les ménages ruraux ciblés par PROGRESA n'ont pas un niveau de pauvreté très élevé. Toutefois, le profil de la pauvreté dépend grandement du sexe du chef de famille. Le niveau de pauvreté et d'inégalités des ménages avec à leur tête un individu male est beaucoup plus faible que celui des ménages ayant une femme à leur tête. En outre, les ménages avec un chef éduqué à leur tête semblent être plus susceptibles d'échapper à la pauvreté et aux inégalités que les ménages avec un chef non-instruit. Ces conclusions apportent des suggestions dans l'élaboration des politiques visant à réduire la pauvreté et les inégalités dans les régions rurales du Mexique. Les politiques visant à réduire l'analphabétisme des membres des ménages dans ces communautés peuvent être efficaces dans la réduction de la pauvreté. Les programmes d'éducation devraient viser les enfants et les adultes, en particulier les chefs de ménages afin de produire un effet immédiat. De même, les politiques visant à assurer le revenu des ménages ayant une femme à leur tête pourrait aider à réduire la pauvreté et les inégalités dans les régions rurales du Mexique. Un exemple de telles politiques peuvent être des réformes visant à garantir la propriété foncière pour les femmes ou à l'amélioration de la productivité du travail pour les ménages avec une femme à leur tête, cette dernière étant moins productive dans des activités demandant un effort physique intensif telles que l'agriculture.

[1] Voir les details sur ce programme en section 9, partie 2 (page 109).

Chapter 1

Improved exact nonparametric confidence bands and tests for distribution functions based on standardized empirical distribution functions

Abstract

Goodness-of-fit tests are of great interest in econometrics. In many procedures, especially in parametric ones, determining the distribution from which the sample comes from may be an important step. The Kolmogorov-Smirnov (KS, henceforth) test is one of the most popular nonparametric goodness-of-fit tests. However, it does not allow to discriminating a lot between distributions that differ mostly through their tails. Weighted KS statistics have been proposed by Anderson and Darling (1952) and Eicker (1979) to improve the performance of the test in the tails but they suffer from important drawbacks.

We propose improved weighted KS statistics to correct these limits. These statistics are obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics. They retain the advantages of the weighted KS statistics but their denominators do not become close to 0 in the tails of distributions as it is the case for the original statistics. We derive exact nonparametric confidence bands (CBs, henceforth) for distribution functions using the weighted and regularized KS statistics. We show that in the continuous case, these CBs are independent of the distribution assumed under the null hypothesis and are conservative for noncontinuous distributions. In the noncontinuous case, we derive monotonicity properties that exploit embeddedness of the image sets of different distributions to narrow the CBs without altering their reliability.

Monte Carlo simulations are performed to study the relative performance of the inference methods and illustrate how to choose the regularization parameter. The results show that the regularized statistics yield more powerful goodness-of-fit tests than the existing ones when applied to distributions with more discrepancy in the tails. Likewise, the CBs for distribution functions based on these regularized statistics are of better performance.

1.1 Introduction

The problem of determining the distribution from which a sample comes from is of great interest in statistics and econometrics. Instead of using an asymptotic law, it is often desirable and even crucial to know the actual distribution of a sample before applying further econometric procedures, in particular parametric ones. Several goodness-of-fit tests have been proposed. They test the null hypothesis that a sample follows a given distribution–which generally needs to be fully specified–against the alternative that the sample does not follow this distribution. Parametric tests have been proposed by Shapiro and Wilk (1965), Lilliefors (1967), Chambers (1983)–probability plots, etc. Likewise, non-parametric procedures have been provided by Snedecor and Cochran (1989)–Khi square test, Anderson and Darling (1952), Kolmogorov (1941), Smirnov (1944), Cramer (1928), Von Mises (1931), etc.

The Kolmogorov-Smirnov (KS, henceforth) test is one of the most popular nonparametric goodness-of-fit tests. It is based on the KS statistic which is the supremum over all observations of the difference between the distribution function assumed under the null hypothesis and the empirical distribution function of the sample. The test rejects the distribution function assumed under the null hypothesis if it is too far from the empirical distribution function, the threshold being defined by the critical point of the KS statistic. The test owes its popularity to a convenient property: the distribution of the KS statistic is independent of the distribution function being tested under the null hypotheses when the latter is continuous. Hence, the critical points of the statistic are not contingent on the assumed distribution and can be used to test any continuous distribution function. These critical points have been tabulated by several authors and are widely published. Inverting the test allows one to build confidence bands (CBs, henceforth) for distribution functions which also benefit from the pivotality of the KS statistic.

Even though the KS statistic is convenient, it has low power to discriminate a lot between distributions that differ mainly through their tails. This property alters the performance of the KS test and CB. In particular, the KS confidence band has often been criticized because of its uniform nature: its width is constant for all observations

and thus, its bounds do not converge to 0 and 1 in the lower and the upper tails of the distribution, as do the distribution functions they bracket. To correct this drawback, we use weighted KS statistics based on the three common principles in econometrics: the Wald, Lagrange multiplier, and likelihood-ratio principles. These statistics have been proposed by Anderson and Darling (1952), Eicker (1979), and Berk and Jones (1979). The Anderson-Darling and the Eicker statistics are standardized versions of the KS statistic where the difference between the theoretical and the empirical distributions is divided by a kind of standard deviation. These statistics allow one to discriminate between distributions that differ mostly through their tails. Using them, we propose finite-sample nonparametric CBs whose widths decrease with observations further from the center of the distribution.

The Anderson-Darling and the Eicker statistics have their own drawbacks. The power of the goodness-of-fit test they yield is smaller than the power of the standard KS test when testing distributions with low dispersion that differ more in the center of the distribution than in the tails. Moreover, the weights in the denominators of those statistics become very close to zero for observations in the tails, leading to erratic behavior of the statistics. We propose improved weighted KS statistics to correct these. These statistics are obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics. They retain the advantages of the weighted KS statistics but do not suffer from instability, improving the performance of the inference. By inversion of the regularized statistics, we build improved exact CBs for distribution functions.

The Berk-Jones statistics uses the supremum, over all observations, of the log-likelihood ratio of the empirical distribution function and the theoretical distribution function as a distance between these two functions. This statistic has been proved to dominate any weighted KS statistic, in the sense of Bahadur and is thus a challenging referral for our inference methods. It has been used by Owen (1995) to propose a CB for distribution functions.

In the continuous case, we show that the distributions of the empirical distribution-based statistics are pivotal and that their critical points do not depend on the distri-

bution function being tested under the null hypothesis. Hence, the corresponding CBs depend on the distribution only through the sample; they are built using the same critical points for all continuous distribution functions, which make them easy to compute. For noncontinuous distribution functions, we derive monotonicity properties which exploit embeddedness of the image sets of different distributions to narrow the CBs without altering their reliability.

We compare the relative performance of the nonparametric and parametric inference methods. Monte Carlo simulations are performed to study the power of the goodness-of-fit tests under various hypotheses. In both studies, we study carefully the choice of the regularization parameter. The results show that regularized statistics yield more powerful goodness-of-fit tests than the existing ones when applied to distributions with more discrepancy in the tails.

The paper is organized as follows. Section 2 presents the Kolmogorov-Smirnov, the Anderson-Darling and the Eicker statistics and derives the expressions to compute them. It also shows how to invert these tests and build the CBs for distribution functions they yield. In section 3, we introduce the regularized statistics and derive explicit expressions to compute them and to build CBs for distribution functions. Sections 4 presents the Owen CB and Section 5 derives some convenient properties of these CBs for continuous cases and monotonicity properties for noncontinuous distribution functions. Section 6 presents Monte Carlo results and Section 7 concludes.

1.2 Distributional properties of goodness-of-fit statistics based on empirical distribution functions

Let's define some notation for the remainder of the paper. Denote $\mathbb{F}$, the set of all distribution functions, $\widetilde{\mathbb{F}}$, the set of continuous distribution functions, $\mathbb{F}_{[a,b]}$ the set of distribution functions with support $[a,b]$, and $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. Let X be a random variable with distribution function $F(x) \in \mathbb{F}$. Denote $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ the order statistics of a sample of n i.i.d. observations on X and $F_n(x)$ the corresponding

empirical distribution function defined as follows: $\forall\ k = 0, \ldots, n$

$$F_n(x) = \frac{k}{n} \text{ for } X_{(k)} \leq x < X_{(k+1)} \tag{1.1}$$

where $(X_{(0)}, X_{(n+1)})$ is the support of $F(x)$, which may be the real line $(-\infty, +\infty)$ and $X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)}$.

Let's consider the following null hypothesis:

$$\mathbf{H_0(F)} : X_1, \ldots,\ X_n \text{ are i.i.d. with distribution function } P[X_i \leq x] = F(x). \tag{1.2}$$

A general statistic to test $H_0(F)$ against its negation $H_1(F)$ is:

$$D\left(F_n, F\right) = \sup_{-\infty<x<+\infty} D_1\left[F_n(x), F(x)\right]. \tag{1.3}$$

where $D_1\left[F_n(x), F(x)\right]$ is a functional of $F_n(x)$ and $F(x)$, which measures the distance between these two functions. In this section, we study interesting properties for statistics of the form $D\left(F_n, F\right)$ when $F(x)$ is continuous and when it is not.

1.2.1 Pivotality and conservativeness

PROPOSITION 2.1. **[Distribution of statistics based on empirical distribution functions when $F(x)$ is continuous]** *Let $X_1, \ldots, X_n$ be n random variables. Let $D\left(F_n, F\right)$ be a statistic of the form:*

$$D\left(F_n, F\right) = \sup_{-\infty<x<+\infty} D_1\left[F_n(x), F(x)\right].$$

If $F(x)$ is a continuous monotonic function, then the following identity holds almost surely:

$$D\left(F_n, F\right) = \sup_{u \in F(\mathbb{R})} D_1\left[H(U_1, \ldots, U_n, u), u\right]$$

where $U_i = F(X_i)$, $i = 1, \ldots, n$ *and*

$$H\left[U_1, \ldots, U_n, u\right] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[U_k \leq u],$$

If, furthermore, $X_1, \ldots, X_n$ *are* n *i.i.d. observations on* X *with continuous distribution function* $F(x) \in \widetilde{\mathbb{F}}$, *then*

$$D\left(F_n, F\right) \overset{a.s}{=} \sup_{0 \leq u \leq 1} D_1\left[H(U_1, \ldots, U_n, u), u\right]$$

where $U_i = F(X_i)$, $i = 1, \ldots, n$ *are i.i.d. with uniform* $U_{[0,1]}$ *distribution.*

Proposition 2.1. states that when $F(x)$ is continuous, the distribution of $D\left(F_n, F\right)$ is independent of $F(x)$. $D\left(F_n, F\right)$ can be rewritten using only uniform statistics. All statistics with general form $D\left(F_n, F\right)$ are pivotal for continuous distribution functions. Hence, the critical points associated to those statistics are also independent of $F(x)$. This property simplifies a lot the implementation of the tests and CBs associated to such statistics. A unique set of critical points is needed to compute these for all continuous distribution functions.

When $F(x)$ is not continuous, the distribution of $D\left(F_n, F\right)$ is different for each $F(x)$ being tested. The associated critical points are also modified by the distribution of the sample. Hence, a new set of critical values need to be computed to implement the tests for each distribution, making the inference methods more difficult to implement. Moreover, in this case, building CBs for $F(x)$ loses all interest because these are usually built to bracket unknown distribution functions using a sample of observations that comes from the distribution under interest. To simplify the implementation of the inference methods and restore the interest of CBs in the case of noncontinuous distribution functions, we propose to exploit the following properties.

PROPOSITION 2.2. **[Conservative nature of continuous case critical points]** *Let* $X_1, \ldots, X_n$ *be* n *i.i.d. observations on* X *and* $F_n(x)$ *the corresponding empirical distribution function. Let* $F(x) \in \widetilde{\mathbb{F}}$ *be a continuous distribution function and* $G(x) \in \mathbb{F}$ *a non-*

continuous one. For any level α, $0 \leq \alpha \leq 1$, *the critical value associated with* $D(F_n, F)$ *for testing the null hypothesis* $H_0(F)$ *as defined by equation (1.2) is larger than or equal to the critical value associated with* $D(F_n, G)$ *for testing the null hypothesis* $H_0(G)$*:*

$$P_G[D(F_n, G) \geq x] \leq P_F[D(F_n, F) \geq x], \ \forall x$$

or equivalently

$$P_F[D(F_n, F) \geq D_\alpha] \leq \alpha \Rightarrow P_G[D(F_n, G) \geq D_\alpha] \leq \alpha, \ \forall D_\alpha.$$

PROPOSITION 2.3. **[Conservative property of continuous case CBs for distribution functions]** *Let* $X_1, \ldots, X_n$ *be* n *i.i.d. observations on* X *and* $F_n(x)$ *the corresponding empirical distribution function. Let* $F(x) \in \widetilde{\mathbb{F}}$ *be a continuous distribution function and* $G(x) \in \mathbb{F}$ *be a noncontinuous one. For any level* α, $0 \leq \alpha \leq 1$, *the confidence band obtained by inverting the test of the null hypothesis* $H_0(F)$ *as defined by equation (1.2) using appropriate critical points with level* α *for* $D(F_n, F)$ *yields a confidence band for* $G(x)$ *with level larger than or equal to* $1 - \alpha$*. Equivalently, if* C_n *is defined as follows:*

$$C_n(\alpha) = \{H \in \mathbb{F} : D(F_n, H) \leq c_\alpha\}$$

where

$$c_\alpha = \inf\{D_\alpha, \ P_F[D(F_n, F) \geq D_\alpha] \leq \alpha\}$$

then

$$P_G[G \in C_n(\alpha)] \geq 1 - \alpha.$$

Propositions 2.2. and 2.3. highlight some interesting properties of the empirical distribution function-based statistics and CBs which simplify their implementation when applied to noncontinuous distributions. Proposition 2.2. states that the critical values of $D(F_n, G)$ for continuous distribution functions $F(x)$ are conservative for noncontinuous functions $G(y)$. Using the appropriate critical values of level α for $F(x)$ provide a test

of level less than or equal to α for $G(y)$. Therefore, rejection of the null hypothesis with such test leads to rejection for the nominal level α. In other words, the result of the test based on $D(F_n, G)$ remains valid and one can use the conservative critical values to compute tests and CBs for noncontinuous distribution functions. Let's remember that those critical points–that applies to continuous distributions–are independent of the function being tested and are thus, identical for all continuous distributions. Note that these propositions hold for any continuous distribution function $F(x)$.

Likewise, the CBs for $G(y)$ built using appropriate critical points for continuous distribution functions will be of level larger than or equal to $1-\alpha$. Using these properties, critical points from continuous distribution functions can be applied to any sample from a general distribution function. The resulting CBs will be of level at least equal to $1-\alpha$.

Even though the conservative critical points provide valid inference for noncontinuous distribution functions, using them alters the performance of the inference. The question is how far the quality of the performed inference is affected? We assess this question using the properties of tests and CBs. Concerning the tests, when the null hypothesis is accepted with level α, the conclusion of the test remains valid for levels less than or equal to α. Conversely, when the null hypothesis is rejected with level α, it will be still rejected for levels larger than α but might be accepted for lower levels. Concerning CBs, the impact of the using conservative critical points can be studied using the level of confidence (accuracy) and the width (precision) of the CBs. Given that the CBs using conservative critical points have a higher level than the targeted one, they will be wider than the CBs with effective level $1-\alpha$. To reduce the width, exact critical values corresponding to the distribution function under interest can be computed. However, by doing this, the CBs will loose one of their major advantages. To avoid this shortcoming, we derive monotonicity properties that can be used to narrow CBs without altering their reliability. These results are based on information about the set of discontinuities of the distribution function.

PROPOSITION 2.4. **[Range monotonicity of critical points]** *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $F_n(x)$ the corresponding empirical distribution function.*

Let $F(x)$ and $G(y)$ be two distribution functions such that $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}})$. For any level α, $0 \leq \alpha \leq 1$, the critical value associated with $D(F_n, F)$ for testing the null hypothesis $H_0(F)$ as defined by equation (1.2) is larger than or equal to the critical value associated with $D(F_n, G)$ for testing the null hypothesis $H_0(G)$:

$$P_F[D(F_n, F) \geq D_\alpha] \leq \alpha \Rightarrow P_G[D(F_n, G) \geq D_\alpha] \leq \alpha, \ \forall D_\alpha.$$

PROPOSITION 2.5. **[Range monotonicity of CBs for distribution functions]** *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $F_n(x)$ the corresponding empirical distribution function. Let $F(x)$ and $G(y)$ be two distribution functions such that $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}})$. For any level α, $0 \leq \alpha \leq 1$, the confidence band obtained by inverting the test of the null hypothesis $H_0(F)$ as defined by equation (1.2) using appropriate critical points with level α for $D(F_n, G)$ yields a confidence band for $G(x)$ with level larger than or equal to $1-\alpha$. Equivalently, if C_n is defined as follows:*

$$C_n(\alpha) = \{G : D(F_n, G) \leq c_\alpha\}$$

where

$$c_\alpha = \inf\{D_\alpha, \ P_F[D(F_n, F) \geq D_\alpha] \leq \alpha\},$$

then

$$P_G[G \in C_n(\alpha)] \geq 1 - \alpha.$$

Propositions 2.4. and 2.5. generalize Propositions 2.2. and 2.3. to all distribution functions. It suggests that CBs can be made narrower by exploiting embeddedness of the image sets of different distributions. When studying a discontinuous distribution $G(y)$, we know that $G(y)$ takes its values in a set V^G which is included in $[0, 1]$. Thus, the conservative CB for a continuous distribution provides a CB for $G(y)$ with level $1 - \delta_1$ greater than or equal to $1 - \alpha$. If additional information about the image set of $G(y)$ is available—in particular, if we know there exists a distribution function with image V^F

such that $V^G \subseteq V^F$—then the critical points for testing $F(x)$ can be used to derive a CB for $G(y)$ with level $1-\delta_2$ such that $1-\alpha \leq 1-\delta_2 \leq 1-\delta_1$. The CB with level $1-\delta_2$ is narrower than the CB with level $1-\delta_1$ while being reliable. Thus, using information about the nature of the discontinuity of the random variable can be useful for providing shorter CBs for $G(y)$. The more is known about the set of discontinuity points of the distribution, the better the inference will be. However, the improvement can be achieved without knowing all discontinuity points of $G(y)$ and their probability masses. Hence, the main advantage of the KS confidence band, that is its independence of the assumed distribution function, is somehow preserved.

Let's consider a special case of embedded image sets. Let X be a random variable with distribution $H(x) \in \mathbb{F}_{[0,1]}$ such that X is a mixture between a continuous variable bounded on $(0,1]$ and a probability mass of $H(0) \equiv p$ at 0. $H(x)$ is continuous on $(0,1]$ with $H(0) = p$ and $H(1) = 1$.

COROLLARY 2.6. **[Range monotonicity with a mass at the lower boundary]** *Let $X_1^2, \cdots, X_n^2$ be n i.i.d. observations on X_2 and $F_n(x)$ the corresponding empirical distribution function. Let $F_1(x)$ and $F_2(x)$ be two distribution functions continuous on $(a,b]$ such that $p_1 = F_1(a) \leq F_2(a) = p_2$. For any level α, $0 \leq \alpha \leq 1$, the confidence band obtained by inverting the test of the null hypothesis $H_0(F_1)$ as defined by equation (1.2) using appropriate critical points with level α for $D(F_n, F_1)$ yields a confidence band for $F_2(x)$ with level larger than or equal to $1-\alpha$. Equivalently, if C_n is defined as follows:*

$$C_n(\alpha) = \{H \in \mathbb{F} : D(F_n, H) \leq c_\alpha\}$$

where

$$c_\alpha = \inf\{D_\alpha,\ P_{F_1}[D(F_n, F_1) \geq D_\alpha] \leq \alpha\},$$

then

$$P_{F_2}[F_2 \in C_n(\alpha)] \geq 1-\alpha.$$

Corollary 2.6. describes the special case where the distribution function being studied is continuous everywhere except at the lower bound of its support. This case is very interesting because such distribution functions are quite frequent in financial studies, and poverty and inequality analysis. Moreover, this case can be easily extended to those where the discontinuity point is at the upper bound of the support or those where both the lower and the upper bounds are discontinuity points.

1.2.2 Special case: the Kolmogorov-Smirnov statistic and confidence band

Let X be a random variable with distribution function $F(x) \in \mathbb{F}$. Denote $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ the order statistics of a sample of n i.i.d. observations on X and $F_n(x)$ the empirical distribution function of the sample. The Kolmogorov-Smirnov (KS, henceforth) statistic is

$$KS = \sup_{-\infty \leq x \leq +\infty} \sqrt{n} \left| F_n(x) - F(x) \right| \tag{1.4}$$

Developing this expression allows to reexpress the KS statistic as follows:

$$KS = \max \left\{ \max_{1 \leq i \leq n} \sqrt{n} \left[\frac{i}{n} - F(X_{(i)}) \right], \max_{1 \leq i \leq n} \sqrt{n} \left[F(X_{(i)}) - \frac{i-1}{n} \right], 0 \right\}$$

This explicit expression is more convenient and can be used to compute the test easily.

The KS is a special case of the statistic $D\left(F_n, F\right)$ where

$$D_1\left[F_n(x), F(x)\right] = \sqrt{n} \left| F_n(x) - F(x) \right|.$$

Hence, all properties derived for this general form of statistics apply to the KS statistic. In particular, when applied to continuous distribution functions, the KS statistic is pivotal and its distribution can be characterized as follows:

$$KS^{cont} = \max \left\{ \max_{1 \leq i \leq n} \left[\frac{i}{n} - U_{(i)} \right], \max_{1 \leq i \leq n} \left[U_{(i)} - \frac{i-1}{n} \right], 0 \right\}$$

where $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are the order statistics of a sample of n i.i.d. observations of an uniform $U_{[0,1]}$ distribution. The critical values used to compute the KS tests are the same for all $F \in \widetilde{\mathbb{F}}$ and thus, do not need to be simulated for each distribution.

Moreover, appropriate KS critical values for testing continuous distribution functions are conservative for noncontinuous distributions. To end, this conservative property extends to the cases of distribution functions with embedded image sets.

It follows that the KS critical point for a level α and a given sample size n is the same for all continuous distribution functions and can be used to test the hypothesis $H_0 : X_i \sim F$ against the alternative one $H_1 : X_i \nsim F$ for all $F \in \widetilde{\mathbb{F}}$. Exact critical points for KS^{cont} can be computed by simulation using the following 3-steps procedure:

1. Generate a sample of n i.i.d. observations from an uniform law $U_{[0,1]}$
2. Compute the Kolmogorov statistic KS^{cont} for this sample using the expression $KS^{cont} = \max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - U_{(i)}\right], \max_{1\leq i\leq n}\left[U_{(i)} - \frac{i-1}{n}\right], 0\right\}$
3. Repeat N times steps 1 and 2—N is the number of replications—and compute the critical value of level α, the $(1-\alpha)^{th}$ fractile.

Tables of the Kolmogorov-Smirnov critical points have been computed and published, for continuous distributions. Having them simplifies the test considerably and makes this goodness-of-fit test more convenient to use than the other exact methods.

The conservative property of the KS statistic has been evoked by Kolmogorov (1941) before being proved by other authors including Noether (1963) and Conover (1972). We provide in Appendix 2, a more convenient proof of this property than those provided in the literature.

The CB for $F(x)$ with level greater than or equal to $1-\alpha$ built inverting the KS test is:

$$C_F^{KS}(\alpha) = \left\{F_0 \in \mathbb{F} : F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq F_0(x) \leq F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}},\ \forall x\right\} \tag{1.5}$$

where $c_{KS}(\alpha)$ satisfies $\Pr[KS_F \leq c_{KS}(\alpha)] \geq 1-\alpha$.

When $F(x)$ is continuous $c_{KS}(\alpha)$ does not depend on it. For a given sample size, the same critical value is used to build CBs for all continuous distributions. Hence, the KS confidence band depends on $F(x)$ only through the sample. This simplifies a lot its computation and makes its popularity. Moreover, owing to the monotonicity of the KS critical points when applied to distributions with embedded image sets, if $F(x)$ and $G(y)$ are two distribution functions such that $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}})$ then the KS confidence band using adequate critical values for $F(x)$ embeds the KS confidence for $G(y)$. Hence, using information about the image set of $G(y)$ one can build narrower CB for the latter without having to use the adequate critical values for it.

1.3 Implementation as a Monte Carlo test

In the preceding section, we studied interesting properties for statistics of the form $D(F_n, F) = \sup_{-\infty<x<+\infty} D_1[F_n(x), F(x)]$. In this section, we show another important advantage of these statistics which make them even more attractive. We show that the test of $H_0(F)$ as defined by equation (1.2) based on these statistics can be implemented using exact randomized test procedures such as Monte Carlo tests (see Dwass (1957), Dufour (1995), Dufour and Kiviet (1998), and Dufour and Khalaf (2001)). Given that $D(F_n, F)$ is pivotal under $H_0(F)$, the Monte Carlo test based on pivotal statistics can be applied.

Given that the distribution of $D(F_n, F)$ is noncontinuous, the standard Monte Carlo test procedure cannot be applied. In this case, a randomized tie-breaking procedure (Dufour, 1995) can be applied to test $H_0(F)$. This procedure is a modified Monte Carlo test adapted for discrete distributions. It can be implemented as follows.

Let D_0 denote the test statistic computed from data and $Đ_0$ the observed value of D_0 based on specific realized data. D_0 is a random variable while $Đ_0$ is fixed. The critical region of the test is $D_0 \geq D_\alpha$ where α is the level of the test and $G(Đ_0) = P(D \geq D_0 \mid D_0 = Đ_0)$ is the realized p-value of the test statistic D_0. Suppose we can generate N i.i.d. replications D_j, $j = 1, \ldots, N$ of $D(F_n, F)$ under $H_0(F)$. The following steps apply:

- Draw $N+1$ i.i.d. variates $W_0, W_1, \ldots, W_N$ independently of D_j.

- Order the pairs (D_j, W_j) following the lexicographic criterion:

$$(D_i, W_i) \geq (D_j, W_j) \Leftrightarrow \{D_i > D_j \text{ or } (D_i = D_j \text{ and } W_i \geq W_j)\}.$$

- Compute an empirical p-value function:

$$\widetilde{p}_N(x) = \frac{N\widetilde{G}_N(x) + 1}{N+1}$$

where

$$\widetilde{G}_N(x) = 1 - \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{[0,\infty)}(x - D_j) + \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{[0]}(D_j - x)\, \mathbb{1}_{[0,\infty)}(W_j - W_0).$$

$\widetilde{p}_N(x)$ is the empirical probability that a value as extreme or more extreme than x is realized if $H_0(F)$ is true. Note that $N\widetilde{G}_N(x)$ is the number of simulated statistics which are greater or equal to x.

- Compute the associated Monte Carlo critical region—which is a randomized critical region—as:

$$\widetilde{p}_N(D_0) \leq \alpha,\ 0 \leq \alpha \leq 1. \tag{1.6}$$

where $\widetilde{p}_N(D_0)$ may be interpreted as an estimate of $G(Đ_0)$. Given $D_0 = Đ_0$, $\widetilde{p}_N(Đ_0)$ can be interpreted as a realized Monte Carlo p-value associated with D_0. Thus if N is chosen such that $\alpha(N+1)$ is an integer, the critical region (1.6) has the same size as the critical region $G(D_0) \leq \alpha$. Moreover,

$$P[\widetilde{p}_N(D_0) \leq \alpha] = \frac{I[\alpha(N+1)]}{N+1},\ 0 \leq \alpha \leq 1$$

where $I[x]$ is the integer part of x.

Thanks to the properties of the Monte Carlo test, the implementation of the studied

goodness-of-fit tests along this procedure is convenient. First, if $F(x)$ is free of nuisance parameters and $\alpha(N+1)$ is an integer then the properties of the critical region are irrespective of the number of replications used. Second, the Monte Carlo test does not need to compute exact critical points for each sample size and each distribution function under test. Besides, the number of replications needed to compute the test is not constraining as its the number of replications needed to simulate valid critical points or to get accurate bootstrap results. Third, confidence intervals can be built from such tests using inversion procedures (see Fieller (1940, 1954) for example).

1.4 Application to the Anderson-Darling, Eicker, and Berk-Jones type statistics and confidence bands

Besides the popularity and the advantages of the Kolmogorov-Smirnov goodness-of-fit test and CB, this inference method suffers from important drawbacks. In fact, the KS statistic is the supremum over x of $D_1\left[F_n(x), F(x)\right] = \sqrt{n}\left|F_n(x) - F(x)\right|$. The latter is often used to test hypotheses of type $H_0 : F(x) = p$ versus $H_1 : F(x) \neq p$. However, $D_1\left[F_n(x), F(x)\right]$ is not standardized and, hence, its distribution is not asymptotically pivotal. Moreover, the KS confidence band for distribution functions is often criticized for its uniform nature. The width of this CB is constant for all observations. Thus its bounds do not converge to 0 and 1 in the lower and upper tails of the distribution, as do the distribution functions they bracket. This property adversely affects the performance of the method in the tails of distributions.

Other inference methods can be used to correct these drawbacks. In fact, $D_1\left[F_n(x), F(x)\right]$ can be improved along three common principles in econometrics: the Lagrange multiplier, Wald, and likelihood-ratio principles. The first one replaces $D_1\left[F_n(x), F(x)\right]$ by a score-type statistic where $D_1\left[F_n(x), F(x)\right]$ is divided its standard deviation estimated under the null hypothesis. The Wald principle standardizes $D_1\left[F_n(x), F(x)\right]$ using an estimation of its standard deviation under H_1 and the last principle replaces $D_1\left[F_n(x), F(x)\right]$ by an evaluation of the ratio between the likelihood of $F_n(x)$ and $F(x)$. Taking the supremum

of the corresponding statistics yields three well-known statistics: the Anderson-Darling, the Eicker and the Berk-Jones statistics. We hereinafter study these statistics and the CBs they induce.

1.4.1 The Anderson-Darling and Eicker statistics and confidence bands

One of the most popular weighted KS statistics has been proposed by Anderson and Darling (1952) and Eicker (1979):

$$AD = \sup_{-\infty<x<+\infty} V_n(x)$$

and

$$E = \sup_{-\infty<x<+\infty} \widehat{V}_n(x)$$

where

$$V_n(x) = \begin{cases} 0 & \text{if } F(x) \in \{0,1\}, \\ \sqrt{n}\left|\frac{F_n(x)-F(x)}{F^{1/2}(x)[1-F(x)]^{1/2}}\right| & \text{otherwise,} \end{cases}$$

and

$$\widehat{V}_n(x) = \begin{cases} 0 & \text{if } F_n(x) \in \{0,1\}, \\ \sqrt{n}\left|\dfrac{F_n(x)-F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2}}\right| & \text{otherwise.} \end{cases}$$

These statistics are standardized versions of the KS statistic. The Anderson-Darling (AD, henceforth) statistic weights each observation by a sort of standard deviation of $F_n(x)-F(x)$, the difference between the empirical distribution function and the theoretical distribution, while the Eicker statistic uses an estimation of this standard deviation. Given that the function $\sqrt{y(1-y)}$ reaches its maximum at $y = \frac{1}{2}$, these statistics give less weight to the observations in the center of the distribution than to the observations in the tails. Hence, the tests they deliver discriminate more between distributions that mostly differ through their tails than the KS test.

Note that $V_n(x)$ and $\widehat{V}_n(x)$ are set to 0 to complete the definition of AD and E but $\widehat{V}_n(x)$ is not continuous in the tails. In fact, $\forall x$, $F(x) = 0 \Rightarrow F_n(x) = 0$ and $F(x) = 1 \Rightarrow F_n(x) = 1$. Hence, $\lim_{F(x)\longrightarrow 0} V_n(x) = \lim_{F(x)\longrightarrow 1} V_n(x) = 0$ while the reverse does not hold: $F_n(x) = 0 \not\Rightarrow F(x) = 0$ and $F_n(x) = 1 \not\Rightarrow F(x) = 1$.

Developing these statistics provides explicit expressions to compute these statistics more easily in practice:

$$AD = \max\left\{0, \max\left\{\sqrt{n}\frac{\frac{i}{n} - F(X_{(i)})}{F^{1/2}(X_{(i)})[1-F(X_{(i)})]^{1/2}} : 0 < F(X_{(i)}) < 1, 1 \le i \le n\right\}, \right.$$
$$\left. \max\left\{\sqrt{n}\frac{F(X_{(i)}) - \frac{i-1}{n}}{F^{1/2}(X_{(i)})[1-F(X_{(i)})]^{1/2}} : 0 < F(X_{(i)}) < 1, 1 \le i \le n\right\}\right\}$$

and

$$E = \max\left\{\max_{1\le i\le n-1} \sqrt{n}\frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]}}, \max_{2\le i\le n} \sqrt{n}\frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{\frac{i-1}{n}[1-\frac{i-1}{n}]}}, 0\right\}.$$

The details of the computation are given in Appendix 2.

Inverting the tests, we propose nonparametric Anderson Darling-type and Eicker-type CBs for distribution functions (see the details of computation in Appendix 2). To our knowledge, expressions for these CBs are not provided in the literature. The Anderson Darling-type CB for $F(x)$ with level greater than or equal to $1-\alpha$ is:

$$C_F^{AD}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(x) \le F_0 \le G_n^U(x), \forall x\right\}$$

where

$$G_n^L(x) = \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} - \sqrt{\Delta(x)}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}, \qquad G_n^U(x) = \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} + \sqrt{\Delta(x)}}{2(1+\frac{c_{AD}^2(\alpha)}{n})},$$

$$\Delta(x) = \left[2F_n(x) + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4F_n^2(x)\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right],$$

and $c_{AD}(\alpha)$ satisfies $Pr(AD \leq c_{AD}(\alpha)) \geq 1-\alpha$.

The Eicker-type CB for $F(x)$ with level greater than or equal to $1-\alpha$ is:

$$C_F^E(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(x) \leq F_0 \leq G_n^U(x)\right\}$$

where

$$G_n^L(x) = \begin{cases} F_n(x) - \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} & \forall\ x \text{ such that } F_n(x) \notin \{0,1\}, \\ 0 & \forall\ x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

$$G_n^U(x) = \begin{cases} F_n(x) + \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} & \forall\ x \text{ such that } F_n(x) \notin \{0,1\}, \\ 1 & \forall\ x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

and $c_E(\alpha)$ satisfies $Pr(E \leq c_E(\alpha)) \geq 1-\alpha$.

Owing to the structure of the weights used by the Anderson-Darling and the Eicker statistics, the widths of the CBs decrease with observations further from the center of the distribution. This property improves the performance of the inference methods in the tails of distributions.

The Anderson-Darling and the Eicker are special cases of the statistic $D(F_n, F)$. Hence, they are pivotal for continuous distribution functions. Moreover, in such case, the expression of these statistics simplifies to the following:

$$AD = \max\left\{0, \max\left\{\sqrt{n}\frac{\frac{i}{n} - U_{(i)}}{U_{(i)}^{1/2}[1-U_{(i)}]^{1/2}} : 0 < U_{(i)} < 1, 1 \leq i \leq n\right\}, \max\left\{\sqrt{n}\frac{U_{(i)} - \frac{i-1}{n}}{U_{(i)}^{1/2}[1-U_{(i)}]^{1/2}} : 0 < U_{(i)} < 1, 1 \leq i \leq n\right\}\right\},$$

and

$$E = \max\left\{\max_{1\leq i\leq n-1}\sqrt{n}\frac{\frac{i}{n} - U_{(i)}}{(\frac{i}{n})^{1/2}(1-\frac{i}{n})^{1/2}},\ \max_{2\leq i\leq n}\sqrt{n}\frac{U_{(i)} - \frac{i-1}{n}}{(\frac{i-1}{n})^{1/2}(1-\frac{i-1}{n})^{1/2}},\ 0\right\}.$$

where $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are the order statistics of a sample of n i.i.d. observations from an uniform $U_{[0,1]}$ distribution. The Anderson-Darling and the Eicker statistics can be rewritten using uniform order statistics. Their distributions are independent of the distribution assumed under the null hypothesis. Likewise, the critical values of the Anderson-Darling and the Eicker statistics for continuous distributions are independent of $F(x)$ and the corresponding CBs uses a single set of these for all continuous distribution functions. Furthermore, these CBs benefit from the same monotonicity properties as the KS confidence band.

1.4.2 The Berk-Jones type statistic and confidence band

The weighted statistics studied above propose improvements of the KS statistic based on two common principles in econometrics: the Wald principle and the Lagrange multiplier one. A third principle is often used to improve procedures in econometrics: the likelihood-ratio principle. Berk and Jones (1979) proposed a statistic based on the empirical distribution function using the likelihood-ratio principle:

$$BJ = \sup_{-\infty \leq x \leq +\infty} K[F_n(x), F(x)] = \max_{1 \leq i \leq n} \max \left\{ K\left(\frac{i-1}{n}, F(X_{(i)})\right), K\left(\frac{i}{n}, F(X_{(i)})\right) \right\}$$

where

$$K(\widehat{p}, p) = \widehat{p} \log\left(\frac{\widehat{p}}{p}\right) + (1-\widehat{p}) \log\left(\frac{1-\widehat{p}}{1-p}\right).$$

Berk and Jones proved that this statistic dominates all weighted KS statistics, in the sense of Bahadur. This statistic is under the form of $D\left[F_n, F\right]$. It is then pivotal when applied to continuous distribution functions and its distribution may be characterized using uniform order statistics as follows:

$$BJ^{cont} = \max_{1 \leq i \leq n} \max \left\{ K\left(\frac{i-1}{n}, U_{(i)}\right), K\left(\frac{i}{n}, U_{(i)}\right) \right\}$$

where $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are the order statistics of a sample of n i.i.d. observations

of an uniform $U_{[0,1]}$ distribution. The critical values of BJ^{cont} are also independent of $F(x)$ and monotonicity properties derived for $D\left(F_n, F\right)$ applies.

Using the Berk-Jones statistic, Owen (1995) proposed the following nonparametric CB for continuous distributions functions with level $1-\alpha$:

$$C_F^O(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(x) \leq F_0(x) \leq G_n^U(x),\ \forall x\right\} \tag{1.7}$$

where

$$G_n^L(x) = \min\left\{p : K[F_n(x), p] \leq c_{BJ}(\alpha)\right\},$$

$$G_n^U(x) = \max\left\{p : K[F_n(x), p] \leq c_{BJ}(\alpha)\right\},$$

$K(\widehat{p}, p) = \widehat{p}\log\left(\frac{\widehat{p}}{p}\right) + (1-\widehat{p})\log\left(\frac{1-\widehat{p}}{1-p}\right)$, and $c_{BJ}(\alpha)$ satisfies $P[BJ > c_{BJ}(\alpha)] \geq 1-\alpha$.

The reasoning behind this CB is quite intuitive. The statistic $nF_n(x)$ follows a binomial law with parameters n and $F(x)$. Thus, $-nK(\widehat{p}, p)$ is the log-likelihood ratio of the probability parameter p based on a binomial observation of $n\widehat{p}$ successes out of n trials. It follows that the Owen's confidence band is computed by performing a likelihood ratio test on the distribution of $F_n(x)$. Only candidates $F(x)$ with sufficiently large likelihood for each observation x belong to the confidence band.

In practice, the Owen confidence band can be built by computing $(n+1)$ values of L_i and H_i for $i = 0, ..., n$ where L_i and H_i are the respective values of $F_n^L(x)$ and $F_n^U(x)$ on the open interval $\left(X_{(i)}, X_{(i+1)}\right)$, and $X_{(0)}$ and $X_{(n+1)}$ are the bounds of the support of $F(x)$. The following procedure can be used:

1. Compute (for $c_{BJ}(\alpha) > 0$)

$$H_i = \begin{cases} 1 - e^{-c_{BJ}(\alpha)} \text{ if } i = 0 \\ 1 \text{ if } i = n \\ \max\limits_{\frac{i}{n} \leq p \leq 1} \left\{p : K[F_n(x), p] \leq c_{BJ}(\alpha)\right\} \text{ if } 2 \leq i \leq n-1 \end{cases}$$

2. Deduce $L_i = 1 - H_{n-i}$ for $0 \leq i \leq n$ (by symmetry)

3. Compute the values of the confidence band for each observation $X_{(i)}$ using $F_n^L(X_{(i)}) \leq F(X_{(i)}) \leq F_n^U(X_{(i)})$ where $F_n^L(X_{(i)}) = \min(L_{i-1}, L_i) = L_{i-1}$ and $F_n^U(X_{(i)}) = \max(H_{i-1}, H_i) = H_i$

Equivalently, one can build the confidence band computing first L_i and deducing $H_i = 1 - L_{n-i}$ as a second stage. Owen (1995) proposed the following polynomial approximation for $c_{BJ}(\alpha)$:

$$c_{BJ}(0.05) = \begin{cases} \frac{1}{n}\left[3.0123 + 0.4835\log(n) - 0.00957\log^2(n) - 0.001488\log^3(n)\right] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}\left[3.0806 + 0.4894\log(n) - 0.02086\log^2(n)\right] & \text{for } 100 < n \leq 1000, \end{cases}$$

and

$$c_{BJ}(0.01) = \begin{cases} \frac{1}{n}[-4.626 - 0.541\log(n) + 0.0242\log^2(n)] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}[-4.71 - 0.512\log(n) + 0.0219\log^2(n)] & \text{for } 100 < n \leq 1000. \end{cases}$$

Studying the Owen's confidence band, Jager and Wellner (2004) found that this CB has a coverage probability lower than the theoretical level of confidence. Their simulations show that for a theoretical confidence level of 95%, the Owen's CB provides a simulated coverage probability from 90 to 93% for sample sizes $n = 2$ to 1000 and $N = 100,000$ replications. According to Jager and Wellner (2004), this shortcoming is due to the polynomial approximations proposed for $c_{BJ}(\alpha)$. These approximations yield values of $c_{BJ}(\alpha)$ much lower than their simulated values. Jager and Wellner (2004) provide the following approximation for the critical points:

$$c_{BJ}(0.05) = \begin{cases} \frac{1}{n}[3.6792 + 0.5720\log n - 0.0567\log^2(n) - 0.0027\log^3(n)] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}[3.7752 + 0.5062\log n - 0.0417\log^2(n) + 0.0016\log^3(n)] & \text{for } 100 < n \leq 1000, \end{cases}$$

and

$$c_{BJ}(0.01) = \begin{cases} \frac{1}{n}[5.3318 + 0.5539 \log n - 0.0370 \log^2(n)] & \text{for } 1 < n \leq 100, \\ \\ \frac{1}{n}[5.6392 + .04018 \log n - 0.0183 \log^2(n)] & \text{for } 100 < n \leq 1000. \end{cases}$$

They show that the new approximated values of $c_{BJ}(\alpha)$ are closer to the simulated values than the Owen's approximated ones. Moreover, the CB based on the Jager-Wellner critical values are of better coverage probability than those based on the Owen critical points with, as a consequence, a larger width. Given that the simulated critical points allow controlling the levels of the test and the CB, we recommend to use these instead of the approximated values.

Owen (1995) shows that the Berk-Jones statistic yields a CB that is narrower in the tails of distributions and wider in the middle than the Kolmogorov-Smirnov CB.

Another interesting property of the Owen's CB is that it can be computed using the same critical points for all samples from continuous distribution functions. This feature simplifies its computation but not as much as the pivotality of the weighted KS statistics simplifies the computation of their corresponding CBs. In fact, the Berk Jones-based CB suffers from an important computational cost. Computing this CB requires one to perform as many optimizations as the number observations–n. Hence, the performance of the inference depend greatly on those of the optimization procedure that is used and building the CB may be highly time consuming when using large samples, which is not the case for the regularized Kolmogorov-Smirnov CBs.

Owing to the monotonicity properties of $D\left(F_n, F\right)$, the Owen CB can be extended to noncontinuous distribution functions. When the sample comes from a noncontinuous distribution function, critical values for distribution functions with embedded image sets are ranked. Hence, the corresponding CBs for those distributions are embedded. First, the Owen CB using adequate critical points of level α for continuous distribution functions provides a CB with level greater than or equal to $1 - \alpha$. Second, using information on the image set of $F(x)$, narrower CBs can be built for this distribution without altering

the reliability of the inference.

There exist other statistics based on the empirical distribution function that can yield exact CBs for continuous distribution functions. Among them are the Anderson-Darling (1952), the Cramer (1928) and Von Mises (1931) statistics. However, even though the goodness-of-fit tests performed with these statistics are easy to compute, the corresponding CBs for distribution functions generally do not have explicit expressions and must be computed numerically.

1.5 Regularized Anderson-Darling and Eicker-type statistics and confidence bands

The CBs presented above perform better than those based on the non-weighted KS statistic. However, the Anderson-Darling and the Eicker statistics suffer from important drawbacks. For observations in the tails of distributions, both $F(x)$ and $F_n(x)$ converge to 0 and 1. Hence, the denominators of those statistics become very close to 0, which leads to an erratic behavior of the ratio. This feature alters the performance of the Anderson Darling-type and the Eicker-type tests and CBs.

1.5.1 Regularization

To solve this problem, we propose improved weighted KS statistics. These statistics are regularized versions of the previous ones where the variance of $F_n(x) - F(x)$ is corrected by adding a positive nonzero regularization term $\zeta_n\left(F_n(x), F(x)\right)$:

$$AD_\zeta = \sup_{-\infty<x<+\infty} \sqrt{n} \left| \frac{F_n(x) - F(x)}{\sqrt{F(x)[1-F(x)] + \zeta_n\left(F_n(x), F(x)\right)}} \right| \tag{1.8}$$

$$E_\zeta = \sup_{-\infty<x<+\infty} \sqrt{n} \left| \frac{F_n(x) - F(x)}{\sqrt{F_n(x)[1-F_n(x)] + \zeta_n\left(F_n(x), F(x)\right)}} \right|$$

The regularization achieves the expected improvement. Shifting the denominator of the statistics by an additional term modifies the weights such that they don't vanish in the tails. This modification avoids the erratic behavior of the statistics and improves the performance of the tests. However, the statistics retain the advantages of weighted KS statistics. Observations in the center of the distribution are less weighted than those in the tails, which enhances the performance of the tests when applied to distributions with more difference through the tails.

Let's assume that $\zeta_n\left(F_n(x), F(x)\right) = \zeta_n > 0$, $\forall x$. Then, the regularized statistics can be computed in practice using the following expression:

$$AD_\zeta = \max\left\{\max_{1\le i\le n}\sqrt{n}\frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})] + \zeta_n}}, \max_{1\le i\le n}\sqrt{n}\frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})] + \zeta_n}}, 0\right\},$$

$$E_\zeta = \max\left\{\max_{1\le i\le n}\sqrt{n}\frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}] + \zeta_n}}, \max_{1\le i\le n}\sqrt{n}\frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{\frac{i-1}{n}(1-\frac{i-1}{n}) + \zeta_n}}, 0\right\}.$$

where $\zeta_n > 0$ $\forall x$.

Inverting the tests, we propose improved nonparametric CBs for distribution functions using the regularized statistics. The ζ-Regularized Anderson Darling-type CB for $F(x)$ with level greater than or equal to $1-\alpha$ is:

$$C_F^{AD_\zeta}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(x) \le F_0 \le G_n^U(x),\ \forall x\right\} \tag{1.9}$$

where

$$G_n^L(x) = \frac{2F_n(x) + \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta}}{2\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)}, \qquad G_n^U(x) = \frac{2F_n(x) + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta}}{2\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)},$$

$$\Delta = \left[2F_n(x) + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(F_n^2(x) - \frac{\zeta_n c^2_{AD_\zeta}(\alpha)}{n}\right),$$

$c_{AD_\zeta}(\alpha)$ satisfies $\Pr[AD_\zeta \leq c_{AD_\zeta}(\alpha)] \geq 1-\alpha$. The ζ-Regularized Eicker-type CB for $F(x)$ with level greater than or equal to $1-\alpha$ is:

$$C_F^{E_\zeta}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(x) \leq F_0(x) \leq G_n^U(x),\ \forall x\right\} \tag{1.10}$$

where

$$G_n^L(x) = F_n(x) - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(x)\left(1-F_n(x)\right)+\zeta_n\right]^{1/2},$$

$$G_n^U(x) = F_n(x) + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(x)\left(1-F_n(x)\right)+\zeta_n\right]^{1/2},$$

and $c_{E_\zeta}(\alpha)$ satisfies $\Pr[E_\zeta \leq c_{E_\zeta}(\alpha)] \geq 1-\alpha$. Owing to the decreasing weights of the underlying statistics, these CBs are nonuniform. Their widths decrease as observations approach the tails of the distribution, even though they do not converge to 0 and 1 in the tails of distributions. Moreover, the regularization resolves the problem of discontinuity of the Eicker CB.

As the initial statistics, the regularized ones are expressed under the general form $D\left(F_n, F\right)$. Consequently, they are pivotal when applied to continuous distribution functions and their distribution can be characterized using uniform order statistics as follows:

$$E_\zeta^{cont} = \max\left\{\max_{1\leq i\leq 1}\sqrt{n}\frac{\frac{i}{n}-U_{(i)}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta_n}}, \max_{1\leq i\leq n}\sqrt{n}\frac{U_{(i)}-\frac{i-1}{n}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta_n}}, 0\right\}$$

and

$$AD_\zeta^{cont} = \max\left\{\max_{1\leq i\leq 1}\sqrt{n}\frac{\frac{i}{n}-U_{(i)}}{\sqrt{U_{(i)}[1-U_{(i)}]+\zeta_n}}, \max_{1\leq i\leq n}\sqrt{n}\frac{U_{(i)}-\frac{i-1}{n}}{\sqrt{U_{(i)}[1-U_{(i)}]+\zeta_n}}, 0\right\}$$

where $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$ are the order statistics of a sample of n i.i.d. observations of an uniform $U_{[0,1]}$ distribution (see Appendix 2 for computation details). Likewise, the critical points of AD_ζ^{cont} and E_ζ^{cont} are independent of $F(x)$ being tested under the

null hypothesis and the corresponding CBs depend on this distribution only through the sample. Adapted critical values do not need to be computed for each distribution under study.

1.5.2 Selection of the regularization parameter ζ_n

In this section, we discuss the choice of the regularization term. Adding this term to the weights used by the Anderson-Darling and the Eicker statistics prevents the denominator of those statistics to become too close to 0 and thus, stabilizes their behavior. However, the regularization term must be specified to compute the regularized statistics and this choice is not obvious. Two key issues need to be considered.

The first one concerns the properties of the statistics: when the regularization term is chosen according to the sample, the pivotality of the tests for continuous distributions might be lost. In this case, the distribution of this term may modify those of the statistics. The properties derived so far may not hold anymore and new critical points may need to be computed for each distribution function. To avoid this problem, we have chosen a regularization term of the form $\zeta_n\left(F_n(x), F(x)\right)$. In this case the studied statistics can still be rewritten as $D\left(F_n, F\right)$ and hence, they are pivotal for continuous distribution functions. So far, we have assumed that $\zeta_n\left(F_n(x), F(x)\right) = \zeta_n > 0$ is a constant function. This choice allows to build CBs with friendly expressions. However, in this case, the optimal value of ζ_n must be estimated. This value depends on the sample but does not have an explicit expression. Hence, the way to determine ζ_n must be chosen carefully in order to preserve the statistics' properties. We propose to estimate the parameter ζ_n and the CBs independently each from the other using a split sample procedure (see Dufour and Jasiak, 2001). The procedure decomposes as follows. First, the initial sample is divided into two independent subsamples using i.i.d. drawings. Second, one sample–the auxiliary sample–is used to estimate the optimal value of the parameter ζ_n. Third, the remaining sample–the estimation sample–is used to perform the tests or to build CBs with the formulas provided in the above sections. The out-of-sample procedure insures

that the auxiliary sample and the estimation sample are independent. The statistics are used conditionally to the value of ζ_n which held their distributions unchanged by the estimation of the parameter and guarantees the validity of the inference. Usually, a small part of the initial sample is used as auxiliary sample–some theoretical studies (see Dufour, 2001) recommends to use up to 10 percent of the sample. However, given that the performance of our inference methods depends a lot on the value of ζ_n, we propose to use at least 20 percent of the initial sample to estimate ζ_n, if the sample size allows us to do so.

Second, once the auxiliary sample is determined, the next step is to define a criterion for chosing ζ_n. The criterion will depend on the objective of the ongoing inference. For example, if the objective is to build CBs for distribution functions with a quite uniform shape, one can choose ζ_n so as to minimize the mean of the widths of the CB over the sample. If, conversely, the distribution is heavy tailed, the observations in the tails are more important than the observations in the center of the distribution. Hence, the criterion may be a weighted mean of the widths of the CB, with larger weights for observations in the tails than those in the center of the distribution. In the case where information about the distribution of the sample is known, one can also choose the minimum value of ζ_n that provides a "sufficiently" powerful test. In fact, given that exact critical points are used, the levels of the tests and those of the corresponding CBs are controlled. Thus, the value of ζ_n that maximizes the power of the goodness of fit tests also minimizes the width of the corresponding CBs (see Pratt, 1961). An example of how to choose the optimal value of ζ_n using this approach is provided in section 6 using Monte Carlo simulations. In a subsequent paper, we will illustrate how to choose ζ_n to perform inference on the mean of random variables. Using a split sample procedure, we will choose ζ_n so as to minimize the width of the confidence intervals for the mean we are interested in. Each procedure being different, the choice of criterion is likely to affect the performance of the inference. This gives room to further improvements.

For the sake of simplicity, we will name the regularization ζ for the remainder of the paper.

1.6 Monte Carlo study

Adding the regularization parameter ζ to the weights used by the Anderson-Darling and Eicker statistics prevents the denominator of those statistics to become too close to 0 and thus, stabilizes their behavior. But, beside this improvement, another issue is of interest: what is the impact of the regularization on the level and the power of the regularized Anderson-Darling and Eicker goodness-of-fit tests? Using exact simulated critical points allows us to control the level of the tests but does not affect the power. Does the regularization improve the power of the tests? Does this effect differ when increasing values of the parameter are used? How to choose the optimal value of ζ? In this section, we will use Monte Carlo simulations to study the effect of adding the regularization parameter ζ to the Anderson-Darling and the Eicker statistics and illustrate how to choose ζ. We will illustrate how the CBs derived from those statistics compare to each other.

1.6.1 Effect of the regularization parameter ζ

We test the null hypothesis $H_0 : X \sim N(0,1)$ vs. the alternative $H_1 : X \sim N(0,1.2)$ using the $\zeta-$regularized Anderson-Darling and Eicker statistics. We compute the level and the power of these tests by Monte Carlo simulations using values of ζ from 0 to $1,000,000$ and sample size $n = 500$. The tests use exact critical values simulated with $N_1 = 3,000,000$ replications and the level and the power of the tests are simulated using $N_2 = 15,000$ replications. Tables 1.1 and 1.2 show the results for the regularized Anderson-Darling and Eicker tests, respectively.

Table 1.1. Effect of the regularization parameter: Critical values, level, and power of the $\zeta-$regularized Anderson-Darling test for different values of ζ

$n = 500$, $N_1 = 3,000,000$ replications for C_{AD}, and $N_2 = 15,000$ replications for the test

$$H0: \ X \sim N(0,1) \qquad \text{vs.} \qquad H1: X \sim N(0,1.2)$$

ζ	C_{AD}	Level (in %)	Power (in %)
0	6.45343825318425	4.97	72.05
0.0001	4.18358963085057	4.83	97.62
0.0005	3.56948140935572	4.85	99.31
0.001	3.42358267535564	5.25	99.53
0.005	3.16539066782623	5.15	99.27
0.07	2.56835999326805	5.16	94.65
0.1	2.42578653035515	5.07	92.90
0.15	2.24054390190001	5.25	90.14
0.2	2.09507654002148	4.88	87.72
0.3	1.87596864573845	5.01	84.33
0.4	1.71567147669466	5.07	81.71
0.5	1.59135382254647	5.13	79.55
0.75	1.37021882434391	5.01	77.28
1	1.22108493720676	5.16	75.83
10	0.42233372755807	5.03	70.22
100	0.13484610150674	5.20	69.21
1000	0.04269341845902	5.05	69.81
1000000	0.00135081892087	4.92	68.45

Table 1.1 shows that the regularization has a major impact on the power of the Anderson-Darling test. While the level of the test is controlled by using exact critical points, the power is low for $\zeta = 0$ but rises quickly when ζ increases before becoming almost constant. However, when ζ is too large, it introduces too much distortion into the distribution of the statistics, which reduces the power of the test and even cancels the

improvement of the regularization. The results show that the improvement is achieved as soon as ζ is high enough to prevent the weight of the statistic from vanishing. The power is 72.05 percent for $\zeta = 0$ and jump to 97.62 for $\zeta = 10^{-4}$ and 99.53 for $\zeta = 10^{-3}$.

Table 1.2 shows similar results for the Eicker test. The regularization achieves the expected improvement for this test too, with an even stronger impact. The power of the test is very low for small values of ζ (6.21 percent for $\zeta = 0$) while it increases sharply for ζ high enough to reach 85.09 for $\zeta = 0.07$ before stabilizing. Table 1.2. also illustrates the noncontinuity of the Eicker statistic. While the power of the test is very low for the non regularized statistic, it is even smaller when the regularization is introduced but ζ is not high enough (for $\zeta < 0.005$).

In conclusion, it appears that the regularization achieves the expected improvement. Moreover, while most of the improvement is achieved as soon as ζ is high enough, using too large values of ζ can hamper the performance of the inference. Hence, we propose to choose the value of ζ that increases "sufficiently" the power of the test. Tables 1.1 and 1.2 show that this value is not the same for the two statistics. The maximum of power is achieved for $\zeta_{AD} = 0.001$ for the Anderson-Darling test and $\zeta_E = 0.07$ for the Eicker one. Moreover, simulations show that for each test, these optimal values depend on the distribution being tested and on the size of the sample. However, as the results show, even if the optimal value is not used, most of the improvement is achieved as soon as reasonably high. We provide other illustrations of how to choose ζ in practice in a subsequent paper on nonparametric confidence intervals for the mean of a bounded random variable.

Table 1.2. Effect of the regularization parameter: critical values, level, and power of the $\zeta-$regularized Eicker test for different values of ζ

$n = 500$, $N_1 = 3,000,000$ replications for C_E, and $N_2 = 15,000$ replications for the test

$$H0: \ X \sim N(0,1) \qquad \text{vs.} \qquad H1 : X \sim N(0, 1.2)$$

ζ	C_E	Level (in %)	Power (in %)
0	4.79920250051836	5.07	6.21
0.0001	16.35442032904740	4.81	0.00
0.0005	7.34715907201899	4.85	0.00
0.001	5.35678751417434	4.89	0.84
0.005	3.50097616169556	4.95	70.51
0.070	2.58974686616846	5.25	85.09
0.1	2.44086016301592	5.13	84.09
0.15	2.24958797778716	5.19	81.90
0.2	2.10106534783129	4.89	80.62
0.3	1.87950819352882	5.03	78.46
0.4	1.71790568033589	4.97	76.47
0.5	1.59320398394615	5.08	74.91
0.75	1.37101349713361	5.06	74.12
1	1.22177934463805	5.15	73.19
10	0.42234366084257	5.05	69.95
100	0.13484748513727	5.20	69.19
1,000	0.04269347073131	5.05	69.81
1,000,000	0.00135081895023	4.92	68.45

1.6.2 Relative performance of the EDF-based goodness-of-fit tests

In this subsection, we compare the performance of the regularized-based tests to those of the other tests we have presented. Two procedures are used.

First, we test the null hypothesis $H_0 : X \sim N(0,1)$ vs. the alternative $H_1 : X \sim N(0,1.2)$ using the empirical distribution function-based statistics. We compute the level and the power of these tests by Monte Carlo simulations for sample size $n = 500$, exact critical values simulated with $N_1 = 3,000,000$ replications, and $\zeta_{AD} = 0.001$ and $\zeta_E = 0.07$. The levels and the powers of the tests are simulated using $N_2 = 15,000$ replications and Table 1.3. shows the results.

Table 1.3. Level and power of the EDF-based tests

$n = 500$, $\zeta_{AD} = 0.001$, $\zeta_E = 0.07$,

$N_1 = 3,000,000$ replications for the critical values, and $N_2 = 15,000$ replications for the test $H_0 : X \sim N(0,1)$ vs. $H_1 : X \sim N(0,1.2)$

	C	Level (in %)	Power (in %)
Kolmogorov-Smirnov	0.06039953611469	4.60	68.44
Eicker	4.79617625286106	5.35	6.25
Eicker$_\zeta$	2.59103603317055	4.64	85.09
Anderson-Darling	6.45272451410767	4.83	71.81
Anderson-Darling$_\zeta$	3.42243384382137	4.83	99.46
Berk-Jones	0.01138040175450	4.57	98.63

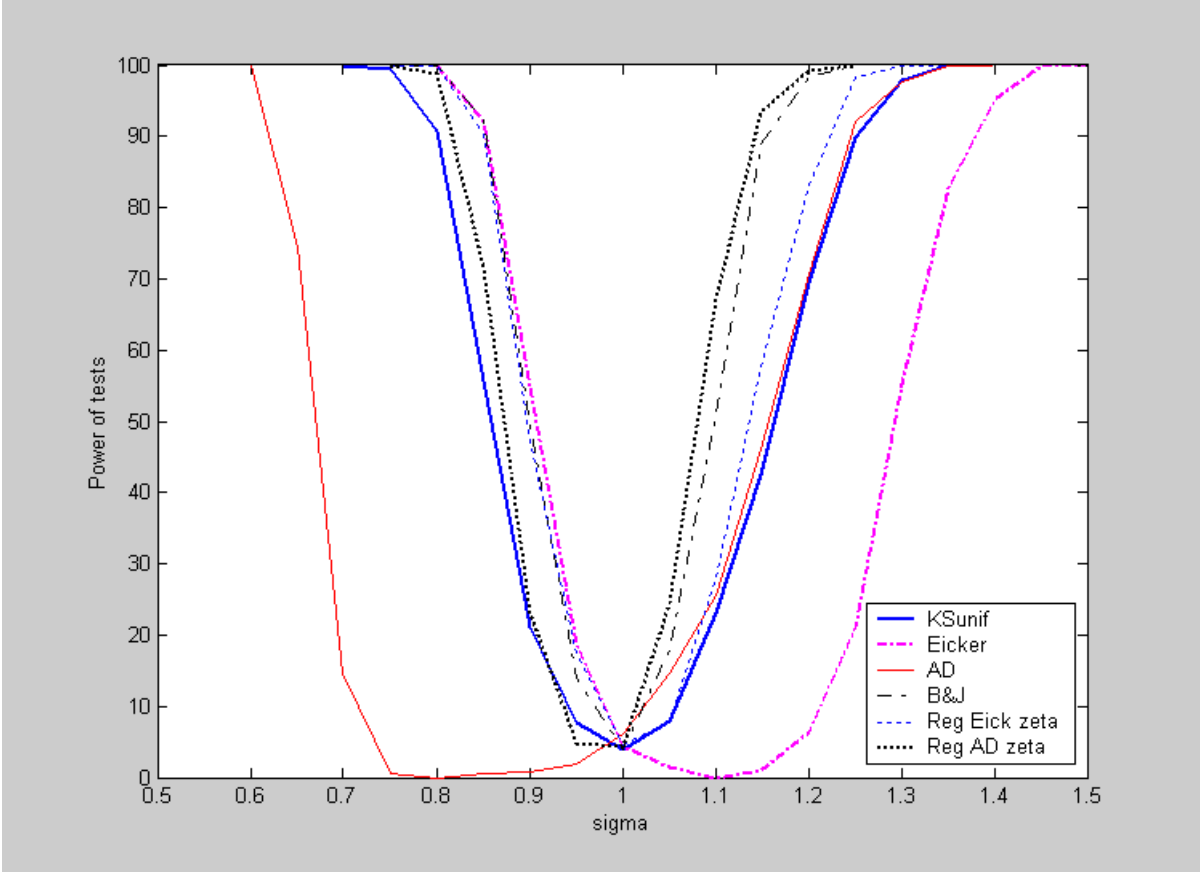
Among all the tests, the $\zeta-$regularized Anderson Darling test achieves the best power (99.46 percent) followed by the likelihood-ratio based test (98.63 percent) and the $\zeta-$regularized Eicker type test (85.09 percent). The Eicker test achieves the less power, which illustrate the erratic behavior of its statistic whereas the Anderson-Darling test has more power than the unweighted Kolmogorov-Statistic.

Second, we test the null hypothesis $H_0 : X \sim N(0,1)$ vs. the alternative $H_1 : X \sim N(0,\sigma)$ for $\sigma = 0.5, 0.55, 0.6, \ldots, 1.5$. We compute the level and the power of these tests by Monte Carlo simulations for the same setting as earlier using the optimal values of ζ ($\zeta_{AD} = 0.001$ and $\zeta_E = 0.07$). Graph1.1 pictures the evolution of the power of tests as σ varies.

Graph 1.1. Level and power of the EDF-based tests

$$H_0 : \ X \sim N(0,1) \qquad \text{vs.} \qquad H_1 : X \sim N(0,\sigma)$$

for $\sigma = 0.5,\ 0.55,\ 0.6, \ldots,\ 1.5$, $n = 500$, $\zeta_{AD} = 0.001$, $\zeta_E = 0.07$, $N_1 = 3,000,000$ replications for the critical values, and $N_2 = 15,000$ replications for the tests



The results shows that tests perform differently when σ is larger than 1 than when σ is smaller than 1. For values of σ greater than 1−i.e., when the sample under study actually comes from a distribution with heavier tails than the one assumed under the null hypothesis, the $\zeta-$regularized Anderson-Darling test yields the best power among the EDF-based tests followed by the Berk-Jones tests and the $\zeta-$regularized Eicker test. The Kolmogorov-Smirnov and the Anderson-Darling tests perform quite similarly and

achieve a better power than the Eicker test, which performs the worst among the studied inference methods.

Conversely, when the analyzed sample comes from a distribution with thinner tails than the one assumed under the null hypothesis, i.e., when σ is smaller than 1, the Eicker-type statistics perform better than the other. The Eicker test achieves the best performance, followed by the $\zeta-$regularized Eicker and the Berk-Jones tests whereas the Anderson-Darling one provides the poorest one. The $\zeta-$regularized Anderson-Darling test performs better than the Kolmogorov-Smirnov one except when the two distributions are very close one to the other.

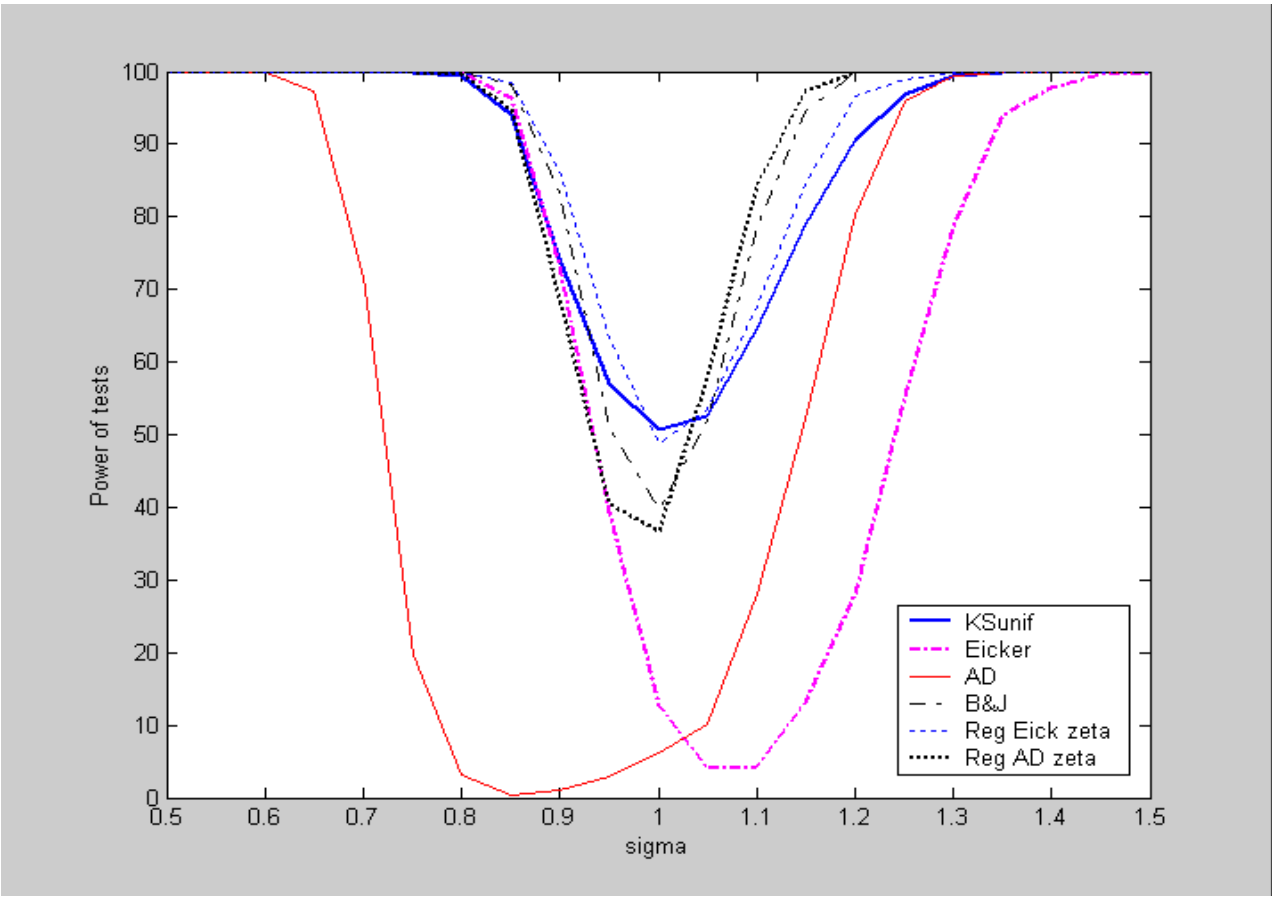
Third, we test the null hypothesis $H_0 : X \sim N(0, 1)$ vs. the alternative $H_1 : X \sim N(0.1, \sigma)$ for $\sigma = 0.5, 0.55, 0.6, \ldots, 1.5$. We compute the level and the power of these tests by Monte Carlo simulations for the same setting as earlier using the optimal values of ζ ($\zeta_{AD} = 0.001$ and $\zeta_E = 0.07$). Graph 1.2 shows the evolution of the power of tests as σ varies.

Graph 1.2. Level and power of the EDF-based tests

$H_0: X \sim N(0,1)$ vs. $H_1: X \sim N(0.1, \sigma)$ for $\sigma = 0.5, 0.55, 0.6, \ldots, 1.5$,

$n = 500$, $\zeta_{AD} = 0.001$, $\zeta_E = 0.07$, $N_1 = 3,000,000$ replications for the

critical values, and $N_2 = 15,000$ replications for the tests



The results shows that for values of σ greater than 1−i.e., when the sample under study actually comes from a distribution with heavier tails than the one assumed under the null hypothesis, the $\zeta-$regularized Anderson-Darling test yields the best power among the EDF-based tests followed by the Berk-Jones tests and the $\zeta-$regularized Eicker test. The Kolmogorov-Smirnov test achieves a better power than the Eicker and the Anderson-Darling one, which achieve the smallest power among the studied inference methods.

When the analyzed sample comes from a distribution with thinner tails than the one assumed under the null hypothesis, i.e., when σ is smaller than 1, the results are somewhat reversed. The $\zeta-$regularized Eicker test then achieves the best performance while the Anderson-Darling statistic provides the poorest one. The other inference methods perform relatively similarly. However, when the two distributions are very close one to the other the Kolmogorov-Smirnov tests performs the best among these methods while the Berk-Jones test dominates when σ is smaller than 0.9.

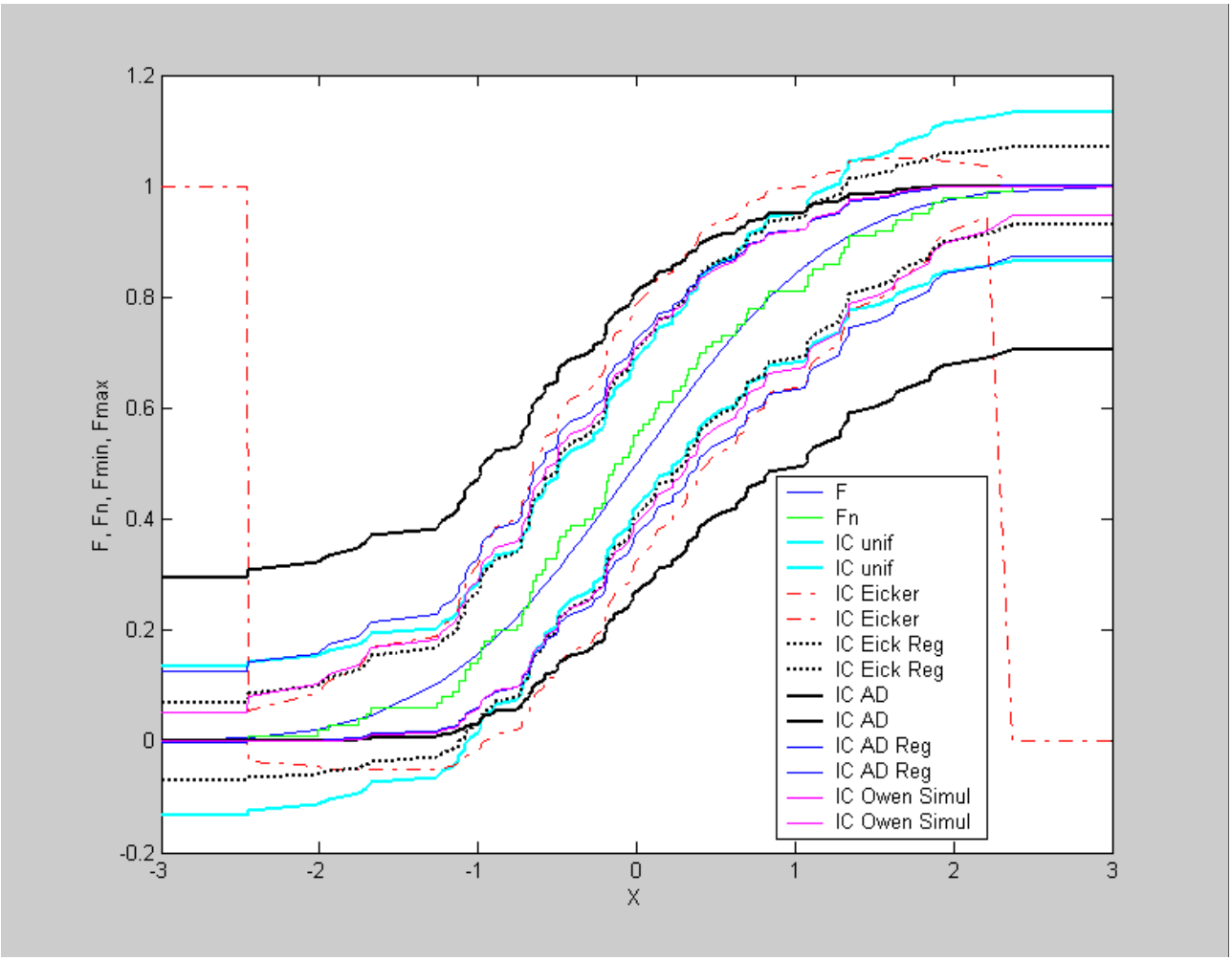
Other interesting conclusions can be driven from Graph 1.1. First, in general, the Eicker-type test allows more discrimination than the Anderson Darling-type tests when the distribution under the alternative hypothesis has heavier tails than those over the null one. Second, as expected, when the actual distribution of the sample is very close to the distribution being tested under the null hypothesis, the weighted statistics perform worse than the uniform KS statistic. Third, the results show that the Anderson-Darling and the Eicker tests are biased. In fact, the power of these tests do not reach their minimum values when $\sigma = 1$ but when it is slightly different from 1.

Last, let's highlight an important advantage of the procedure we use. Usually, regularized statistics are biased due to the distortion the regularization term introduces to the distribution of the initial statistic. By computing the critical values by simulation, we offset this shortcoming. In fact, using the exact critical points controls the level of the tests and the CBs, avoiding the bias. Likewise, regularizing the statistics also offsets suppresses the bias of the initial Eicker and Anderson-Darling statistics.

1.6.3 Performance of confidence bands for distribution functions

As a last illustration, we compare the relative performance of the CBs based on the tests we derived. We build CBs for the Normal $N(0, 1)$ distribution using a sample of $n = 100$. While the small number of observations will probably hamper the performance of the inference methods, it will insure to have a graph clear enough to compare CBs easily. Graph 1.3 depicts the results.

Graph 1.3. Empirical distribution function-based confidence bands for the distribution function $N(0,1)$
$n = 100$, $\zeta_{AD} = 0.001$, and $\zeta_E = 0.07$



Graph 1.3 shows that for observations in the center of the distribution, the Kolmogorov-Smirnov CB has the smallest width among the KS-based CBs, closely followed by the $\zeta-$regularized Eicker-type CB. The Owen and the $\zeta-$regularized Anderson Darling-type CB achieve the best following performance. The Anderson Darling-type and the Eicker-type CBs perform the worse among the studied inference methods.

In the tails of the distribution, the ranking of the performance of the inference methods changes a lot. The uniform CB performs worse than most of the weighted CBs. The $\zeta-$regularized Anderson Darling-type CB performs the best followed by the $\zeta-$regularized Eicker-type CB. The Eicker statistic provides a CB whose width tends to zero as observations go further from the center of the distribution. However, the Eicker CB suffers from discontinuity at the first and last observations of the sample. For all values of x lower than the first observation ($x < X_{(1)}$) or greater or equal to the last observation ($x \geq X_{(n)}$), the CB becomes the non informative $[0, 1]$ interval. At the opposite, the Anderson-Darling CB is always continuous. Its width converges to $c_{AD}^2(\alpha)/(n + c_{AD}^2(\alpha))$ in the tail of the sample which is to compare to 2 $c_{KS}(\alpha)$, the constant width of the Kolmogorov-Smirnov CB. Adding the regularization term to the Eicker statistic corrects the discontinuity problem of the Eicker CB. Nevertheless, for ζ_E different from zero, the width of the corresponding CB does not converge to zero anymore but performs well, with a width—equal to $2c_E^2(\alpha)\zeta^{1/2}/n^{1/2}-$ around half the width of the uniform CB for $n = 100$. Likewise, adding the regularization term to the Anderson-Darling statistic improves a lot the performance of the CB, in particular in the tails. We simulate the critical values of the statistics for sample sizes from 50 to 1000 using $N = 1,000,000$ replications. The results (see Table 1.4.) shows that for n smaller than 150, the KS confidence band performs better than the Anderson-Darling CB in the very end of the tails of continuous distributions but the Anderson-Darling CB becomes better in the tails for n greater than 150. The statistics using regularization parameters yield CBs with smaller widths than those without regularization and than the uniform CB.

In conclusion, we see that as expected, weighted KS statistics achieve better performance than the unweighted one at some points of the distribution and for large and moderately large samples but do not clearly dominate the latter. Conversely, the regularized statistics dominates the other KS-based ones in the tails of distributions for all sample sizes. Moreover, it appears that the $\zeta-$regularized Anderson Darling-type CB is the best CB among the KS-based one: its width in the center of the distribution is very

close to that of the uniform one while it achieves the best width in the tails.

Table 1.4. Simulated critical points of empirical distribution-based statistics and width of the corresponding confidence bands in the tails of distributions $n = 50, 100, \ldots, 1000$ using $N = 1,000,000$ replications

Table 1.4 a. Critical Points

	n		
	50	100	150
KS	0.188335597	0.134056205	0.109638527
E	4.522280174	4.666516217	4.70725121
E_ζ	2.793607145	2.673939389	2.635861357
AD	6.43881938	6.457275098	6.474963046
AD_ζ	4.14789771	3.780490269	3.681454542
BJ	0.104133743	0.053727502	0.036421672

	n			
	200	250	500	1000
KS	0.095178969	0.085207603	0.060368076	0.042771814
E	4.739536779	4.763956841	4.798134574	4.823671809
E_ζ	2.619270234	2.610421227	2.587466879	2.580100277
AD	6.458800396	6.45095347	6.454634534	6.444683756
AD_ζ	3.600576974	3.540729297	3.424467947	3.352876861
BJ	0.027593217	0.022293228	0.011369705	0.005804824

Table 1.4b. Width of CBs in the tails of distributions

n	KS	E	E_ζ	AD	AD_ζ	BJ
50	0.38	1.00	0.21	0.45	0.37	0.10
100	0.27	1.00	0.14	0.29	0.23	0.05
150	0.22	1.00	0.11	0.22	0.17	0.04
200	0.19	1.00	0.10	0.17	0.14	0.03
250	0.17	1.00	0.09	0.14	0.13	0.02
500	0.12	1.00	0.06	0.08	0.08	0.01
1000	0.09	1.00	0.04	0.04	0.06	0.01

When comparing the KS based CBs to the Owen one—using simulated critical points which allow to control the level of the test and CBs, it appears that in the center of the distribution, the Owen CB perform worse than the uniform CB and the $\zeta-$regularized Anderson Darling-type CB but better than the $\zeta-$regularized Eicker-type CB. Conversely, in the tails of the distribution, the Owen CB overcomes all the other CBs. However, the Owen critical has a major drawback: it is very computationally demanding. Building the Owen CB requires to perform as many optimizations as there are observations, which is time demanding and condition the performance of the inference method to that of the optimization method used.

1.7 Conclusion

The Kolmogorov-Smirnov test is one of the most popular nonparametric goodness-of-fit tests. It allows to test the null hypothesis that a sample follows a given distribution against the alternative that the sample does not follow this distribution, without any hypothesis on the law of the studied sample. However, the Kolmogorov-Smirnov test does not allow to discriminate a lot between distributions that differ mostly through their tails. The CB it allows to build is uniform: its width is constant for all observations and do not converge to 0 and 1 as do the distribution functions it brackets. To correct

this drawback, we study weighted KS statistics based on the three common principles in econometrics: the Wald, likelihood-ratio, and Lagrange multiplier principles.

Weighted Kolmogorov-Smirnov statistics have been proposed by Anderson and Darling (1952), and Eicker (1979). However, these statistics suffer from important drawbacks too. For observations in the tails of distributions, the weights in the denominators of those statistics become very close to zero, leading to erratic behavior of the statistic.

We propose regularized weighted statistics to correct these limits. These statistics are obtained by adding a regularization term in the denominator of the Anderson-Darling and the Eicker statistics. They retain the advantages of the weighted KS statistics but do not suffer from instability, improving the performance of the inference.

We show that in the continuous case, the distributions of the empirical distribution-based statistics of the form of the Kolmogorov-Smirnov and the weighted Kolmogorov-Smirnov statistics are pivotal and that their critical points do not depend on the distribution function being tested under the null hypothesis. Inverting these statistics, we propose exact CBs for distribution functions, which inherit the properties of the statistics they are based on. We show that for all continuous distribution functions, these CBs depend on the distribution only through the sample. A unique set of critical points is needed to build CBs for all continuous distributions, which make them easy to compute. For noncontinuous distribution functions, we derive monotonicity properties that exploit embeddedness of the image sets of different distributions to narrow the CBs without altering their reliability. We study a statistic based on the likelihood-ratio principle: the Berk-Jones statistic and the Owen CB, which is derived from it. We show that these statistics and CBs follow the same properties as the Kolmogorov-Smirnov based ones.

We study the performance of these inference methods using Monte Carlo simulations. The results show that the regularized statistics deliver the best performance among the studied inference methods. The regularization increases the power of the tests and the corresponding CBs are thinner than the other ones. Compared to the weighted statistics, the Owen's band yields generally better results. Nevertheless, it suffers from an important computational shortcoming. In fact, its computation requires to perform as

many optimization as the number of observations of the sample we use. This leads to an important loss of time whereas the computation of the Kolmogorov Smirnov based bands is almost time free.

1.8 Appendix 1: Proofs of propositions and corollaries

PROOF OF PROPOSITION 2.1. For a continuous monotonic, the empirical distribution function is

$$\begin{aligned} F_n(x) &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}[X_k \leq x] \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}[F(X_k) \leq F(x)] \\ &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}[U_k \leq F(x)] \\ &= H(U_1, \ldots, U_n, F(x)) \end{aligned}$$

where $U_i = F(X_i)$. Hence,

$$\begin{aligned} D\left[F_n(x), F(x)\right] &= \sup_{-\infty<x<+\infty} D_1\left[H(U_1, \ldots, U_n, F(x)), F(x)\right] \\ &= \sup_{u\in F(\mathbb{R})} D_1\left[H(U_1, \ldots, U_n, u), u\right]. \end{aligned}$$

If $X_1, \ldots, X_n$ are n i.i.d. observations on X with distribution function $F(x) \in \widetilde{\mathbb{F}}$ then $F(x)$ takes all values between 0 and 1 and $F(X)$ follows a uniform distribution on $[0, 1]$: $F(X) \sim U_{[0,1]}$. Hence, we can rewrite

$$\begin{aligned} D\left[F_n(x), F(x)\right] &= \sup_{-\infty<x<+\infty} D_1\left[H(U_1, \ldots, U_n, F(x)), F(x)\right] \\ &= \sup_{0\leq u\leq 1} D_1\left[H(U_1, \ldots, U_n, u), u\right]. \end{aligned}$$

where $U_i = F(X_i) \sim U_{[0,1]}$ and U_k, $k = 1, .., n$ are i.i.d. The distribution of the statistic and its critical points do not depend on $F(x)$.

PROOF OF PROPOSITION 2.2. Proposition 2.1. states that for continuous distribu-

tion functions, the statistic $D\left(F_n(x), F(x)\right)$

$$D\left(F_n(x), F(x)\right) = \sup_{-\infty<x<+\infty} D_1\left(F_n(x), F(x)\right).$$

is pivotal and can be reexpressed as follows:

$$D\left(F_n(x), F(x)\right) = \sup_{0\leq u\leq 1} D_1\left[H(U_1, \ldots, U_n, u), u\right].$$

If $G(y)$ is a noncontinuous distribution function,

$$\begin{aligned} D\left(G_n(x), G(x)\right) &= \sup_{-\infty<x<+\infty} D_1\left(H\left[G(X_1), \ldots, G(X_n), G(x)\right], G(x)\right) \\ &= \sup_{\upsilon\in G(\overline{\mathbb{R}})} D_1\left(H\left[G(X_1), \ldots, G(X_n), \upsilon\right], \upsilon\right) \\ &\leq D\left(F_n(x), F(x)\right) \end{aligned}$$

because $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}}) = [0, 1]$. The critical points associated with $D\left(F_n, F\right)$ are larger than or equal to the corresponding one for D_G. Critical values associated with continuous distribution functions are conservative for noncontinuous distributions.

Proof of Proposition 2.3. Proposition 2.2. states that $D\left[F_n(x), F(x)\right]$ is greater than or equal to $D\left[G_n(x), G(x)\right]$. Hence, for a given level α, the critical point associated to $D\left[F_n(x), F(x)\right]$ is larger than those associated to $D\left[G_n(x), G(x)\right]$ or equivalently, the critical point with level α associated to $D\left[F_n(x), F(x)\right]$ represents a critical point for $D\left[G_n(x), G(x)\right]$ with $\beta \leq \alpha$. Therefore, the CB for $G(y)$ using the appropriate critical point for $D\left[F_n(x), F(x)\right]$ will be of level $1 - \beta \geq 1 - \alpha$.

Proof of Proposition 2.4. In the proof of Proposition 2.1., we showed that

$$\begin{aligned} D\left[F_n(x), F(x)\right] &= \sup_{-\infty<x<+\infty} D_1\left[H(U_1, \ldots, U_n, F(x)), F(x)\right] \\ &= \sup_{u\in F(\overline{\mathbb{R}})} D_1\left[H(U_1, \ldots, U_n, u), u\right]. \end{aligned}$$

Given that $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}})$, when taking the supremum over $G(\overline{\mathbb{R}})$, the result will be

smaller than when taking the supremum over $F(\overline{\mathbb{R}})$. Hence,

$$D\left[F_n(x), F(x)\right] \geq \sup_{v \in G(\overline{\mathbb{R}})} D_1\left[H(U_1, \ldots, U_n, u), u\right] = D\left[G_n(x), G(x)\right].$$

PROOF OF PROPOSITION 2.5. Proposition 2.4. states that for a given level α, the critical point associated to $D\left[F_n(x), F(x)\right]$ is larger than those associated to $D\left[G_n(x), G(x)\right]$. In other words, the critical point with level α associated to $D\left[F_n(x), F(x)\right]$ represents a critical point for $D\left[G_n(x), G(x)\right]$ with $\beta \leq \alpha$. Hence, using the same reasoning as in the proof of Proposition 2.3., it follows that the CB for $G(y)$ using the appropriate critical point for $F(x)$ will be of level $1 - \beta \geq 1 - \alpha$.

PROOF OF COROLLARY 2.6. This proposition is a special case of Proposition 2 5. In this case, $p_1 = F_1(a) \leq F_2(a) = p_2$. This implies that $F_2(\overline{\mathbb{R}}) = [p_2, 1] \subseteq [p_1, 1] = F_1(\overline{\mathbb{R}})$ and applying Proposition 2.5. yields the result.

1.9 Appendix 2: Details of computation

1.A2.1. Explicit expression of the Kolmogorov-Smirnov statistic.

The Kolmogorov-Smirnov statistic for $F(x)$ is:

$$\begin{aligned} KS &= \sup_{-\infty<x<+\infty} \sqrt{n} \mid F_n(x) - F(x) \mid \\ &= \max\left\{ \sup_{-\infty<x<+\infty} \sqrt{n}\left[F_n(x) - F(x)\right], \sup_{-\infty<x<+\infty} \sqrt{n}\left[F(x) - F_n(x)\right] \right\} \\ &= \max\left\{ \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \sqrt{n}\left[\frac{i}{n} - F(x)\right], \right. \\ &\quad \left. \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \sqrt{n}\left[F(x) - \frac{i}{n}\right] \right\}. \end{aligned}$$

$F(x)$ is non decreasing and $l_i(p) = \frac{i}{n} - p$, $0 \le p \le 1$ is non increasing in p. Hence $l_i(F(x))$ is non increasing in x and

$$\begin{aligned} KS &= \max\left\{ \max_{0\le i\le n} \sqrt{n}\left[\frac{i}{n} - F(X_{(i)})\right], \max_{0\le i\le n} \sqrt{n}\left[F(X_{(i+1)}) - \frac{i}{n}\right] \right\} \\ &= \max\left\{ \max_{1\le i\le n} \sqrt{n}\left[\frac{i}{n} - F(X_{(i)})\right], \max_{1\le i\le n} \sqrt{n}\left[F(X_{(i)}) - \frac{i-1}{n}\right], 0 \right\}. \end{aligned}$$

1.A2.2. Distribution of empirical distribution function-based statistics for continuous distribution functions

Let $F(x) \in \widetilde{\mathrm{F}}$. Then, the random variable $F(X)$ takes all the values between 0 and 1 and follows a uniform distribution on $[0,1]$: $F(X) \sim U_{[0,1]}$. Hence, $F(X_{(i)}) = U_{(i)}$ where $U_{(i)}$ is the i^{th} ordered realization of an uniform $U_{[0,1]}$ distribution. KS can be reexpressed in the form of $D\left(F_n(x), F(x)\right)$ and Proposition 2.3. applies to it. Moreover, the expression of KS simplifies to:

$$\begin{aligned} KS^{cont} &= \max\left\{ \max_{1\le i\le n} \sqrt{n}\left[\frac{i}{n} - F(X_{(i)})\right], \max_{1\le i\le n} \sqrt{n}\left[F(X_{(i)}) - \frac{i-1}{n}\right], 0 \right\} \\ &= \max\left\{ \max_{1\le i\le n} \left[\frac{i}{n} - U_{(i)}\right], \max_{1\le i\le n} \left[U_{(i)} - \frac{i-1}{n}\right], 0 \right\}. \end{aligned}$$

The distribution of KS^{cont} is independent of $F(x)$ and so does its critical points, and the CB for $F(x)$ it implies.

1.A2.3. Kolmogorov-Smirnov CB for distribution functions

Let $c_{KS}(\alpha)$ such that $\Pr[KS_F \leq c_{KS}(\alpha)] \geq 1-\alpha$ i.e.,

$$\Pr[\sup_{-\infty<x<+\infty} \sqrt{n} \mid F_n(x) - F(x) \mid \leq c_{KS}(\alpha)] \geq 1-\alpha.$$

Then, with probability greater than or equal to $1-\alpha$

$$\mid F_n(x) - F(x) \mid \leq \frac{c_{KS}(\alpha)}{n} \qquad \forall\ x$$

$\Leftrightarrow$

$$F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq F(x) \leq F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}},\ \forall x.$$

1.A2.4. Explicit expression of the Anderson-Darling and the Eicker statistics

We develop the expressions of the statistics.

For the Anderson-Darling statistic:

$$\begin{aligned} AD &= \sup_{-\infty<x<+\infty} V_n(x) \\ &= \max\left\{\sup\left\{\sqrt{n}\left|\frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]}}\right| : -\infty < x < +\infty \text{ st } 0 < F(x) < 1\right\}, 0\right\} \end{aligned}$$

$$\begin{aligned} &= \max\ \left\{\sup\left\{\sqrt{n}\left|\frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]}}\right| : -\infty < x < +\infty \text{ st } 0 < F(x) < 1\right\}\right. \\ &\left.\sup\left\{\sqrt{n}\left|\frac{F(x)-F_n(x)}{\sqrt{F(x)[1-F(x)]}}\right| : -\infty < x < +\infty \text{ st } 0 < F(x) < 1\right\}, 0\right\} \end{aligned}$$

$$= \max \left\{ \max_{0\le i\le n} \sup \left\{ \sqrt{n} \left| \frac{\frac{i}{n} - F(x)}{\sqrt{F(x)[1-F(x)]}} \right| : X_{(i)} \le x < X_{(i+1)} \text{ st } 0 < F(x) < 1 \right\}, \right.$$
$$\left. \max_{0\le i\le n} \sup \left\{ \sqrt{n} \left| \frac{F(x) - \frac{i}{n}}{\sqrt{F(x)[1-F(x)]}} \right| : X_{(i)} \le x < X_{(i+1)} \text{ st } 0 < F(x) < 1 \right\}, 0 \right\}$$

Define $l_i(p) = \frac{\frac{i}{n}-p}{[p(1-p)]^{1/2}}$ where $0 < p < 1$.

$$l_i'(p) = \frac{-[p(1-p)]^{1/2} - \frac{1}{2}(\frac{i}{n}-p)(1-2p)[p(1-p)]^{-1/2}}{[p(1-p)]}$$

$$= [p(1-p)]^{-3/2}\left[-\frac{i}{2n} - p(\frac{1}{2}-\frac{i}{n})\right] = c_0 * h_i(p)$$

where $c_0 = [p(1-p)]^{-3/2} \ge 0 \ \forall p$ and $h_i(p) = -\frac{i}{2n} - p(\frac{1}{2} - \frac{i}{n})$.

We know that $-\frac{1}{2}\frac{i}{n} \le 0 \ \forall i \ge 0$ and $\begin{cases} -p(\frac{1}{2}-\frac{i}{n}) \le 0 \text{ for } 0 \le i \le n/2 \\ -p(\frac{1}{2}-\frac{i}{n}) \ge 0 \text{ for } n/2 \le i \le n \end{cases}$. Thus, when $0 \le i \le n/2$, $h_i(p) \le 0 \ \forall \ p$ while when i is such that $n/2 < i \le n$, $h_i(p) \le 0$ if $p < \frac{i}{2i-n}$. Moreover $\frac{i}{2i-n} \ge 1$ when $n/2 \le i \le n$. Hence $p < \frac{i}{2i-n} \ \forall \ i$, $n/2 \le i \le n$; $h_i(p) \le 0 \ \forall \ p$, i and $l_i(p)$ is always non increasing. As a consequence,

1) $\max\limits_{0\le i\le n} \sup \left\{ \sqrt{n}\frac{\frac{i}{n}-F(x)}{\sqrt{F(x)[1-F(x)]}} : X_{(i)} \le x < X_{(i+1)} \text{ st } 0 < F(x) < 1 \right\}$
$= \max \left\{ \sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]}} : 0 \le i \le n \text{ st } 0 < F(X_{(i)}) < 1 \right\}$

2) $\max\limits_{0\le i\le n} \sup \left\{ \sqrt{n}\frac{F(x)-\frac{i}{n}}{\sqrt{F(x)[1-F(x)]}} : X_{(i)} \le x < X_{(i+1)} \text{ st } 0 < F(x) < 1 \right\}$
$= \max \left\{ \sqrt{n}\frac{F(X_{(i+1)})-\frac{i}{n}}{\sqrt{F(X_{(i+1)})[1-F(X_{(i+1)})]}} : 0 \le i \le n \text{ st } 0 < F(X_{(i)}) < 1 \right\}$
$= \max \left\{ \sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]}} : 1 \le i \le n+1 \text{ st } 0 < F(X_{(i)}) < 1 \right\}$

The statistic AD can then be computed using

$$AD = \max \left\{ \max \left\{ \sqrt{n}\frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]}} : 0 \le i \le n \text{ st } 0 < F(X_{(i)}) < 1 \right\}, \right.$$
$$\left. \max \left\{ \sqrt{n}\frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]}} : 1 \le i \le n+1 \text{ st } 0 < F(X_{(i)}) < 1 \right\}, 0 \right\}$$

$\Leftrightarrow$

$$= \max \left\{ \max \left\{ \sqrt{n} \frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{F(X_{(i)})[1 - F(X_{(i)})]}} : 1 \leq i \leq n \text{ st } 0 < F(X_{(i)}) < 1 \right\}, \right.$$
$$\left. \max \left\{ \sqrt{n} \frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{F(X_{(i)})[1 - F(X_{(i)})]}} : 1 \leq i \leq n \text{ st } 0 < F(X_{(i)}) < 1 \right\}, 0 \right\}.$$

For the Eicker statistic:

$$E = \sup_{-\infty < x < +\infty} \widehat{V}_n(x)$$
$$= \max \left\{ \sup \left\{ \sqrt{n} \left| \frac{F_n(x) - F(x)}{\sqrt{F_n(x)[1 - F_n(x)]}} \right| : -\infty < x < +\infty \text{ st } 0 < F_n(x) < 1 \right\}, 0 \right\}$$

$$= \max \left\{ \sup \left\{ \sqrt{n} \frac{F_n(x) - F(x)}{\sqrt{F_n(x)[1 - F_n(x)]}} : -\infty < x < +\infty \text{ st } 0 < F_n(x) < 1 \right\}, \right.$$
$$\left. \sup \left\{ \sqrt{n} \frac{F(x) - F_n(x)}{\sqrt{F_n(x)[1 - F_n(x)]}} : -\infty < x < +\infty \text{ st } 0 < F_n(x) < 1 \right\}, 0 \right\}$$

$$= \max \left\{ \max_{0 \leq i \leq n} \sup \left\{ \sqrt{n} \frac{\frac{i}{n} - F(x)}{\sqrt{\frac{i}{n}[1 - \frac{i}{n}]}} : X_{(i)} \leq x < X_{(i+1)}, 0 < \frac{i}{n} < 1 \right\}, \right.$$
$$\left. \max_{0 \leq i \leq n} \sup \left\{ \sqrt{n} \frac{F(x) - \frac{i}{n}}{\sqrt{\frac{i}{n}[1 - \frac{i}{n}]}} : X_{(i)} \leq x < X_{(i+1)}, 0 < \frac{i}{n} < 1 \right\}, 0 \right\}$$

Define $l_i(p) = \frac{\frac{i}{n} - p}{[\frac{i}{n}(1 - \frac{i}{n})]^{1/2}}$ where $0 < p < 1$. $l_i(p)$ is always non increasing. Then,

$$E = \max\left\{\max\left\{\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]}} : 0\le i\le n,\ 0<\frac{i}{n}<1\right\},\right.$$
$$\left.\max\left\{\sqrt{n}\frac{F(X_{(i+1)})-\frac{i}{n}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]}} : 0\le i\le n,\ \text{st } 0<\frac{i}{n}<1\right\},0\right\}$$

$$= \max\left\{\max_{1\le i\le n-1}\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]}}, \max_{2\le i\le n}\sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{\frac{i-1}{n}[1-\frac{i-1}{n}]}}, 0\right\}.$$

1.A2.5. Anderson Darling-type CB for distribution functions

Let $c_{AD}(\alpha)$ be such that $\Pr[AD \le c_{AD}(\alpha)] = \Pr[\sup_{-\infty<x<+\infty} V_n(x) \le c_{AD}(\alpha)] \ge 1-\alpha$ i.e.,

$$\Pr\left[\sup\left\{0, \sup\left\{\sqrt{n} \mid \frac{F_n(x)-F(x)}{(F(x)[1-F(x)])^{1/2}} \mid : -\infty<x<+\infty \text{ st } F(x)\notin\{0,1\}\right\}\right\} \le c_{AD}(\alpha)\right]$$
$$\ge 1-\alpha\ .$$

It is obvious that $V_n(x)\ge 0\ \forall x$.Then, the above equality yields that with probability greater than or equal to $1-\alpha$

$$\frac{[F_n(x)-F(x)]^2}{F(x)[1-F(x)]} \le \frac{c_{AD}^2(\alpha)}{n} \qquad \forall\ x$$

for x such that $F(x)\notin\{0,1\}$, i.e. $\forall\ x$,

$$F_n^2(x) - 2F_n(x)F(x) + F^2(x) \le \frac{c_{AD}^2(\alpha)}{n}\left(F(x)-F^2(x)\right)$$

$$\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)F^2(x) - \left(2F_n(x)+\frac{c_{AD}^2(\alpha)}{n}\right)F(x) + F_n^2(x) \le 0$$

This condition is satisfied if and only if $\forall\ x$

$$F_n^L(x) \le F(x) \le F_n^U(x)$$

where $F_n^L(x) = \frac{2F_n(x)+\frac{c_{AD}^2(\alpha)}{n}-\sqrt{\Delta}}{2\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)}$; $F_n^U(x) = \frac{2F_n(x)+\frac{c_{AD}^2(\alpha)}{n}+\sqrt{\Delta}}{2\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)}$, and $\Delta(x) = \left[2F_n(x) + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4F_n^2(x)\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right]$.

1.A2.6. Eicker-type CB for distribution functions

Let $c_E(\alpha)$ be such that $\Pr[E \leq c_E(\alpha)] = \Pr[\sup_{-\infty<x<+\infty} \widehat{V}_n(x) \leq c_E(\alpha)] \geq 1-\alpha$, i.e.,

$$\Pr\left[\sup\left\{0, \sup\left\{\sqrt{n} \mid \frac{F_n(x)-F(x)}{(F_n(x)[1-F_n(x)])^{1/2}} \mid: -\infty < x < +\infty \text{ st } F_n(x) \notin \{0,1\}\right\}\right\} \leq c_E(\alpha)\right]$$
$$\geq \quad 1-\alpha \ .$$

It is obvious that $\widehat{V}_n(x) \geq 0 \ \forall x$.Then, the above equality yields that with probability greater than or equal to $1-\alpha$

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

for x such that $F_n(x) \notin \{0,1\}$, i.e.

$$F_n(x) - \frac{c_E(\alpha)}{\sqrt{n}}\left(F_n(x)[1-F_n(x)]\right)^{1/2} \leq F(x) \leq F_n(x) + \frac{c_E(\alpha)}{\sqrt{n}}\left(F_n(x)[1-F_n(x)]\right)^{1/2}.$$

1.A2.7. Distribution of the Anderson-Darling and the Eicker statistics for continuous distribution functions

Let $F(x) \in \widetilde{\mathbf{F}}$. Then, the random variable $F(X)$ takes all the values between 0 and 1 and follows a uniform distribution on $[0,1] : F(X) \sim U_{[0,1]}$. Hence, $F(X_{(i)}) = U_{(i)}$ where $U_{(i)}$ is the i^{th} ordered realization of an uniform $U_{[0,1]}$ distribution

Hence the expression of AD in Theorem 2.5. is equivalent to the following:

$$AD^{cont} = \max\ \left\{\max\left\{\sqrt{n}\frac{\frac{i}{n}-U_{(i)}}{\sqrt{U_{(i)}[1-U_{(i)}]}} : 1 \leq i \leq n \text{ st } 0 < U_{(i)} < 1\right\},\right.$$
$$\left.\max\left\{\sqrt{n}\frac{U_{(i)}-\frac{i-1}{n}}{\sqrt{U_{(i)}[1-U_{(i)}]}} : 1 \leq i \leq n \text{ st } 0 < U_{(i)} < 1\right\}, 0\right\}$$

and those of the statistic E is equivalent to:

$$E^{cont} = \max\left\{ \max_{1\le i\le 1} \sqrt{n}\frac{\frac{i}{n} - U_{(i)}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta}}, \max_{1\le i\le n} \sqrt{n}\frac{U_{(i)} - \frac{i-1}{n}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta}}, 0 \right\}$$

These statistics and their corresponding distributions do not depend on $F(x)$. Neither does their critical points and the CBs using these critical points.

1.A2.8. Explicit expression of the ζ-Regularized Anderson-Darling and Eicker statistics

We develop the expressions of the ζ-regularized statistics.

For the ζ-regularized *Anderson-Darling statistic:*

$$AD_\zeta = \sup_{-\infty<x<+\infty} \sqrt{n}\mid \frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]+\zeta}} \mid$$

$$\begin{aligned} &= \max\ \{ \sup_{-\infty<x<+\infty} \sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]+\zeta}}, \\ &\quad \sup_{-\infty<x<+\infty} \sqrt{n}\frac{F(x)-F_n(x)}{\sqrt{F(x)[1-F(x)]+\zeta}}\} \end{aligned}$$

$$\begin{aligned} &= \max\{ \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \sqrt{n}\frac{\frac{i}{n}-F(x)}{\sqrt{F(x)[1-F(x)]+\zeta}}, \\ &\quad \max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \sqrt{n}\frac{F(x)-\frac{i}{n}}{\sqrt{F(x)[1-F(x)]+\zeta}}\} \end{aligned}$$

Define $l_i(p) = \frac{\frac{i}{n}-p}{[p(1-p)+\zeta]^{1/2}}$ where $0 \le p \le 1$.

$$\begin{aligned} l_i'(p) &= \frac{-[p(1-p)+\zeta]^{1/2} - \frac{1}{2}(\frac{i}{n}-p)(1-2p)[p(1-p)+\zeta]^{-1/2}}{[p(1-p)+\zeta]} \\ &= [p(1-p)+\zeta]^{-3/2}\left[-\zeta - \frac{i}{2n} - p(\frac{1}{2}-\frac{i}{n})\right] = c_0 h_i(p) \end{aligned}$$

where $c_0 = [p(1-p)+\zeta]^{-3/2} \geq 0 \ \forall p$ and $h_i(p) = -\zeta - \frac{i}{2n} - p(\frac{1}{2} - \frac{i}{n})$

We know that $-\zeta - \frac{1}{2}\frac{i}{n} \leq 0 \ \forall i \geq 0$ and $\begin{cases} -p(\frac{1}{2} - \frac{i}{n}) \leq 0 \text{ for } 0 \leq i \leq n/2 \\ -p(\frac{1}{2} - \frac{i}{n}) \geq 0 \text{ for } n/2 \leq i \leq n \end{cases}$.

Thus, when $0 \leq i \leq n/2$, $h_i(p) \leq 0 \ \forall \ p$ while when i is such that $n/2 < i \leq n$, $h_i(p) \leq 0$ if $p < \frac{i}{2i-n}$.

Moreover $\frac{i}{2i-n} \geq 1$ when $n/2 \leq i \leq n$. Hence $p < \frac{i}{2i-n} \ \forall \ i$, $n/2 \leq i \leq n$; $h_i(p) \leq 0 \ \forall$ p, i and $l_i(p)$ is always non increasing. As a consequence,

1) $\max\limits_{0\leq i\leq n} \sup\limits_{X_{(i)}\leq x<X_{(i+1)}} \sqrt{n}\frac{\frac{i}{n}-F(x)}{\sqrt{F(x)[1-F(x)]+\zeta}} = \max\limits_{0\leq i\leq n} \sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}$

2) $\max\limits_{0\leq i\leq n} \sup\limits_{X_{(i)}\leq x<X_{(i+1)}} \sqrt{n}\frac{F(x)-\frac{i}{n}}{\sqrt{F(x)[1-F(x)]+\zeta}} = \max\limits_{0\leq i\leq n} \sqrt{n}\frac{F(X_{(i+1)})-\frac{i}{n}}{\sqrt{F(X_{(i+1)})[1-F(X_{(i+1)})]+\zeta}}$
$= \max\limits_{1\leq i\leq n+1} \sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}$

The statistic AD_ζ can then be computed using

$$\begin{aligned} AD_\zeta &= \max\{\max_{0\leq i\leq n}\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}, \\ &\qquad \max_{1\leq i\leq n+1}\sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}\} \\ &= \max\{\max_{1\leq i\leq n}\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}, \\ &\qquad \max_{1\leq i\leq n}\sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})]+\zeta}}, 0\}. \end{aligned}$$

For the ζ-regularized *Eicker statistic:*

$$\begin{aligned} E_\zeta &= \sup_{-\infty<x<+\infty} \sqrt{n}\,|\,\frac{F_n(x)-F(x)}{\sqrt{F_n(x)[1-F_n(x)]+\zeta}}\,| \\ &= \max\{\sup_{-\infty<x<+\infty} \sqrt{n}\frac{F_n(x)-F(x)}{\sqrt{F_n(x)[1-F_n(x)]+\zeta}}, \\ &\quad \sup_{-\infty<x<+\infty} \sqrt{n}\frac{F(x)-F_n(x)}{\sqrt{F_n(x)[1-F_n(x)]+\zeta}}\} \\ &= \max\{\max_{0\le i\le n} \sup_{X_{(i)}\le x<X_{(i+1)}} \sqrt{n}\frac{\frac{i}{n}-F(x)}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta}}, \\ &\quad \max_{1\le i\le n+1} \sup_{X_{(i-1)}\le x<X_{(i)}} \sqrt{n}\frac{F(x)-\frac{i-1}{n}}{\sqrt{\frac{i-1}{n}(1-\frac{i-1}{n})+\zeta}}\} \end{aligned}$$

Define $l_i(p)=\frac{\frac{i}{n}-p}{[\frac{i}{n}(1-\frac{i}{n})+\zeta]^{1/2}}$ where $0\le p\le 1$. $l_i(p)$ is always non increasing. Then,

$$\begin{aligned} E_\zeta &= \max\{\max_{0\le i\le n} \sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta}}, \\ &\quad \max_{1\le i\le n+1} \sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{\frac{i-1}{n}(1-\frac{i-1}{n})+\zeta}}\} \\ &= \max\{\max_{1\le i\le n} \sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}]+\zeta}}, \\ &\quad \max_{1\le i\le n} \sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{\sqrt{\frac{i-1}{n}(1-\frac{i-1}{n})+\zeta}}, 0\} \end{aligned}$$

1.A2.9. ζ-Regularized Anderson Darling-type CB for distribution functions

Let $c_{AD_\zeta}(\alpha)$ such that $\Pr[AD_\zeta \le c_{AD_\zeta}(\alpha)] \ge 1-\alpha$ i.e.,

$$P[\sup_{-\infty<x<+\infty} \sqrt{n}\,|\,\frac{F_n(x)-F(x)}{\sqrt{F(x)[1-F(x)]+\zeta}}\,|\le c_{AD_\zeta}(\alpha)] \ge 1-\alpha\ .$$

Then, with probability greater than or equal to $1-\alpha$

$$\frac{[F_n(x)-F(x)]^2}{F(x)[1-F(x)]+\zeta(x)} \leq \frac{c^2_{AD_\zeta}(\alpha)}{n} \qquad \forall\ x$$

i.e., $\forall\ x$,

$$\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)F^2(x) - \left(2F_n(x)+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)F(x) + F_n^2(x) - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n} \leq 0$$

This condition is satisfied if and only if $\forall\ x$

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

where $F_n^L(x) = \frac{2F_n(x)+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta}}{2\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)}$; $F_n^U(x) = \frac{2F_n(x)+\frac{c^2_{AD_\zeta}(\alpha)}{n}+\sqrt{\Delta}}{2\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)}$, and $\Delta = \left[2F_n(x)+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(F_n^2(x)-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$.

1.A2.10. ζ-Regularized Eicker-type CB for distribution functions

Let $c_{E_\zeta}(\alpha)$ such that $\Pr[E_\zeta \leq c_{E_\zeta}(\alpha)] \geq 1-\alpha$, i.e.,

$$\Pr[\sup_{-\infty<x<+\infty}\sqrt{n}\mid \frac{F_n(x)-F(x)}{\sqrt{F_n(x)[1-F_n(x)]+\zeta}} \mid \leq c_{E_\zeta}(\alpha)] \geq 1-\alpha\ .$$

Then, with probability greater than or equal to $1-\alpha$

$$-c_{E_\zeta}(\alpha) \leq \frac{\sqrt{n}[F_n(x)-F(x)]}{\sqrt{F_n(x)[1-F_n(x)]+\zeta}} \leq c_{E_\zeta}(\alpha) \qquad \forall\ x$$

and

$$F_n(x) - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(x)\left(1-F_n(x)\right)+\zeta\right]^{1/2} \leq F(x) \leq F_n(x) + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(x)\left(1-F_n(x)\right)+\zeta\right]^{1/2}.$$

1.A2.11. Distribution of the ζ-Regularized Anderson-Darling and the Eicker statistics for continuous distribution functions

When $F(x)$ is continuous, the random variable $F(X)$ has its values between 0 and 1 and follows a uniform distribution on $[0,1] : F(X) \sim U_{[0,1]}$. Hence, $F(X_{(i)}) = U_{(i)}$ where $U_{(i)}$ is the i^{th} ordered realization of an uniform $U_{[0,1]}$ distribution

Hence the expression of AD_ζ in Theorem 3.1. is equivalent to the following:

$$AD_\zeta^{cont} = \max\left\{ \max_{1\le i\le 1} \sqrt{n} \frac{\frac{i}{n} - U_{(i)}}{\sqrt{U_{(i)}[1-U_{(i)}] + \zeta}}, \max_{1\le i\le n} \sqrt{n} \frac{U_{(i)} - \frac{i-1}{n}}{\sqrt{U_{(i)}[1-U_{(i)}] + \zeta}}, 0 \right\}$$

Similarly, the expression of E_ζ can be rewritten:

$$E_\zeta^{cont} = \max\left\{ \max_{1\le i\le 1} \sqrt{n} \frac{\frac{i}{n} - U_{(i)}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}] + \zeta}}, \max_{1\le i\le n} \sqrt{n} \frac{U_{(i)} - \frac{i-1}{n}}{\sqrt{\frac{i}{n}[1-\frac{i}{n}] + \zeta}}, 0 \right\}.$$

Chapter 2

Improved nonparametric inference for the mean of a bounded random variable with application to poverty measures

Abstract

Despite the growing interest in poverty and inequality studies and the large standard errors found in many empirical studies, most of the work in this area remains descriptive and neglects statistical inference. Two types of inference procedures for poverty measures have been considered: asymptotic distributions and bootstrapping. These methods can be quite unreliable, even with fairly large samples, but no study has proposed provably valid finite-sample nonparametric inference methods for such problems.

In this paper, we develop such inference methods for the Foster, Greer and Thorbecke (FGT, 1984) poverty measures. We first observe that the poverty indicators can be interpreted as the expectation of a bounded random variable which is itself a functional of a distribution function. Using projection techniques, we derive finite-sample nonparametric confidence intervals for the mean from confidence bands for the distribution of the underlying variable. We investigate methods based on improved standardized Kolmogorov-Smirnov statistics and a likelihood-ratio criterion. We then apply these procedures to the FGT poverty measures.

Monte Carlo simulations show that asymptotic and bootstrap confidence intervals can fail to provide reliable inference, while the proposed methods are robust and yield shorter confidence intervals. As an illustration, we analyze the profile of poverty of Mexico in 1998. The results show that the widths of the asymptotic confidence intervals are often too small to be realistic while those of the bootstrap can be ten times larger than the widths delivered by exact methods. The study shows that the poverty profile of Mexican households depends greatly on the type of households' head: poverty levels among households with a male head or an educated head is much smaller than poverty levels among other households. Hence, policies aimed at reducing illiteracy and at securing the income of households with a female head could help reduce poverty in rural Mexico.

2.1 Introduction

In recent decades, there has been growing interest in poverty and inequality studies. However, most of the work in this area is descriptive and does not use rigorous statistical inference methods, despite the large standard errors found in many empirical analyses. Two types of inference procedures for poverty and inequality measures have been considered: asymptotic distributions and bootstrap methods; see Beran (1988), Kakwani (1993), Rongve (1997), Mills and Zandvakili (1997), Dardanoni and Forcina (1999), Biewen (2002), Davidson and Duclos (2000), Zheng (2001), and Davidson and Flachaire (2007). Most of these studies recommend the use of bootstrap inference rather than asymptotic approximations, because the latter are quite unreliable in finite samples. Bootstrap inference has a better performance, but is still unsatisfactory. Despite these problems, no study has proposed finite-sample nonparametric inference methods for poverty and inequality measures.

In this paper, we develop such inference methods for the popular Foster, Greer and Thorbecke (1984) (FGT, henceforth) poverty measures. On observing that poverty measures can be interpreted as the expectation of a bounded random variable—a mixture of a continuous bounded variable and a probability mass at the poverty line—we propose that exact nonparametric inference methods for the mean of a bounded random variable be applied to them.

At first sight, this problem appears to have no solution. According to Bahadur and Savage (1956), nonparametric inference cannot be performed for the mean of a random variable when observations are independent and identically distributed (i.i.d.) from an unknown distribution function with finite mean [see Dufour (2003) for more details]. However, in our case, the bounded nature of the random variable provides a sufficient restriction to allow nonparametric inference. Such nonparametric confidence intervals (CIs, henceforth) for the mean of a bounded random variable have been provided by Anderson (1969), Hora and Hora (1990), and Fishman (1991). Sutton and Young (1997) have compared these methods to asymptotic and bootstrap CIs using Beta distributions.

They showed that asymptotic and bootstrap CIs have very bad coverage probability in finite samples, while exact methods are strongly reliable but yield wider intervals than the former.

This paper provides two types of contributions. The first—a purely statistical contribution—consists of proposing finite-sample nonparametric CIs for the mean of a bounded random variable. We show that CIs for the mean can be derived from confidence bands (CBs, henceforth) for distribution functions using projection techniques. For general discussions of projection-based inference, see Dufour (1990), Abdelkhalek and Dufour (1998), and Dufour and Taamouti (2005). We then build improved CIs for the mean using CBs for distribution functions based on regularized weighted Kolmogorov-Smirnov statistics and likelihood-ratio type statistics proposed in Diouf and Dufour (2005). Our study focuses on the question of building CIs for the mean of a bounded random variable but our methodology is far from being as restricting as it appears. Solving the problem for the mean of Y allows to solve the problem for any moment of Y by replacing the original data by a function of these data. For example, if we are interested in building CIs for the moment of order 2, we can transform the original data using the square function and compute the empirical distribution function corresponding to those transformed data. The CIs we propose in this paper then provide valid CIs for the mean of the new data which are CIs for the second moment of the original data. All kinds of transformations can be studied. Continuous ones are handled using the same CIs as those presented in this paper while for noncontinuous ones, interesting monotonicity properties are provided to solve the problem.

The second contribution is econometric and consists of developing finite-sample nonparametric CIs for FGT measures. We re-express the poverty measures as the mean of a mixture of a continuous bounded random variable and a probability mass at the poverty line, and show that inference methods for the mean of a bounded variable apply to these. We build improved CIs with explicit expressions that are easy to compute. Monte Carlo simulations show that asymptotic and bootstrap CIs can fail to provide reliable inference, even with fairly large samples, e.g. when the distribution presents a high probability of

assuming the value zero—which is quite frequent in practice. By contrast, exact inference methods are robust to the underlying distribution and the sample size. The proposed CIs have coverage probability typically larger than the nominal level while remaining informative.

Finally, the methods are illustrated using household survey data to analyze the profile of poverty of Mexico in 1998. The results show that in addition to being unreliable, the widths of the asymptotic CIs are often too small to be realistic while the bootstrap can fail even in precision, delivering CIs whose widths can be ten times larger than those of the exact methods. The study shows that on average, rural households targeted by PROGRESA[1] do not have a very high level of poverty. However, the poverty profile depends greatly on the type of households' head. The level of poverty among households with a male head is much smaller than the level of poverty among households with a female head. Moreover, households with an educated head appear to be more prone to escape poverty than households with a non-educated head. These conclusions provide hints for designing policies to reduce poverty in rural Mexico. Policies aimed at reducing illiteracy of households members in these communities can be effective in reducing poverty. Education programs should target both children and adults, in particular households' heads to have short-term effects. Likewise, policies aimed at securing the income of households with a female head could help reduce poverty in rural Mexico. An example of such policies can be reforms aimed at securing land ownership for female or at improving labor productivity for households with a female head, the latter being less productive for physically intensive activities such as farming.

The paper is organized as follows. Section 2 summarizes the relevant literature on CIs for the mean of a bounded random variable. Section 3 describes a projection principle that allows CIs for the mean of a bounded random variable to be built from CBs for distribution functions. It also derives a general expression for such CIs. Section 4 proposes CIs for the mean of a continuous bounded random variable using the projection princi-

[1] See details about this program in section 9, page 109.

ple. CIs based on unweighted, weighted, and regularized weighted Kolmogorov-Smirnov statistics are considered. A likelihood-ratio type statistic is also used. Extension of these CIs to bounded noncontinuous variables are proposed in section 5. Section 6 proposes two approaches to estimate the regularization parameter of the proposed statistics. Section 7 applies the inference methods to the FGT poverty measures. Section 8 presents Monte Carlo simulations of the CIs for poverty measures using income from Singh-Maddala distribution. Section 9 illustrates the inference methods analyzing the profile of poverty of Mexico in 1998 with data from PROGRESA. We conclude in section 10.

2.2 Confidence intervals for the mean of a bounded random variable

Several inference methods for the mean of a bounded random variable have been proposed. Asymptotic procedures such as asymptotic distributions and bootstrap methods are popular and widely used. Some exact procedures have also been provided by Anderson (1969), Hora and Hora (1990), and Fishman (1991). Other studies have also proposed one-sided nonparametric inference methods for the mean of a censored variable [see Breth (1976)] and the mean of a nonnegative random variable [see Breth, Maritz and Williams (1978) and Kaplan (1987)]. In this section, we present the asymptotic and exact inference methods for the mean of a bounded random variable.

2.2.1 Asymptotic methods

Let X be a random variable with distribution function $F(x)$ and mean $E(X) = \mu$. Assume that n i.i.d observations $X_1, \ldots, X_n$ on X are available and let $F_n(x)$ be the corresponding empirical distribution function. Let W be the t-statistic:

$$W = \frac{\widehat{\mu} - \mu}{\left[\widehat{V}(\widehat{\mu})\right]^{1/2}}$$

where $\widehat{\mu}$ is an estimate of μ that is often $\overline{X}$ (the sample mean) and $\widehat{V}[\widehat{\mu}]$ is the estimated variance of the estimator.

Asymptotic inference

Assuming that W is asymptotically $N(0,1)$ as $n \to \infty$, an asymptotic CI for μ with level $1-\alpha$ is:

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

where $z_{(p)}$ is the p^{th} percentile of the standard normal distribution.

This CI is easy to compute, but can perform poorly in finite samples, because of the underlying asymptotic approximations [see Sutton and Young (1997) and Davidson and Flachaire (2007)].

Bootstrap inference

The simplest and most popular bootstrap CI for the mean is based on the percentile-t method:

$$\widehat{\mu} - D^{W}_{(1-\frac{\alpha}{2})} \left[\widehat{V}\left(\widehat{\mu}\right)\right]^{1/2} \leq \mu \leq \widehat{\mu} - D^{W}_{(\frac{\alpha}{2})} \left[\widehat{V}\left(\widehat{\mu}\right)\right]^{1/2}$$

where $D^{W}_{(p)}$ is the p^{th} percentile of the bootstrap distribution of W; see DiCiccio and Efron (1996) and Horowitz (2001).

This method performs better in finite samples than the asymptotic CI [Sutton and Young (1997) and Davidson and Flachaire (2007)]. However, its performance is still unsatisfactory and deteriorates when the distribution of the variable presents patterns such as heavy tails, multiple outliers or probability masses. Improved bootstrap inference methods have been provided but they are difficult to use. The causes of the bootstrap failure must be known for an adequate correction method to be chosen. Moreover, the bootstrap inference method involves a resampling procedure that is computationally demanding.

2.2.2 Exact methods

Anderson (1969)

Let X be a random variable with an unknown continuous cumulative distribution function $F(x)$ with finite support $[a, b]$ ($a < b$, $F(a) = 0$ and $F(b) = 1$) and mean $\mu = E(X)$. Denote $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ the order statistics of a sample of n available i.i.d observations on X and $F_n(x)$ the empirical distribution function of the sample such that $\forall\ k = 0, ..., n$

$$F_n(x) = \frac{k}{n} \text{ for } X_{(k)} \leq x < X_{(k+1)} \tag{2.1}$$

where $X_{(0)} = a$, and $X_{(n+1)} = b$ may be infinite. Note that this definition of $F_n(x)$ holds for non continuous distributions with observations $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ that might be equal ($X_{(1)} = X_{(2)}$ for some i, j). We will use frequently the empirical distribution function throughout this paper.

Anderson (1969) proposes the following CI with level $1 - \alpha$ for μ :

$$\begin{aligned} &\frac{1}{n}\left[\sum_{j=1}^{n-s-1} X_{(j)} + (s+1)\ X_{(n-s)}\right] - \gamma\ \left[X_{(n-s)} - a\right] \leq \mu \\ &\leq \frac{1}{n}\left[(r+1)\ X_{(r+1)} + \sum_{j=r+2}^{n} X_{(j)}\right] + \beta\ \left[b - X_{(r+1)}\right] \end{aligned} \tag{2.2}$$

where $r = I[n\beta]$, $s = |n\gamma|$, and $I[k]$ is the integer part of k. β and γ are such that:

$$P[F_n(x) - \beta\ \leq F(x) \leq F_n(x) + \gamma,\ \forall x] \geq 1 - \alpha.$$

The Anderson CI is nonparametric and robust to sample size. However, it is restricted to continuous bounded random variables. Moreover, it is based on the Kolmogorov-Smirnov CB for distribution functions and thus, inherits some drawbacks from the latter[2].

[2] These properties will be studied in detail later in this paper.

Hora and Hora (1990)

Hora and Hora (1990) propose the interval:

$$\overline{X}_n - c_{KS}(\alpha) \leq \mu \leq \overline{X}_n + c_{KS}(\alpha)$$

where $c_{KS}(\alpha)$ is the $(1-\alpha)^{th}$ percentile of the Kolmogorov-Smirnov statistic, $\overline{X}_n$ is the sample mean of X, and X is a continuous random variable bounded on $[0,1]$.

In practice, the Hora and Hora CI is easy to compute and nonparametric. However, owing to its dependence on the mean, this CI is sensitive to outliers and may perform badly when applied to atypical distributions with heavy tails or probability masses.

Fishman (1991)

Using Hoeffding's (1963) inequality, Fishman (1991) derives the following CI for the mean $\mu = E(X_i)$ of n i.i.d random variables $X_1, ..., X_n$ such that $Pr[0 \leq X_i \leq 1] = 1$:

$$\Pr[\mu_L(\overline{X}_n, n, \alpha) \leq \mu \leq \mu_U(\overline{X}_n, n, \alpha)] \geq 1 - \alpha$$

where

$$\begin{aligned}
\mu_L &= \begin{cases} \left\{t : 0 < t \leq \overline{X}_n \leq 1 \text{ and } e^{nf(\overline{X}_n - t, t)} = \alpha/2\right\} \text{ if } \overline{X}_n > 0, \\ 0 \text{ if } \overline{X}_n = 0, \end{cases} \\
\mu_U &= \begin{cases} \left\{t : 0 \leq \overline{X}_n \leq t < 1 \text{ and } e^{nf(t - \overline{X}_n, 1-t)} = \alpha/2\right\} \text{ if } \overline{X}_n < 1, \\ = 1 \text{ if } \overline{X}_n = 1. \end{cases}
\end{aligned}$$

According to Hoeffding (1963),

$$\Pr[\overline{X}_n - \mu \geq \varepsilon] \leq e^{nf(\varepsilon, \mu)}$$

for $0 < \varepsilon < 1 - \mu$, where $f(\varepsilon, \mu) = (\varepsilon + \mu)\ln[\mu(\varepsilon + \mu)^{-1}] + (1 - \varepsilon - \mu)\ln[(1-\mu)(1-\varepsilon-\mu)^{-1}]$.

The Fishman (1990) CI applies to bounded random variables with support $[0,1]$, but

can be generalized to any domain $[a, b]$. It is more general than the Anderson (1969) and Hora and Hora (1990) CIs, which do not apply to discontinuous bounded variables. However, the bounds of the Fishman interval are not defined explicitly. They are computed as the zero of a function. Consequently, the accuracy of this inference method relies largely on the accuracy of the iterative procedure used to derive μ_L and μ_U. Furthermore, this CI depends on the sample mean $\overline{X}_n$, which is very sensitive to outliers. These properties undermine the performance of the Fishman CI.

The relative performance of the asymptotic and exact procedures described in this section is examined by Sutton and Young (1997). They investigate the accuracy (confidence level) and the precision (width) of the CIs using Beta distributions and Monte Carlo simulations. Their results show that asymptotic and standard bootstrap procedures are not reliable in small samples, yielding CIs with coverage probability lower than the nominal confidence level $1 - \alpha$. Moreover, the precision of both CIs is reduced when the distribution presents a high probability of assuming the value zero. Conversely, exact methods yield coverage probabilities greater than $1 - \alpha$, but at the cost of wider CIs than those from asymptotic methods. The Anderson CI is shown to achieve the best width among exact CIs.

2.3 Projection methods for building confidence intervals for the mean of a bounded random variable

In this section we propose a projection approach for building CIs for the mean of a bounded random variable from CBs for distribution functions. We define some notation for the remainder of the paper. Denote:

$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$;

$\Gamma(.)$: a functional $\Gamma[F] : \mathcal{L} \to \overline{\mathbb{R}}$ defined on a space $\mathcal{L}$ of functions;

$\mathcal{F}$: a space of distribution functions;

$\widetilde{\mathcal{F}}$: a space of continuous distribution functions;

$\mathcal{F}_{[a,b]}$: a space of distribution functions with support $[a, b]$ (for finite numbers $a < b$);

$\widetilde{\mathcal{F}}_{[a,b]}$: a space of continuous distribution functions with support $[a, b]$ (for finite numbers $a < b$);

Let X be a random variable with distribution function $F(x) \in \mathcal{F}$. Denote $X_{(0)} \leq X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)}$ the order statistics of a sample of n available i.i.d observations on X, where $[X_{(0)}, X_{(n+1)}]$ is the support of $F(x)$. Denote $F_n(x)$ the empirical distribution function of the sample such that $\forall\ k = 0, ..., n$

$$F_n(x) = \frac{k}{n} \text{ for } X_{(k)} \leq x < X_{(k+1)}.$$

DEFINITION 3.1 *A functional $\Gamma[F] : \mathcal{L} \to \overline{\mathbb{R}}$ is monotonic nondecreasing if and only if*

$$F_1(x) \leq F_2(x),\ \forall x \Rightarrow \Gamma[F_1] \leq \Gamma[F_2],\ \forall F_1, F_2 \in \mathcal{L}.$$

PROPOSITION 3.2 **[Projection principle]** *Let $\Gamma[G]$ be a monotonic nondecreasing functional on $\mathcal{L}$, let $\mathcal{F}$ be a space of distribution functions included in $\mathcal{L}$, and let X be a random variable with distribution function $F(x) \in \mathcal{F}$. If*

$$C_F(\alpha) = \left\{F_0 \in \mathcal{L} : G_n^L(x) \leq F_0(x) \leq G_n^U(x),\ \forall x\right\}$$

is a confidence band with level $1 - \alpha$ for $F(x)$ such that $G_n^L \in \mathcal{L}$ and $G_n^U \in \mathcal{L}$, then

$$C_{\Gamma[F]}(\alpha) = \left\{\Gamma_0 \in \overline{\mathbb{R}} : \Gamma[G_n^L] \leq \Gamma_0 \leq \Gamma[G_n^U]\right\}$$

is a confidence interval with level $1 - \alpha$ for $\Gamma[F]$.

The proof of this proposition and all other proofs for this paper are provided in Appendix 2.

Two important conclusions arise from Proposition 3.2. First, this proposition allows the derivation of CIs for any monotonic functional of a distribution function from CBs for this distribution using projection techniques. In particular, this method can be applied to any (centered and non-centered) moment of X, specifically, to the mean of a bounded random variable $\Gamma[F] = \int_a^b x \; dF(x)$. Second, Proposition 3.2 states that any CB for a distribution function can be used, including nonparametric CBs. Hence, all available inference methods for distribution functions can yield CIs for the mean of a bounded variable. To provide improved nonparametric finite-sample inference methods for the mean, we can investigate such procedures for distribution functions.[3]

It is important to note that if $F(x)$ has a non bounded support, the CI for $\Gamma[F]$, $C_{\Gamma[F]}(\alpha)$, can be unbounded. This implies that $\Gamma[F_n^L]$ or $\Gamma[F_n^U]$ can be infinite. In the case of the mean of a random variable bounded on $[a, b]$ for some finite a and b, $C_{\Gamma[F]}(\alpha)$ is bounded. The remainder of this paper studes this particular case. A generalization of our inference methods to non bounded variables is provided in a next paper (Diouf and Dufour, 2006).

The following corollary applies Proposition 3.2 to the mean of a bounded random variable.

PROPOSITION 3.3. *Let $\mathcal{L}$ be a space of functions such that the Stieltjes integral $\Gamma[G] = \int_a^b y \; dG(y)$ is finite, let $\mathcal{F}_{[a,b]}$ be a space of distribution functions with support $[a, b]$ for finite numbers $a < b$ included in $\mathcal{L}$, and let X be a random variable with distribution function $F(x) \in \mathcal{F}_{[a,b]}$. If*

$$C_F(\alpha) = \left\{ F_0 \in \mathcal{L} : G_n^L(x) \leq F_0(x) \leq G_n^U(x), \; \forall x \right\}$$

is a confidence band with level $1 - \alpha$ for $F(x)$ such that $G_n^L \in \mathcal{L}$ and $G_n^U \in \mathcal{L}$, then

[3] Other versions of this principle can be found for specific functionals $\Gamma[F]$ in Dufour (1990, 1997), Abdelkhalek and Dufour (1998), Dufour and Neifar (2004), and Dufour and Taamouti (2005, 2007).

$$C_\mu(\alpha) = \left\{ \mu_0 \in \mathbb{R} : b - aG_n^U(a) - \int_a^b G_n^U(x)dx \leq \mu_0 \leq b - aG_n^L(a) - \int_a^b G_n^L(x)dx \right\}$$

is a confidence interval with level $1-\alpha$ *for* μ. *Moreover*

$$\widetilde{C}_\mu(\alpha) = \left\{ \mu_0 \in \mathbb{R} : b - a\widetilde{F}_n^U(a) - \int_a^b \widetilde{F}_n^U(x)dx \leq \mu_0 \leq b - a\widetilde{F}_n^L(a) - \int_a^b \widetilde{F}_n^L(x)dx \right\}$$

where $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, *and* $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$ *defines a confidence interval with level* $1-\alpha$ *for* μ *which is tighter than* $C_\mu(\alpha)$.

In the case where $F(x)$ is continuous ($F(x) \in \widetilde{\mathcal{F}}_{[a,b]}$), $\widetilde{F}_n^L(a) = \widetilde{F}_n^U(a) = F(a) = 0$ and the CIs for μ simplifies to

$$C_\mu(\alpha) = \left\{ \mu_0 \in \mathbb{R} : b - \int_a^b G_n^U(x)dx \leq \mu_0 \leq b - \int_a^b G_n^L(x)dx \right\}$$

and

$$\widetilde{C}_\mu(\alpha) = \left\{ \mu_0 \in \mathbb{R} : b - \int_a^b \widetilde{F}_n^U(x)dx \leq \mu_0 \leq b - \int_a^b \widetilde{F}_n^L(x)dx \right\}.$$

Proposition 3.3 justifies a general approach—which will be used throughout this paper—for building CIs for the mean of a bounded random variable using CBs for distribution functions. Two CIs can be derived from each CB: one using the whole bands and one using their restricting parts. The latter accounts for the property of the distribution functions which, by definition, always have values between 0 and 1 and is thus, thinner ($\widetilde{C}_\mu(\alpha) \subseteq C_\mu(\alpha)$). Note that $\widetilde{F}_n^L(x)$ and $\widetilde{F}_n^U(x)$ have values between 0 and 1 but are not necessarily distribution functions. In fact, they might never attain these values. For better results, we will use the thinner CI: $\widetilde{C}_\mu(\alpha)$ for the remaining of the paper. However, the CIs for the mean will retain the properties (in particular, the drawbacks) of the underlying CBs for distribution functions, owing to the use of projection techniques. This feature will be used to improve them. Finally, Proposition 3.3 can be used to derive an explicit expression for CIs for the mean of a bounded random variable. Proposition

3.4 does so.

PROPOSITION 3.4 **[General expression for CIs for the mean of a bounded random variable]** *Let $\mathcal{L}$ be a space of functions such that the Stieltjes integral $\Gamma[G] = \int_a^b y\, dG(y)$ is finite, let $\mathcal{F}_{[a,b]}$ be a space of distribution functions with support $[a,b]$ for finite numbers $a < b$ included in $\mathcal{L}$, and let X be a random variable with distribution function $F(x) \in \mathcal{F}_{[a,b]}$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics of a sample of n i.i.d. observations on X. If*

$$C_F(\alpha) = \left\{F_0 \in \mathcal{L} : G_n^L(x) \leq F_0(x) \leq G_n^U(x),\ \forall x\right\}$$

is a confidence band for $F(x)$ with level $1-\alpha$ where $G_n^L \in \mathcal{L}$, $G_n^U \in \mathcal{L}$, and $G_n^L(x)$ and $G_n^U(x)$ are step functions with jumps only at $X_{(1)}, \ldots, X_{(n)}$ then

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{2.3}$$

is a confidence interval for μ with level $1-\alpha$ where

$$\begin{aligned}
\mu_L &= [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)},\\
\mu_U &= \left[1 - \widetilde{F}_n^L(X_{(n)})\right]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right]X_{(k)},\\
X_{(0)} &= a,\ X_{(n+1)} = b,\ \widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\},\ and\ \widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\},\ \forall x.
\end{aligned}$$

Proposition 3.4 states that if a CB for $F(x)$ exists and if the lower and the upper bounds this CB are constant between $X_{(k)}$ and $X_{(k+1)}$ for all k, then one can use the Rieman integral to derive a general expression for CIs for the mean. This theorem is an application of the projection principle to stepwise CBs. However, it is not as restrictive as it seems. In fact, we know that all functions can be bounded by its closest lower and upper stepwise correspondent functions, using a given set of observations. Then, if CBs are not stepwise, we can use the estimation sample to derive a lower stepwise bound

for $F_n^L(x)$ and an upper one for $F_n^U(x)$ These bounds will define a CB for $F(x)$ of level greater than or equal to $1-\alpha$ that can be projected to the space of mean to yield a CI for the mean of level greater than or equal to $1-\alpha$.

Let k_l and $k_u \in \{1,\ldots,n\}$ such that

$$\widetilde{F}_n^L(x) = \begin{cases} G_n^L(x), & \forall x \geq X_{(k_l)} \\ 0, & \forall x < X_{(k_l)} \end{cases} \qquad \textit{and} \qquad \widetilde{F}_n^U(x) = \begin{cases} G_n^U(x), & \forall x \leq X_{(k_u)} \\ 1, & \forall x > X_{(k_u)}. \end{cases}$$

k_l and k_u represent the thresholds from which $F_n^L(X_{(k)}) \geq 0$ and $F_n^U(X_{(k)}) \leq 1$, and start to be binding.

If these numbers exist, then the bounds of $\widetilde{C}_\mu(\alpha)$ can be simplified to:

$$\mu_L = [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{ku+1}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)}$$

and

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=k_l}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right]X_{(k)}. \tag{2.4}$$

Moreover, if $k_u < n$ and $k_l < n$ then

$$\mu_L = [1-G_n^U(X_{(n)})]X_{(n+1)} + \left[1-G_n^U(X_{(k_u)})\right]X_{(k_u)} + \sum_{k=1}^{k_u}\left[G_n^U(X_{(k)}) - G_n^U(X_{(k-1)})\right]X_{(k)}$$

and

$$\mu_U = G_n^L(X_{(k_l)})X_{(k_l)} + [1-G_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=k_l+1}^{n}\left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right]X_{(k)}.$$

We will use the approach developed in this section repeatedly throughout the paper.

To derive CIs for the mean, we will use the improved finite-sample nonparametric CBs we proposed in Diouf and Dufour (2005). Those are obtained by inverting goodness-of-fit tests based on weighted Kolmogorov-Smirnov statistics and regularized versions of these statistics that provide better test power and narrower CBs for distributions. We will also use a likelihood-ratio based CB proposed by Owen (1995) using the Berk and Jones (1979) statistic.

2.4 Nonparametric confidence intervals for the mean of a bounded random variable

In this section, we will propose CIs based on empirical distribution functions (EDF, henceforth). Our study focuses on the question of building CIs for the mean of a bounded random variable using Proposition 3.3 and 3.4. However, our methodology is far from being as restrictive as it appears. Solving the problem for the mean of Y allows one to solve the problem for any moment of Y by replacing the original data by a function of these data—$exp(Y)$, Y^{β}, etc. For example, if we are interested in building CIs for the moment of order 2, we can transform the original data using the square function and compute the empirical distribution function corresponding to those transformed data. The CIs we propose in this section then provide valide CIs for the mean of the new data which are CIs for the second moment of the original data. All kinds of transformations can be studied. Continuous ones will be handled using the same CIs as those presented below while for noncontinuous ones, the next section provides interesting monotonicity properties that allow to solve the problem.

2.4.1 Three principles for building confidence intervals

CIs for the mean of a bounded random variable we proposed earlier use CBs for distribution functions based on EDFs. Those CBs can be built inverting goodness-of-fit tests for which the statistics of test involve EDFs. Several examples of those can be found in the

literature, among which the most popular is the Kolmogorov-Smirnov (KS, henceforth) statistic.

Let's consider the test of $H_0 : F(x) = p$ versus $H_1 : F(x) \neq p$. A common statistic for this test is $D_1 [F_n(x), F(x)] = \sqrt{n} \mid F_n(x) - F(x) \mid$. Taking the supremum of $D_1 [F_n(x), F(x)]$ over all x yields the following KS statistic:

$$KS_F = \sup_{-\infty \leq x \leq +\infty} \sqrt{n} \mid F_n(x) - F(x) \mid .$$

This statistic can be used to test hypotheses of type:

$$H_0(F) : X_1, \ldots, \ X_n \text{ are i.i.d. with distribution function } P[X_i \leq x] = F(x). \qquad (2.5)$$

versus the negative of $H_0(F)$. However, $D_1 [F_n(x), F(x)]$ is not standardized and hence, its distribution is not asymptotically pivotal.

Other inference methods have used statistics where $D_1 [F_n(x), F(x)]$ is improved along three common principles in econometrics: the Lagrange multiplier, Wald, and likelihood-ratio principles. The first one replaces $D_1 [F_n(x), F(x)]$ by a score-type statistic where $D_1 [F_n(x), F(x)]$ is divided its standard deviation estimated under the null hypothesis. This statistic had been proposed by Anderson and Darling (1952):

$$AD = \sup_{-\infty < x < +\infty} V_n(x)$$

where

$$V_n(x) = \begin{cases} 0 & \text{if } F(x) \in \{0, 1\}, \\ \sqrt{n} \left| \frac{F_n(x) - F(x)}{F^{1/2}(x)[1 - F(x)]^{1/2}} \right| & \text{otherwise.} \end{cases}$$

The second one standardizes $D_1 [F_n(x), F(x)]$ using an estimation of its standard deviation under H_1. The corresponding Wald-type statistic had been proposed by Eicker (1979):

$$E = \sup_{-\infty < x < +\infty} \widehat{V}_n(x)$$

where

$$\widehat{V}_n(x) = \begin{cases} 0 & \text{if } F_n(x) \in \{0,1\}, \\ \sqrt{n}\left|\dfrac{F_n(x)-F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2}}\right| & \text{otherwise.} \end{cases}$$

The last improvement replaces $D_1(x)$ by a likelihood ratio-type statistic. Such statistic had been proposed by Berk and Jones(1979):

$$BJ = \sup_{-\infty \le x \le +\infty} K[F_n(x), F(x)]$$

where

$$K(\widehat{p}, p) = \widehat{p}\log(\frac{\widehat{p}}{p}) + (1-\widehat{p})\log(\frac{1-\widehat{p}}{1-p}).$$

These statistics will be used to derive CIs for the mean as well as improvements of the score-type and Wald-type statistics using regularized versions of these.

2.4.2 Confidence intervals based on the Kolmogorov-Smirnov statistic

In the rest of the paper, we define X as a random variable with distribution $F(x) \in \widetilde{\mathcal{F}}_{[a,b]}$ and a finite mean μ. We also let $X_{(1)} \le \cdots \le X_{(n)}$ be the order statistics of a sample of n i.i.d observations available on X, $X_{(0)} = a$, $X_{(n+1)} = b$, and $F_n(x)$ the corresponding empirical distribution function. The Kolmogorov-Smirnov statistic is:

$$\begin{aligned} KS_F &= \sup_{-\infty \le x \le +\infty} \sqrt{n} \mid F_n(x) - F(x) \mid \\ &= \max\left\{ \max_{1 \le k \le n}\left\{\frac{k}{n} - F(X_{(k)})\right\}, \max_{1 \le k \le n}\left\{F(X_{(k)}) - \frac{k-1}{n}\right\}, 0\right\}. \end{aligned}$$

A symmetric CB for $F(x)$ with level $1-\alpha$ can be obtained by inverting the KS goodness-of-fit test:

$$C_F^{KS}(\alpha) = \left\{F_0 \in \mathcal{F} : F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \le F_0(x) \le F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}},\ \forall x\right\}$$

where $c_{KS}(\alpha)$ satisfies $Pr(KS_F \leq c_{KS}(\alpha)) \geq 1-\alpha$.

Applying Proposition 3.3, we can derive from $C_F^{KS}(\alpha)$ the two following projection-based CIs for μ :

$$C_\mu^{KS}(\alpha) = \left\{ \mu_0 \in \mathbb{R} : \overline{X}_n - \left[X_{(n+1)} - X_{(0)}\right] \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + \left[X_{(n+1)} - X_{(0)}\right] \frac{c_{KS}(\alpha)}{\sqrt{n}} \right\} \tag{2.6}$$

where $\overline{X}_n$ is the sample mean of X, and

$$\widetilde{C}_\mu^{KS}(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{2.7}$$

where

$$\mu_L = \frac{1}{n}\left[\sum_{j=1}^{n-s-1} X_{(j)} + (s+1)\ X_{(n-s)}\right] - \frac{c_{KS}(\alpha)}{\sqrt{n}}\ \left[X_{(n-s)} - a\right]$$

$$\mu_U = \frac{1}{n}\left[(r+1)\ X_{(r+1)} + \sum_{j=r+2}^{n} X_{(j)}\right] + \frac{c_{KS}(\alpha)}{\sqrt{n}}\ \left[b - X_{(r+1)}\right]$$

$r = I[n\frac{c_{KS}(\alpha)}{\sqrt{n}}]$, $s = \left|n\frac{c_{KS}(\alpha)}{\sqrt{n}}\right|$, and $I[k]$ is the integer part of k.

The $C_\mu^{KS}(\alpha)$ CI for the mean of X—equation *(2.6)*—is a generalization of the Hora-Hora CI to the mean of random variables bounded on $[a,b]$. Setting $b=1$ and $a=0$ provides the original Hora-Hora CI:

$$C_\mu^{HH}(\alpha) = \left\{ \mu_0 \in \mathbb{R} : \overline{X}_n - \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + \frac{c_{KS}(\alpha)}{\sqrt{n}} \right\}$$

We provide in Appendix 2 a proof of the generalized Hora-Hora CI which, setting $b=1$ and $a=0$ also gives a proof the original CI which have not been clearly given by the literature. Likewise, the second CI, $\widetilde{C}_\mu^{KS}(\alpha)$, corresponds to the Anderson CI for the mean of a continuous bounded random variable defined in equation (2.2). We prove in Appendix 2 that the Anderson CI is a projection of the KS confidence band where all constraints about distribution functions are exploited. Hence, the Hora-Hora and the Anderson CIs are both some special cases of Proposition 3.3 using the projection of the

KS CB for distributions onto the space of mean. However, the Hora-Hora CI is dominated by the Anderson CI. The latter excludes the parts of the Kolmogorov-Smirnov CB that are not effective i.e., the parts of the bands below 0 or above 1 and is then shorter than the Hora-Hora CI.

Even if the KS statistic can be used to build a well-behaved CI for the mean–the Anderson CI— it is not based on the three common principles in econometrics: the Wald, likelihood-ratio, and Lagrange multiplier principles. By the properties of the projection method, the Anderson CI inherits some characteristics from the KS confidence band. Thus, using the properties of the latter, we can determine the limits of the Anderson method and improve it along these lines. In particular, the KS confidence band for distribution functions is often criticized for its uniform nature. The width of this CB is constant for all observations and thus its bounds do not converge to 0 and 1 in the lower and upper tails of the distribution, as do the distribution functions they bracket. This adversely affects its performance in the tails and can easily lead to a large projected CI for the mean. Weighted KS statistics based on Wald, likelihood-ratio, and Lagrange multiplier improvements have been proposed to allow more discrimination between distributions in the tails. We use these statistics to build CIs for the mean and show by Monte Carlo simulations that those CIs have better properties than the Anderson CI.

2.4.3 Confidence intervals based on weighted Kolmogorov-Smirnov statistics

We will now propose improved CIs for the mean based on nonuniform Kolmogorov-Smirnov CBs for distributions. Two weighted Kolmogorov-Smirnov statistics have been proposed by Anderson and Darling (1952),

$$AD = \sup_{-\infty<x<+\infty} V_n(x)$$

and Eicker (1979),

$$E = \sup_{-\infty<x<+\infty} \widehat{V}_n(x)$$

where

$$V_n(x) = \begin{cases} 0 & \text{if } F(x) \in \{0,1\}, \\ \sqrt{n}\left|\frac{F_n(x)-F(x)}{F^{1/2}(x)[1-F(x)]^{1/2}}\right| & \text{otherwise,} \end{cases}$$

and

$$\widehat{V}_n(x) = \begin{cases} 0 & \text{if } F_n(x) \in \{0,1\}, \\ \sqrt{n}\left|\dfrac{F_n(x)-F(x)}{F_n^{1/2}(x)[1-F_n(x)]^{1/2}}\right| & \text{otherwise.} \end{cases}$$

In practice, these statistics can be computed as follows [see Diouf and Dufour (2005a)]:

$$AD = \max\left\{0, \max\left\{\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{F^{1/2}(X_{(i)})[1-F(X_{(i)})]^{1/2}} : 0 < F(X_{(i)}) < 1,\ 1 \le i \le n\right\}\right.$$
$$\left.\max\left\{\sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{F^{1/2}(X_{(i)})[1-F(X_{(i)})]^{1/2}} : 0 < F(X_{(i)}) < 1,\ 1 \le i \le n\right\}\right\},$$

$$E = \max\left\{\max_{1\le i\le n-1}\sqrt{n}\frac{\frac{i}{n}-F(X_{(i)})}{(\frac{i}{n})^{1/2}(1-\frac{i}{n})^{1/2}},\ \max_{2\le i\le n}\sqrt{n}\frac{F(X_{(i)})-\frac{i-1}{n}}{(\frac{i-1}{n})^{1/2}(1-\frac{i-1}{n})^{1/2}},\ 0\right\}.$$

These two statistics can be used to derive finite-sample nonparametric CBs for distribution functions whose widths decrease with observations further from the center of the distribution. The Eicker and Anderson-Darling statistics are better at discriminating between distributions that essentially differ in their tails than the uniform (unweighted) KS statistic and, in consequence, they provide narrower CBs in the tails.

However, the Anderson-Darling (AD) and Eicker statistics have their own drawbacks. The power of the goodness-of-fit tests they yield is less than the power of the standard KS goodness-of-fit test when testing distributions with low dispersion that differ more in the center of the distribution than in the tails. Moreover, the weights in the denominator of those statistics become very close to 0 for observations in the tails, leading to erratic

behavior of the statistics. To solve this problem, we proposed regularized statistics where the variance of $F_n(x) - F(x)$ in the denominator of the statistics is corrected by adding a nonzero positive regularization term ζ :

$$AD_\zeta = \sup_{-\infty<x<+\infty} \sqrt{n} \left| \frac{F_n(x) - F(x)}{\sqrt{F(x)[1-F(x)] + \zeta}} \right|,$$

$$E_\zeta = \sup_{-\infty<x<+\infty} \sqrt{n} \left| \frac{F_n(x) - F(x)}{\sqrt{F_n(x)[1-F_n(x)] + \zeta}} \right|.$$

For a constant $\zeta \geq 0$, the regularized statistics can be computed using the following expressions:

$$AD_\zeta = \max \left\{ \max_{1\leq i\leq n} \sqrt{n} \frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{F(X_{(i)})[1-F(X_{(i)})] + \zeta}}, \max_{1\leq i\leq n} \sqrt{n} \frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{F(X_{(i)})[1-F(X_{(i)})] + \zeta}}, 0 \right\},$$

$$E_\zeta = \max \left\{ \max_{1\leq i\leq n} \sqrt{n} \frac{\frac{i}{n} - F(X_{(i)})}{\sqrt{\frac{i}{n}[1-\frac{i}{n}] + \zeta}}, \max_{1\leq i\leq n} \sqrt{n} \frac{F(X_{(i)}) - \frac{i-1}{n}}{\sqrt{\frac{i-1}{n}(1-\frac{i-1}{n}) + \zeta}}, 0 \right\}.$$

The regularization achieves the expected improvement. First, we showed by Monte Carlo simulation that using a constant positive ζ considerably improves the power of the regularized Anderson-Darling and Eicker goodness-of-fit tests. The test power is low for small ζ but rises quickly when ζ increases before becoming almost constant. Even if the value of ζ that maximizes the power of the test is not known, most of the improvement is achieved as soon as ζ is high enough. An example of how to choose the optimal value of ζ in practice is provided for inference methods on poverty measures in section 7. Second, computing the critical values of the statistics by simulation eliminates the bias to which these kinds of (regularized) statistics are subject. Finally, regularizing prevents the erratic behavior of the original Anderson-Darling and Eicker statistics in the tails of distributions and provides nonuniform CBs for distribution functions that decline in width as observations approach the tail of the distributions. Monte Carlo simulations show that the widths of these CBs are smaller than the widths of the Anderson-Darling

and the Eicker CBs. Hence, their projections are expected to provide better CIs for the mean than the original CBs. Theorems 5.1 to 5.4 propose finite-sample nonparametric CIs for the mean of a continuous bounded random variable using the standard and the regularized Anderson-Darling and Eicker CBs for distributions. Analytic expressions are derived, in which ζ is assumed constant.

PROPOSITION 4.1. **[Anderson Darling-type CI for the mean of a bounded random variable]** *Let X be a random variable with distribution function $F(x) \in \mathcal{F}_{[a,b]}$ and a finite mean μ. Assume that n ordered i.i.d. observations $X_{(1)} \leq \cdots \leq X_{(n)}$ on X are available. Suppose that the following confidence band of type Anderson-Darling with level $1-\alpha$ is valid for the space of distributions $\mathcal{F}_{[a,b]}$:*

$$C_F^{AD}(\alpha) = \left\{F_0 \in \mathcal{F} : G_n^L(x) \leq F_0 \leq G_n^U(x),\, \forall x\right\}$$

where

$$\begin{aligned} G_n^L(x) &= \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} - \sqrt{\Delta(x)}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}, \\ G_n^U(x) &= \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} + \sqrt{\Delta(x)}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}, \\ \Delta(x) &= \left[2F_n(x) + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4F_n^2(x)\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right], \end{aligned}$$

and $c_{AD}(\alpha)$ satisfies $Pr(AD \leq c_{AD}(\alpha)) \geq 1-\alpha$. Then the following confidence interval:

$$\begin{aligned} \widetilde{C}_\mu^{AD}(\alpha) = \Bigg\{ &\mu_0 \in \mathbb{R} : [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)} \leq \mu_0 \\ &\leq [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right]X_{(k)}\Bigg\} \end{aligned} \tag{2.8}$$

where $X_{(0)} = a$, $X_{(n+1)} = b$, and $\forall x$, $\widetilde{F}_n^L(x) = \max\left\{G_n^L(x), 0\right\}$ and $\widetilde{F}_n^U(x) = \min\left\{G_n^U(x), 1\right\}$ is a confidence interval for μ with level $1-\alpha$.

COROLLARY 4.1BIS **[Explicit expression for the Anderson Darling-type CI]** *Under the hypothesis of Proposition 4.1., the effective part of the Anderson-Darling confidence band for $F(x)$ is:*

$$\widetilde{F}_n^L(X_{(k)}) = G_n^L(X_{(k)}),\ \forall k = 1, \ldots, n+1$$

and

$$\widetilde{F}_n^U(X_{(k)}) = G_n^U(X_{(k)}),\ \forall k = 1, \ldots, n+1$$

and an equivalent expression of the Anderson-Darling confidence interval for the mean of a bounded random variable is:

$$\widetilde{C}_\mu^{AD}(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\}$$

where

$$\mu_L = \sum_{k=1}^{n} \left[\frac{\frac{2}{n} + \sqrt{\Delta(k)} - \sqrt{\Delta(k-1)}}{2(1 + \frac{c_{AD}^2(\alpha)}{n})} \right] X_{(k)},$$

$$\mu_U = \left(1 - \left(1 + \frac{c_{AD}^2(\alpha)}{n}\right)^{-1}\right) X_{(n+1)} + \sum_{k=1}^{n} \left[\frac{\frac{2}{n} - \left(\sqrt{\Delta(k)} - \sqrt{\Delta(k-1)}\right)}{2(1 + \frac{c_{AD}^2(\alpha)}{n})} \right] X_{(k)},$$

$$\Delta(k) = \left[2\frac{k}{n} + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4(\frac{k}{n})^2 \left[1 + \frac{c_{AD}^2(\alpha)}{n}\right],$$

and $c_{AD}(\alpha)$ satisfies $Pr[AD \leq c_{AD}(\alpha)] \geq 1 - \alpha$.

PROPOSITION 4.2 **[Eicker-type CI for the mean of a bounded random variable]** *Let X be a random variable with distribution function $F(x) \in \mathcal{F}_{[a,b]}$ and a finite mean μ. Assume that n ordered i.i.d. observations $X_{(1)} \leq \cdots \leq X_{(n)}$ on X are available. Suppose that the following confidence band of type Eicker with level $1-\alpha$ is valid for the space of distributions $\mathcal{F}_{[a,b]}$:*

$$C_F^E(\alpha) = \left\{F_0 \in \mathcal{F} : G_n^L(x) \leq F_0 \leq G_n^U(x)\right\}$$

where

$$G_n^L(x) = \begin{cases} F_n(x) - \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} & \text{for such that } F_n(x) \notin \{0,1\}, \\ 0 & \text{for } x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

$$G_n^U(x) = \begin{cases} F_n(x) + \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1-F_n(x)]^{1/2} & \text{for } x \text{ such that } F_n(x) \notin \{0,1\}, \\ 1 & \text{for } x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

and $c_E(\alpha)$ satisfies $Pr(E \leq c_E(\alpha)) \geq 1-\alpha$. Then the following confidence interval:

$$\widetilde{C}_\mu^E(\alpha) = \left\{ \mu_0 \in \mathbb{R} : [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)} \leq \mu_0 \right.$$
$$\left. \leq [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)} \right\} \tag{2.9}$$

where $X_{(0)} = a$, $X_{(n+1)} = b$,.and $\forall x$, $\widetilde{F}_n^L(x) = \max\left\{G_n^L(x), 0\right\}$ and $\widetilde{F}_n^U(x) = \min\left\{G_n^U(x), 1\right\}$ is a confidence interval for μ with level $1-\alpha$.

COROLLARY 4.2BIS **[explicit expression for the Eicker-type CI]** *Under the hypothesis of Proposition 4.2., the effective part of the Eicker confidence bounds for $F(x)$ is:*

$$\widetilde{F}_n^L(X_{(k)}) = \begin{cases} \frac{k}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} & \forall k = k_E^L, \ldots, \ n-1 \\ 0 \qquad \forall k = 0, \ldots, \ k_E^L - 1, \ n, \ n+1, \end{cases}$$

$$\widetilde{F}_n^U(X_{(k)}) = \begin{cases} \frac{k}{n} + \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \ \forall k = 1, \ldots, k_E^U \\ 1 \ \forall k = 0, \ k_E^U + 1, \ldots, \ n, \ n+1. \end{cases}$$

where $k_E^L = I\left[nc_E^2(\alpha)\left(n + c_E^2(\alpha)\right)^{-1}\right] + 1$, $k_E^U = I\left[\frac{n^2}{\left(n + c_E^2(\alpha)\right)}\right]$, and $I[\kappa]$ is the integer part of κ; and an equivalent expression of the Eicker confidence interval for the mean of

a bounded random variable is:

$$\widetilde{C}_{\mu}^{E}(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\}$$

where

$$\mu_L = \left[1 - \frac{k_E^U}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k_E^U}{n}]^{1/2}[1 - \frac{k_E^U}{n}]^{1/2}\right] X_{(k_E^U+1)} + \sum_{k=1}^{k_E^U}\left[\frac{1}{n} + \frac{c_E(\alpha)}{\sqrt{n}}\left(\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)} - \sqrt{\frac{k-1}{n}\left(1-\frac{k-1}{n}\right)}\right)\right] X_{(k)},$$

$$\mu_U = \left[\frac{k_E^L}{n} - \frac{c_E(\alpha)}{\sqrt{n}}\left(\frac{k_E^L}{n}\right)^{1/2}\left(1-\frac{k_E^L}{n}\right)^{1/2}\right] X_{(k_E^L)} + X_{(n+1)} + \sum_{k=k_E^L+1}^{n}\left[\frac{1}{n} - \frac{c_E(\alpha)}{\sqrt{n}}\left(\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)} - \sqrt{\frac{k-1}{n}\left(1-\frac{k-1}{n}\right)}\right)\right] X_{(k)},$$

and $c_E(\alpha)$ *satisfies* $Pr(E \leq c_E(\alpha)) \geq 1 - \alpha$.

PROPOSITION 4.3. [$\zeta-$**Regularized Anderson Darling-type CI for the mean of a bounded random variable**] *Let* X *be a random variable with distribution function* $F(x) \in \mathcal{F}_{[a,b]}$ *and a finite mean* μ. *Assume that* n *ordered i.i.d. observations* $X_{(1)} \leq \cdots \leq X_{(n)}$ *on* X *are available. Suppose that the following confidence band of type* ζ*-regularized Anderson-Darling with level* $1 - \alpha$ *is valid for the space of distributions* $\mathcal{F}_{[a,b]}$:

$$C_F^{AD_\zeta}(\alpha) = \left\{F_0 \in \mathcal{F} : G_n^L(x) \leq F_0 \leq G_n^U(x),\ \forall x\right\}$$

where

$$G_n^L(x) = \frac{2F_n(x) + \frac{c_{AD_\zeta}^2(\alpha)}{n} - \sqrt{\Delta(x)}}{2\left(1 + \frac{c_{AD_\zeta}^2(\alpha)}{n}\right)},$$

$$G_n^U(x) = \frac{2F_n(x) + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta(x)}}{2\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)},$$

$$\Delta(x) = \left[2F_n(x) + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(F_n^2(x) - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right),$$

and $c_{AD_\zeta}(\alpha)$ *satisfies* $\Pr[AD_\zeta \leq c_{AD_\zeta}(\alpha)] \geq 1-\alpha$. *Then, the following confidence interval:*

$$\widetilde{C}_\mu^{AD_\zeta}(\alpha) = \left\{\mu_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)} \leq \mu_0 \right. \tag{2.10}$$

$$\left. \leq [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right]X_{(k)}\right\}$$

where $X_{(0)} = a$, $X_{(n+1)} = b$, *and* $\forall x$, $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$ *and* $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$ *is a confidence interval for* μ *with level* $1-\alpha$.

COROLLARY 4.3BIS **[Explicit expression for the $\zeta-$Regularized Anderson Darling-type CI]** *Under the hypothesis of Proposition 4.3., the effective bounds of the* ζ*-regularized Anderson-Darling confidence band for* $F(x)$ *are:*

$$\widetilde{F}_n^L(x) = \begin{cases} G_n^L(x)\ \forall x \geq X_{(k^L_{AD_\zeta})} \\ 0\ \forall x < X_{(k^L_{AD_\zeta})}, \end{cases}$$

$$\widetilde{F}_n^U(x) = \begin{cases} G_n^U(x)\ \forall x \leq X_{(k^U_{AD_\zeta})} \\ 1\ \forall x > X_{(k^U_{AD_\zeta})} \end{cases}$$

where $k^L_{AD_\zeta} = I\left[n^{1/2}\zeta^{1/2}c_{AD_\zeta}(\alpha)\right] + 1$, $k^U_{AD_\zeta} = I\left[n - c_{AD_\zeta}(\alpha)(n\zeta)^{1/2}\right]$, $I[\kappa_0] \equiv$*the integer part of* κ_0*; and an equivalent expression of the* ζ*-regularized Anderson-Darling confidence interval for the mean of a bounded random variable is:*

$$\widetilde{C}_{\mu}^{AD_{\zeta}}(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{2.11}$$

where

$$\mu_L = \frac{1}{2(1+\frac{c^2_{AD_{\zeta}}(\alpha)}{n})} \left\{ \left(2 + \frac{c^2_{AD_{\zeta}}(\alpha)}{n} - 2\frac{k^U_{AD_{\zeta}}}{n} - \sqrt{\Delta(k^U_{AD_{\zeta}})} \right) X_{(k^U_{AD_{\zeta}}+1)} \right.$$
$$\left. + \sum_{k=1}^{k^U_{AD_{\zeta}}} \left(\frac{2}{n} + \sqrt{\Delta(k)} - \sqrt{\Delta(k-1)} \right) X_{(k)} \right\},$$

$$\mu_U = [1 - \widetilde{F}^L_n(X_{(n)})]X_{(n+1)} + \frac{1}{2(1+\frac{c^2_{AD_{\zeta}}(\alpha)}{n})} \left\{ \left(\frac{2k^L_{AD_{\zeta}}}{n} + \frac{c^2_{AD_{\zeta}}(\alpha)}{n} - \sqrt{\Delta(k^L_{AD_{\zeta}})} \right) X_{(k^L_{AD_{\zeta}})} \right.$$
$$\left. + \sum_{k=k^L_{AD_{\zeta}}+1}^{n} \left[\frac{2}{n} - \sqrt{\Delta(k)} + \sqrt{\Delta(k-1)} \right] X_{(k)}, \right.$$

$$\widetilde{F}^L_n(X_{(n)}) = \begin{cases} \left(2+\frac{c^2_{AD_{\zeta}}(\alpha)}{n} - \sqrt{\Delta(n)} \right) \left(2(1+\frac{c^2_{AD_{\zeta}}(\alpha)}{n}) \right)^{-1} & \text{if } k^L_{AD_{\zeta}} \leq n, \\ 0 \text{ otherwise,} \end{cases}$$

$$\Delta(k) = \left[2\frac{k}{n} + \frac{c^2_{AD_{\zeta}}(\alpha)}{n} \right]^2 - 4 \left[1 + \frac{c^2_{AD_{\zeta}}(\alpha)}{n} \right] \left(\frac{k^2}{n^2} - \frac{\zeta c^2_{AD_{\zeta}}(\alpha)}{n} \right),$$

and $c_{AD_{\zeta}}(\alpha)$ *satisfies* $\Pr[AD_{\zeta} \leq c_{AD_{\zeta}}(\alpha)] \geq 1 - \alpha$.

PROPOSITION 4.4. [$\zeta-$**Regularized Eicker-type CI for the mean of a bounded random variable**] *Let* X *be a random variable with distribution function* $F(x) \in \mathcal{F}_{[a,b]}$ *and a finite mean* μ. *Suppose that a random sample of* X *is available and denote* $X_{(1)} \leq ... \leq X_{(n)}$ *the ordered observations. Suppose that the following confidence band of type* ζ*-regularized Eicker with level* $1 - \alpha$ *is valid for the space of distributions* $\mathcal{F}_{[a,b]}$:

$$C^{E_{\zeta}}_F(\alpha) = \left\{ F_0 \in \mathcal{F} : G^L_n(x) \leq F_0(x) \leq G^U_n(x),\ \forall x \right\}$$

where

$$G_n^L(x) = F_n(x) - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}} \left[F_n(x)\left(1 - F_n(x)\right) + \zeta\right]^{1/2},$$

$$G_n^U(x) = F_n(x) + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}} \left[F_n(x)\left(1 - F_n(x)\right) + \zeta\right]^{1/2},$$

and $c_{E_\zeta}(\alpha)$ *satisfies* $\Pr[E_\zeta \leq c_{E_\zeta}(\alpha)] \geq 1 - \alpha$. *Then the following confidence interval:*

$$\widetilde{C}_\mu^{E_\zeta}(\alpha) = \left\{ \mu_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)} \leq \mu_0 \right. \tag{2.12}$$
$$\left. \leq [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)} \right\}$$

where $X_{(0)} = a$, $X_{(n+1)} = b$, *and* $\forall x$, $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$ *and* $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$ *is a confidence interval for* μ *with level* $1 - \alpha$.

COROLLARY 4.4BIS **[Explicit expression for the ζ-Regularized Eicker-type CI]** *Under the hypothesis of Proposition 4.4, the effective bounds of the* ζ*-regularized Eicker confidence band for* $F(x)$ *are:*

$$\widetilde{F}_n^L(x) = \begin{cases} G_n^L(x)\ \forall x \geq X_{(k_{E_\zeta}^L)} \\ 0\ \forall x < X_{(k_{E_\zeta}^L)}, \end{cases} \quad \textit{and} \quad \widetilde{F}_n^U(x) = \begin{cases} G_n^U(x)\ \forall x \leq X_{(k_{E_\zeta}^U)} \\ 1\ \forall x > X_{(k_{E_\zeta}^U)} \end{cases}$$

where $k_{E_\zeta}^L = I\left[k_2\right] + 1$, $k_{E_\zeta}^U = I\left[k_3\right]$, $k_2 = \left[nc_{E_\zeta}^2(\alpha) + \sqrt{\Delta^L}\right]\left[2\left(n + c_{E_\zeta}^2(\alpha)\right)\right]^{-1}$, $k_3 = \left[\left(2n + c_{E_\zeta}^2(\alpha)\right)n - \sqrt{\Delta^U}\right]\left[2\left(n + c_{E_\zeta}^2(\alpha)\right)\right]^{-1}$, $\Delta^L = n^2c_{(1-\alpha)}^4 + 4\left(n + c_{E_\zeta}^2(\alpha)\right)n^2c_{E_\zeta}^2(\alpha)\zeta$, $\Delta^U = \left(2n + c_{E_\zeta}^2(\alpha)\right)^2 n^2 - 4\left(n + c_{E_\zeta}^2(\alpha)\right)\left(n - c_{E_\zeta}^2(\alpha)\zeta\right)n^2$, *and* $I\left[\kappa\right]$ *is the integer part of* κ; *and an equivalent expression of the* ζ*-regularized Eicker confidence interval for the mean of a bounded random variable is:*

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{2.13}$$

where

$$\begin{aligned}\mu_L &= \left[1-k_{E_\zeta}^U n^{-1}-c_{E_\zeta}(\alpha)n^{-1/2}\left[k_{E_\zeta}^U n^{-1}\left(1-k_{E_\zeta}^U n^{-1}\right)+\zeta\right]^{1/2}\right]X_{(k_{E_\zeta}^U+1)} \\ &\quad +\sum_{k=1}^{k_{E_\zeta}^U}\left[\frac{1}{n}+c_{E_\zeta}(\alpha)n^{-1/2}\left(\sqrt{\frac{k}{n}[1-\frac{k}{n}]+\zeta}-\sqrt{\frac{k-1}{n}[1-\frac{k-1}{n}]+\zeta}\right)\right]X_{(k)},\end{aligned}$$

$$\begin{aligned}\mu_U &= \left[k_{E_\zeta}^L n^{-1}-c_{E_\zeta}(\alpha)n^{-1/2}\left(k_{E_\zeta}^L n^{-1}\left(1-k_{E_\zeta}^L n^{-1}\right)+\zeta\right)^{1/2}\right]X_{(k_{E_\zeta}^L)} \\ &\quad +\sum_{k=k_{E_\zeta}^L+1}^{n}\left[\frac{1}{n}-c_{E_\zeta}(\alpha)n^{-1/2}\left(\sqrt{\frac{k}{n}[1-\frac{k}{n}]+\zeta}-\sqrt{\frac{k-1}{n}[1-\frac{k-1}{n}]+\zeta}\right)\right]X_{(k)},\end{aligned}$$

$$\widetilde{F}_n^L(X_{(n)})=\begin{cases}1-c_{E_\zeta}(\alpha)\zeta^{1/2}n^{-1/2} \text{ if } k_{E_\zeta}^L\le n,\\ 0 \text{ if } k_{E_\zeta}^L>n,\end{cases}$$

and $c_{E_\zeta}(\alpha)$ *satisfies* $Pr[E_\zeta \le c_{E_\zeta}(\alpha)] \ge 1-\alpha$.

2.4.4 Confidence interval based on likelihood ratio-type statistics

We propose an improved finite-sample nonparametric CI for the mean by applying Theorem 3.4 to the Owen (1995) CB for distribution functions. This CB is based on the Berk and Jones (B-J, 1979) likelihood ratio-type statistic:

$$BJ=\sup_{-\infty\le x\le+\infty}K[F_n(x),F(x)]$$

where

$$K(\widehat{p},p)=\widehat{p}\log(\frac{\widehat{p}}{p})+(1-\widehat{p})\log(\frac{1-\widehat{p}}{1-p}).$$

PROPOSITION 4.5. **[Berk Jones-type CI for the mean of a bounded random variable]** *Let X be a random variable with distribution function $F(x) \in \mathcal{F}_{[a,b]}$ and a finite mean μ. Assume that n i.i.d. observations on X are available and denote $X_{(1)} \leq \cdots \leq X_{(n)}$ the ordered observations. Suppose that the following confidence band of type Owen with level $1-\alpha$ is valid for the space of distributions $\mathcal{F}_{[a,b]}$:*

$$C_F^O(\alpha) = \left\{ F_0 \in \mathcal{F} : \widetilde{F}_n^L(x) \leq F_0(x) \leq \widetilde{F}_n^U(x) \right\},$$

where $\widetilde{F}_n^L(x) = \min\{p : K[F_n(x), p] \leq c_{BJ}(\alpha)\}$, $\widetilde{F}_n^U(x) = \max\{p : K[F_n(x), p] \leq c_{BJ}(\alpha)\}$, and $c_{BJ}(\alpha)$ satisfies $P[BJ > c_{BJ}(\alpha)] \geq 1-\alpha$. Then the following confidence interval

$$\begin{aligned} \widetilde{C}_\mu^{BJ}(\alpha) \;=\; & \left\{ \mu_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(X_{(n)})] X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)}) \right] X_{(k)} \leq \mu_0 \right. \\ & \left. \leq [1 - \widetilde{F}_n^L(X_{(n)})] X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)}) \right] X_{(k)} \right\} \end{aligned} \tag{2.14}$$

where $X_{(0)} = a$, and $X_{(n+1)} = b$, is a confidence interval for μ with level $1-\alpha$.

It is important to highlight that the original CB proposed by Owen was derived for continuous distribution functions. In the above theorem, we apply it to general distributions. This generalization and other properties of the Berk-Jones statistic–such as its pivotality–and its corresponding CB are discussed in Diouf and Dufour (2005a). The procedure used to derive Owen's CB also applies in noncontinuous cases. In the next section we show that the results obtained in the continuous case remain valid in the discrete case, using monotonicity properties. Note that the Owen's CB for a continuous distribution function is such that $\widetilde{F}_n^L(X_{(0)}) = 0$, $\widetilde{F}_n^L(X_{(n)}) = e^{-c_{BJ}(\alpha)} < 1$, $\widetilde{F}_n^U(X_{(0)}) = 1 - e^{-c_{BJ}(\alpha)}$, and $\widetilde{F}_n^L(X_{(n)}) = 1$. Hence, $\widetilde{F}_n^L(x)$ and $\widetilde{F}_n^U(x)$ are the effective parts of the Owen CB for all x and are used directly to build the CI for the mean.

Comparing the Owen CB to the Kolmogorov-Smirnov based CBs showed that the

Berk Jones-type CB performs similarly to the CBs based on the regularized statistics (see Diouf and Dufour, 2005a). However, for each observation, the bounds of Owen's CB are defined as the extremum—minimum or maximum—of a function. Its computation therefore requires twice as many optimizations as the number of observations.

Owen (1995) proposed an analytic approximation for $c_{BJ}(0.05)$ and $c_{BJ}(0.01)$:

$$c_{BJ}(0.05) = \begin{cases} \frac{1}{n}\left[3.0123 + 0.4835\log(n) - 0.00957\log^2(n) - 0.001488\log^3(n)\right] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}\left[3.0806 + 0.4894\log(n) - 0.02086\log^2(n)\right] & \text{for } 100 < n \leq 1000, \end{cases}$$

$$c_{BJ}(0.01) = \begin{cases} \frac{1}{n}[-4.626 - 0.541\log(n) + 0.0242\log^2(n)] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}[-4.71 - 0.512\log(n) + 0.0219\log^2(n)] & \text{for } 100 < n \leq 1000. \end{cases}$$

Jager and Wellner (2004) found that computing the Owen approximations of $c_{BJ}(\alpha)$ yield a CB with coverage probability lower than the nominal level. They propose the following correction:

$$c_{BJ}(0.05) = \begin{cases} \frac{1}{n}[3.6792 + 0.5720\log n - 0.0567\log^2(n) - 0.0027\log^3(n)] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}[3.7752 + 0.5062\log n - 0.0417\log^2(n) + 0.0016\log^3(n)] & \text{for } 100 < n \leq 1000, \end{cases}$$

$$c_{BJ}(0.01) = \begin{cases} \frac{1}{n}[5.3318 + 0.5539\log n - 0.0370\log^2(n)] & \text{for } 1 < n \leq 100, \\ \frac{1}{n}[5.6392 + .04018\log n - 0.0183\log^2(n)] & \text{for } 100 < n \leq 1000. \end{cases}$$

In our Monte Carlo investigations, we simulate the critical value of the distribution of BJ to sidestep this discussion.

2.5 Properties of confidence intervals in the continuous and noncontinuous cases

We proposed in the last section several CIs for the mean of a bounded random variable. In this section, we will study some interesting properties of those CIs in the continuous and the noncontinuous cases. For continuous $F(x)$, the distributions of the Kolmogorov Smirnov, Anderson-Darling, Eicker, regularized Anderson-Darling and Eicker, and Berk-Jones statistics are independent of the distribution being assumed under the null hypothesis and so do their critical points. Hence, the CBs for distribution functions they yield and the corresponding projected CIs for the mean depend on $F(x)$ only through the sample. The same critical points are used to build the CBs for all continuous distributions which makes them easier to compute.

One may wonder what happens in the discrete case? For noncontinuous distributions, the test distributions depend on $F(x)$, so new critical values may need to be computed in each case making CIs more difficult to compute. We will show that for those distributions, the CIs we proposed remain valid in the sense that they provide conservative CIs for the discrete variables. Moreover, we will propose important properties that can be exploited.

PROPOSITION 5.1. **[Conservative nature of continuous case critical points]** *Let $X_1, \ldots, X_n$ be n i.i.d. observations on X and $F_n(x)$ be the corresponding empirical distribution function. Let $F(x) \in \widetilde{\mathcal{F}}$ be a continuous distribution function and $G(x) \in \mathcal{F}$ be a noncontinuous one. For any level α, $0 \leq \alpha \leq 1$, the critical value associated with KS_F for testing the null hypothesis $H_0(F)$ as defined by the equation (2.5) is larger than or equal to the critical value associated with KS_G for testing the null hypothesis $H_0(G)$ and similarly for the statistics AD, E, AD_ζ, E_ζ, and BJ.*

PROPOSITION 5.2. **[Conservative property of continuous case CIs for the mean of a bounded random variable]** *Let X and Y be two bounded random variables with respective distribution functions $F(x) \in \widetilde{\mathcal{F}}_{[a,b]}$ and $G(y) \in \mathcal{F}_{[a,b]}$ whose means are finite. Let $Y_1, \ldots, Y_n$ be n i.i.d. observations on Y and $G_n(x)$ be the corresponding empirical*

distribution function. For any level α, $0 \leq \alpha \leq 1$, *the Anderson confidence interval obtained using appropriate critical points for testing the null hypothesis* $H_0(F)$ *as defined by the equation* (2.5) *yields a confidence interval for the mean of* Y *with level larger than or equal to* $1 - \alpha$, *and similarly for the confidence intervals based on* AD, E, AD_ζ, E_ζ, *and* BJ.

Proposition 5.1. shows that CBs for continuous distribution functions can be applied to general distributions and CIs for the mean may also be derived in the case where the random variable has bounded support $[a, b]$. We refer the reader to Diouf and Dufour (2005a) for a detailed discussion of these properties of the statistics and the CBs for continuous distributions. Using these properties, Proposition 5.2. states that the CIs for the mean of a continuous bounded random variable—computed with the conservative KS percentiles—can be applied to any sample from a distribution function with bounded support. The bounds so obtained will be a CI for the mean of the variable under examination with level at least equal to $1 - \alpha$. We derive a monotonicity property that can be used to reduce the width of the intervals without altering their reliability. These results are based on information about the set of discontinuities of the distribution function.

PROPOSITION 5.3. **[Range monotonicity of critical points]** *Let* $X_1, \ldots, X_n$ *be* n *i.i.d. observations on* X *and* $F_n(x)$ *be the corresponding empirical distribution function. Let* $F(x)$ *and* $G(y)$ *be two distribution functions such that* $G(\overline{\mathbb{R}}) \subseteq F(\overline{\mathbb{R}})$. *For any level* α, $0 \leq \alpha \leq 1$, *the critical value associated with* KS_F *for testing the null hypothesis* $H_0(F)$ *as defined by equation (2.5) is larger than or equal to the critical value associated with* KS_G *for testing the null hypothesis* $H_0(G)$ *and similarly for the statistics* AD, E, AD_ζ, E_ζ, *and* BJ.

PROPOSITION 5.4. **[Range monotonicity of CIs for the mean of a bounded random variable]** *Let* X *and* Y *be two bounded random variables with respective distribution functions* $F(x) \in \widetilde{\mathcal{F}}_{[a,b]}$ *and* $G(y) \in \mathcal{F}_{[a,b]}$ *whose means are finite such that* $G(\mathbb{R}) \subseteq F(\mathbb{R})$. *Let* $Y_1, \ldots, Y_n$ *be* n *i.i.d. observations on* Y *and* $G_n(x)$ *be the corresponding empirical distribution function. For any level* α, $0 \leq \alpha \leq 1$, *the Anderson confidence interval*

obtained using appropriate critical points for testing the null hypothesis $H_0(F)$ *as defined by equation* (2.5) *yields a confidence interval for the mean of* Y *with level larger than or equal to* $1-\alpha$, *and similarly for the confidence intervals based on* AD, E, AD_ζ, E_ζ, *and* BJ.

Proposition 5.4. generalizes Proposition 5.2. to all distribution functions with existing mean. It suggests that CIs for the mean can be made narrower by exploiting embeddedness of the image sets of different distributions. When studying a discontinuous distribution $G(y)$, we know that $G(y)$ takes its values in a set V^G which is included in $[0,1]$. Thus, the conservative CI for a continuous bounded random variable provides a CI for $E(Y)=\mu_Y$ with level $1-\delta_1$ greater than or equal to $1-\alpha$. If additional information about the image set of $G(y)$ is available—in particular, if we know there exists a distribution function with image V^F such that $V^G \subseteq V^F$—then the critical points for testing $F(x)$ can be used to derive a CI for μ_Y with level $1-\delta_2$ such that $1-\alpha \leq 1-\delta_2 \leq 1-\delta_1$. The CI with level $1-\delta_2$ is narrower than the CI with level $1-\delta_1$ while being reliable. Thus, using information about the nature of the discontinuity of the random variable can be useful for providing shorter CIs for μ_Y.

Consider a special case of Proposition 5.4.[4] Let Y be a random variable with distribution $G(y) \in \widetilde{\mathcal{F}}$ and X be the variable $X = (\frac{z-Y}{z})^\alpha \; I_{[0\leq Y\leq z]}$ with distribution $F(x) \in \mathcal{F}_{[0,1]}$, where z is deterministic. X is a mixture between a continuous variable bounded on $(0,1]$ and a probability mass at $x=0$. Hence, $F(x)$ is continuous on $(0,1]$ with $F(0) = Prob(Y>z) \equiv p$ and $F(1)=1$. Its corresponding Kolmogorov-Smirnov statistic is (see Appendix 2 for the proof):[5]

$$KS_F = \max_{v\in[p,1]} \left| \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[F(X_k)\leq v] - v \right|.$$

COROLLARY 5.5. **[Range monotonicity with a mass at the lower boundary]**

[4]The interest of this special case will appear later when applying the inference methods to poverty measures.

[5]In practice, the true value of p is generally unknown, but can be estimated.

Let X_1 and X_2 *be two random variables with respective distribution functions* $F_1(x) \in \mathcal{F}_{[a,b]}$ *and* $F_2(x) \in \mathcal{F}_{[a,b]}$, *continuous on* $(a, b]$, *whose means are finite such that* $p_1 = F_1(a) \leq F_2(a) = p_2$. *Let* $X_1^2, \ldots, X_n^2$ *be* n *i.i.d. observations on* X_2 *and* $F_n(x)$ *be the corresponding empirical distribution function. For any level* α, $0 \leq \alpha \leq 1$, *the Anderson confidence interval obtained using appropriate critical points for testing the null hypothesis* $H_0(F_1)$ *as defined by equation* (2.5) *yields a confidence interval for the mean of* X_2 *with level larger than or equal to* $1 - \alpha$, *and similarly for the confidence intervals based on* AD, E, AD_ζ, E_ζ, *and* BJ.

2.6 Choosing the values of parameter ζ

In this section, we assess the choice of the regularization parameter ζ. Regularizing the Anderson-Darling and the Eicker statistics improves the quality of the inference. Positive values of ζ considerably increase the power of the regularized Anderson-Darling and Eicker goodness-of-fit tests: the test power is low for small ζ but rises quickly when ζ increases before becoming almost constant.

In practice, if ζ is not chosen independently of the sample that is used to estimate the CIs, the distributions of the statistics may be modified and the properties we have derived so far will have to be reinvestigated and new critical points simulated. To avoid this, the optimal value of ζ may be chosen on an auxiliary sample independent from the estimation sample of the CIs using a split-sample procedure or other approaches. We illustrate two ways of choosing ζ in sections 8 and 9.

In section 8, we investigate the performance of the CIs using Monte Carlo simulations. In this case, we know the distribution from which the sample is. Then, we can choose the minimum value of ζ that provides a "sufficiently" powerful test. In fact, given that we compute the critical points of the statistics by simulation, we control the level of the test and the corresponding CIs. Thus, maximizing the power of the goodness of fit test allows to minimize the width of the CIs (see Pratt, 1961). The optimal value of ζ so obtained depends on the sample size and, to a lesser extent, on the distribution function.

However, even if the value of ζ that maximizes the power of the test is not used, most of the improvement is achieved as soon as ζ is high enough. We use this procedure to choose the optimal value of ζ that will be used to perform inference on the Foster, Greer, and Thorbecke poverty measures.

In section 9, we use PROGRESA data (from Mexico) to analyze the profile of poverty of Mexican households. In this case, no information about the distribution of the variable is available and the total size of the sample is fixed. A split-sample procedure (see Dufour and Jasiak, 1993) can be used to estimate the parameter ζ and the CIs independently each from other. The procedure decomposes as follows. First, the initial sample is split into two independent subsamples using i.i.d. drawings. Second, one sample–the auxiliary sample–is used to estimate (by trial and error) the values of the parameter that minimize the width of the Anderson-Darling and the Eicker CIs. Third, the remaining sample–the estimation sample–is used to estimate CIs with the formulas provided in the last two sections. This out of sample procedure guarantees that the auxiliary sample and the estimation sample are independent and insures the validity of the inference, the distribution of the statistics being held unchanged by the estimation of the parameters. Ideally, one would use a small part of the initial sample as auxiliary sample–some theoretical studies (see Dufour and Jasiak, 1993) recommends to use up to 10 percent of the sample. However, the width of the nonparametric CIs from the regularized statistics depends "critically" of the value of ζ. So does the performance of the inference methods. We thus propose to use up to 20 percent of the initial sample to estimate ζ, whenever the sample size allows us to do so.

For both procedures, the optimal value of ζ for the Anderson-Darling statistic is very likely to be different from those for the Eicker statistic. Moreover, the critical points of both statistics need to be simulated at each step of the procedures, for the value of ζ being tested and the current choice of sample size and other involved parameters.

2.7 Application to the Foster, Greer, and Thorbecke poverty measures

Using the generalization of our inference methods to higher moments, we apply them to the popular Foster, Greer and Thorbecke (1984) poverty measures. These are defined by

$$P_\delta(Y,z) = \int_0^z \left(\frac{z-y}{z}\right)^\delta \, dF(y)$$

where $\delta > 0$, Y is a welfare indicator (which is generally income or expenditures) with continuous distribution function $F_Y(y)$ of support $[0,+\infty)$, and z is the poverty line. Rewriting them yields:

$$P_\delta(Y,z) = \int_0^{+\infty} \left(\frac{z-y}{z}\right)^\delta \mathbb{1}[y \leq z] \, dF(y) = E[X]$$

where

$$X = \left(\frac{z-Y}{z}\right)^\delta \mathbb{1}[0 \leq Y \leq z]$$

is a random variable bounded on $[0,1]$ with cumulative distribution

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

and a probability mass $G_X(0) = 1 - F_Y(z)$ at 0.

The FGT poverty measure is the mean of a bounded random variable. Procedures adapted to bounded random variables can then be applied to them. We perform inference on these measures using inference methods proposed in this paper. Monte Carlo simulations are done to study the performance of the CIs.

Let

$$\widehat{P}_\delta = \frac{1}{n}\sum_{i=1}^n \left(\frac{z-Y_i}{z}\right)^\delta \mathbb{1}[Y_i \leq z] = \frac{1}{n}\sum_{i=1}^n X_i$$

be an unbiased estimator of P_δ with variance $\widehat{V}(\widehat{P}_\delta) = \frac{1}{n}[\widehat{P}_{2\delta} - \widehat{P}_\delta^2]$. The asymptotic $N(0,1)$ t-statistic $W = \left(\widehat{P}_\delta - P_\delta\right) \big/ \widehat{V}\left[\widehat{P}_\delta\right]$ can be used to test $H_0 : P_\delta = P_\delta^0$ against the alternative $H_1 : P_\delta \neq P_\delta^0$ (see Kakwani, 1993). The corresponding asymptotic and bootstrap CIs with level $1-\alpha$ for P_δ are, respectively:

$$\begin{aligned} C_{P_\delta}^A(\alpha) &= \left\{p_0 \in \mathbb{R} : \widehat{P}_\delta - z_{(1-\frac{\alpha}{2})} * \left[\widehat{V}(\widehat{P}_\delta)\right]^{1/2} \leq p_0 \leq \widehat{P}_\delta + z_{(1-\frac{\alpha}{2})} * \left[\widehat{V}(\widehat{P}_\delta)\right]^{1/2}\right\}, \\ C_{P_\delta}^B(\alpha) &= \left\{p_0 \in \mathbb{R} : \widehat{P}_\delta - D_{(1-\frac{\alpha}{2})}^W * \left[\widehat{V}(\widehat{P}_\delta)\right]^{1/2} \leq p_0 \leq \widehat{P}_\delta - D_{(\frac{\alpha}{2})}^W * \left[\widehat{V}(\widehat{P}_\delta)\right]^{1/2}\right\}, \end{aligned}$$

where $z_{(p)}$ and $D_{(p)}^W$ are the p^{th} percentiles of the standard normal distribution and the bootstrap distribution of W, respectively.

Similarly, rewriting $\widehat{P}_\delta = \frac{1}{n}\sum_{i=1}^{n} X_i$ where $X_i = \left(\frac{z-Y_i}{z}\right)^\delta \mathbb{1}_{[Y_i \leq z]}$, $\forall i$, the Hora and Hora and Fishman CIs for P_δ are

$$\begin{aligned} C_{P_\delta}^{H\&H}(\alpha) &= \left\{p_0 \in \mathbb{R} : \widehat{P}_\delta - c_{KS}(\alpha) \leq p_0 \leq \widehat{P}_\delta + c_{KS}(\alpha)\right\}, \\ \widetilde{C}_{P_\delta}^F(\alpha) &= \left\{p_0 \in \mathbb{R} : \mu_L(\overline{X}_n, n, \alpha) \leq p_0 \leq \mu_U(\overline{X}_n, n, \alpha)\right\}, \end{aligned}$$

where $c_{KS}(\alpha)$ is the $(1-\alpha)^{th}$ Kolmogorov-Smirnov percentile value,

$$\mu_L = \begin{cases} \left\{t : 0 < t \leq \widehat{P}_\delta \leq 1 \text{ and } e^{nf(\widehat{P}_\delta - t, t)} = \alpha/2\right\} \text{ if } \widehat{P}_\delta > 0, \\ 0 \text{ if } \widehat{P}_\delta = 0, \end{cases}$$

$$\mu_U = \begin{cases} \left\{t : 0 \leq \widehat{P}_\delta \leq t < 1 \text{ and } e^{nf(t - \widehat{P}_\delta, 1-t)} = \alpha/2\right\} \text{ if } \widehat{P}_\delta < 1, \\ = 1 \text{ if } \widehat{P}_\delta = 1, \end{cases}$$

and $f(s,u) = (s+u)\ln[u(s+u)^{-1}] + (1-s-u)\ln[(1-u)(1-s-u)^{-1}]$ for $0 < s < 1-u$.

Finally, Theorem 3.4 provides the following empirical distribution-based CIs for P_δ :

$$\widetilde{C}_{P_\delta}(\alpha) = \left\{p_0 \in \mathbb{R} : \mu_L \leq p_0 \leq \mu_U\right\},$$

where

$$\mu_L = [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)},$$

$$\mu_U = [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)},$$

and $X_{(1)} < \cdots < X_{(n)}$ are the order statistics of a sample of n i.i.d. observations on X, $X_{(0)} = 0$, $X_{(n+1)} = 1$, $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, and $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$, $\forall x$, where G_n^L and G_n^U are the lower and the upper bounds of the CB with level $1 - \alpha$ for $G_X(x)$, respectively. These CBs can be the Kolmogorov-Smirnov, Anderson-Darling, Eicker, regularized Anderson-Darling and Eicker, or Owen CBs for distribution functions. We demonstrated that the continuous conservative critical points of the underlying empirical distribution-based statistics can be used to derive conservative CIs for P_δ. Moreover, the property of range monotonicity with a mass at the lower boundary we derived in Corollary 5.5 allows the construction of narrower CIs for P_δ, adjusted for the pattern that exhibits $G_X(x)$.

2.8 Monte Carlo study

In this section, we use Monte Carlo simulations to study the performance of the CIs for the poverty measure P_2 (P_δ, for $\delta = 2$). We suppose that the income Y comes from a mixture:

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

where $SM(a, b, c)$ is the Singh-Maddala distribution with cumulative distribution function $F(y) = 1 - [1 + ay^b]^{-c}$. This distribution has been proven by Brachman, Stich, and Trede (1996) to mimic the income of several developed countries, such as Germany, well. Following Davidson and Flachaire (2007), we set $a = 100$, $b = 2.8$, and $c = 1.7$. We assume that the poverty line, z, is half the median of the $SM(a, b, c)$ distribution and that $P_0 = 0.1$. The true value of P_2 is $P_2^0 = 0.013017\ P_0$, for our setup. CIs

with level 95% are simulated for sample sizes $n = 50, 100, 200$ and $n = 500, 1000$ using $N = 10,000$ and $N = 500$ replications, respectively. For a $SM(100, 2.8, 1.7)$, the probability of having observations greater than z is $\eta = 0.89$. Thus, the probability that $X = 0$ is $p = (1 - P_0) + \eta P_0$. The results of the simulations are presented in Table 2.1 and 2.2. Table 2.1 illustrates the choice of the regularization parameter ζ and Table 2.2 provides the coverage probability and the average width of the simulated CIs for $p = 0.989$–which corresponds to $\eta = 0.89$–and $p = 0$–which corresponds to the continuous conservative case.

2.8.1 Choice of ζ

We study the choice of the regularization term ζ adapted to the Singh-Maddala distribution and the sample sizes of the simulations. We choose ζ such as to maximize the power of the test $H_0 : X \sim SM(100, 2.8, 1.7)$ versus $H_1 : X \sim SM(100, 2.89, 1.7)$. We compute the power of the test for $n = 500$ observations—the median sample size of our simulations—using $N = 10,000$ replications. The critical points of the statistics of E, E_ζ, A, and A_ζ are simulated for a level $\alpha = 5\%$ and values of ζ from 0.005 to $1,000,000$.

Table 2.1 provides the level and the power of the corresponding goodness-of-fit tests. The results show that for $\zeta = 0$, the Anderson-Darling and Eicker tests have a very low test power, less than 4.2% for AD and less than 19% for E. The power increases considerably when ζ is different from zero: it is low for small ζ but increases quickly when ζ increases before becoming almost constant. We choose $\zeta = 0.07$, the lowest value of ζ that increases the test power "sufficiently."

Table 2.1 : Choice of ζ : simulated level and power of the Kolmogorov-Smirnov based tests for different values of ζ
$H0: \ X \sim SM[100, 2.8, 1.7]$ vs $H1 : X \sim SM[100, 2.89, 1.7]$
$n = 500$, and $N = 10,000$ replications

	level (in %)			
ζ	E	E_ζ	AD	AD_ζ
0.005	5.11	4.52	4.59	5.09
0.070	5.14	5.08	4.77	5.01
0.100	4.95	4.95	4.84	4.96
0.150	5.40	5.08	5.16	5.30
0.200	4.75	4.58	4.98	4.71
0.300	5.11	5.24	5.29	5.04
0.400	5.28	5.06	4.74	5.22
0.500	4.82	5.18	5.17	5.13
0.750	4.95	5.23	4.58	5.21
1.000	5.16	4.98	5.23	5.08
10.000	5.02	4.61	4.88	4.61
100.000	5.17	5.20	5.20	5.20
1,000.000	5.12	5.38	4.78	5.38
1,000,000.000	5.11	4.94	5.03	4.94

	Power (in %)			
ζ	E	E_ζ	AD	AD_ζ
0.005	18.32	45.79	4.12	51.20
0.070	17.96	62.88	3.56	61.16
0.100	18.29	62.89	3.89	61.80
0.150	18.41	63.37	3.90	63.42
0.200	17.86	62.69	4.00	61.88
0.300	18.81	64.44	3.90	63.63
0.400	17.78	61.94	3.72	62.00
0.500	17.52	63.34	3.84	62.92
0.750	18.31	63.36	3.96	63.28
1.000	17.50	62.91	3.87	62.74
10.000	18.04	61.73	4.03	61.69
100.000	18.22	62.57	3.73	62.57
1,000.000	17.37	63.40	3.93	63.40
1,000,000.000	17.92	62.32	3.57	62.32

2.8.2 Results

Table 2.2 provides the coverage probability and the average width of the CIs for P_2. The first part of Table 2.1 shows that the asymptotic and bootstrap CIs are not reliable, even for fairly large samples. The coverage probability of the asymptotic CI is 31.85 percent for $n = 50$ and increases to 84.20 percent for $n = 1000$ while those of the bootstrap-t CI goes from 25.03 percent for $n = 50$ to 92.10 percent for $n = 500$, reaching the nominal level of confidence–95 percent–only at $n = 1000$. As n increases, the proportion of zero values becomes lower and lower, improving the coverage probability of the bootstrap CI. However, when the bootstrap fails, it yields bad precision in addition to its poor coverage probability. By contrast, the exact nonparametric CIs are strongly reliable. Exact nonparametric CIs provide coverage probability typically larger than the nominal level for all sample size, even when the adjusted (nonconservative) critical values are

used.

Amongst the exact CIs, the regularized Anderson Darling-type and Eicker-type CIs perform the best. When continuous conservative critical points are used, Fishman's CI achieves the best width because it accounts for the noncontinuity of the distribution. However, the computation of this CI is time consuming. When critical points are adjusted to take into account the discontinuity of X, the regularized Eicker-type CI achieves the best width followed by the regularized Anderson Darling-type CI. The adjustment to the discrete framework has a great impact on the width of CIs, reducing it by a factor of approximately 3. The Anderson Darling-type and Eicker-type CIs are worse than the regularized CIs, as expected. Anderson's CI performs as well as the regularized Anderson Darling-type one for $n = 500$ and $n = 1000$. Likewise, for these sample sizes, the Berk Jones-type CI yields precision comparable to those based on the regularized statistics but building it requires to compute 500 and 1000 optimizations, respectively. Then, when considering the reliability, the constancy in the performance, and the easiness of computation, the regularized Anderson Darling-type and Eicker-type CIs provides the best CIs for P_2, even though the optimal value for ζ are not used.

Table 2.2: Simulated confidence intervals for the FGT poverty measure $P_2(Y, z)$

$$\text{with } Y = \begin{cases} z \text{ with probability } 1 - P_0 = 0.1 \\ SM(100, 2.8, 1.7) \text{ with probability } P_0 = 0.9 \end{cases}, \zeta = 0.07,$$

$N = 10,000$ replications for $n = 50, 100, 200$, and

$N = 500$ replications for $n = 500, 1000$

		Coverage probability (in %)				
	n	50	100	200	500	1000
Asymptotic	$p = 0$	31.85	46.03	61.87	77.00	84.20
	$p = 0.989$	-	-	-	-	-
Bootstrap-t	$p = 0$	25.03	39.13	62.78	92.10	96.40
	$p = 0.989$	-	-	-	-	-
Fishman	$p = 0$	94.07	93.76	95.24	99.30	100.00
	$p = 0.989$	-	-	-	-	-
Hora & Hora	$p = 0$	100.00	100.00	100.00	100.00	100.00
	$p = 0.989$	100.00	100.00	100.00	100.00	100.00
Anderson	$p = 0$	100.00	100.00	100.00	100.00	100.00
	$p = 0.989$	99.89	99.88	99.98	99.90	100.00
Eicker	$p = 0$	100.00	100.00	100.00	100.00	100.00
	$p = 0.989$	99.91	99.82	99.86	100.00	100.00
E_ζ	$p = 0$	100.00	100.00	100.00	100.00	100.00
	$p = 0.989$	99.90	99.88	99.98	99.90	100.00
AD	$p = 0$	99.99	100.00	100.00	100.00	100.00
	$p = 0.989$	99.84	99.95	99.99	100.00	100.00
AD_ζ	$p = 0$	100.00	100.00	100.00	100.00	100.00
	$p = 0.989$	100.00	100.00	100.00	100.00	100.00
BJ	$p = 0$	99.99	99.99	100.00	100.00	100.00
	$p = 0.989$	99.61	99.79	99.92	99.90	99.80

		Width				
	n	50	100	200	500	1000
Asymptotic	$p = 0$	0.005	0.004	0.004	0.003	0.002
	$p = 0.989$	-	-	-	-	-
Bootstrap-t	$p = 0$	7.421	7.076	60.935	4.119	0.011
	$p = 0.989$	-	-	-	-	-
Fishman	$p = 0$	0.076	0.041	0.023	0.011	0.007
	$p = 0.989$	-	-	-	-	-
Hora & Hora	$p = 0$	0.377	0.268	0.190	0.121	0.086
	$p = 0.989$	0.098	0.058	0.038	0.022	0.015
Anderson	$p = 0$	0.190	0.135	0.096	0.062	0.044
	$p = 0.989$	0.050	0.030	0.020	0.012	0.008
Eicker	$p = 0$	0.946	0.898	0.824	0.691	0.596
	$p = 0.989$	0.942	0.895	0.821	0.690	0.595
E_ζ	$p = 0$	0.107	0.073	0.051	0.032	0.023
	$p = 0.989$	0.038	0.025	0.017	0.011	0.008
AD	$p = 0$	0.455	0.296	0.175	0.079	0.042
	$p = 0.989$	0.306	0.183	0.102	0.045	0.024
AD_ζ	$p = 0$	0.167	0.105	0.067	0.039	0.026
	$p = 0.989$	0.047	0.029	0.019	0.012	0.008
BJ	$p = 0$	0.103	0.056	0.031	0.015	0.009
	$p = 0.989$	0.056	0.034	0.020	0.011	0.007

2.9 Empirical illustration

In this section, we analyze the profile of poverty of rural Mexican households using our inference methods. We employ data that have been collected as part of the targeting and evaluation program of PROGRESA.[6] A census of households in a set of 506 rural

[6]PROGRESA is a health, education, and nutrition program of the Mexican government aimed to reduce poverty in targeted rural communities. Details about this program and the data they collect can

communities has been conducted in 1997, 1998, and 1999 and the data processed to insure comparability. Data about households' characteristics are extracted from the November 1997 survey and expenditure aggregate is constructed using the March 1998 survey.[7] The poverty line is set to 159 pesos, the per capita expenditure of the median household in the full set of households.

Interestingly, these data allow one to analyze poverty in Mexico both at the national and regional levels. First, using the census as a whole, we build CIs for the level of poverty P_2 of rural households in Mexico. Then, drawing samples randomly from the census, we study the profile of poverty of PROGRESA targeted communities and compare the performance of our improved nonparametric inference techniques to the existing ones. Furthermore, we analyze the determinants of poverty in rural areas in Mexico for the involved communities using various household characteristics including the gender and the level of education of the household's head. We compare the poverty profile of households with a female head to those of households which head is a male, and the poverty profile of households with an educated head to those with a non-educated head.

A split-sample approach is chosen to estimate the regularization term ζ.

The results show that in addition to being unreliable, the widths of the asymptotic CIs are often too small to be realistic while the bootstrap can fail even in precision, delivering CIs which widths can be ten times larger than those of the exact methods. The analysis of profile of poverty shows that on average, rural households targeted by PROGRESA do not have a very high level of poverty. However, the poverty profile depends greatly on the type of households' head. The level of poverty among households with a male head is much smaller than the level of poverty among households with a female head. Moreover, households with an educated head appear to be more prone to escape poverty than households with a non-educated head. These conclusions raise questions about the equity in the distribution of Mexican wealth and provide hints for

be obtained on the website of IFPRI (International Food Policy Research Institute): www.ifpri.org.

[7] The data set excludes households in the expenditure survey that had not been interviewed in November 1997 and 10 communities with fewer than 10 households with expenditure information, leaving 20544 households in 496 communities (see Demombynes, Elbers, Lanjouw and Lanjouw, 2007)

designing policies to reduce poverty in rural Mexico. Policies aimed at reducing illiteracy of households members in these communities can be effective in reducing poverty. Those education programs should target both children and adults, in particular households' heads to have short-term effects. Likewise, policies aimed at securing the income of households with a female head could help reduce poverty in rural Mexico. An example of such policies can be reforms aimed at securing land ownership for female or at improving labor productivity for households with a female head, the latter being less productive for physically intensive activities such as farming.

2.9.1 Analysis of the poverty profile of rural households in Mexico

We use the census data to build CIs for the level of poverty P_2 of rural households in Mexico. The following split sample procedure is used to choose the value of ζ. First, we draw randomly without replacement an auxiliary sample of 20 percent of the original data set to estimate ζ for the regularized Anderson-Darling and Eicker statistics. For different values of ζ, we simulate the critical points of the statistics and compute the corresponding CI for P_2. Then, we choose ζ so as to minimize the width of each CI. Second, we build CIs for the FGT poverty measure P_2 using the remaining 80 percent of the sample–the estimation sample. By construction, the built auxiliary and the estimation samples are independent and our inference methods can be applied.

Table 2.3 presents the critical point of the $\zeta-$regularized Anderson-Darling and Eicker CIs and the width of the CIs using different values of ζ on the auxiliary sample. The smallest widths are achieve by $\zeta_{AD} = 0.004$ for the regularized Anderson Darling-type CI and $\zeta_E = 0.0062$ for the regularized Eicker-type CI. We use these values for the rest of this analysis.

Table 2.3: Choice of ζ_E and ζ_{AD} based on an auxiliary sample of $n_1 = 1000$

ζ	c_{E_ζ}	width (in %)	ζ	c_{AD_ζ}	width (in %)
0.0001	9.4615631	3.299	0.0001	3.2947588	2.065
0.0004	4.9269052	2.110	0.0004	2.9433350	1.825
0.0008	3.8322506	1.832	0.0008	2.7829854	1.743
0.0010	3.5789932	1.769	0.0010	2.7307002	1.722
0.0030	2.7472731	1.593	0.0020	2.5686677	1.678
0.0040	2.5956398	1.578	0.0030	2.4592033	1.662
0.0050	2.4723399	1.568	0.0035	2.4167582	1.661
0.0060	2.3799549	1.567	0.0039	2.3817260	1.658
0.00610	2.3709240	1.567	0.0040	2.3730547	1.657
0.00615	2.3674749	1.567	0.0045	2.3356472	1.657
0.0062	2.3615190	1.566	0.0050	2.3010155	1.658
0.0065	2.3356573	1.566	0.0055	2.2770130	1.666
0.0070	2.2972038	1.566	0.0060	2.2569682	1.677
0.0075	2.2704249	1.571	0.0070	2.2036625	1.684
0.0080	2.2361952	1.572	0.0100	2.0612441	1.696
0.0090	2.1752713	1.575	0.0300	1.5690714	1.756
0.0100	2.1205124	1.578	0.0500	1.3455539	1.802
0.0300	1.5741812	1.666	0.0700	1.1872093	1.815
0.0600	1.2460575	1.732	0.1000	1.0268528	1.822
0.07	1.1742951	1.743	0.5000	0.4938216	1.840
0.08	1.1116034	1.750	2	0.2510734	1.845

Table 2.4 provides the estimation results. The lines Asymp$_r$ and Bootstrap$_r$ give the asymptotic and bootstrap-t CIs estimated on the residual sample and the lines Asymp and Bootstrap refer to CIs estimated with the whole sample. The regularized Anderson Darling-type CI–based on AD_ζ–achieves the best width among all CIs followed by the

asymptotic and bootstrap CIs. However, we have shown that these asymptotic methods can be unreliable when applied to distributions of the kind involved in poverty studies, even with fairly large samples. The width of the regularized Eicker-type CI (based on E_ζ) is relatively close to those of the others: it is 0.201 for all households, 0.04 higher than the width of the asymptotic CI and 0.059 more than the width of the bootstrap CI. However, this CI performs far better than the other nonparametric CIs, including the Berk Jones-type CI.

According to the AD_ζ CI, rural Mexican households have a level of poverty between 0.499 percent and 0.566 percent with a level of confidence of 95 percent. This range of poverty level seems relatively low. However, it does not reflect the underlying situation of the country. When we study the level of poverty of households taking in account some characteristics of those households, it appears that poverty is unevenly distributed among populations. Still according to the regularized Anderson Darling-type CI, households with a male head are much less poor than households with a female head. Poverty levels range from 0.405 percent to 0.469 percent for households with male head and from 1.166 percent to 1.538 percent for households with female head, a difference of 0.761 percent and 1.069 percent for the lower and the upper bounds, respectively. This difference may be due to the fact that Mexican households are in large part farmers. Households with a male head are likely to raise more revenue from harvest than others and thus, they are more prone to escape from poverty. This feature might also be related to the land successional law in rural areas which usually prioritize men against women. Likewise, Tables 2.4.b and 2.4.c show that, the level of education of households' head has a dampening effect on poverty. The level of poverty of households whose head has no education ranges from 0.915 percent to 1.107 percent, which is respectively 0.624 percent and 0.757 percent larger than the bounds of the poverty level for households with educated leader. Education appears to be determinant in improving the financial situation of the households and allow them not to live in poverty.

Comparing the performance of the various inference methods, we see that the asymptotic and the bootstrap CIs achieve the smaller width among all the methods. However,

we have shown that this method can be unreliable when applied to distributions of the kind involved in poverty and inequality studies, even with fairly large samples. Among the nonparametric CIs, the regularized Anderson-Darling and Eicker statistics yield the best inference, with widths very close to those of the asymptotic CI.

Table 2.4: Mexican households: Confidence intervals for the FGT poverty measure $P_2(Y, z)$ for different types of households' heads

$n = 20485$, $\zeta_{AD} = 0.004$, $\zeta_E = 0.0062$

Table 2.4a: All households

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.450	0.613	0.163
Asymp	0.462	0.624	0.161
Bootstrap_r	0.445	0.627	0.182
Bootstrap	0.476	0.619	0.142
Fishman	0.415	0.695	0.279
Hora Hora	-1.463	2.549	4.012
Anderson	0.003	2.549	2.546
Eicker	0.354	0.732	0.378
E_ζ	0.432	0.632	0.201
AD	0.360	0.833	0.473
AD_ζ	0.499	0.566	0.066
BJ	0.044	2.688	2.644

Table 2.4b: Households with an educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.244	0.395	0.151
Asymp	0.254	0.404	0.150
Bootstrap_r	0.256	0.419	0.163
Bootstrap	0.266	0.432	0.166
Fishman	0.215	0.477	0.262
Hora-Hora	-0.807	1.464	2.271
Anderson	0.003	1.464	1.461
Eicker	0.147	0.511	0.364
E_ζ	0.227	0.413	0.186
AD	0.177	0.636	0.459
AD_ζ	0.291	0.350	0.059
BJ	0.039	1.370	1.331

Table 2.4c: Households with a non-educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.806	1.210	0.405
Asymp	0.825	1.224	0.399
Bootstrap_r	0	1.230	0.388
Bootstrap	0.842	1.292	0.451
Fishman	0.717	1.408	0.691
Hora-Hora	-1.639	3.689	5.328
Anderson	0.031	3.689	3.658
Eicker	0	1.507	0.964
E_ζ	0.753	1.262	0.509
AD	0.602	1.794	1.192
AD_ζ	0.915	1.107	0.192
BJ	0.163	3.639	3.476

Table 2.4d: Households with a female head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.919	1.765	0.847
Asymp	0.971	1.805	0.834
Bootstrap_r	0.941	1.829	0.888
Bootstrap	1.014	1.847	0.833
Fishman	0.812	2.184	1.373
Hora Hora	-3.025	5.801	8.827
Anderson	0.002	5.801	5.800
Eicker	0.506	2.270	1.764
E_ζ	0.858	1.825	0.967
AD	0.633	3.132	2.500
AD_ζ	1.166	1.538	0.372
BJ	0.087	6.782	6.696

Table 2.4e: Households with a male head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.360	0.512	0.152
Asymp	0.368	0.518	0.151
Bootstrap_r	0.369	0.529	0.160
Bootstrap	0.384	0.531	0.147
Fishman	0.323	0.590	0.267
Hora Hora	-0.701	1.587	2.288
Anderson	0.022	1.587	1.565
Eicker	0.256	0.630	0.374
E_ζ	0.338	0.535	0.197
AD	0.276	0.731	0.454
AD_ζ	0.405	0.469	0.064
BJ	0.091	1.420	1.329

Comparing the performance of the various inference methods, we see that the asymptotic and the bootstrap CIs achieve the smaller width among all the methods. However, we have shown that this method can be unreliable when applied to distributions of the kind involved in poverty and inequality studies, even with fairly large samples. Among the nonparametric CIs, the regularized Anderson-Darling and Eicker statistics yield the best inference, with widths very close to those of the asymptotic CI.

2.9.2 Analysis of the profile of poverty of the Mexican households targeted by PROGRESA

We use subsamples of $n = 500$ and 1000 to perform inference for the level of poverty of PROGRESA-targeted households and illustrate the relative performance of the improved CIs compared to those of the other methods on samples of such sizes. We implement the same procedure as previously. First, we draw randomly without replacement a sample of n observations from the census. Second, we apply a split sample procedure to choose ζ and estimate the CIs: we use an auxiliary sample of twenty percent (20%) of each subsample to estimate ζ and use the independent remaining sample to estimate the CIs.

Tables 2.5 and 2.7 present the critical points of the $\zeta-$regularized Anderson-Darling and Eicker statistics and the width of the corresponding CIs using different values of ζ and the auxiliary samples. For $n = 500$, the smallest widths are achieved by $\zeta_{AD} = 0.45$ for the regularized Anderson Darling-type CI and $\zeta_E = 0.039$ for the regularized Eicker-type CI. For $n = 1000$, the smallest widths are achieve by $\zeta_{AD} = 0.5$ and $\zeta_E = 0.05$. These values will be used to perform inference for P_2 in the remaining of this subsection.

Table 2.5: Choice of ζ_E and ζ_{AD} based on an auxiliary sample of $n_1 = 100$

ζ	c_{E_ζ}	width (in %)	ζ	c_{AD_ζ}	width (in %)
0.0001	26.3000000	19.085	0.0001	4.40795891	17.391
0.0004	13.1500000	10.501	0.0010	3.15955927	10.366
0.0010	8.3167902	7.406	0.0030	2.71751268	8.346
0.0020	5.8808588	5.897	0.0070	2.42001073	7.307
0.0030	4.8017011	5.255	0.0100	2.31582510	7.062
0.0040	4.1583951	4.886	0.0400	1.70609029	6.001
0.0070	3.1434512	4.337	0.070	1.36385347	5.463
0.0100	2.6300000	4.082	0.100	1.16889614	5.227
0.0200	1.8596908	3.745	0.20	0.84814868	4.92147
0.0300	1.5184312	3.620	0.25	0.75047690	4.78842
0.0350	1.4057941	3.582	0.30	0.68926165	4.75870
0.0370	1.3672719	3.570	0.40	0.60397513	4.73807
0.0380	1.3491615	3.564	0.45	0.56616351	4.69061
0.0390	1.3317523	3.559	0.50	0.54091567	4.70103
0.0400	1.3220528	3.569	0.55	0.51486184	4.67863
0.0410	1.3060056	3.564	0.60	0.49171955	4.65577
0.0420	1.3086985	3.600	1	0.38420230	4.63463
0.0440	1.2892718	3.615	10	0.12147042	4.558
0.0450	1.2823184	3.628	40	0.06140494	4.594
0.0500	1.2477436	3.683	100	0.03832555	4.542

Table 2.6: Mexican householdsin PROGRESA: Confidence intervals for $P_2(Y, z)$ for different types of households' heads

$n = 500$, $\zeta_{AD} = 0.45$, $\zeta_E = 0.039$

Table 2.6a: All households

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.132	1.257	1.390
Asymp	-0.017	1.200	1.217
Bootstrap_r	-0.009	9.908	9.917
Bootstrap	0.167	11.797	11.631
Fishman	0.071	2.055	1.985
Hora-Hora	-0.995	2.178	3.172
Anderson	4.9E-05	2.178	2.173
Eicker	0	1.938	1.938
E_ζ	7.7E-06	2.104	2.104
AD	0.069	5.333	5.264
AD_ζ	4.9E-06	2.422	2.422
BJ	0.069	2.188	2.120

Table 2.6b: Households with an educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.313	1.079	1.392
Asymp	-0.166	1.125	1.291
Bootstrap_r	-0.029	4.838	4.866
Bootstrap	0.077	7.265	7.188
Fishman	0.022	2.229	2.207
Hora-Hora	-1.133	2.092	3.225
Anderson	0	2.092	2.092
Eicker	0	1.800	1.800
Eicker_ζ	0	1.855	1.855
AD	0.036	6.721	6.685
AD_ζ	0	2.208	2.208
BJ	0.026	2.303	2.278

Table 2.6c: Households with a non-educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.674	2.657	3.331
Asymp	-0.521	2.274	2.795
Bootstrap_r	0	63.382	63.634
Bootstrap	-0.010	44.883	44.893
Fishman	0.017	4.920	4.903
Hora-Hora	-3.064	4.817	7.882
Anderson	0.005	4.817	4.812
Eicker	0	3.680	3.680
Eicker_ζ	0.001	4.626	4.625
AD	0.055	15.092	15.037
AD_ζ	0.002	5.708	5.706
BJ	0.036	5.106	5.070

Table 2.6d: Households with a female head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.019	0.165	0.185
Asymp	-0.948	3.696	4.644
Bootstrap_r	0.014	0.555	0.541
Bootstrap	-0.437	56.411	56.848
Fishman	3.2E-05	10.682	10.679
Hora-Hora	-6.764	9.511	16.275
Anderson	3.6E-05	9.511	9.508
Eicker	0.017	40.349	40.332
E_ζ	0	5.179	5.179
AD	0.075	31.291	31.216
AD_ζ	0	10.033	10.033
BJ	0.053	9.471	9.418

Table 2.6e: Households with a male head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.156	1.392	1.548
Asymp	-0.123	1.121	1.244
Bootstrap_r	0.074	11.075	11.001
Bootstrap	0.054	47.155	47.102
Fishman	0.038	2.007	1.969
Hora-Hora	-1.142	2.140	3.283
Anderson	0	2.140	2.140
Eicker	0	1.808	1.808
E_ζ	0	2.182	2.182
AD	0.044	5.690	5.646
AD_ζ	0	2.459	2.459
BJ	0.035	2.136	2.101

Tables 2.6 and 2.8 show the estimated CIs for P_2 corresponding to $n = 500$ and $n = 1000$, respectively. The lines Asymp_r and Bootstrap_r give the asymptotic and bootstrap CIs estimated on the residual sample and the lines Asymp and Bootstrap refer to CIs estimated on the whole sample. We use CIs based on simulated critical points to compare the Berk Jones-type CI to the other ones.

For both samples, the asymptotic CI achieves the best width among all CIs but given the size of the samples, its validity is questionable. The performance of the bootstrap method is not consistent throughout the subsamples: the bootstrap CI sometimes delivers the second best width but usually provides a very poor precision with a width over 50 times larger than those of the nonparametric CIs. Nonparametric methods provide CIs with width very close to those of the asymptotic CI, while being strongly reliable as proved by the Monte Carlo simulation. For $n = 500$, the performance of the regularized methods is similar to that of the other nonparametric methods but it improves when $n = 1000$. The regularized Eicker CI becomes the best nonparametric method or one of the best, depending on the studied subsample. However, the performance of the regularized statistic is not clearly above those of the other nonparametric approaches as was the case when the census was used. On some subsamples, The Eicker and the Fishman CIs provide results comparable those based on the regularized CIs. These results emphasize two features: (1) the importance of estimating the regularization term with large enough number of observations and (2) the link between the magnitude of the improvement and the underlying distribution of the observations. When chosen over an auxiliary sample of 100 or 200 observations, the values of the parameters are higher than those chosen when using an auxiliary sample of 1000 observations. Thus, though still improving significantly the initial methods–especially for the Anderson-Darling statistic–the magnitude of this improvement is lowered. Of course, this improvement depends a lot on the underlying distribution. The more the distribution exhibits heavy tails, the higher is the improvement achieved by regularized statistics.

For $n = 100$, the CI based on E shows that 95 percent of rural households targeted by PROGRESA have a level of poverty between 0 and 1.94 percent. However, the incidence

of poverty differs depending on the characteristics of the head. Still according to the Eicker-type CI, the incidence of poverty among households with a non-educated head is more than twice those of the households with an educated head. A similar picture is depicted by the regularized Eicker-type CI when the gender of the households' head is accounted for: households with a male head appear to have half less poor than households with a female head. Similar conclusions arises for $n = 1000$. A male head with a minimum level of education increases highly the likelihood of escaping from poverty.

2.10 Conclusion

In this paper, we propose finite-sample nonparametric CIs based on empirical distribution functions for the mean of a bounded random variable. We develop an innovative methodology to derive CIs for any monotonic functional of a distribution function from CBs for this distribution using projection techniques. Two CIs can be derived from each CB: one using the whole bands and one using their restricting parts. The latter accounts for the property of the distribution functions which, by definition, always have values between 0 and 1 and is thus, thinner. We apply this method to the mean of a bounded random variable and provide explicit expressions that are easy to compute.

We prove that the Anderson's (1969) CI for the mean a bounded random variable is an application of our general methodology using the Kolmogorov-Smirnov CB. We employ standardized Kolmogorov-Smirnov statistics improved along the three common principles in econometrics: the Wald, likelihood-ratio, and Lagrange multiplier principles.

Table 2.7: Choice of ζ_E and ζ_{AD} based on an auxiliary sample of $n_1 = 200$

ζ	C_{E_ζ}	width (in %)	ζ	C_{AD_ζ}	width (in %)
0.0001	20.9925446	13.601	0.0001	4.1608955	9.274
0.0010	6.6547755	5.085	0.0010	2.9710255	5.618
0.0050	3.0160473	3.155	0.0100	2.1509049	4.164
0.0070	2.7526871	3.116	0.0700	1.2028817	3.513
0.0100	2.4304286	3.048	0.1000	1.0494495	3.485
0.0300	1.6107040	2.981	0.3000	0.6431794	3.396
0.035	1.5066083	2.97118	0.4000	0.5642136	3.397
0.04	1.4203946	2.96373	0.4500	0.5325776	3.389
0.043	1.3752673	2.95998	0.5000	0.5053236	3.380
0.045	1.3474672	2.95772	0.5500	0.4832031	3.380
0.047	1.3212876	2.95564	0.6000	0.4636555	3.380
0.0500	1.2847303	2.953	0.6500	0.4463966	3.380
0.055	1.2409542	2.968	0.7000	0.4295599	3.372
0.060	1.2129317	3.003	0.7500	0.4157600	3.373
0.065	1.1867258	3.034	0.8000	0.4025083	3.369
0.070	1.1621481	3.062	0.9000	0.3802611	3.369
0.090	1.0771866	3.152	1	0.3601782	3.360
1	0.3594028	3.323	2	0.2553817	3.347
2	0.2560654	3.338	10	0.1151254	3.351
3	0.2100206	3.348	100	0.0364016	3.347
100	0.0364098	3.347	1000	0.0115271	3.350

Table 2.8: Mexican households in PROGRESA: Confidence intervals for $P_2(Y, z)$ for different types of households' heads
$n = 1000$, $\zeta_{AD} = 0.5$, and $\zeta_E = 0.05$

Table 2.8a: All households

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.331	1.313	0.982
Asymp	0.364	1.261	0.897
Bootstrap_r	0.447	2.428	1.981
Bootstrap	0.455	1.515	1.060
Fishman	0.263	1.840	1.577
Hora-Hora	-1.194	2.818	4.012
Anderson	0.019	2.818	2.799
Eicker	0.038	1.858	1.820
E_ζ	0.151	1.948	1.796
AD	0.212	3.558	3.346
AD_ζ	0.149	2.071	1.922
BJ	0.106	3.185	3.080

Table 2.8b: Households with an educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	-0.046	0.758	0.804
Asymp	0.059	0.953	0.894
Bootstrap_r	0.043	2.441	2.398
Bootstrap	0.147	1.753	1.606
Fishman	0.076	1.616	1.539
Hora-Hora	-0.629	1.641	2.271
Anderson	0.005	1.641	1.637
Eicker	0	1.545	1.545
E_ζ	0.001	1.457	1.456
AD	0.072	4.005	3.933
AD_ζ	0.001	1.594	1.593
BJ	0.083	1.693	1.610

Table 2.8c: Households with a non-educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.574	3.218	2.644
Asymp	0.457	2.582	2.125
Bootstrap_r	0	5.196	4.391
Bootstrap	0.628	4.506	3.877
Fishman	0.308	4.273	3.965
Hora-Hora	-1.144	4.184	5.328
Anderson	0.233	4.184	3.950
Eicker	0	4.140	4.129
E_ζ	0.333	4.546	4.213
AD	0.303	9.417	9.114
AD_ζ	0.362	4.973	4.612
BJ	0.377	4.413	4.035

Table 2.8d: Households with a female head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.274	6.855	6.580
Asymp	0.193	5.215	5.023
Bootstrap_r	1.288	13.938	12.650
Bootstrap	0.907	14.360	13.453
Fishman	0.353	8.899	8.546
Hora-Hora	-1.709	7.117	8.827
Anderson	0.225	7.117	6.893
Eicker	0	7.976	7.976
E_ζ	0.293	8.470	8.177
AD	0.359	20.293	19.933
AD_ζ	0.459	8.871	8.412
BJ	0.423	9.073	8.649

Table 2.8e: Households with a male head

	Confidence Intervals (in %)		
	min	max	width
Asymp_r	0.118	0.848	0.730
Asymp	0.181	0.947	0.765
Bootstrap_r	0.185	1.316	1.131
Bootstrap	0.266	1.349	1.083
Fishman	0.123	1.545	1.421
Hora-Hora	-0.580	1.708	2.288
Anderson	0.040	1.708	1.668
Eicker	0.004	1.519	1.515
E_ζ	0.031	1.608	1.577
AD	0.115	3.457	3.342
AD_ζ	0.028	1.757	1.730
BJ	0.143	1.612	1.469

These statistics have been proposed by Anderson and Darling (1952), Eicker (1979), and Berk and Jones (1979). For further improvement, we use regularized Anderson-Darling and Eicker statistics proposed in Diouf and Dufour (2005a) where the denominators of the statistics are corrected by adding a positive nonzero regularization term $\zeta(x)$. This regularization prevents the denominator of these statistics to become too close to 0 in the tails of the distributions, which would lead to erratic behavior of the statistics. The weighted Kolmogorov-Smirnov statistics yield CBs which width decreases with observations further from the center of the distribution and thus provide thinner CIs for the mean than the Anderson's CI.

Our study focuses on the question of building CIs for the mean of a bounded random variable but our methodology is not as restricting as it appears. Solving the problem for the mean of Y allows one to solve the problem for any moment of Y by replacing the original data by a function of these data such as $exp(Y)$ and Y^{β}. The CIs we propose in this paper then provide valid ones for the mean of the new data, which are CIs for the corresponding moment of the original data. All kinds of transformations can be studied. Continuous ones will be handled using the same intervals as those proposed while for noncontinuous ones, interesting monotonicity properties allow to solve the problem.

We apply these inference methods to the Foster, Greer and Thorbecke (1984) poverty measures. On observing that poverty measures can be interpreted as the expectation of a bounded random variable—a mixture of a continuous bounded variable and a probability mass at the poverty line—we propose that exact nonparametric inference methods for the mean of a bounded random variable be applied to them.

Monte Carlo simulations show that asymptotic and bootstrap CIs can fail to provide reliable inference, even with fairly large samples, e.g. when the distribution presents a high probability of assuming the value zero—which is quite frequent in practice. By contrast, exact inference methods are robust to the underlying distribution and the sample size. The proposed CIs have coverage probability typically larger than the nominal level while remaining informative. The CIs based on the regularized statistics provide the best width among the exact methods.

An illustration using household survey data from Mexico confirms these results. While the width of the asymptotic CI is often unrealistic, the standard bootstrap can fail even in precision, delivering CIs of width ten times larger than those of the exact methods. The study shows that on average, rural households targeted by PROGRESA do not have a very high level of poverty. However, the poverty profile is uneven from a group of households to another. Households with a male head and households with an educated head appear to be more prone to escape poverty than households with a female head or a non educated head. These conclusions provide hints for designing policies to reduce poverty in rural Mexico. Policies aimed at reducing illiteracy, both for children and adults, in these communities can be effective in reducing poverty. Likewise, policies aimed at securing the income of households with a female head by for example (1) securing land ownership for female or (2) improving labor productivity for households with a female head, could help reduce poverty in rural Mexico.

2.11 Appendix 1: Simulated critical points of the statistics

Table A1. Simulated 95^{th} percentile of the distribution of the Kolmogorov-Smirnov based statistics
for $n = 500$ and $N = 1,000,000$ replications

	Critical points			
ζ	E	E_ζ	AD	AD_ζ
0.005	4.80102	3.52808	6.46743	3.18485
0.070	4.80040	2.58235	6.47020	2.57257
0.100	4.79360	2.44064	6.41938	2.42047
0.150	4.80039	2.24627	6.43338	2.22224
0.200	4.79787	2.10995	6.47507	2.10377
0.300	4.79652	1.86659	6.44419	1.86415
0.400	4.79972	1.72506	6.47749	1.71657
0.500	4.80187	1.57861	6.44769	1.57977
0.750	4.79959	1.36796	6.43938	1.36707
1.000	4.79817	1.22237	6.45349	1.22005
10.000	4.79646	0.42397	6.45115	0.42408
100.000	4.79009	0.13478	6.45162	0.13479
1,000.000	4.79356	0.04240	6.45573	0.04240
1,000,000.000	4.79652	0.00135	6.43729	0.00135

Table A2. Simulated 95^{th} percentile of the distribution of the statistics for $\zeta = 0.07$ and $N = 3,000,000$ replications

		Critical points of the statistics				
	n	50	100	200	500	1000
	$p = 0$	0.18830	0.13403	0.09513	0.06038	0.04280
KS	$p = 0.89$	0.10497	0.07000	0.05000	0.03186	0.02240
	$p = 0.989$	0.04900	0.02900	0.01900	0.01100	0.00726
	$p = 0$	4.52172	4.65730	4.74005	4.79734	4.82078
E	$p = 0.89$	3.53697	3.89181	4.03122	4.09756	4.11948
	$p = 0.989$	1.45895	1.47990	1.57515	3.63403	3.96574
	$p = 0$	2.79486	2.67435	2.61896	2.59007	2.58192
E_ζ	$p = 0.89$	1.94266	1.87663	1.84708	1.84184	1.84028
	$p = 0.989$	0.97456	0.88081	0.85356	0.81307	0.81568
	$p = 0$	6.44400	6.46184	6.46283	6.44509	6.45195
AD	$p = 0.89$	4.71387	4.72593	4.73024	4.72705	4.73197
	$p = 0.989$	4.67778	4.69788	4.71145	4.72273	4.72018
	$p = 0$	2.56615	2.56437	2.56485	2.56812	2.57179
AD_ζ	$p = 0.89$	1.89825	1.83230	1.83339	1.83460	1.83706
	$p = 0.989$	1.21833	1.01972	0.94482	0.86489	0.82060
	$p = 0$	0.10414	0.05374	0.02760	0.01138	0.00581
BJ	$p = 0.89$	0.07945	0.04193	0.02194	0.00920	0.00474
	$p = 0.989$	0.05467	0.03157	0.01680	0.00783	0.00415

Table A3. Simulated critical points of order 95 percent for the distribution of the statistics with the optimal values of ζ for the E_ζ and AD_ζ
$N = 3,000,000$ replications

Table A3a: Critical points for the whole sample using $\zeta_{AD} = 0.004$ and $\zeta_E = 0.0062$

	Characteristics of the households' head				
	Female	Male	Educated	Non-educated	All
n_r	2054	17431	13479	6006	19485
n	2168	18317	14178	6307	20485
p	0.9493	0.9166	0.9827	0.9535	0.9737
KS_n	0.493389384	0.338851519	0.291594774	0.473502205	0.358909643
E	4.124905337	4.145787028	4.145149225	4.139179162	4.14495824
E_ζ	2.501772599	2.212303346	2.104179478	2.439923824	2.252535994
AD	4.731275042	4.741005986	4.72736826	4.753315289	4.73462176
AD_ζ	2.541923946	2.346763514	2.250207895	2.526654076	2.379458442
BJ	0.002181669	0.000271334	0.000343604	0.000778607	0.000244325

Table A3b: Critical points for the subsample of n=500 using $\zeta_{AD} = 0.45$ and $\zeta_E = 0.039$

	Characteristics of the households' head				
	Female	Male	Educated	Non-educated	All
n_r	41	359	282	118	400
n	53	447	359	141	500
p	0.9493	0.9166	0.9827	0.9535	0.9737
KS_n	0.592422376	0.347008748	0.305547568	0.467941715	0.354681929
E	1.749763640	3.958754864	3.682018883	3.664690633	3.992906840
E_ζ	1.643870433	1.395120081	1.194585231	1.788845505	1.460821026
AD	4.701031976	4.739227068	4.724159189	4.717678892	4.722350336
AD_ζ	0.867694822	0.506517449	0.446277892	0.721830287	0.540496511
BJ	0.062177097	0.009300228	0.010994130	0.028165890	0.008466719

Table A3c: Critical points for the subsample of n=1000 using $\zeta_{AD} = 0.5$ and $\zeta_E = 0.05$

	Characteristics of the households' head				
	Female	Male	Educated	Non-educated	All
n_r	88	712	558	242	800
n	116	884	698	302	1000
p	0.9493	0.9166	0.9827	0.9535	0.9737
KS_n	0.475329978	0.340136976	0.299950534	0.462943395	0.348564201
E	3.626525226	4.059987327	4.001971001	4.005076739	3.880275064
E_ζ	1.596639583	1.283268215	1.123470499	1.585207457	1.322617712
AD	4.720719420	4.733591949	4.723195428	4.727067310	4.721280048
AD_ζ	0.653461255	0.472736968	0.407026067	0.647307799	0.489618262
BJ	0.034128232	0.004908066	0.005997227	0.014066270	0.013412891

2.12 Appendix 2: Proofs of theorems and propositions

PROOF OF PROPOSITION 3.2. The proof follows by projection. Considering the CBs of $F(x)$, $C_F(\alpha)$, and the property of the functional Γ :

$$\begin{aligned} G_n^L(x) &\leq F_0(x)\forall x \Longrightarrow \Gamma[G_n^L] \leq \Gamma[F_0], \\ F_0(x) &\leq G_n^U(x)\ \forall x \Longrightarrow \Gamma[F_0] \leq \Gamma[G_n^U]. \end{aligned}$$

If $C_F(\alpha)$ exists then with probability $1-\alpha$, $\Gamma[G_n^L] \leq \Gamma[F_0] \leq \Gamma[G_n^U]$, $\forall\ F_0 \in C_F(\alpha)$. Moreover, all $\Gamma[F_0]$ such that $F_0 \in C_F(\alpha)$ belong to the CI of $\Gamma[F]$.

PROOF OF PROPOSITION 3.3. This corollary is an application of Proposition 3.2 with the functional $\Gamma[F(x)] = \int_a^b x\ dF(x) \equiv \mu$. Computing $\Gamma[F(x)]$ using integration by parts provides

$$\mu = bF(b) - aF(a) - \int_a^b F(x)dx = b - aF(a) - \int_a^b F(x)dx$$

where the second equality follows because $F(b) = 1$ for any distribution function $F(x) \in \mathcal{F}_{[a,b]}$, by definition. Then if $C_F(\alpha)$ exists, $b - aG_n^U(a) - \int_a^b G_n^U(x)dx \leq \mu \leq b - aG_n^L(a) - \int_a^b G_n^L(x)dx$ with probability $1-\alpha$, i.e.

$$C_\mu(\alpha) = \left\{\mu_0 \in \mathbb{R} : b - aG_n^U(a) - \int_a^b G_n^U(x)dx \leq \mu_0 \leq b - aG_n^L(a) - \int_a^b G_n^L(x)dx\right\}$$

is a CI with level $1-\alpha$ for μ. Given that distribution functions always have values between 0 and 1,

$$\widetilde{C}_F(\alpha) = \left\{F_0 \in \mathcal{L} : \widetilde{F}_n^L(x) \leq F_0(x) \leq \widetilde{F}_n^U(x),\ \forall x\right\}$$

represents a CB with level $1-\alpha$ for $F(x)$.Applying the same procedure as for $C_F(\alpha)$ provide the following CI with level $1-\alpha$ for μ :

$$\widetilde{C}_\mu(\alpha) = \left\{\mu_0 \in \mathbb{R} : b - a\widetilde{F}_n^U(a) - \int_a^b \widetilde{F}_n^U(x)dx \leq \mu_0 \leq b - a\widetilde{F}_n^L(a) - \int_a^b \widetilde{F}_n^L(x)dx]\right\}.$$

$\widetilde{C}_F$ has the same level as C_F but is thinner for each observation, hence $\widetilde{C}_\mu(\alpha)$ is better than $C_\mu(\alpha)$.

PROOF OF PROPOSITION 3.4. Let $\widetilde{F}_n^L(x)$ and $\widetilde{F}_n^U(x)$ be such that

$$\forall x \qquad \widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\} \qquad \textit{and} \qquad \widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}.$$

By definition, all distribution functions $F(x)$ satisfy the inequality $0 \leq F(x) \leq 1$, $\forall x$. Then, $\widetilde{F}_n^L(x)$ and $\widetilde{F}_n^U(x)$ can be considered as the effective (or restrictive) part of the CB $C_F(\alpha)$. Corollary 3.3 then provides the following bounds for μ :

$$\mu_L = b - a\widetilde{F}_n^U(a) - \int_a^b \widetilde{F}_n^U(x)dx \leq \mu \leq b - a\widetilde{F}_n^L(a) - \int_a^b \widetilde{F}_n^L(x)dx = \mu_U. \tag{2.15}$$

Using the lower bound:

$$\begin{aligned}
\mu_L &= b - a\widetilde{F}_n^U(a) - \int_a^b \widetilde{F}_n^U(x)\ dx = b - a\widetilde{F}_n^U(a) - \sum_{k=0}^{n}[X_{(k+1)} - X_{(k)}]\widetilde{F}_n^U(X_{(k)}) \\
&= b - a\widetilde{F}_n^U(a) + \sum_{k=0}^{n} X_{(k)}\widetilde{F}_n^U(X_{(k)}) - \sum_{k=1}^{n+1} X_{(k)}\widetilde{F}_n^U(X_{(k-1)}) \\
&= b - a\widetilde{F}_n^U(a) + X_{(0)}\widetilde{F}_n^U(X_{(0)}) - X_{(n+1)}\widetilde{F}_n^U(X_{(n)}) + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)} \\
&= [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)}.
\end{aligned}$$

Let $k_u \in \{1, \cdots, n\}$ such that

$$\widetilde{F}_n^U(x) = \begin{cases} G_n^U(x),\ \forall x \leq X_{(k_u)} \\ \quad 1,\ \forall x > X_{(k_u)}. \end{cases}$$

We can rewrite

$$\mu_L = [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{k_u+1} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)}$$

Moreover, if $k_u < n$ then

$$\begin{aligned}\mu_L &= [1 - G_n^U(X_{(n)})]X_{(n+1)} + \left[1 - G_n^U(X_{(k_u)})\right] X_{(k_u)} \\ &\quad + \sum_{k=1}^{k_u} \left[G_n^U(X_{(k)}) - G_n^U(X_{(k-1)})\right] X_{(k)}\end{aligned}$$

Similarly for the upper bound,

$$\begin{aligned}\mu_U &= b - a\widetilde{F}_n^L(a) - \int_a^b \widetilde{F}_n^L(x)\, dx = b - a\widetilde{F}_n^L(a) - \sum_{k=0}^{n}[X_{(k+1)} - X_{(k)}]\widetilde{F}_n^L(X_{(k)}) \\ &= b - a\widetilde{F}_n^L(a) + \sum_{k=0}^{n} X_{(k)}\widetilde{F}_n^L(X_{(k)}) - \sum_{k=1}^{n+1} X_{(k)}\widetilde{F}_n^L(X_{(k-1)}) \\ &= b - a\widetilde{F}_n^L(a) + \widetilde{F}_n^L(X_{(0)})X_{(0)} - \widetilde{F}_n^L(X_{(n)}) * X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)} \\ &= [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)}.\end{aligned}$$

Let $k_l \in \{1, \cdots, n\}$ such that

$$\widetilde{F}_n^L(x) = \begin{cases} G_n^L(x),\ \forall x \geq X_{(k_l)}, \\ 0,\ \forall x < X_{(k_l)}. \end{cases}$$

μ_U simplifies to

$$\mu_U = [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=k_l}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)}$$

Moreover, if $k_l < n$ then we can rewrite

$$\mu_U = G_n^L(X_{(k_l)}) * X_{(k_l)} + [1 - G_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=k_l+1}^{n} \left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right] X_{(k)}$$

Proof of the expressions of the Kolmogorov-Smirnov projection-based confidence intervals These two CIs are obtained by applying Propositions 3.3 and 3.4. on the KS confidence band for distribution functions. Let X be a random variable with a continuous distribution function $F(x)$ whose support is $[a, b]$. Assume that n i.i.d. observations $X_n, \cdots, X_n$ on X are available and let $X_{(0)} = a \leq X_{(1)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = b$ be the corresponding order statistics.

First, let's show that the generalization of the Hora-Hora CI for continuous random variables bounded on a finite interval $[a, b]$ is a projection of the KS confidence band where the constraint that $0 \leq F(x) \leq 1$ is not accounted for. The mean μ of X is:

$$\mu = \int_a^b x \; dF(x) = b - \int_a^b F(x) \; dx$$

where the last equality follows on integration by part. Let $c_{KS}(\alpha)$ be the $(1 - \alpha)^{th}$ percentile of the Kolmogorov-Smirnov (KS) statistic. The KS confidence band for $F(x)$ with level $1 - \alpha$ is

$$F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq F(x) \leq F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}}, \; \forall x$$

where $F_n(x)$ is the empirical distribution function of the sample as defined by equation (*2.1*). Taking the integral $\int_a^b$, we get

$$\int_a^b F_n(x) \; dx - (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}} \leq \int_a^b F(x) \; dx \leq \int_a^b F_n(x) \; dx + (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}}$$

$\Leftrightarrow$

$$b - \int_a^b F_n(x)\ dx - (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}} \leq \mu \leq b - \int_a^b F_n(x)\ dx + (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}}.$$

Further,

$$\begin{aligned}
\int_a^b F_n(x)\ dx &= \frac{1}{n}[X_{(2)} - X_{(1)}] + \frac{2}{n}[X_{(3)} - X_{(2)}] + \cdots + \frac{n-1}{n}[X_{(n)} - X_{(n-1)}] + [b - X_{(n)}] \\
&= \sum_{k=1}^{n-1}\frac{k}{n}[X_{(k+1)} - X_{(k)}] + [b - X_{(n)}] \\
&= -\frac{1}{n}X_{(1)} + \sum_{k=2}^{n-1}[\frac{k-1}{n}X_{(k)} - \frac{k}{n}X_{(k)}] + \frac{n-1}{n}X_{(n)} + [b - X_{(n)}] \\
&= -\frac{1}{n}X_{(1)} - \frac{1}{n}\sum_{k=2}^{n-1}X_{(k)} - \frac{1}{n}X_{(n)} + b = -\frac{1}{n}\sum_{k=1}^{n}X_{(k)} + b
\end{aligned}$$

hence,

$$b - \int_a^b F_n(x)\ dx = \frac{1}{n}\sum_{k=1}^{n}X_{(k)} = \overline{X}$$

and

$$\overline{X} - (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}} \leq \mu \leq \overline{X} + (b-a)\frac{c_{KS}(\alpha)}{\sqrt{n}}.$$

Setting $b = 1$ and $a = 0$ yields the Hora-Hora CI for the mean of a continuous random variable X, bounded on $[0, 1]$.

Second, let's show that the Anderson CI is a projection of the KS confidence band where all constraints about distribution functions are exploited. Let X be a random variable with an unknown continuous cumulative distribution function $F(x)$ with finite support $[a, b]$ ($a < b$, $F(a) = 0$ and $F(b) = 1$). Denote $X_{(1)} \leq \cdots \leq X_{(n)}$ the order statistics of the sample, $X_{(0)} = a$, and $X_{(n+1)} = b$. The mean of X is :

$$\mu = \int_a^b x\ dF(x) = b - \int_a^b F(x)\ dx$$

and the Kolmogorov-Smirnov (KS) confidence band for distribution functions by:

$$P[F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \ \leq F(x) \leq F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}}, \ \forall x] = 1 - \alpha$$

where β and γ are the adequate percentiles of the KS distribution. Then, projecting the KS confidence band yields:

$$b - \int_a^b \min\{F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}}, 1\}dx \leq \mu \leq b - \int_a^b \max\{F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}}, 0\}dx$$

Let $r = I\ [n\frac{c_{KS}(\alpha)}{\sqrt{n}}]$ and $s = |n\frac{c_{KS}(\alpha)}{\sqrt{n}}|$. Computing the left-hand side:

$$\begin{aligned}
\mu &\geq b - [\int_a^{X_{(n-s)}} F_n(x) + \frac{c_{KS}(\alpha)}{\sqrt{n}}\ dx + [b - X_{(n-s)}] * 1] \\
&\geq X_{(n-s)} - \frac{c_{KS}(\alpha)}{\sqrt{n}}[X_{(n-s)} - a] - \frac{1}{n}\sum_{k=1}^{n-s}(k-1)[X_{(k)} - X_{(k-1)}] \\
&\geq X_{(n-s)} - \frac{c_{KS}(\alpha)}{\sqrt{n}}[X_{(n-s)} - a] - \frac{1}{n}\sum_{k=1}^{n-s}(k-1)X_{(k)} + \frac{1}{n}\sum_{k=0}^{n-s-1} kX_{(k)} \\
&\geq X_{(n-s)} - \frac{c_{KS}(\alpha)}{\sqrt{n}}[X_{(n-s)} - a] + \frac{1}{n}\sum_{k=1}^{n-s-1} X_{(k)} + \frac{1}{n}X_{(n-s)} - \frac{1}{n}(n-s)X_{(n-s)} \\
&\geq \frac{1}{n}[\sum_{k=1}^{n-s-1} X_{(k)} + (n+1-(n-s))X_{(n-s)}] - \frac{c_{KS}(\alpha)}{\sqrt{n}}[X_{(n-s)} - a] \\
&\geq \frac{1}{n}[\sum_{k=1}^{n-s-1} X_{(k)} + (s+1)X_{(n-s)}] - \frac{c_{KS}(\alpha)}{\sqrt{n}}[X_{(n-s)} - a]
\end{aligned}$$

which is the lower bound of of equation *(2.7)*.

Similarly computing the right-hand side:

$$\begin{aligned}
\mu &\leq b - 0 * [X_{(r+1)} - a] - \int_{X_{(r+1)}}^b F_n(x) - \frac{c_{KS}(\alpha)}{\sqrt{n}}\ dx \\
&\leq b + \frac{c_{KS}(\alpha)}{\sqrt{n}}[b - X_{(r+1)}] - \frac{1}{n}\sum_{k=r+2}^{n+1}(k-1)[X_{(k)} - X_{(k-1)}
\end{aligned}$$

$$
\begin{aligned}
&\leq b + \frac{c_{KS}(\alpha)}{\sqrt{n}}[b - X_{(r+1)]} - \frac{1}{n}\sum_{k=r+2}^{n+1}(k-1)X_{(k)} + \frac{1}{n}\sum_{k=r+1}^{n} kX_{(k)} \\
&\leq b + \frac{c_{KS}(\alpha)}{\sqrt{n}}[b - X_{(r+1)}] - \frac{1}{n}\sum_{k=r+2}^{n} X_{(k)} + \frac{b}{n} - \frac{(r+1)X_{(r+1)}}{n} - \frac{n+1}{n}b \\
&\leq \frac{1}{n}[(r+1)X_{(r+1)} + \sum_{k=r+2}^{n} X_{(k)}] + \frac{c_{KS}(\alpha)}{\sqrt{n}}[b - X_{(r+1)}]
\end{aligned}
$$

which corresponds to the upper bound of equation *(2.7)*.

PROOF OF PROPOSITION 4.1. Proposition 4.1 is a direct application of the general expression of the CI for the mean of a continuous bounded random variable using the Anderson-Darling CB for distribution functions.

PROOF OF COROLLARY 4.1BIS. Developing the expression of the Anderson-Darling CI for the mean of a bounded random variable yields Corollary 4.1bis. This CI is:

$$
\begin{aligned}
C_{\mu}^{AD}(\alpha) = \Big\{ \mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U \text{ where } \mu_L &= [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} \\
&+ \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)}, \\
\mu_U &= [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right]X_{(k)}, \Big\}
\end{aligned}
$$

$\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$,

$$
G_n^L(x) = \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} - \sqrt{\Delta(x)}}{2(1 + \frac{c_{AD}^2(\alpha)}{n})},
$$

$$
G_n^U(x) = \frac{2F_n(x) + \frac{c_{AD}^2(\alpha)}{n} + \sqrt{\Delta(x)}}{2(1 + \frac{c_{AD}^2(\alpha)}{n})},.
$$

$\Delta(x) = \left[2F_n(x) + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4F_n^2(x)\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right]$, and $c_{AD}(\alpha)$ satisfies $Pr(AD \leq c_{AD}(\alpha)) \geq 1 - \alpha$. It is easy to prove that $G_n^L(X_{(0)}) = 0$, and $F_n^U(X_{(0)}) = \frac{c_{AD}^2(\alpha)}{n}\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right]^{-1} > 0$. Moreover,

$$
\begin{aligned}
G_n^L(X_{(n)}) &= \frac{2+\frac{c_{AD}^2(\alpha)}{n}-\sqrt{\left[2+\frac{c_{AD}^2(\alpha)}{n}\right]^2-4\left[1+\frac{c_{AD}^2(\alpha)}{n}\right]}}{2(1+\frac{c_{AD}^2(\alpha)}{n})} \\
&= \frac{2+\frac{c_{AD}^2(\alpha)}{n}-\sqrt{4+\frac{c_{AD}^4(\alpha)}{n^2}+\frac{4c_{AD}^2(\alpha)}{n}-4-\frac{4c_{AD}^2(\alpha)}{n}}}{2(1+\frac{c_{AD}^2(\alpha)}{n})} \\
&= \left(1+\frac{c_{AD}^2(\alpha)}{n}\right)^{-1} < 1
\end{aligned}
$$

and, similarly, $G_n^U(X_{(n)}) = 1$. Thus, the effective part of the Anderson-Darling CB is $\widetilde{F}_n^L(X_{(k)}) = \max\{G_n^L(X_{(k)}), 0\} = G_n^L(X_{(k)})$ and $\widetilde{F}_n^U(X_{(k)}) = \min\{G_n^U(X_{(k)}), 1\} = G_n^U(X_{(k)})$, $\forall k = 1, \ldots, n+1$ and

$$
\begin{aligned}
\mu_L &= [1-1]\, X_{(n+1)} + +\sum_{k=1}^{n}\left[G_n^U(X_{(k)}) - G_n^U(X_{(k-1)})\right] X_{(k)} \\
&= \sum_{k=1}^{n}\left[\frac{2\frac{k}{n}+\frac{c_{AD}^2(\alpha)}{n}+\sqrt{\Delta(X_{(k)})}}{2(1+\frac{c_{AD}^2(\alpha)}{n})} - \frac{2\frac{k-1}{n}+\frac{c_{AD}^2(\alpha)}{n}+\sqrt{\Delta(X_{(k-1)})}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}\right] X_{(k)} \\
&= \sum_{k=1}^{n}\left[\frac{\frac{2}{n}+\sqrt{\Delta(k)}-\sqrt{\Delta(k-1)}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}\right] X_{(k)}
\end{aligned}
$$

$$
\begin{aligned}
\mu_U &= [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)} \\
&= [1-\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)^{-1}]X_{(n+1)} + \sum_{k=1}^{n}\left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right] X_{(k)} \\
&= [1-\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)^{-1}]X_{(n+1)} \\
&\quad +\sum_{k=1}^{n}\left[\frac{2\frac{k}{n}+\frac{c_{AD}^2(\alpha)}{n}-\sqrt{\Delta(X_{(k)})}}{2(1+\frac{c_{AD}^2(\alpha)}{n})} - \frac{2\frac{k-1}{n}+\frac{c_{AD}^2(\alpha)}{n}-\sqrt{\Delta(X_{(k-1)})}}{2(1+\frac{c_{AD}^2(\alpha)}{n})}\right] X_{(k)} \\
&= [1-\left(1+\frac{c_{AD}^2(\alpha)}{n}\right)^{-1}]X_{(n+1)} + \sum_{k=1}^{n}\left[\frac{\frac{2}{n}-\left(\sqrt{\Delta(k)}-\sqrt{\Delta(k-1)}\right)}{2(1+\frac{c_{AD}^2(\alpha)}{n})}\right] X_{(k)}
\end{aligned}
$$

PROOF OF PROPOSITION 4.2. Proposition 4.2 is a direct application of the general expression of the CI for the mean of a continuous bounded random variable using the Eicker CB for distributions functions.

PROOF OF COROLLARY 4.2BIS. Developing the expression of the Eicker CI for the mean of a bounded random variable yields Corollary 4.2bis. The latter is:

$$\begin{aligned} C_{\mu}^{E}(\alpha) = \Big\{ & \mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U \text{ where } \mu_L = [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} \\ & + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)}) \right] X_{(k)}, \\ & \mu_U = [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)}) \right] X_{(k)}, \Big\} \end{aligned}$$

$\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$,

$$G_n^L(x) = \begin{cases} F_n(x) - \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1 - F_n(x)]^{1/2} & \forall\ x \text{ such that } F_n(x) \notin \{0,1\} \\ 0 & \forall\ x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

$$G_n^U(x) = \begin{cases} F_n(x) + \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(x)[1 - F_n(x)]^{1/2} & \forall\ x \text{ such that } F_n(x) \notin \{0,1\} \\ 1 & \forall\ x \text{ such that } F_n(x) \in \{0,1\}, \end{cases}$$

and $c_E(\alpha)$ satisfies $Pr(E \leq c_E(\alpha)) \geq 1 - \alpha$.

Using $G_n^L(x)$:

$$\begin{aligned} G_n^L(X_{(k)}) &\geq 0 \\ &\Leftrightarrow \frac{k}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \geq 0 \\ &\Leftrightarrow k \geq c_E(\alpha)\sqrt{n}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \\ &\Leftrightarrow k^2 \geq c_E^2(\alpha)n\frac{k}{n}\left(1-\frac{k}{n}\right) \\ &\Leftrightarrow k^2 \geq c_E^2(\alpha)k - c_E^2(\alpha)\frac{k^2}{n} \\ &\Leftrightarrow \left[n+c_E^2(\alpha)\right]k^2 - nc_E^2(\alpha)k \geq 0 \\ &\Leftrightarrow k \geq \frac{nc_E^2(\alpha)}{n+c_E^2(\alpha)} = K_E^L \end{aligned}$$

Thus, $G_n^L(X_{(k)}) \geq 0\ \forall k = k_E^L, ..., n-1$ where $k_E^L = I[nc_E^2(\alpha)\left(n+c_E^2(\alpha)\right)^{-1}] + 1$ where $I[\kappa]$ is the integer part of κ. The effective lower bound of the Eicker CB is

$$\widetilde{F}_n^L(X_{(k)}) = \begin{cases} \frac{k}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} & \forall k = k_E^L, \ldots,\ n-1 \\ 0 & \forall k = 0, \ldots,\ k_E^L - 1,\ n,\ n+1. \end{cases}$$

where $k_E^L = I[nc_E^2(\alpha)\left(n+c_E^2(\alpha)\right)^{-1}] + 1$.

Similarly, for $G_n^U(x)$:

$$\begin{aligned} G_n^U(X_{(k)}) &\leq 1 \\ &\Leftrightarrow \frac{k}{n} + \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2} \leq 1 \\ &\Leftrightarrow \frac{c_E(\alpha)}{n^{3/2}}[k]^{1/2}[n-k]^{1/2} \leq [1-\frac{k}{n}] \\ &\Leftrightarrow c_E^2(\alpha)k[n-k] \leq n(n^2-2nk+k^2) \\ &\Leftrightarrow \left(n+c_E^2(\alpha)\right)k^2 - n\left(2n+c_E^2(\alpha)\right)k + n^3 \geq 0. \end{aligned}$$

This is the case for all $k \leq \vartheta_1$ and $k \geq \vartheta_2$ such that $\vartheta_1 = \frac{n\left(2n+c_E^2(\alpha)\right)-\sqrt{\Lambda}}{2\left(n+c_E^2(\alpha)\right)}$, $\vartheta_2 =$

$\frac{n\left(2n+c_E^2(\alpha)\right)+\sqrt{\Lambda}}{2\left(n+c_E^2(\alpha)\right)}$,

$$\begin{aligned}\Lambda &= n^2\left(2n+c_E^2(\alpha)\right)^2-4n^3\left(n+c_E^2(\alpha)\right)\\ &= n^2\left[4n^2+4nc_E^2(\alpha)+c_E^4(\alpha)\right]-4n^4-4n^3c_E^2(\alpha)\\ &= n^2c_E^4(\alpha).\end{aligned}$$

Developing yields $\vartheta_1 = \frac{n^2}{\left(n+c_E^2(\alpha)\right)}$ and $\vartheta_2 = \frac{2n^2+2nc_E^2(\alpha)}{2\left(n+c_E^2(\alpha)\right)} = n$. Then, $F_n^U(X_{(k)}) \leq 1$ for $k \leq \vartheta_1$ or equivalently $k = 1, ..., k_E^U$ where $k_E^U = I[\frac{n^2}{\left(n+c_E^2(\alpha)\right)}]$ is the integer part of ϑ_1. Hence, the effective upper bound of the Eicker CB is

$$\widetilde{F}_n^U(X_{(k)}) = \begin{cases} \frac{k}{n}+\frac{c_E(\alpha)}{\sqrt{n}}[\frac{k}{n}]^{1/2}[1-\frac{k}{n}]^{1/2}\ \forall k = 1, ..., k_E^U \\ \quad 1\ \forall k = 0,\ k_E^U+1, \ldots,\ n,\ n+1. \end{cases}$$

where $k_E^U = I[\frac{n^2}{\left(n+c_E^2(\alpha)\right)}]$. It follows that

$$\mu_L = [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{k_E^U+1}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right]X_{(k)}$$

$$\begin{aligned}&= [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}_n^U(X_{(k_E^U+1)}) - \widetilde{F}_n^U(X_{(k_E^U)})\right]X_{(k_E^U+1)}\\ &+\sum_{k=1}^{k_E^U}\left[F_n^U(X_{(k)}) - F_n^U(X_{(k-1)})\right]X_{(k)}.\end{aligned}$$

Our former results show that $\widetilde{F}_n^U(X_{(n)}) = 1$, $\widetilde{F}_n^U(X_{(k_E^U+1)}) = 1$, $\widetilde{F}_n^U(X_{(k_E^U)}) = \frac{k_E^U}{n} + \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k_E^U}{n}]^{1/2}[1-\frac{k_E^U}{n}]^{1/2}$. Hence,

$$\begin{aligned}&\mu_L = [1-1]X_{(n+1)} + \left[1-\frac{k_E^U}{n}-\frac{c_E(\alpha)}{\sqrt{n}}[\frac{k_E^U}{n}]^{1/2}[1-\frac{k_E^U}{n}]^{1/2}\right]X_{(k_E^U+1)}\\ &+\sum_{k=1}^{k_E^U}\left[\frac{k}{n}+\frac{c_E(\alpha)}{\sqrt{n}}\left(\frac{k}{n}\right)^{1/2}\left(1-\frac{k}{n}\right)^{1/2}-\frac{k-1}{n}-\frac{c_E(\alpha)}{\sqrt{n}}\left(\frac{k-1}{n}\right)^{1/2}\left(1-\frac{k-1}{n}\right)^{1/2}\right]X_{(k)}\end{aligned}$$

$$\mu_L = \left[1 - \frac{k_E^U}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k_E^U}{n}]^{1/2}[1 - \frac{k_E^U}{n}]^{1/2}\right] X_{(k_E^U+1)}$$
$$+ \sum_{k=1}^{k_E^U} \left[\frac{1}{n} + \frac{c_E(\alpha)}{\sqrt{n}} \left(\sqrt{\frac{k}{n}\left(1 - \frac{k}{n}\right)} - \sqrt{\frac{k-1}{n}\left(1 - \frac{k-1}{n}\right)}\right)\right] X_{(k)}.$$

Similarly for the upper bound:

$$\mu_U = [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=k_E^L}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)}$$

$$= [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}_n^L(X_{(k_E^L)}) - \widetilde{F}_n^L(X_{(k_E^L-1)})\right] X_{(k_E^L)}$$
$$+ \sum_{k=k_E^L+1}^{n} \left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right] X_{(k)}.$$

Our former results show that $\widetilde{F}_n^L(X_{(n)}) = 0$, $\widetilde{F}_n^L(X_{(k_E^L-1)}) = 0$, $\widetilde{F}_n^L(X_{(k_E^L)}) = \frac{k_E^L}{n} - \frac{c_E(\alpha)}{\sqrt{n}}[\frac{k_E^L}{n}]^{1/2}[1 - \frac{k_E^L}{n}]^{1/2}$. Hence,

$$\mu_U = [1 - 0]X_{(n+1)} + \left[\frac{k_E^L}{n} - \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k_E^L}{n}\right)^{1/2} \left(1 - \frac{k_E^L}{n}\right)^{1/2} - 0\right] X_{(k_E^L)}$$
$$+ \sum_{k=k_E^L+1}^{n} \left[\frac{k}{n} - \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k}{n}\right)^{1/2} \left(1 - \frac{k}{n}\right)^{1/2} - \frac{k-1}{n} + \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k-1}{n}\right)^{1/2} \left(1 - \frac{k-1}{n}\right)^{1/2}\right] X_{(k)}$$

$$\mu_U = \left[\frac{k_E^L}{n} - \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k_E^L}{n}\right)^{1/2} \left(1 - \frac{k_E^L}{n}\right)^{1/2} - 0\right] X_{(k_E^L)} + X_{(n+1)}$$
$$+ \sum_{k=k_E^L+1}^{n} \left[\frac{k}{n} - \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k}{n}\right)^{1/2} \left(1 - \frac{k}{n}\right)^{1/2} - \frac{k-1}{n} + \frac{c_E(\alpha)}{\sqrt{n}} \left(\frac{k-1}{n}\right)^{1/2} \left(1 - \frac{k-1}{n}\right)^{1/2}\right] X_{(k)}$$

$$\mu_U = \left[\frac{k_E^L}{n} - \frac{c_E(\alpha)}{\sqrt{n}}\left(\frac{k_E^L}{n}\right)^{1/2}\left(1-\frac{k_E^L}{n}\right)^{1/2}\right] X_{(k_E^L)} + X_{(n+1)} + \sum_{k=k_E^L+1}^{n}\left[\frac{1}{n} - \frac{c_E(\alpha)}{\sqrt{n}}\left(\sqrt{\frac{k}{n}\left(1-\frac{k}{n}\right)} - \sqrt{\frac{k-1}{n}\left(1-\frac{k-1}{n}\right)}\right)\right] X_{(k)}.$$

PROOF OF PROPOSITION 4.3. Proposition 4.3 is a direct application of the general expression of the CI for the mean of a continuous bounded random variable on the case of the ζ-regularized Anderson-Darling CI.

PROOF OF COROLLARY 4.3BIS. Developing the expression of the ζ-regularized Anderson-Darling CI for the mean of a bounded random variable yields Corollary 4.3bis. The latter is:

$$C_\mu^{AD_\zeta}(\alpha) = \Big\{\mu_0 \in \mathbb{R} : \mu_L \le \mu_0 \le \mu_U \text{ where } \mu_L = [1-\widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)}$$
$$\text{and } \mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)}\Big\}$$

for $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$,

$$G_n^L(x) = \frac{2F_n(x) + \frac{c_{AD_\zeta}^2(\alpha)}{n} - \sqrt{\Delta(x)}}{2\left(1+\frac{c_{AD_\zeta}^2(\alpha)}{n}\right)},$$

$$G_n^U(x) = \frac{2F_n(x) + \frac{c_{AD_\zeta}^2(\alpha)}{n} + \sqrt{\Delta(x)}}{2\left(1+\frac{c_{AD_\zeta}^2(\alpha)}{n}\right)},.$$

$\Delta(x) = \left[2F_n(x) + \frac{c_{AD_\zeta}^2(\alpha)}{n}\right]^2 - 4\left[1+\frac{c_{AD_\zeta}^2(\alpha)}{n}\right]\left(F_n^2(x) - \frac{\zeta c_{AD_\zeta}^2(\alpha)}{n}\right)$, and $c_{AD_\zeta}(\alpha)$ satisfies $\Pr[AD_\zeta \le c_{AD_\zeta}(\alpha)] \ge 1-\alpha$..

Using $G_n^L(x)$:

$$G_n^L(X_{(k)}) \ge 0$$

$$\Leftrightarrow 2\frac{k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(k)}\geq 0$$

$$\Leftrightarrow \left(2\frac{k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)^2 \geq \left[2\frac{k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(\frac{k^2}{n^2}-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$$

$$\Leftrightarrow \left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(\frac{k^2}{n^2}-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)\geq 0$$

$$\Longleftrightarrow \frac{k^2}{n^2}-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\geq 0$$

$$\Longleftrightarrow k^2 \geq \zeta c^2_{AD_\zeta}(\alpha)n$$

$$\Rightarrow k \geq n^{1/2}\zeta^{1/2}c_{AD_\zeta}(\alpha) \text{ or } k \leq -n^{1/2}\zeta^{1/2}c_{AD_\zeta}(\alpha).$$

Let $\kappa_0 = n^{1/2}\zeta^{1/2}c_{AD_\zeta}(\alpha)$. Given that k is always positive, $F^L_n(X_{(k)}) \geq 0$ for $k \geq \kappa_0$ or equivalently for $k = k^L_{AD_\zeta}, ..., n$ where $k^L_{AD_\zeta} = I[\kappa_0]+1$ and $I[\kappa_0] \equiv$the integer part of κ_0. Hence,

$$\widetilde{F}^L_n(x) = \max\{G^L_n(x), 0\} = \begin{cases} G^L_n(x)\ \forall x \geq X_{(k^L_{AD_\zeta})} \\ 0\ \forall x < X_{(k^L_{AD_\zeta})} \end{cases} \quad \text{where } k^L_{AD_\zeta} = I[\kappa_0]+1.$$

Using $G^U_n(x)$:

$$G^U_n(X_{(k)}) \leq 1$$

$$\Leftrightarrow 2\frac{k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}+\sqrt{\Delta(k)} \leq 2\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)$$

$$\Leftrightarrow \left[2\frac{k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(\frac{k^2}{n^2}-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$$

$$\leq \left[2\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)-2\frac{k}{n}-\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2$$

$$\Leftrightarrow 4\frac{k^2}{n^2}+\frac{c^4_{AD_\zeta}(\alpha)}{n^2}+4\frac{c^2_{AD_\zeta}(\alpha)}{n^2}k-\frac{4}{n^2}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k^2+\frac{4}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\zeta c^2_{AD_\zeta}(\alpha)$$
$$\leq 4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)^2+\frac{4}{n^2}k^2+\frac{c^4_{AD_\zeta}(\alpha)}{n^2}-\frac{8}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k+4\frac{c^2_{AD_\zeta}(\alpha)}{n^2}k-4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\frac{c^2_{AD_\zeta}(\alpha)}{n}$$

$$\Leftrightarrow 4\frac{k^2}{n^2}-\frac{4}{n^2}k^2-\frac{4}{n^2}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k^2+4\frac{c^2_{AD_\zeta}(\alpha)}{n^2}k+\frac{8}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k-4\frac{c^2_{AD_\zeta}(\alpha)}{n^2}k$$
$$+\frac{4}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\zeta c^2_{AD_\zeta}(\alpha)-4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)^2-\frac{c^4_{AD_\zeta}(\alpha)}{n^2}+\frac{c^4_{AD_\zeta}(\alpha)}{n^2}+4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\frac{c^2_{AD_\zeta}(\alpha)}{n}\leq 0$$

$$\Leftrightarrow -\frac{4}{n^2}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k^2+\frac{8}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k$$
$$+\frac{4}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\zeta c^2_{AD_\zeta}(\alpha)-4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)^2+4\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\frac{c^2_{AD_\zeta}(\alpha)}{n}\leq 0$$

$$\Leftrightarrow -\frac{1}{n^2}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k^2+\frac{2}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k$$
$$+\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\left[\frac{1}{n}\zeta c^2_{AD_\zeta}(\alpha)-1-\frac{c^2_{AD_\zeta}(\alpha)}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\leq 0$$

$$\Leftrightarrow \frac{1}{n^2}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k^2-\frac{2}{n}\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)k$$
$$-\left(1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right)\left[\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}-1\right]\geq 0$$

$$\Leftrightarrow \left(n + c^2_{AD_\zeta}(\alpha)\right) k^2 - 2n \left(n + c^2_{AD_\zeta}(\alpha)\right) k \\ - n \left(n + c^2_{AD_\zeta}(\alpha)\right) \left[\zeta c^2_{AD_\zeta}(\alpha) - n\right] \geq 0.$$

This is the case for all $k \leq \kappa_1$ and $k \geq \kappa_2$ where $\kappa_1 = \frac{2n\left(n+c^2_{AD_\zeta}(\alpha)\right)-\sqrt{\delta}}{2\left(n+c^2_{AD_\zeta}(\alpha)\right)}$ and $\kappa_2 = \frac{2n\left(n+c^2_{AD_\zeta}(\alpha)\right)+\sqrt{\delta}}{2\left(n+c^2_{AD_\zeta}(\alpha)\right)}$ where $\delta = 4n\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)^2 \zeta c^2_{AD_\zeta}(\alpha)$. Developing these expressions yields $\kappa_1 = n - c_{AD_\zeta}(\alpha)(n\zeta)^{1/2}$ and $\kappa_2 = n + c_{AD_\zeta}(\alpha)(n\zeta)^{1/2} \geq n$. Hence, $G^U_n(X_{(k)}) \leq 1$ $\forall k \leq \kappa_1$ or equivalently $\forall k = 0, 1, ..., k^U_{AD_\zeta}$ where $k^U_{AD_\zeta} = I[\kappa_1]$ and $I[\kappa_1]$ is the integer part of κ_1. Then,

$$\widetilde{F}^U_n(x) = \min\left\{G^U_n(x), 1\right\} = \begin{cases} G^U_n(x) \ \forall x \leq X_{(k^U_{AD_\zeta})} \\ 1 \ \forall x > X_{(k^U_{AD_\zeta})} \end{cases}$$

where $k^U_{AD_\zeta} = I[\kappa_1] = I[n - c_{AD_\zeta}(\alpha)(n\zeta)^{1/2}]$. It follows that

$$\mu_L = [1 - \widetilde{F}^U_n(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{k^U_{AD_\zeta}+1} \left[\widetilde{F}^U_n(X_{(k)}) - \widetilde{F}^U_n(X_{(k-1)})\right] X_{(k)}$$

$$\mu_L = [1 - \widetilde{F}^U_n(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}^U_n(X_{(k^U_{AD_\zeta}+1)}) - \widetilde{F}^U_n(X_{(k^U_{AD_\zeta})})\right] X_{(k^U_{AD_\zeta}+1)} \\ + \sum_{k=1}^{k^U_{AD_\zeta}} \left[G^U_n(X_{(k)}) - G^U_n(X_{(k-1)})\right] X_{(k)}.$$

We showed that $\widetilde{F}^U_n(X_{(n)}) = 1$ (because $\kappa_1 = n - c_{AD_\zeta}(\alpha)(n\zeta)^{1/2} < n$), $\widetilde{F}^U_n(X_{(k^U_{AD_\zeta}+1)}) = 1$, and $\widetilde{F}^U_n(X_{(k^U_{AD_\zeta})}) = \frac{2\frac{k^U_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta(k^U_{AD_\zeta})}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}$ where $\Delta(k^U_{AD_\zeta}) = \left[\frac{2k^U_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(\left(\frac{k^U_{AD_\zeta}}{n}\right)^2 - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$ and $\Delta(n) = \left[2 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(1 - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$.

Then,

$$\mu_L = [1-1]\,X_{(n+1)} + \left[1 - \frac{2\frac{k^U_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta(k^U_{AD_\zeta})}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\right] X_{(k^U_{AD_\zeta}+1)}$$
$$+ \sum_{k=1}^{k^U_{AD_\zeta}} \left[\frac{\frac{2k}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta(k)}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})} - \frac{\frac{2(k-1)}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta(k-1)}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\right] X_{(k)}$$

or

$$\mu_L = \frac{1}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\left\{\left(2 + \frac{2c^2_{AD_\zeta}(\alpha)}{n} - 2\frac{k^U_{AD_\zeta}}{n} - \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta(k^U_{AD_\zeta})}\right) X_{(k^U_{AD_\zeta}+1)}\right.$$
$$\left. + \sum_{k=1}^{k^U_{AD_\zeta}} \left(\frac{2}{n} + \sqrt{\Delta(k)} - \sqrt{\Delta(k-1)}\right) X_{(k)}\right\}$$

$$\mu_L = \frac{1}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\left\{\left(2 + \frac{c^2_{AD_\zeta}(\alpha)}{n} - 2\frac{k^U_{AD_\zeta}}{n} - \sqrt{\Delta(k^U_{AD_\zeta})}\right) X_{(k^U_{AD_\zeta}+1)}\right.$$
$$\left. + \sum_{k=1}^{k^U_{AD_\zeta}} \left(\frac{2}{n} + \sqrt{\Delta(k)} - \sqrt{\Delta(k-1)}\right) X_{(k)}\right\}.$$

Similarly, for the upper bound:

$$\mu_U = [1 - \widetilde{F}^L_n(X_{(n)})]X_{(n+1)} + \sum_{k=k^L_{AD_\zeta}}^{n} \left[\widetilde{F}^L_n(X_{(k)}) - \widetilde{F}^L_n(X_{(k-1)})\right] X_{(k)}$$

$$= [1 - \widetilde{F}^L_n(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}^L_n(X_{(k^L_{AD_\zeta})}) - \widetilde{F}^L_n(X_{(k^L_{AD_\zeta}-1)})\right] X_{(k^L_{AD_\zeta})}$$
$$+ \sum_{k=k^L_{AD_\zeta}+1}^{n} \left[G^L_n(X_{(k)}) - G^L_n(X_{(k-1)})\right] X_{(k)}.$$

The precedent results show that
$\widetilde{F}_n^L(X_{(n)}) = \begin{cases} \left(2+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(n)}\right)\left(2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})\right)^{-1} & \text{if } k^L_{AD_\zeta} \leq n \\ 0 \text{ otherwise} \end{cases}$, $\widetilde{F}_n^L(X_{(k^L_{AD_\zeta}-1)}) =$
0, and $\widetilde{F}_n^L(X_{(k^L_{AD_\zeta})}) = \left(2\frac{k^L_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta(k^L_{AD_\zeta})}\right)\left(2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})\right)^{-1}$ where $\Delta(k^L_{AD_\zeta}) = \left[\frac{2k^L_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(\left(\frac{k^L_{AD_\zeta}}{n}\right)^2 - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$ and $\Delta(n) = \left[2+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1+\frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(1-\frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$. Then,

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[\frac{\frac{2k^L_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta(k^L_{AD_\zeta})}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\right] X_{(k^L_{AD_\zeta})} + \sum_{k=k^L_{AD_\zeta}+1}^{n}\left[\frac{\frac{2k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(k)}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})} - \frac{\frac{2(k-1)}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(k-1)}}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\right] X_{(k)}$$

or

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \frac{1}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\left\{\left(\frac{2k^L_{AD_\zeta}}{n} + \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta(k^L_{AD_\zeta})}\right) X_{(k^L_{AD_\zeta})} + \sum_{k=k^L_{AD_\zeta}+1}^{n}\left[\frac{2k}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(k)}-\frac{2(k-1)}{n}-\frac{c^2_{AD_\zeta}(\alpha)}{n}+\sqrt{\Delta(k-1)}\right] X_{(k)}$$

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \frac{1}{2(1+\frac{c^2_{AD_\zeta}(\alpha)}{n})}\left\{\left(\frac{2k^L_{AD_\zeta}}{n}+\frac{c^2_{AD_\zeta}(\alpha)}{n}-\sqrt{\Delta(k^L_{AD_\zeta})}\right) X_{(k^L_{AD_\zeta})} + \sum_{k=k^L_{AD_\zeta}+1}^{n}\left[\frac{2}{n}-\sqrt{\Delta(k)}+\sqrt{\Delta(k-1)}\right] X_{(k)}.\right.$$

Proof of Proposition 4.4. Proposition 4.4 is a direct application of the general expression of the CI for the mean of a continuous bounded random variable on the case of the ζ-regularized Eicker CB for distribution functions.

PROOF OF COROLLARY 4.4BIS. Corollary 4.4bis is obtained developing the expression of the ζ-regularized Eicker CI for the mean of a bounded random variable. This CI is:

$$C_{\mu}^{E_\zeta}(\alpha) = \Big\{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U \text{ where } \mu_L = [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)}$$
$$\text{and } \mu_U = [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(X_{(k)}) - \widetilde{F}_n^L(X_{(k-1)})\right] X_{(k)} \Big\}$$

for $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\}$, $\widetilde{F}_n^U(x) = \min\{G_n^U(x), 1\}$,

$$G_n^L(x) = F_n(x) - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}} \left[F_n(x)\left(1 - F_n(x)\right) + \zeta\right]^{1/2},$$

$$G_n^U(x) = F_n(x) + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}} \left[F_n(x)\left(1 - F_n(x)\right) + \zeta\right]^{1/2},$$

and $c_{E_\zeta}(\alpha)$ satisfies $\Pr[E_\zeta \leq c_{E_\zeta}(\alpha)] \geq 1 - \alpha$.

Using $G_n^L(x)$:

$$\begin{aligned} G_n^L(X_{(k)}) \geq 0 \\ &\Leftrightarrow \frac{k}{n} - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}} \sqrt{\frac{k}{n}[1 - \frac{k}{n}] + \zeta} \geq 0 \\ &\Leftrightarrow k^2 - c_{E_\zeta}^2(\alpha)[k - \frac{k^2}{n} + n\zeta] \geq 0 \\ &\Leftrightarrow \left(n + c_{E_\zeta}^2(\alpha)\right) k^2 - n c_{E_\zeta}^2(\alpha) k - n^2 c_{E_\zeta}^2(\alpha)\zeta \geq 0. \end{aligned}$$

This is the case for all $k \leq k_1$ and $k \geq k_2$ where $k_1 = \frac{n c_{E_\zeta}^2(\alpha) - \sqrt{\Delta^L}}{2\left(n + c_{E_\zeta}^2(\alpha)\right)}$ and $k_2 = \frac{n c_{E_\zeta}^2(\alpha) + \sqrt{\Delta^L}}{2\left(n + c_{E_\zeta}^2(\alpha)\right)}$ with $\Delta^L = n^2 c_{(1-\alpha)}^4 + 4\left(n + c_{E_\zeta}^2(\alpha)\right) n^2 c_{E_\zeta}^2(\alpha)\zeta$. However, it is easy to see that k_1 is always negative. So, $G_n^L(X_{(k)}) \geq 0\ \forall k \geq k_2$ or equivalently $\forall k = k_{E_\zeta}^L, ..., n$ where $k_{E_\zeta}^L = I[k_2] + 1$ and we define $I[k] \equiv$the integer part of k.

Hence, $\widetilde{F}_n^L(x) = \max\{G_n^L(x), 0\} = \begin{cases} G_n^L(x)\ \forall x \geq X_{(k_{E_\zeta}^L)} \\ 0\ \forall x < X_{(k_{E_\zeta}^L)} \end{cases}$ where $k_{E_\zeta}^L = I[k_2] + 1$, $k_2 =$

$\left[nc^2_{E_\zeta}(\alpha)+\sqrt{\Delta^L}\right]\left[2\left(n+c^2_{E_\zeta}(\alpha)\right)\right]^{-1}$, and $\Delta^L = n^2c^4_{E_\zeta}(\alpha)+4\left(n+c^2_{E_\zeta}(\alpha)\right)n^2c^2_{E_\zeta}(\alpha)\zeta$.

Similarly using $G^U_n(x)$:

$$\begin{aligned}
G^U_n(X_{(k)}) &\leq 1 \\
&\Leftrightarrow \frac{k}{n}+\frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\sqrt{\frac{k}{n}[1-\frac{k}{n}]+\zeta}\leq 1 \\
&\Leftrightarrow \frac{c^2_{E_\zeta}(\alpha)}{n}\left(\frac{k}{n}-\frac{k^2}{n^2}+\zeta\right)\leq 1-\frac{2k}{n}+\frac{k^2}{n^2} \\
&\Leftrightarrow c^2_{E_\zeta}(\alpha)k-\frac{c^2_{E_\zeta}(\alpha)}{n}k^2+nc^2_{E_\zeta}(\alpha)\zeta\leq n^2-2nk+k^2 \\
&\Leftrightarrow \left(n+c^2_{E_\zeta}(\alpha)\right)k^2-\left(2n^2+nc^2_{E_\zeta}(\alpha)\right)k+n^3-n^2c^2_{E_\zeta}(\alpha)\zeta\geq 0 \\
&\Leftrightarrow \left(n+c^2_{E_\zeta}(\alpha)\right)k^2-\left(2n+c^2_{E_\zeta}(\alpha)\right)nk+\left(n-c^2_{E_\zeta}(\alpha)\zeta\right)n^2\geq 0.
\end{aligned}$$

This is the case for all $k\leq k_3$ and $k\geq k_4$ where $k_3=\frac{\left(2n+c^2_{E_\zeta}(\alpha)\right)n-\sqrt{\Delta^U}}{2\left(n+c^2_{E_\zeta}(\alpha)\right)}$, $k_4=\frac{\left(2n+c^2_{E_\zeta}(\alpha)\right)n+\sqrt{\Delta^U}}{2\left(n+c^2_{E_\zeta}(\alpha)\right)}$, and $\Delta^U=\left(2n+c^2_{E_\zeta}(\alpha)\right)^2n^2-4\left(n+c^2_{E_\zeta}(\alpha)\right)\left(n-c^2_{E_\zeta}(\alpha)\zeta\right)n^2$. However, it can be proved that k_4 is always greater than n. So, $G^U_n(X_{(k)})\leq 1\ \forall k\leq k_3$ or equivalently $\forall k=0,1,...,k^U_{E_\zeta}$ where $k^U_{E_\zeta}=I[k_3]$ and $I[k]$ is defined as above.

Hence, $\widetilde{F}^U_n(x)=\min\{G^U_n(x),1\}=\begin{cases} G^U_n(x)\ \forall x\leq X_{(k^U_{E_\zeta})} \\ 1\ \forall x>X_{(k^U_{E_\zeta})}\end{cases}$ where $k^U_{E_\zeta}=I[k_3]$, $k_3=\left[\left(2n+c^2_{E_\zeta}(\alpha)\right)n-\sqrt{\Delta^U}\right]\left[2\left(n+c^2_{E_\zeta}(\alpha)\right)\right]^{-1}$, and
$\Delta^U=\left(2n+c^2_{E_\zeta}(\alpha)\right)^2n^2-4\left(n+c^2_{E_\zeta}(\alpha)\right)\left(n-c^2_{E_\zeta}(\alpha)\zeta\right)n^2$.

$\widetilde{F}^L_n(x)$ and $\widetilde{F}^U_n(x)$ represent the effective part of the ζ-regularized Eicker CB for distribution functions. It follows that

$$\begin{aligned}
\mu_U &= [1-\widetilde{F}^L_n(X_{(n)})]X_{(n+1)}+\sum_{k=1}^{n}\left[\widetilde{F}^L_n(X_{(k)})-\widetilde{F}^L_n(X_{(k-1)})\right]X_{(k)} \\
&= [1-\widetilde{F}^L_n(X_{(n)})]X_{(n+1)}+\sum_{k=k^L_{E_\zeta}}^{n}\left[\widetilde{F}^L_n(X_{(k)})-\widetilde{F}^L_n(X_{(k-1)})\right]X_{(k)}
\end{aligned}$$

$$= [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}_n^L(X_{(k_{E_\zeta}^L)}) - \widetilde{F}_n^L(X_{(k_{E_\zeta}^L-1)})\right] X_{(k)} + \sum_{k=k_{E_\zeta}^L+1}^{n} \left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right] X_{(k)}$$

$$= [1 - \widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}_n^L(X_{(k_{E_\zeta}^L)}) - \widetilde{F}_n^L(X_{(k_{E_\zeta}^L-1)})\right] X_{(k_{E_\zeta}^L)} + \sum_{k=k_{E_\zeta}^L+1}^{n} \left[G_n^L(X_{(k)}) - G_n^L(X_{(k-1)})\right] X_{(k)}.$$

Given our results, $\widetilde{F}_n^L(X_{(k_{E_\zeta}^L-1)}) = 0$, $\widetilde{F}_n^L(X_{(k_{E_\zeta}^L)}) = G_n^L(X_{(k_{E_\zeta}^L)}) = k_{E_\zeta}^L n^{-1} - c_{E_\zeta}(\alpha)n^{-1/2}\left[k_{E_\zeta}^L n^{-1}\left(1 - k_{E_\zeta}^L n^{-1}\right) + \zeta\right]^{1/2}$, and $\widetilde{F}_n^L(X_{(n)}) = \begin{cases} 1 - c_{E_\zeta}(\alpha)\zeta^{1/2}n^{-1/2} \text{ if } k_{E_\zeta}^L \leq n \\ 0 \text{ if } k_{E_\zeta}^L > n \end{cases}$. Then

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[k_{E_\zeta}^L n^{-1} - c_{E_\zeta}(\alpha)n^{-1/2}\left(k_{E_\zeta}^L n^{-1}\left(1 - k_{E_\zeta}^L n^{-1}\right) + \zeta\right)^{1/2}\right] X_{(k_{E_\zeta}^L)} + \sum_{k=k_{E_\zeta}^L+1}^{n} \left[\frac{k}{n} - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left(\frac{k}{n}[1 - \frac{k}{n}] + \zeta\right)^{1/2} - \frac{k-1}{n} + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left(\frac{k-1}{n}[1 - \frac{k-1}{n}] + \zeta\right)^{1/2}\right] X_{(k)}$$

or

$$\mu_U = [1-\widetilde{F}_n^L(X_{(n)})]X_{(n+1)} + \left[k_{E_\zeta}^L n^{-1} - c_{E_\zeta}(\alpha)n^{-1/2}\left(k_{E_\zeta}^L n^{-1}\left(1 - k_{E_\zeta}^L n^{-1}\right) + \zeta\right)^{1/2}\right] X_{(k_{E_\zeta}^L)} + \sum_{k=k_{E_\zeta}^L+1}^{n} \left[\frac{1}{n} - c_{E_\zeta}(\alpha)n^{-1/2}\left(\sqrt{\frac{k}{n}[1 - \frac{k}{n}] + \zeta} - \sqrt{\frac{k-1}{n}[1 - \frac{k-1}{n}] + \zeta}\right)\right] X_{(k)}$$

where $\widetilde{F}_n^L(X_{(n)}) = \begin{cases} 1 - c_{E_\zeta}(\alpha)\zeta^{1/2}n^{-1/2} \text{ if } k_{E_\zeta}^L \leq n \\ 0 \text{ if } k_{E_\zeta}^L > n. \end{cases}$.

Similarly, for the lower bound:

$$\mu_L = [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(X_{(k)}) - \widetilde{F}_n^U(X_{(k-1)})\right] X_{(k)}$$

$$= [1 - \widetilde{F}_n^U(X_{(n)})]X_{(n+1)} + \left[\widetilde{F}_n^U(X_{(k_{E_\zeta}^U+1)}) - \widetilde{F}_n^U(X_{(k_{E_\zeta}^U)})\right] X_{(k_{E_\zeta}^U+1)} + \sum_{k=1}^{k_{E_\zeta}^U} \left[G_n^U(X_{(k)}) - G_n^U(X_{(k-1)})\right] X_{(k)}.$$

Our early results show that, $\widetilde{F}_n^U(X_{(k_{E_\zeta}^U+1)}) = 1$, $\widetilde{F}_n^U(X_{(k_{E_\zeta}^U)}) = G_n^U(X_{(k_{E_\zeta}^U)}) = k_{E_\zeta}^U n^{-1} + c_{E_\zeta}(\alpha)n^{-1/2}\left[k_{E_\zeta}^U n^{-1}\left(1 - k_{E_\zeta}^U n^{-1}\right) + \zeta\right]^{1/2}$,and $\widetilde{F}_n^U(X_{(n)}) = 1$ (because $G_n^U(X_{(n)}) = 1 + c_{E_\zeta}(\alpha)\zeta^{1/2}n^{-1/2} \geq 1$). Then

$$\begin{aligned}\mu_L &= [1-1]X_{(n+1)} + \left[1 - k_{E_\zeta}^U n^{-1} - c_{E_\zeta}(\alpha)n^{-1/2}\left[k_{E_\zeta}^U n^{-1}\left(1 - k_{E_\zeta}^U n^{-1}\right) + \zeta\right]^{1/2}\right] X_{(k_{E_\zeta}^U+1)} \\ &\quad + \sum_{k=1}^{k_{E_\zeta}^U}\left[\frac{k}{n} + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left(\frac{k}{n}[1-\frac{k}{n}] + \zeta\right)^{1/2} - \frac{k-1}{n} - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left(\frac{k-1}{n}[1-\frac{k-1}{n}] + \zeta\right)^{1/2}\right] X_{(k)}\end{aligned}$$

or

$$\begin{aligned}\mu_L &= \left[1 - k_{E_\zeta}^U n^{-1} - c_{E_\zeta}(\alpha)n^{-1/2}\left[k_{E_\zeta}^U n^{-1}\left(1 - k_{E_\zeta}^U n^{-1}\right) + \zeta\right]^{1/2}\right] X_{(k_{E_\zeta}^U+1)} \\ &\quad + \sum_{k=1}^{k_{E_\zeta}^U}\left[\frac{1}{n} + c_{E_\zeta}(\alpha)n^{-1/2}\left(\sqrt{\frac{k}{n}[1-\frac{k}{n}] + \zeta} - \sqrt{\frac{k-1}{n}[1-\frac{k-1}{n}] + \zeta}\right)\right] X_{(k)}\end{aligned}$$

where $\widetilde{F}_n^L(X_{(n)}) = \begin{cases} 1 - c_{E_\zeta}(\alpha)\zeta^{1/2}n^{-1/2} \text{ if } k_{E_\zeta}^L \leq n \\ 0 \text{ if } k_{E_\zeta}^L > n. \end{cases}$.

Proof of Proposition 4.5. This theorem is an application of Proposition 3.4 to the Owen (1995) CB for distribution functions where $\forall x$, $\widetilde{F}_n^L(x) = \max\{\widetilde{F}_n^L(x), 0\}$ and $\widetilde{F}_n^U(x) = \min\{\widetilde{F}_n^U(x), 1\}$ because $\widetilde{F}_n^L(X_{(0)}) = 0$, $\widetilde{F}_n^U(X_{(0)}) = 1 - e^{-\lambda_n} < 1$, $\widetilde{F}_n^L(X_{(n)}) = e^{-\lambda_n} > 0$, and $\widetilde{F}_n^U(X_{(n)}) = 1$.

Proof of Proposition 5.1. We refer the reader to Diouf and Dufour (2005a) for a complete proof.

Proof of Proposition 5.2. Proposition 5.1. implies that the Kolmogorov-Smirnov CB obtained using appropriate critical points for $F(x)$ yields a CB for $G(y)$ with level

larger than or equal to $1-\alpha$, and similarly for the Anderson-Darling, Eicker, regularized Anderson-Darling and Eicker, and Owen CBs. By projection, so too are the corresponding CIs for the mean, when the involved distributions have bounded support.

PROOF OF PROPOSITION 5.3. We refer the reader to Diouf and Dufour (2005a) for a complete proof.

PROOF OF PROPOSITION 5.4. A similar proof to those of Proposition 5.2. applies for Proposition 5.4. In fact, Proposition 5.3. implies that the Kolmogorov-Smirnov CB obtained using appropriate critical points for $F(x)$ yields a CB for $G(y)$ with level larger than or equal to $1-\alpha$, and similarly for the Anderson-Darling, Eicker, regularized Anderson-Darling and Eicker, and Owen CBs[see Diouf and Dufour (2005a) for the proof]. By projection, so too are the corresponding CIs for the mean, when the involved distributions have bounded support.

PROOF OF THE KOLMOGOROV-SMIRNOV STATISTIC FOR A DISTRIBUTION WITH A PROBABILITY MASS AT THE LOWER BOUND. Let Y be a random variable with continuous distribution function $G(y)$. Define $X = (\frac{z-Y}{z})^{\alpha}\ \mathbb{1}[0 \leq Y \leq z]$, a mixture between a bounded continuous variable and a probability mass $F(0) = 1 - G(z)$ at 0 with distribution function

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

or equivalently,

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ p & \text{if } x = 0, \\ p + \int_0^x h(u)\, du & \text{if } x > 0 \end{cases}$$

where $p = 1 - F(z)$ and $h(x)$ is an adequate density function. Hence, $F_X(X)$ is a mixture:

$$F(X) = \begin{cases} p & \text{with probability } p, \\ U & \text{with probability } 1 - p \end{cases}$$

where $U \sim U_{(p,1]}$. The corresponding Kolmogrov-Smirnov statistic is:

$$KS_F = \max_{0<x\leq 1} \mid F_n(x) - F(x) \mid = \max \left\{ |\widehat{p} - p|, \max_{0<x\leq 1} \mid F_n(x) - F(x) \mid \right\}$$

$$= \max \left\{ |\widehat{p} - p|, \max_{0<x\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[X_k \leq x] - F(x) \mid \right\}$$

$$= \max \left\{ |\widehat{p} - p|, \max_{0<x\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[F(X_k) \leq F(x)] - F(x) \mid \right\}$$

$$= \max \left\{ |\widehat{p} - p|, \max_{p<v\leq 1} \mid \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[F(X_k) \leq v] - v \mid \right\}$$

$$= \max_{p\leq v\leq 1} \left| \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}[F(X_k) \leq v)] - v \right|.$$

Proof of corollary 5.5. This corollary is an application of Proposition 5.4. Let V_1 and V_2 be the sets of values of $F_1(x)$ and $G_2(x)$, respectively. Then $V_1 = [p_1, 1]$ and $V_2 = [p_2, 1]$ with $V_2 \subseteq V_1$. Hence, the percentiles of the statistics KS_{F_1}, AD_{F_1}, E_{F_1}, AD_{ζ,F_1}, E_{ζ,F_1}, and BJ_{F_1} are conservative for the percentile values of KS_{F_2}, AD_{F_2}, E_{F_2}, AD_{ζ,F_2}, E_{ζ,F_2}, and BJ_{F_2}. Hence, the CBs for distribution functions and the corresponding CIs for the mean–when the variable has a bounded support $[a, b]$–using the appropriate critical points for $F(x)$ yield CBs and CIs for $G(y)$ with level larger than or equal to $1 - \alpha$.

Chapter 3

Finite-sample nonparametric inference for inequality measures

Abstract

Inference studies for poverty and inequality measures show that asymptotic and bootstrap inference methods can be quite unreliable when applied to those measures (see Davidson and Flachaire, 2007). In a preceding paper, we proposed improved finite-sample nonparametric confidence intervals for the Foster, Greer and Thorbecke (1984) poverty measures using confidence bands for distribution functions and projection techniques. We showed that these confidence intervals are robust and perform better than asymptotic ones.
In this paper, we propose improved finite-sample confidence intervals for inequality measures. We propose a generalized projection principle to derive confidence intervals for the mean of a random variable from confidence bands for distribution functions which tails are bounded by a Pareto distribution. Reexpressing the inequality measures as a function of the means of a bounded random variable and a non bounded one, we apply the inference methods to those. Monte Carlo simulations show that the corresponding confidence intervals yield very reliable and good performance. We illustrate how to use the inference methods analyzing inequalities among Mexican rural households in 1998 using PROGESA data sets. The results show that while the level of inequality among households targeted by PROGRESA is fairly low, this level almost three times higher for households with a female head and almost twice higher for households with a non educated head.

3.1 Introduction

Inference studies for inequality measures show that asymptotic and bootstrap methods do not perform well when applied to these measures. Davidson and Flachaire (2007) showed that asymptotic approximations provide a poor approximation to the real distributions of statistics, for small and even fairly large samples. Using a Singh-Maddala distribution, they show that the i.i.d. bootstrap confidence interval do not perform well for the Theil inequality index. The heavy tail of the Singh-Maddala distribution alters the performance of the standard bootstrap and a modified bootstrap procedure, which is more adapted to heavy tails must be used to improve the results.

Other papers have studied the performance of asymptotic and bootstrap inference methods for poverty and inequality measures; see Beran (1988), Kakwani (1993), Dardanoni and Forcina (1999), Biewen (2002), Davidson and Duclos (2000), and Cowell and Flachaire (2002). Most of these studies recommend using bootstrap procedures rather than asymptotic ones but also acknowledge the limits of the i.i.d. bootstrap procedure. Bootstrap often fails to provide reliable inference when applied to distributions with heavy tails or probability masses. Hence, adequate bootstrap procedures must be used to provide good performance. However, the origin of the failure of the bootstrap must be identified to correct the drawback, which is not obvious when data come from an unknown distribution function.

In this paper, we propose nonparametric inference methods for the mean of random variables and apply them to inequality measures. We show that inequality measures can be reexpressed as a function of the mean of two random variables: a bounded random variable and an unbounded one. Using projection techniques we proposed in a preceding paper (Diouf and Dufour (2005b)), we build confidence intervals for the mean of the bounded part of the inequality measures. Then, we propose a generalization of these projection techniques to nonbounded random variables when the tails of the corresponding distribution is bounded by a Pareto distribution. We apply these methods to derive confidence intervals for inequality measures using confidence bands for the underlying distributions. Empirical distribution function-based statistics using the three common

principles in econometrics are used to build confidence bands: the Wald, the score and the likelihood-ratio principles (see Diouf and Dufour (2005a)).

We propose finite-sample confidence intervals for the most popular inequality measures: the generalized entropy class of indexes—which include the Theil index, the Lorenz curve, the Gini index and the Atkinson class of indexes. According to Bahadur and Savage (1956), nonparametric inference cannot be performed for the mean of a random variable when observations are independent and identically distributed (i.i.d.) from an unknown distribution function with finite mean (see Dufour (2003) for more details). To avoid this impossibility theorem, we suppose that the tail of the distribution of the sample is bounded by a Pareto distribution and consider two cases: the case where the parameters of the Pareto distribution are known and the case where the parameters are unknown. In this last case, we build a joint confidence region for the parameters of the Pareto distribution using Chen(1996) and the Bonferroni inequality.

Monte Carlo simulations are performed to study the performance of these methods for the Theil index. The results show that the standard bootstrap procedure and the alternative proposed by Davidson and Flachaire (2007), as well as the asymptotic method can fail in providing reliable confidence intervals for the Theil index while nonparametric inference methods are strongly reliable and provide informative confidence intervals. The regularized statistics deliver the best width among the latter.

At last, the profile of inequality of Mexican households involved in PROGRESA is assessed using the Gini index. The results show that there are more inequalities among households with a female head or a non-educated head. Hence, in addition to implementing policies that would help reduce poverty among households with a female head or a non-educated head, authorities plan policies targeted to the most vulnerable among those households to help them catch up with other households and get insured against negative shocks that would increase inequality further.

The remainder of the paper is organized as follows. Section 2 and 3 presents the desirable properties for inequality measures and the most popular inequality measures. Section 4 provides asymptotic and bootstrap confidence intervals for the generalized entropy

class of index. Section 5 proposes finite-sample nonparametric confidence intervals for these inequality measures using projection techniques, when income is bounded. In section 6, we propose a generalization of the projection techniques to nonbounded random variables whose tails of distribution are bounded by a Pareto distribution. Section 7 proposes nonparametric confidence intervals for the most popular inequality measures (the Theil index, the Lorenz curve, the Gini index, the Mean Logarithmic Deviation, the Logarithmic Variation index, and the Atkinson Class of index). Section 8 presents Monte Carlo results. Section 9 analyzes the profile of inequality of rural Mexican households targeted by PROGRESA using the studied inference methods. Section 10 concludes.

3.2 Desirable properties for inequality measures

Inequality studies have more and more been acknowledged to be complementary to poverty analysis. Both studies are mostly performed simultaneously to better assess the profile of poverty of households in a given community. Several inequality measures can be used, depending on the notion of inequality the study is intended to be assessed. These inequality measures must satisfy a set of suitable properties to be considered as reliable measures of inequality. In this section, we present the most important axioms that need to be filled by inequality measures.

Let $\mathcal{F}$ be a space of distribution functions with support $\aleph$. Let's consider a community which households' income is a random variable Y with distribution function $F(y) \in \mathcal{F}$. An inequality measure is a functional $I : \mathcal{F} \longmapsto \mathbb{R}$ defined on the space of distribution functions $\mathcal{F}$ (see Cowell and Flachaire(2002) and Cowell (2003)). To be reliable, inequality measures may satisfy a set of desirable properties that ensure their coherence. Among the most important of these properties are the following.

DEFINITION 2.1. **[Transfer Principle]** *Let F_1 and F_2 be two distributions functions. If*

$$I(F_1) < I(F_2),$$

then F_2 is a mean-preserving spread of F_1,

i.e., if Y_1 and Y_2 are two random variables with distribution functions $F_1(y_1)$ and $F_2(y_2)$, respectively, then

$$Y_2 = Y_1 + Z$$

where Z is a random variable with distribution function $H(z)$ such that $\int z\, dH(z) = 0$.

Definition 2.1. characterizes a very important property for inequality measures. It defines coherence in the ranking of the level of inequality of communities. Let's consider two communities: community 1 and community 2. Let Y_1 and Y_2 be the income of these communities, whose distributions of income F_1 and F_2 are such that $E(Y_1) = E(Y_2)$ and $V(Y_1) \leq V(Y_2)$. Hence, for inequality measures that satisfy the transfer principle, the level of inequality in community 1 is larger than the inequality in community 2.

DEFINITION 2.2. [T-Independence] *Let $\tau : \mathbb{R} \longrightarrow \mathbb{R}$ be a strictly monotonic continuous function defined on $\mathbb{R}$. Let $\aleph^{(\tau)}$ be the set*

$$\aleph^{(\tau)} = \{\tau(y) : y \in \aleph\} \cap \aleph$$

and let $F^{(\tau)} \in \mathcal{F}$ be the $\tau-$transformed distribution function such that

$$F^{(\tau)}(x) = F[\tau^{-1}(y)],\ y \in \aleph^{(\tau)}$$

*Let T be a set of transformation functions τ. The inequality measure I is T-*independent *if and only if for all $\tau \in T$,*

$$I(F^{(\tau)}) = I(F)$$

Definition 2.2 implies a very interesting property for inequality measures. For a given set of admissible transformations τ, all communities whose income distribution is a $\tau-$transformation of $F(y)$ yield the same level of inequality. Let's consider two special cases of this property:

- *Scale independence:* if T is the set of functions τ such that $\tau(y) = ay$, $a > 0$ then the inequality measure I is homogeneous of degree 0.

- *Translation independence: If* T is the set of functions τ such that $\tau(y) = y + b$, then the inequality measure I is invariant by translation.

DEFINITION 2.3. **[Additive separability]** *The inequality measure I is additively separable if there are two functions*

$$\Phi : X \longrightarrow \mathbb{R} \qquad \textit{and} \qquad \Psi : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

such that

$$I(F) = \Psi[\mu(F), \int \Phi(y)\ dF(y)]$$

where $\mu(F) = \int y\ dF(y)$.

The functions Φ and Ψ characterize inequality measures. Φ is named the income-evaluation function and Ψ the cardinalisation function. They can be used to define many of the desirable properties of inequality measures such as the following decomposability.

DEFINITION 2.4. **[Decomposability]** *The inequality measure I is decomposable if and only if it can be rewritten as*

$$I(F) = \int f[y, \mu(F)]\ dF(y)$$

where $f(y, z) : \mathbb{R}^2 \longrightarrow \mathbb{R}$ *is a function monotonically increasing in its first argument* y.

Cowell and Victoria-Feser (1996) showed that a decomposable inequality measure satisfies the transfer principle, due to the monotonicity of $f(y, z)$ with respect to y.

3.3 The inequality measures

This section presents the most popular inequality measures. Let Y be a random variable that represents the income of households in a given community and let $F(y)$ the distribution function of Y. Let $\mu = \int y\ dF(y)$ be the mean of Y.

- The *Generalized Entropy class* of index:

$$I_E^{\delta} = \int \frac{1}{\delta(\delta-1)}[(\frac{y}{\mu})^{\delta} - 1]\ dF(y)$$

where $\delta \in \mathbb{R}\backslash\{0,1\}$.

For $\delta = 0$ and $\delta = 1$, the Generalized Entropy becomes the mean logarithmic deviation and the Theil index, respectively. The *Mean Logarithmic Deviation* corresponds to the Generalized Entropy class of index with $\delta = 0$. It is:

$$\begin{aligned} I_E^0 &= -\int \log(\frac{y}{\mu})\ dF(y) \\ &= log(\mu) - \int log(y)\ dF(y) \end{aligned}$$

The *Theil index* (Theil, 1967) corresponds to the generalized entropy class of index with $\delta = 1$. It is one of the most popular inequality measures:

$$\begin{aligned} I_E^1 &= \int \frac{y}{\mu}\ \log(\frac{y}{\mu})\ dF(y) \\ &= \frac{1}{\mu}\int y\ log(y)\ dF(y) - log(\mu) \end{aligned}$$

- The *Atkinson class* of measures:

$$I_A^{\varepsilon} = \begin{cases} 1 - [\int (\frac{y}{\mu})^{1-\varepsilon}\ dF(y)]^{\frac{1}{1-\varepsilon}} & \text{if } \varepsilon > 0 \text{ and } \varepsilon \neq 1 \\ 1 - e^{-I_E^0} & \text{if } \varepsilon = 1 \end{cases}$$

- The *Logarithmic variation index*:

$$I_{LV} = \int [\log(\frac{y}{\mu})]^2 \; dF(y)$$

- The *Gini index*: many expressions of this index are proposed in the literature. The most useful of them is the following one

$$I_G = 1 - 2R(F)$$

where $R(F) = \frac{1}{\mu}\int_0^1 C(F;q)\; dq$, $C(F;q) = \int_0^{Q(F;q)} y\; dF(y)$, and $Q(F;q) = \inf\{y \mid F(y) \geq q\}$ for $q \in [0,1]$. $C(F;q)$ is the cumulative income function and $Q(F;q)$ is the quantile function.

- The *Lorenz curve*:

$$\begin{aligned} L(p) &= \frac{E\{Y \mid Y \leq F_Y^{-1}(p)\}}{E(Y)} \\ &= \frac{1}{E(Y)} \int_0^{F_Y^{-1}(p)} y\; dF_Y(y) \end{aligned}$$

where $p \in (0,1)$.

With the exception of the Gini index, all inequality measures defined above are additively separable. Moreover, Cowell (2003) shows that a continuous inequality measure I is scale invariant, decomposable and satisfies the principle of transfer if and only if it is ordinally equivalent to the generalized entropy class for some δ. In other words, an inequality measure that achieves the same ranking of communities as the generalized entropy class satisfies three of the most important suitable axioms for inequality measures: the scale invariance, the decomposability and the transfer principle.

3.4 Asymptotic confidence intervals for the generalized entropy class of index

Let Y be a random variable that represents the income of a community's households. Let $F(y)$ be the distribution function of Y and let's consider the test of the hypothesis $H_0 : I_E^\delta = I_0$ versus the alternative $H_1 : I_E^\delta \neq I_0$. The t-statistic for this test is:

$$W = \frac{\widehat{I}_E^\delta - I_0}{\left[\widehat{V}\left(\widehat{I}_E^\delta\right)\right]^{1/2}}$$

where $\widehat{I}_E^\delta$ is an estimation of I_E^δ and $\widehat{V}\left(\widehat{I}_E^\delta\right)$ is the estimated variance of $\widehat{I}_E^\delta$. Using this statistic, asymptotic and bootstrap confidence intervals (CIs, henceforth) can be built for the Generalized Entropy Index.

3.4.1 Confidence intervals when $\delta \neq 0, 1$

Let $Y_1, \ldots, Y_n$ be n i.i.d. observations on Y with distribution function $F(y)$. Let $\widehat{I}_E^\delta$ and $\widehat{V}[\widehat{I}_E^\delta]$ be the statistics:

$$\widehat{I}_E^\delta = \frac{1}{\delta(\delta-1)}\left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^\delta}{\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right]^\delta} - 1\right)$$

and

$$\widehat{V}[\widehat{I}_E^\delta] = \frac{1}{(\delta^2-\delta)^2(n-1)}\widehat{\mu}_1^{-2\delta}\left(\widehat{\mu}_{2\delta} - \widehat{\mu}_\delta^2\right)$$

where $\delta \neq 0, 1$, $\widehat{\mu}_\varrho = \frac{1}{n}\sum_{i=1}^{n} Y_i^\varrho$ (see Cowell, 1989 for a general expression using a subgroup decomposition).

Assuming that W is asymptotically $N(0,1)$ as $n \to \infty$, an asymptotic CI for I_E^δ with level $1-\alpha$ is:

$$\widehat{I}_E^\delta - z_{(1-\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^\delta\right)\right]^{1/2} \leq I_E^\delta(y) \leq \widehat{I}_E^\delta + z_{(1-\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^\delta\right)\right]^{1/2} \tag{3.1}$$

where $z_{(p)}$ is the p^{th} percentile of the standard normal distribution.

Similarly, a bootstrap CI for I_E^δ with level $1-\alpha$ is:

$$\widehat{I}_E^\delta - D^W_{(1-\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^\delta\right)\right]^{1/2} \leq I_E^\delta \leq \widehat{I}_E^\delta - D^W_{(\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^\delta\right)\right]^{1/2} \tag{3.2}$$

where $D^W_{(p)}$ is the p^{th} percentile of the bootstrap distribution of W.

3.4.2 Confidence interval for the Theil index

The t-statistic corresponding the Theil index is:

$$W = \frac{\widehat{I}_E^1 - I_0}{\left[\widehat{V}\left(\widehat{I}_E^1\right)\right]^{1/2}}$$

where

$$\widehat{I}_E^1 = \frac{\frac{1}{n}\sum_{i=1}^n Y_i \log(Y_i)}{\frac{1}{n}\sum_{i=1}^n Y_i} - \log\left(\frac{1}{n}\sum_{i=1}^n Y_i\right),$$

$$\widehat{V}\left(\widehat{I}_E^1\right) = \frac{1}{n(n-1)}\sum_{i=1}^n (X_i - \overline{x})^2,$$

$$X_i = \frac{Y_i \log(Y_i)}{\frac{1}{n}\sum_{i=1}^n Y_i} - \frac{Y_i}{\frac{1}{n}\sum_{i=1}^n Y_i}\left[\frac{\frac{1}{n}\sum_{i=1}^n Y_i \log(Y_i)}{\frac{1}{n}\sum_{i=1}^n Y_i} + 1\right] + 1,$$

and $\overline{x} = \frac{1}{n}\sum_{i=1}^n X_i$ (see Biewen and Jenkins (2003) for more details when observations are weighted).

An asymptotic CI for I_E^1 with level $1-\alpha$ is:

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

where $z_{(p)}$ is the p^{th} percentile of the standard normal distribution.

Likewise, a bootstrap CI for I_E^1 with level $1-\alpha$ is:

$$\widehat{I}_E^1 - D^W_{(1-\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^1\right)\right]^{1/2} \leq I_E^1 \leq \widehat{I}_E^1 - D^W_{(\frac{\alpha}{2})} * \left[\widehat{V}\left(\widehat{I}_E^1\right)\right]^{1/2}$$

where $D^W_{(p)}$ is the p^{th} percentile of the bootstrap distribution of the statistic W.

3.5 Nonparametric confidence intervals for generalized entropy class of index when income is bounded

Let Y be a random variable that represents the income of a community's households. Let's suppose that Y is bounded over $[0, \overline{y}]$ with continuous distribution function $F(y)$ where $\overline{y}$ can be as large as necessary ($Y \in [0, \overline{y}]$). Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of a sample of n i.i.d. observations on Y. Denote $F_n(y)$ the empirical distribution function of the sample such that $\forall\ k = 0, \ldots, n$

$$F_n(y) = \frac{k}{n} \text{ for } Y_{(k)} \leq y < Y_{(k+1)}. \tag{3.3}$$

Denote for the remainder of this section:

$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$;

$\Lambda(.)$: a functional $\Lambda[F] : \mathcal{L} \to \overline{\mathbb{R}}$ defined on a space $\mathcal{L}$ of functions;

$\mathcal{F}$: a space of distribution functions;

$\widetilde{\mathcal{F}}$: a space of continuous distribution functions.

In this section, we propose nonparametric confidence intervals for inequality measures that belong to the generalized entropy class of index, when households' income is bounded. To do so we reexpress the generalized entropy inequality measures as a function of the means of two bounded random variables. Then, we use nonparametric confidence intervals for the mean of a bounded random variable that was proposed by Diouf and Dufour (2005b) to build CIs for the two means involved in the inequality measures. To end, we derive CIs for inequality measures from those and use the Bonferroni inequality

to compute its level of confidence based on the levels of confidence of the underlying CIs.

3.5.1 Nonparametric confidence intervals when $\delta \neq 0, 1$

For $\delta \neq 0, 1$, the generalized entropy measure is

$$I_E^{\delta}(y) = \int \frac{1}{\delta(\delta-1)}[(\frac{y}{\mu})^{\delta} - 1] \; dF(y)$$

where $\mu = \int y \; dF(y)$. It can be reformulated as:

$$I_E^{\delta}(y) = \frac{1}{\delta(\delta-1)} \left(\frac{\int y^{\delta} \; dF(y)}{\left[\int y \; dF(y)\right]^{\delta}} - 1 \right) = \frac{1}{\delta(\delta-1)} \left(\frac{\Lambda_{\delta}(F)}{\Lambda_1^{\delta}(F)} - 1 \right)$$

where $\Lambda_{\delta}(F) = \int y^{\delta} \; dF(y)$ is the non centered moment of order δ of Y and $\Lambda_1(F) = \int y \; dF(y)$ is the mean of Y.

Given that Y is bounded over $[0, \overline{y}]$, Y^{δ} is also bounded over $[0, \overline{y}^{\delta}]$. In a former paper, we showed that nonparametric CIs for the mean of a bounded random variable can be derived from confidence bands for distribution functions using projections techniques. We proposed confidence intervals based on Wald, Score and likelihood-ratio improvements of the Kolmogorov-Smirnov statistic. We use these to build confidence intervals for Λ_{δ}.

Let $\mathcal{L}$ be a space of functions such that the Stieltjes integrals $\Lambda_{\delta}(G) = \int y^{\delta} \; dG(y)$, $\delta \neq 0$ are finite. Let $G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ be two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(n)}$ that define the following confidence band for $F(y)$ with level $1 - \alpha$:

$$C_F(\alpha) = \left\{ F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y), \; \forall y \right\}.$$

Following Diouf and Dufour (2005b), a nonparametric CI for $\Lambda_1[F]$ with level $1 - \alpha$

is:

$$\widetilde{C}_{\Lambda_1}(\alpha) = \left\{ \mu_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)}) \right] Y_{(k)} \leq \mu_0 \right.$$
$$\left. \leq [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)}) \right] Y_{(k)} \right\} \tag{3.4}$$

where $Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, and $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ and $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$.

Likewise, a nonparametric CI for $\Lambda_\delta[F]$ with level $1 - \alpha$ is:

$$\widetilde{C}_{\Lambda_\delta}(\alpha) = \left\{ \lambda_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)}^\delta + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)}) \right] Y_{(k)}^\delta \leq \lambda_0 \right.$$
$$\left. \leq [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)}^\delta + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)}) \right] Y_{(k)}^\delta \right\} \tag{3.5}$$

where $Y_{(0)}$, $Y_{(n+1)}$, $\widetilde{F}_n^L(y)$, and $\widetilde{F}_n^U(y)$ are defined as before.

Using these confidence intervals, Result 5.1. proposes a nonparametric confidence interval for I_E^δ. The level of this latter can be computed using the following Bonferroni inequality:

$$\Pr(E_1 \cap E_2) \geq 1 - \Pr(\overline{E}_1) - \Pr(\overline{E}_2)$$

where E_1 and E_2 are two given events.

RESULT 5.1. **[Nonparametric CIs for the generalized entropy Index with bounded income distribution]** *Let Y represent the income of a community's households. Let $F(y) \in \mathcal{L}$ be the distribution function of Y with support $[0, \overline{y}]$ and let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be n ordered i.i.d. observations on Y. Suppose that the following confidence band for $F(y)$ with level $1 - \alpha$ is valid for the space of distributions $\mathcal{L}$:*

$$C_F(\alpha) = \left\{ F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y \right\}$$

where $G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(n)}$.

Then the level of the following confidence interval for the Generalized Entropy Measure I_E^δ *is greater than or equal to* $1-\alpha$:

$$C_{I_E^\delta}(\alpha) = \{I_0 \in \mathbb{R} : I_{\delta,\min} \leq I_0 \leq I_{\delta,\max}\}$$

where

$$I_{\delta,\min} = \frac{1}{\delta(\delta-1)} \min_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{ \frac{\Lambda_\delta}{\Lambda_1^\delta} - 1 \right\},$$

$$I_{\delta,\max} = \frac{1}{\delta(\delta-1)} \max_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{ \frac{\Lambda_\delta}{\Lambda_1^\delta} - 1 \right\},$$

$$\Lambda_{\delta,\min} = [1 - \widetilde{F}_n^U(Y_{(n)})] Y_{(n+1)}^\delta + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)}) \right] Y_{(k)}^\delta,$$

$$\Lambda_{\delta,\max} = [1 - \widetilde{F}_n^L(Y_{(n)})] Y_{(n+1)}^\delta + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)}) \right] Y_{(k)}^\delta,$$

$Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, *and* $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ *and* $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$.

Result 5.1. proposes a general methodology to build CIs for inequality measures that belong to the class of generalized entropy indexes for δ different from 0 and 1. These CIs are derived from CBs for the underlying distribution with step-function bounds. We have proposed and studied interesting examples of such type of CBs in a former paper, which can be used to perform inference for I_E^δ. Those CBs for $F(y)$ with level $1-\alpha$ are:

- the *Kolmogorov-Smirnov* CB:

$$C_F^{KS}(\alpha) = \left\{ F_0 \in \mathbb{F} : F_n(y) - \frac{c_{KS}(\alpha)}{\sqrt{n}} \leq F_0(y) \leq F_n(y) + \frac{c_{KS}(\alpha)}{\sqrt{n}},\ \forall y \right\} \qquad (3.6)$$

  where $c_{KS}(\alpha)$ satisfies $\Pr[KS_F \leq c_{KS}(\alpha)] \geq 1-\alpha$, and $KS = \sup_{-\infty \leq y \leq +\infty} \sqrt{n} \mid F_n(y) - F(y) \mid$.

- the *Anderson Darling-type* CB:

$$C_F^{AD}(\alpha) = \left\{ F_0 \in \mathbb{F} : G_n^L(y) \leq F_0 \leq G_n^U(y),\ \forall y \right\}$$

where

$$G_n^L(y) = \frac{2F_n(y) + \frac{c_{AD}^2(\alpha)}{n} - \sqrt{\Delta(y)}}{2(1 + \frac{c_{AD}^2(\alpha)}{n})},$$
$$G_n^U(y) = \frac{2F_n(y) + \frac{c_{AD}^2(\alpha)}{n} + \sqrt{\Delta(y)}}{2(1 + \frac{c_{AD}^2(\alpha)}{n})},$$

$$\Delta(y) = \left[2F_n(y) + \frac{c_{AD}^2(\alpha)}{n}\right]^2 - 4F_n^2(y)\left[1 + \frac{c_{AD}^2(\alpha)}{n}\right],$$

$c_{AD}(\alpha)$ satisfies $Pr\left[AD \leq c_{AD}(\alpha)\right] \geq 1 - \alpha$, $AD = \sup_{-\infty<y<+\infty} V_n(y)$, and $V_n(y) =$

$$\begin{cases} 0 & \text{if } F(y) \in \{0,1\}, \\ \sqrt{n}\left|\frac{F_n(y)-F(y)}{F^{1/2}(y)[1-F(y)]^{1/2}}\right| & \text{otherwise.} \end{cases}.$$

- the *Eicker-type* CB:

$$C_F^E(\alpha) = \left\{F_0 \in \text{Ŧ} : G_n^L(y) \leq F_0 \leq G_n^U(y)\right\}$$

where

$$G_n^L(y) = \begin{cases} F_n(y) - \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(y)[1 - F_n(y)]^{1/2} & \forall\ y \text{ such that } F_n(y) \notin \{0,1\}, \\ 0 & \forall\ y \text{ such that } F_n(y) \in \{0,1\}, \end{cases}$$

$$G_n^U(y) = \begin{cases} F_n(y) + \frac{c_E(\alpha)}{\sqrt{n}} F_n^{1/2}(y)[1 - F_n(y)]^{1/2} & \forall\ y \text{ such that } F_n(y) \notin \{0,1\}, \\ 1 & \forall\ y \text{ such that } F_n(y) \in \{0,1\}, \end{cases}$$

$c_E(\alpha)$ satisfies $Pr\left[E \leq c_E(\alpha)\right] \geq 1 - \alpha$,

$$E = \sup_{-\infty<y<+\infty} \widehat{V}_n(y),$$

and

$$\widehat{V}_n(y) = \begin{cases} 0 & \text{if } F_n(y) \in \{0,1\}, \\ \sqrt{n}\left|\frac{F_n(y)-F(y)}{F_n^{1/2}(y)[1-F_n(y)]^{1/2}}\right| & \text{otherwise.} \end{cases}$$

- the ζ-*Regularized Anderson Darling-type* CB:

$$C_F^{AD_\zeta}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(y) \leq F_0 \leq G_n^U(y),\ \forall y\right\} \tag{3.7}$$

where

$$\begin{aligned} G_n^L(y) &= \frac{2F_n(y) + \frac{c^2_{AD_\zeta}(\alpha)}{n} - \sqrt{\Delta}}{2\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)}, \\ G_n^U(y) &= \frac{2F_n(y) + \frac{c^2_{AD_\zeta}(\alpha)}{n} + \sqrt{\Delta}}{2\left(1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right)}, \end{aligned}$$

$\Delta = \left[2F_n(y) + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]^2 - 4\left[1 + \frac{c^2_{AD_\zeta}(\alpha)}{n}\right]\left(F_n^2(y) - \frac{\zeta c^2_{AD_\zeta}(\alpha)}{n}\right)$, $c_{AD_\zeta}(\alpha)$ satisfies $\Pr[AD_\zeta \leq c_{AD_\zeta}(\alpha)] \geq 1-\alpha$, and $AD_\zeta^R = \sup_{-\infty<y<+\infty} \sqrt{n}\left|\frac{F_n(y)-F(y)}{\sqrt{F(y)[1-F(y)]+\zeta_n}}\right|$.

- the ζ-*Regularized Eicker-type* CB:

$$C_F^{E_\zeta}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\} \tag{3.8}$$

where

$$G_n^L(y) = F_n(y) - \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(y)\left(1 - F_n(y)\right) + \zeta\right]^{1/2},$$

$$G_n^U(y) = F_n(y) + \frac{c_{E_\zeta}(\alpha)}{\sqrt{n}}\left[F_n(y)\left(1 - F_n(y)\right) + \zeta\right]^{1/2},$$

$c_{E_\zeta}(\alpha)$ satisfies $\Pr[E_\zeta \leq c_{E_\zeta}(\alpha)] \geq 1-\alpha$, and $E_\zeta^R = \sup_{-\infty<y<+\infty} \sqrt{n}\left|\frac{F_n(y)-F(y)}{\sqrt{F_n(y)[1-F_n(y)]+\zeta_n}}\right|$.

- the *Berk Jones-type CB* (Owen, 1995)**:**

$$C_F^{O}(\alpha) = \left\{F_0 \in \mathbb{F} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\} \tag{3.9}$$

where

$$G_n^L(y) = \min\left\{p : K[F_n(y), p] \leq c_{BJ}(\alpha)\right\},$$

$$G_n^U(y) = \max\left\{p : K[F_n(y), p] \leq c_{BJ}(\alpha)\right\},$$

$K(\widehat{p}, p) = \widehat{p}\log\left(\frac{\widehat{p}}{p}\right) + (1-\widehat{p})\log\left(\frac{1-\widehat{p}}{1-p}\right)$, $c_{BJ}(\alpha)$ satisfies $P[BJ > c_{BJ}(\alpha)] \geq 1-\alpha$, and $BJ = \sup_{-\infty \leq y \leq +\infty} K[F_n(y), F(y)]$.

These CBs are convenient to use. A unique set of critical points is needed to build CBs for all continuous distribution functions, and hence, to build CIs for I_E^δ. When $F(y)$ is not continuous, the critical points of the statistics are conservative: using critical points adapted to continuous distribution functions provides CBs for $F(y)$ with level of confidence larger or equal to the theoretical level. Moreover, further information about the nature of the discontinuity of Y may allow to improve the performance of these inference methods. Using embeddedness of image sets of distribution functions, less conservative critical points can be computed, which reduces the width of CIs without altering their reliability.

Other CIs for the mean of a bounded random variable have been proposed: asymptotic and bootstrap CIs—which can be quite reliable when applied to small, and even large samples and to distribution functions with heavy tails or probability mass— and finite-sample nonparametric CIs—which have been proposed by Hora and Hora (1990) and Fishman (1991). We will compare the performance of these inference methods on inequality measures using Monte Carlo simulations.

3.5.2 Nonparametric confidence intervals when $\delta = 1$

The Theil index can be reexpressed as follows:

$$\begin{aligned} I_E^1 &= \int \frac{y}{\mu}\, \log(\frac{y}{\mu})\, dF(y) = \frac{1}{\mu}\int y\ log(y)\ dF(y) - log(\mu) \\ &= \frac{\Upsilon(F)}{\Lambda_1(F)} - log(\Lambda_1(F)) \end{aligned}$$

where $\mu = \Lambda_1(F) = \int y\ dF(y) \neq 0$ is the mean of Y and $\Upsilon(F) = \int y\ log(y)\ dF(y)$ is the mean of $Y\ log(Y)$. Given that Y is bounded between $[0, \overline{y}]$, $Y\ log(Y)$ is also bounded. In fact, the function $y\ log(y)$ is strictly decreasing between 0 and $1/e$ and strictly increasing between $1/e$ and $+\infty$. If $\overline{y} \leq \frac{1}{e}$ then $Ylog(Y)$ is bounded on $[\overline{y}\log\overline{y}, 0]$. If, on the contrary, $\overline{y} \geq \frac{1}{e}$ then $Ylog(Y)$ is bounded on $[-\frac{1}{e}, \overline{y}\log\overline{y}]$. In both cases, $Ylog(Y)$ is bounded. We assume for the remainder of the paper that $Ylog(Y) \in [\upsilon_1, \upsilon_2]$ where $\upsilon_1 < \upsilon_2$. Hence, the Theil index is a function of the means of two bounded random variables. Following Diouf and Dufour (2005b), we propose nonparametric CIs for these two means, which we can be used to build CIs for the Theil index.

Let $\mathcal{L}$ be a space of functions such that the Stieltjes integrals $\Lambda_1(G) = \int_0^{\overline{y}} y\ dG(y)$ and $\Upsilon(F) = \int_{\upsilon_1}^{\upsilon_2} y\ log(y)\ dG(y)$ are finite. Let $G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ be two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(n)}$ that define the following confidence band for $F(y)$ with level $1-\alpha$:

$$C_F(\alpha) = \left\{F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\}.$$

A nonparametric CI for $\Lambda_1[F]$ with level $1-\alpha$ is:

$$\widetilde{C}_{\Lambda_1}(\alpha) = \left\{\mu_0 \in \mathbb{R} : [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)})\right]Y_{(k)} \leq \mu_0 \right.$$
$$\left. \leq [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)})\right]Y_{(k)}\right\}$$

where $Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, and $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ and $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$.

When applying the same procedure to $\Upsilon[F]$, let $Z = YlogY$ and $H(z)$ be the distribution function of Z. Let $\breve{G}_n^L(y) \in \mathcal{L}$ and $\breve{G}_n^U(y) \in \mathcal{L}$ be two step functions with jumps only at $Z_{(1)}, \ldots, Z_{(n)}$ that define the following confidence band for $H(z)$ with level $1-\alpha$:

$$C_H(\alpha) = \left\{H_0 \in \mathcal{L} : \breve{G}_n^L(z) \leq H_0(z) \leq \breve{G}_n^U(z),\ \forall z\right\}.$$

A nonparametric CI for $\Upsilon[F]$ with level $1-\alpha$ is:

$$\widetilde{C}_{\Upsilon}(\alpha) = \left\{ \upsilon_0 \in \mathbb{R} : [1 - \widetilde{H}_n^U(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{H}_n^U(Z_{(k)}) - \widetilde{H}_n^U(Z_{(k-1)}) \right] Z_{(k)} \leq \upsilon_0 \right.$$
$$\left. \leq [1 - \widetilde{H}_n^L(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{H}_n^L(Z_{(k)}) - \widetilde{H}_n^L(Z_{(k-1)}) \right] Z_{(k)} \right\}$$

where $Z_{(0)} = \upsilon_1$, $Z_{(n+1)} = \upsilon_2$, and $\forall z$, $\widetilde{H}_n^L(z) = \max\left\{ \breve{G}_n^L(z), 0 \right\}$ and $\widetilde{H}_n^U(z) = \min\left\{ \breve{G}_n^U(z), 1 \right\}$

Using these confidence intervals, Result 5.2 proposes finite-sample nonparametric confidence intervals for I_E^1 which level can be derived using the Bonferroni inequality.

RESULT 5.2. **[Nonparametric CIs for the Theil index with bounded income distribution]** *Let Y represent the income of a community's households and $F(y) \in \mathcal{L}$ be the distribution function of Y with support $[0, \overline{y}]$. Let $Z = YlogY$ and $H(z)$ be its distribution function, which support is $[\upsilon_1, \upsilon_2]$. Let $Y_{(1)} \leq ... \leq Y_{(n)}$ be n ordered i.i.d. observations on Y and $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the corresponding ordered values of Z. Suppose that the following confidence band for $F(y)$ with level $1 - \alpha_1$ is valid for the space of distributions $\mathcal{L}$:*

$$C_F(\alpha_1) = \left\{ F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y \right\}$$

and that the following confidence band for $H(z)$ with level $1 - \alpha_2$ is valid for the space of distributions $\mathcal{L}$:

$$C_H(\alpha_2) = \left\{ H_0 \in \mathcal{L} : \breve{G}_n^L(z) \leq H_0(z) \leq \breve{G}_n^U(z),\ \forall z \right\}.$$

where $G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(n)}$ and $\breve{G}_n^L(y) \in \mathcal{L}$ and $\breve{G}_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Z_{(1)}, \ldots, Z_{(n)}$.

Then a confidence interval for the Theil index I_E^1 with level greater than or equal to $1 - \alpha$ is:

$$C_{I_E^1}(\eta) = \{ I_0 \in \mathbb{R} : I_{1,\min} \leq I_0 \leq I_{1,\max} \}$$

where

$$I_{1,\min} = \min\left\{\frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

$$I_{1,\max} = \max\left\{\frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

$$\Lambda_{1,\min} = [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)})\right]Y_{(k)},$$

$$\Lambda_{1,max} = [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)})\right]Y_{(k)},$$

$$\Upsilon_{\min} = [1 - \widetilde{H}_n^U(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{H}_n^U(Z_{(k)}) - \widetilde{H}_n^U(Z_{(k-1)})\right]Z_{(k)},$$

$$\Upsilon_{\max} = [1 - \widetilde{H}_n^L(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n}\left[\widetilde{H}_n^L(Z_{(k)}) - \widetilde{H}_n^L(Z_{(k-1)})\right]Z_{(k)},$$

$Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, $Z_{(0)} = \upsilon_1$, $Z_{(n+1)} = \upsilon_2$, $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ *and* $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$, *and* $\forall z$, $\widetilde{H}_n^L(z) = \max\left\{\breve{G}_n^L(z), 0\right\}$ *and* $\widetilde{H}_n^U(y) = \min\left\{\breve{G}_n^U(z), 1\right\}$, *and* $\alpha = \alpha_1 + \alpha_2$.

Result 5.2. allows to build nonparametric CIs for the Theil index using CBs for distribution functions, in particular those that had been cited earlier. The CIs so built have the same properties as those noted for the generalized entropy index.

3.6 Finite-sample confidence intervals for the mean of a random variable

We proposed finite-sample nonparametric CIs for inequality measures that belong to the generalized entropy class of indexes, when households' income is bounded. When Y is not bounded, Bahadur and Savage (1956) show that nonparametric CIs cannot be built

without further information about the distribution of Y. Hence the projection principle used earlier provides CIs for the inequality measures that are too wide to convey any information. To avoid this problem, we assume that the tails of the distribution of Y satisfy some regularity conditions: the rate of decline of each tail is bounded by those of a Pareto distribution. Under this hypothesis, we propose CIs for the mean of a lower bounded random variable–which can be easily extended to an upper bounded random variable– and for the mean of an unbounded random variable. Applying these CIs for the mean, we build CIs for inequality measures with unbounded households' income.

3.6.1 Confidence intervals for the mean of a lower bounded random variable

Let W be a random variable that follows a Pareto distribution $P(w_0, \gamma)$ with density function

$$g(w) = \begin{cases} \frac{\gamma w_0^\gamma}{w^{\gamma+1}} \text{ for } w \geq w_0 \\ 0 \text{ otherwise} \end{cases}$$

and cumulative distribution function

$$G(w) = 1 - \left(\frac{w_0}{w}\right)^\gamma \text{ for } w \geq w_0$$

where $\gamma > 0$ is the shape parameter and $w_0 > 0$ is the scale parameter.

For $k < \gamma$, the moments of order k of W are:

$$E(W^k) = \frac{\gamma w_0^k}{\gamma - k} \tag{3.10}$$

and the mean of W is $E(W) = \frac{\gamma w_0}{\gamma - 1}$ when $\gamma > 1$.

Let Y be a random variable that is lower bounded: $Y \in [\underline{y}, +\infty)$ with continuous distribution function $F(y)$ and mean $E(Y) = \mu$. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of a sample of n i.i.d. observations on Y. To insure that we can build nonparametric CIs for μ, we suppose that the rate of decrease of the right tail of $F(y)$ is bounded by the

rate of decrease of the tail of a Pareto distribution:

HYPOTHESIS (I) *The right tail of* $F(y)$*—the distribution of* Y*— is bounded by a Pareto* $P(w_0, \gamma)$ *distribution of type I with cumulative distribution function*

$$G(w) = 1 - \left(\frac{w_0}{w}\right)^{\gamma} \text{ for } w \geq w_0$$

where $w_0 > 0$ *is the scale parameter and* $\gamma > 1$ *is the shape parameter, i.e.*

$$G(y) \leq F(y) \ \forall y \geq \overline{y}$$

for some threshold $\overline{y} = w_0$.

The mean of Y is:

$$\begin{aligned} \mu &= \int_{\underline{y}}^{+\infty} y \ dF(y) = \int_{\underline{y}}^{\overline{y}} y \ dF(y) + \int_{\overline{y}}^{+\infty} y \ dF(y) \\ &= E[Y \mid \underline{y} \leq Y \leq \overline{y}] \ \Pr(\underline{y} \leq Y \leq \overline{y}) + E[Y \mid Y \geq \overline{y}] \ \Pr(Y \geq \overline{y}) \\ &= I_B \ \Pr(\underline{y} \leq Y \leq \overline{y}) + I_{LB} \ \Pr(Y \geq \overline{y}) \end{aligned} \tag{3.11}$$

where $\overline{y} \in [\underline{y}, +\infty)$, $I_B = E[Y \mid \underline{y} \leq Y \leq \overline{y}]$ is the mean of a bounded random and $I_{LB} = E[Y \mid Y \geq \overline{y}]$. Hence, μ is a weighted sum of the mean of a bounded random variable, I_B, and the mean of an unbounded random variable, I_{LB}, which contains the exploding part of Y. To build CIs for μ, we use CIs for I_B and I_{LB}.

Following Diouf and Dufour (2005b), nonparametric CIs for I_B can be built using CBs for the distribution function of Y_B where $Y_B = Y \mid \underline{y} \leq Y \leq \overline{y}$. Let's define some notation for the remainder of the paper. Denote:

$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$;

$\Gamma(.)$: a functional $\Gamma[F] : \mathcal{L} \to \overline{\mathbb{R}}$ defined on a space $\mathcal{L}$ of functions;

$\mathcal{F}$: a space of distribution functions;

$\widetilde{\mathcal{F}}$: a space of continuous distribution functions.

Let $\mathcal{L}$ be a space of functions such that the Stieltjes integrals $\Gamma[G] = \int_{\underline{y}}^{\overline{y}} y \ dG(y)$ is

finite. Let $\breve{G}_n^L \in \mathcal{L}$ and $\breve{G}_n^U(y) \in \mathcal{L}$ be two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(m)}$ where $m = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)} \leq \overline{y}\right]$. Let suppose that these functions define the following confidence band for $F_{Y_B}(y) = F_{Y|\underline{y} \leq Y \leq \overline{y}}(y)$ with level $1 - \alpha_2$:

$$C_{F_{Y_B}}(\alpha_2) = \left\{ F_0 \in \mathcal{L} : \breve{G}_n^L(y) \leq F_0(y) \leq \breve{G}_n^U(y),\ \forall y \right\}$$

Following Diouf and Dufour (2005b), a nonparametric CI for I_B with level $1 - \alpha_2$ is:

$$\begin{aligned} \widetilde{C}_{I_B}(\alpha_2) \quad = \quad & \left\{ \mu_0 \in \mathbb{R} : [1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)}) \right] Y_{(k)} \leq \mu_0 \right. \\ & \left. \leq \left[1 - \widehat{F}_n^L(Y_{(m)}) \right] \overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)}) \right] Y_{(k)} \right\} \end{aligned} \tag{3.12}$$

where $Y_{(0)} = \underline{y}$, $Y_{(n+1)} = \overline{y}$, $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$, *and* $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}\ \forall y$.

The methodology used before does not apply to I_{LB}. The random variable $Y_{LB} = Y \mid Y \geq \overline{y}$ is not bounded. Hence, following Bahadur and Savage (1956), informative nonparametric CIs cannot be built for I_{LB} without further information about the distribution of Y_{LB}. We provide this information by assuming that Y satisfies the hypothesis (i). Result 6.1. provides a bound for I_{LB} under this hypothesis.

RESULT 6.1. *Let Y be a lower bounded random variable with continuous distribution function $F(y)$ and μ. Let $Y_{LB} = Y \mid Y \geq \overline{y}$ with distribution function $F_{Y_{LB}}(y_{LB})$ and W be a random variable which follows a Pareto distribution $P(\overline{y}, \gamma)$ of type I with cumulative distribution function*

$$G(w) = 1 - \left(\frac{\overline{y}}{w} \right)^{\gamma} \text{ for } w \geq \overline{y}$$

where $\overline{y} > 0$ and $\gamma > 1$. If

$$G(w) \leq F_{Y_{LB}}(w),\ \forall w$$

then,

$$E(Y_{LB}) \leq \frac{\gamma \overline{y}}{\gamma - 1}.$$

Confidence intervals when all parameters are known

Let's define a set of assumptions:

ASSUMPTION 1.1: *Let $\mathcal{L}$ be a space of functions such that the Stieltjes integral $\Gamma[G] = \int_{\underline{y}}^{+\infty} y\, dG(y)$ is finite, $\mathcal{F}_{[\underline{y},+\infty)}$ a space of distribution functions included in $\mathcal{L}$ with support $[\underline{y},+\infty)$ for finite number $\underline{y}$. Let Y be a random variable with distribution function $F(y) \in \mathcal{F}_{[\underline{y},+\infty)}$ such that*

$$F(y) \geq 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma}, \forall y \geq \overline{y}$$

where $\gamma > 1$ and $\overline{y} > 0$ are known and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ the order statistics of a sample of n i.i.d. observations on Y.

ASSUMPTION 2: *Let*

$$C_F(\alpha_1) = \left\{F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\}$$

be a confidence band for $F(y)$ with level $1-\alpha_1$ where $G_n^L \in \mathcal{L}$, $G_n^U \in \mathcal{L}$, and $G_n^L(y)$ and $G_n^U(y)$ are step functions with jumps only at $Y_{(1)}, \cdots, Y_{(n)}$ and

$$C_{F_{Y_B}}(\alpha_2) = \left\{F_0 \in \mathcal{L} : \breve{G}_n^L(y) \leq F_0(y) \leq \breve{G}_n^U(y),\ \forall y\right\}$$

a confidence band for $F_{Y|\underline{y}\leq Y\leq\overline{y}}(y)$ with level $1-\alpha_2$ where $\breve{G}_n^L \in \mathcal{L}$, $\breve{G}_n^U \in \mathcal{L}$, and $\breve{G}_n^L(y)$ and $\breve{G}_n^U(y)$ are step functions with jumps only at $Y_{(1)}, \ldots, Y_{(m)}$ where $m = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)\leq\overline{y}}\right]$.

We use hypothesis (i) and result 6.1. to propose a general methodology to build nonparametric confidence intervals for the mean of a lower bounded random variable..

PROPOSITION 6.2. **[Nonparametric CIs for the mean of a lower bounded random variable with a Pareto-bounded tail of distribution when parameters**

are known] *Let*

$$\mu_L = \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y}) \right] + \overline{y}[1 - \widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \left(\left[1 - \widehat{F}_n^L(Y_{(m)}) \right] \overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y}) \right] + \frac{\gamma \overline{y}}{\gamma - 1}[1 - \widetilde{F}_n^L(\overline{y})],$$

where $Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ *and* $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$ $\forall y$, *and* $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ *and* $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$ $\forall y$. *Under assumptions 1.1 and 2, the following confidence interval for* μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{3.13}$$

has level greater than or equal to $1 - \alpha$ *where* $\alpha = \alpha_1 + \alpha_2$.

Proposition 6.2. proposes a procedure to build nonparametric CIs for the mean of a lower bounded random variable whose right tail is bounded by a Pareto distribution. These CIs are built using nonparametric CBs for $F(y)$ and for the conditional distribution function of $Y \mid \underline{y} \leq Y \leq \overline{y}$, which can be derived from CBs for $F(y)$. For that purpose, the empirical distribution function-based CBs presented in section 5 can be applied. Note that this procedure can be easily extended to upper bounded random variables by considering the opposite variable $-Y$. Also note that the lower bound of any CI for the mean of a bounded random variable defines a one-sided—upper—CI for the mean of a lower bounded random variable. Hence, nonparametric CIs we proposed for the mean of a bounded random variable can be applied to lower bounded random variables for which the mean exists.

CIs proposed by Proposition 6.2. are derived under the hypothesis that the right tail of $F(y)$ is bounded by a Pareto distribution, which provides a good upper bound for I_{LB} but not for the lower bound of the CIs. To improve the performance of CIs, a stronger hypothesis needs to be set:

HYPOTHESIS (II) *The right tail of $F(Y)$—the distribution of Y— is a Pareto $P(w_0, \gamma)$ distribution of type I with cumulative distribution function*

$$G(w) = 1 - \left(\frac{w_0}{w}\right)^{\gamma} \text{ for } w \geq w_0$$

where $w_0 > 0$ is the scale parameter and $\gamma > 1$ is the shape parameter, i.e.

$$F(y) = 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma} \ \forall y \geq \overline{y}$$

for some threshold $\overline{y} = w_0$.

Let's define the following assumption:

ASSUMPTION 1.2: *Let $\mathcal{L}$ be a space of functions such that the Stieltjes integral $\Gamma[G] = \int_{\underline{y}}^{+\infty} y \, dG(y)$ is finite, $\mathcal{F}_{[\underline{y},+\infty)}$ be a space of distribution functions included in $\mathcal{L}$ with support $[\underline{y}, +\infty)$ for finite number $\underline{y}$, and Y be a random variable with distribution function $F(y) \in \mathcal{F}_{[\underline{y},+\infty)}$ such that*

$$F(y) = 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma}, \forall y \geq \overline{y}$$

where $\gamma > 1$ and $\overline{y} > 0$ are known and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ the order statistics of a sample of n i.i.d. observations on Y.

Under hypothesis (ii), the following proposition provides a general expression for nonparametric CIs for the mean of a lower bounded random variable with a Pareto tail.

PROPOSITION 6.3. **[Nonparametric CIs for the mean of a lower bounded**

random variable with a Pareto tail when parameters are known] *Let*

$$\mu_L = \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y}) \right] + \frac{\gamma \overline{y}}{\gamma - 1} [1 - \widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \left(\left[1 - \widehat{F}_n^L(Y_{(m)}) \right] \overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y}) \right] + \frac{\gamma \overline{y}}{\gamma - 1} [1 - \widetilde{F}_n^L(\overline{y})],$$

where $Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ *and* $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$ $\forall y$, *and* $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ *and* $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$ $\forall y$. *Under assumptions 1 and 2, the following confidence interval for* μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{3.14}$$

has level greater than or equal to $1 - \alpha_1 - \alpha_2$.

Confidence Intervals when $\overline{y}$ and γ are unknown

CIs proposed in the last subsection are valid when the parameters $\overline{y}$ and γ of the Pareto distribution are fully known. If those parameters are unknown, CIs must be built for them before considering CIs for μ. We first consider the benchmark case where both the threshold $\overline{y}$ and the shape parameter γ are unknown. Then, we consider the case where $\overline{y}$ is known but the shape parameter γ is unknown and tackle the choice of $\overline{y}$. We study CIs under the hypothesis (ii).

Confidence intervals for $\overline{y}$ and γ

Let W be a random variable that follows a Pareto distribution $P(\overline{y}, \gamma)$ with density

function $g(w)$ such that

$$g(w) = \begin{cases} \frac{\gamma \overline{y}^{\gamma}}{w^{\gamma+1}} \text{ for } w \geq \overline{y} \\ 0 \text{ otherwise} \end{cases}$$

where $\overline{y} > 0$ and $\gamma > 1$. Let $W_{(1)}, \ldots, W_{(n)}$ be the order statistics of a sample of n i.i.d. observations on W. Chen (1996) proposed the following joint confidence region for the $\overline{y}$ and γ with level $1 - \alpha$:

$$C_{\overline{y},\gamma}(\alpha) = \left\{(\overline{y}_0, \gamma_0) \in \mathbb{R}^2 : \overline{y}_l \leq \overline{y}_0 \leq \overline{y}_u \text{ and } \gamma_l(\overline{y}_0) \leq \gamma_0 \leq \gamma_u(\overline{y}_0)\right\} \tag{3.15}$$

where for $3 \leq k \leq n$,

$$\overline{y}_l = W_{(1)} \exp\left(\frac{\sum_{i=2}^{k-1} \ln\left(\frac{W_{(1)}}{W_{(i)}}\right) + (n-k+1)\ln\left(\frac{W_{(1)}}{W_{(k)}}\right)}{n(k-1)F_{\frac{1+\sqrt{1-\alpha}}{2}}(2k-2,2)} \right),$$

$$\overline{y}_u = W_{(1)} \exp\left\{ \frac{\sum_{i=2}^{k-1} \ln\left(\frac{W_{(1)}}{W_{(i)}}\right) + (n-k+1)\ln\left(\frac{W_{(1)}}{W_{(k)}}\right)}{n(k-1)F_{\frac{1-\sqrt{1-\alpha}}{2}}(2k-2,2)} \right\},$$

$$\gamma_l(\overline{y}_0) = \frac{-\chi^2_{\frac{1+\sqrt{1-\alpha}}{2}}(2k)}{2\left[n \ \ln \overline{y}_0 - \sum_{i=1}^{k-1} \ln W_{(i)} - (n-k+1)\ln W_{(k)}\right]},$$

$$\gamma_u(\overline{y}_0) = \frac{-\chi^2_{\frac{1-\sqrt{1-\alpha}}{2}}(2k)}{2\left[n \ \ln \overline{y}_0 - \sum_{i=1}^{k-1} \ln W_{(i)} - (n-k+1)\ln W_{(k)}\right]}$$

where $F_p(\eta_1, \eta_2)$ is the p^{th} percentile of the Fisher distribution with η_1 and η_2 degrees of freedom and $\chi^2_p(\eta)$ is the p^{th} percentile of the χ^2 distribution with η degrees of freedom.

Using the Chen's joint confidence region for $\overline{y}_0$ and γ, we propose the following one-sided and two-sided CIs for γ.

COROLLARY 6.4. **[One-sided CI for the shape parameter γ]** *Let W be a random*

variable with a Pareto distribution $P(\overline{y}_0, \gamma)$ *with density function* $g(w)$ *such that*

$$g(w) = \begin{cases} \frac{\gamma \overline{y}^{\gamma}}{w^{\gamma+1}} \text{ for } w \geq \overline{y} \\ 0 \text{ otherwise} \end{cases}$$

where $\overline{y} > 0$ *is known and* $\gamma > 0$*. Let* $W_{(1)}, \ldots, W_{(n)}$ *be the order statistics of a sample of* n *i.i.d. observations on* W*. For any* $k \in [3, n]$*, an upper one-sided confidence interval for* γ *with level* $1 - \alpha$ *is*

$$C_{\gamma}(\alpha) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0\}$$

where

$$\gamma_l = \frac{-\chi^2_{\alpha}(2k)}{2\left[\sum_{i=1}^{k-1} ln\left(\frac{\overline{y}}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{\overline{y}}{W_{(k)}}\right)\right]}$$

where $\chi^2_p(\eta)$ *is the* p^{th} *percentile of the* χ^2 *distribution with* η *degrees of freedom.*

COROLLARY 6.5. **[Two-sided CI for the shape parameter** γ**]** *Let* W *be a random variable with a Pareto distribution* $P(\overline{y}, \gamma)$ *with density function* $g(w)$ *such that*

$$g(w) = \begin{cases} \frac{\gamma \overline{y}^{\gamma}}{w^{\gamma+1}} \text{ for } w \geq \overline{y} \\ 0 \text{ otherwise} \end{cases}$$

where $\overline{y} > 0$ *is known and* $\gamma > 0$*. Let* $W_{(1)}, \ldots, W_{(n)}$ *be the order statistics of a sample of* n *i.i.d. observations on* W*.*

For any $k \in [3, n]$*, a confidence interval for* γ *with level* $1 - \alpha$ *is*

$$C_{\gamma}(\alpha) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

where

$$\gamma_l = \frac{-\chi^2_{\frac{\alpha}{2}}(2k)}{2\left[\sum_{i=1}^{k-1} ln\left(\frac{\overline{y}}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{\overline{y}}{W_{(k)}}\right)\right]}$$

and

$$\gamma_u = \frac{-\chi^2_{1-\frac{\alpha}{2}}(2k)}{2\left[\sum_{i=1}^{k-1} ln\left(\frac{\overline{y}}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{\overline{y}}{W_{(k)}}\right)\right]}$$

where $\chi^2_p(\eta)$ is the p^{th} percentile of the χ^2 distribution with η degrees of freedom.

Confidence intervals when $\overline{y}$ and γ are unknown

In this subsection, we propose CIs for μ when $\overline{y}$ and γ are unknown. These CIs are benchmark. We discuss later the challenges of their implementation.

PROPOSITION 6.6. **[Nonparametric CI for the mean of a lower bounded random variable with a Pareto tail when $\overline{y}$ and γ are unknown]** *Let*

$$C_{\overline{y},\gamma}(\alpha_3) = \left\{(\overline{y}_0, \gamma_0) \in \mathbb{R}^2 : \overline{y}_l \leq \overline{y}_0 \leq \overline{y}_u \text{ and } \gamma_l \leq \gamma_0 \leq \gamma_u\right\}$$

be a confidence region for $\overline{y}$ and γ with level $1-\alpha_3$ and

$$\mu_L = \left([1-\widehat{F}^U_n(Y_{(m)})]\overline{y_l} + \sum_{k=1}^{m}\left[\widehat{F}^U_n(Y_{(k)}) - \widehat{F}^U_n(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}^L_n(\overline{y_l}) - \widetilde{F}^U_n(\underline{y})\right] + \frac{\overline{y_l}}{1-\frac{1}{\gamma_u}}[1-\widetilde{F}^U_n(\overline{y}_u)],$$

$$\mu_U = \left(\left[1-\widehat{F}^L_n(Y_{(m)})\right]\overline{y}_u + \sum_{k=1}^{m}\left[\widehat{F}^L_n(Y_{(k)}) - \widehat{F}^L_n(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}^U_n(\overline{y}_u) - \widetilde{F}^L_n(\underline{y})\right] + \frac{\overline{y}_u}{1-\frac{1}{\gamma_l}}[1-\widetilde{F}^L_n(\overline{y}_l)],$$

where $Y_{(0)} = \underline{y}$, $\widetilde{F}^L_n(y) = \max\{G^L_n(y), 0\}$ and $\widetilde{F}^U_n(y) = \min\{G^U_n(y), 1\}$ $\forall y$, and $\widehat{F}^L_n(y) = \max\{\breve{G}^L_n(y), 0\}$ and $\widehat{F}^U_n(y) = \min\{\breve{G}^U_n(y), 1\}$ $\forall y$. Under assumptions 1.2 and 2, the following confidence interval for μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\}$$

has level greater than or equal to $1-\alpha$ *where* $\alpha = \alpha_1 + \alpha_2 + \alpha_3$.

To build this CI, the sample $Y_{(1)}, \cdots, Y_{(n)}$ must be split into two parts: one—$Y_{(1)}, \ldots, Y_{(m)}$—on which the parameters will be inferred and another—$Y_{(m+1)}, \cdots, Y_{(n)}$—on which CIs for I_B will be built. The choice of m raises similar questions to the choice of $\overline{y}$, in the case where $\overline{y}$ is predetermined.

Confidence intervals when γ is unknown

A more realistic case than the benchmark one is the case where is where γ is unknown but the threshold $\overline{y}$ from which $F(y)$ becomes a Pareto distribution is known. In such case, $\overline{y}$ must be chosen prior to the computation of the CIs for μ. Let $m = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)\leq\overline{y}}\right]$ and $n_1 = n - m$. The m first observations of Y are used to build CIs for I_B while the n_1 remaining observations are used to build the CI for γ. $\overline{y}$ must be large enough to minimize the distortion on the distribution of Y. But, if $\overline{y}$ is too large, the mass of probability in the tail of the distribution of Y is small. Hence, the number of observations larger than $\overline{y}$ (n_1) may be too small to provide a CI for γ which performance won't alter that of the CI for μ. Hence, the performance of the CIs for γ is likely to deliver a poor performance, which alters the performance of the CIs for μ. The more n is large, the more n_1 can be chosen large without altering the performance of I_B too. Moreover, note that the choice of $\overline{y}$ determines the relative weight of I_B and I_{LB}. CIs for I_B are built using CBs for distribution functions and projection techniques while CIs for I_{LB} are built using CIs for γ. Putting more weight on the part of μ for which a more accurate CI is achieved improves the performance of the overall inference.

In our Monte Carlo simulations, $F(y)$ is known. In this case, a convenient idea is to set $\overline{y}$ equal to a percentile of $F(y)$, i.e. $\overline{y} = F^{-1}(p)$ where $p \in (0,1)$. Doing so allows to control the probability mass in the tail of the distribution. We recommend to choose $p \geq 0.95$ so as to alter the least possible the real distribution of Y.

If $F(y)$ is unknown, one can set $\overline{y}$ arbitrarily by choosing the subsample of Y that will be used for the inference on γ. In this subsection, we study CIs for μ when the shape parameter γ is larger than 1 and unknown but the threshold $\overline{y}$ from which $F(y)$ becomes

a Pareto distribution is fully known.

Proposition 6.7. proposes nonparametric CIs for μ when $\overline{y}$ is predetermined and γ is unknown.

PROPOSITION 6.7. **[Nonparametric CI for the mean of a lower bounded random variable with a Pareto tail when γ is unknown]** *Let*

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

be a confidence interval for γ with level $1-\alpha_3$ and

$$\mu_L = \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y}) \right] + \frac{\overline{y}}{1 - \frac{1}{\gamma_u}} [1 - \widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \left(\left[1 - \widehat{F}_n^L(Y_{(m)}) \right] \overline{y} + \sum_{k=1}^{m} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)}) \right] Y_{(k)} \right) \left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y}) \right] + \frac{\overline{y}}{1 - \frac{1}{\gamma_l}} [1 - \widetilde{F}_n^L(\overline{y})],$$

where $Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ and $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$ $\forall y$, and $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ and $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$ $\forall y$. Under assumptions 1.2 and 2, the following confidence interval for μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\}$$

has level greater than or equal to $1-\alpha$ where $\alpha = \alpha_1 - \alpha_2 - \alpha_3$.

3.6.2 Confidence interval for the mean of an unbounded random variable

In this section, we propose a generalization of the CIs for the mean of a lower bounded random variable to all random variables. This procedure applies to the left tail of $F(y)$ similar techniques as those applied to the right tail of $F(y)$.

Let Y be a random variable with continuous distribution function $F(y)$ and mean $E(Y) = \mu$. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of a sample of n i.i.d. observations on Y. Rewriting the mean of Y, we split it into three parts that involve the means of three different random variables:

$$\begin{aligned} E(Y) &= \int_{-\infty}^{+\infty} y \; dF(y) \\ &= \int_{-\infty}^{\underline{y}} y \; dF(y) + \int_{\underline{y}}^{\overline{y}} y \; dF(y) + \int_{\overline{y}}^{+\infty} y \; dF(y) \end{aligned}$$

$$\Leftrightarrow E(Y) = E[Y \mid Y \leq \underline{y}] \;\Pr(Y \leq \underline{y}) + E[Y \mid \underline{y} \leq Y \leq \overline{y}] \;\Pr(\underline{y} \leq Y \leq \overline{y})$$
$$+E[Y \mid Y \geq \overline{y}] \;\Pr(Y \geq \overline{y})$$

$$\Leftrightarrow E(Y) = I_{UB} \Pr(Y \leq \underline{y}) + I_B \Pr(\underline{y} \leq Y \leq \overline{y}) + I_{LB} \Pr(Y \geq \overline{y})$$

where I_B is the mean of a bounded random variable $Y_B = Y \mid \underline{y} \leq Y \leq \overline{y}$, I_{UB} is the mean of an upper-bounded random variable $Y_{UB} = Y \mid Y \leq \overline{y}$, and I_{LB} is the mean of a lower-bounded random variable $Y_{LB} = Y \mid Y \geq \underline{y}$.

Following Bahadur and Savage (1956), additional information about the distribution of Y_{LB} and Y_{UB} is needed to build nonparametric CIs for I_{UB} and I_{LB}. We provide this information by assuming that the tails of $F(y)$ are bounded by Pareto distributions as stated by the following hypotheses. Hypotheses (i) and (ii) are the same as before; we remind them here while hypotheses (iii) and (iv) relates to the left tail of $F(y)$.

HYPOTHESIS (I) *The right tail of $F(y)$ is bounded by a Pareto $P(\overline{y}, \gamma)$ distribution*

of type I: with cumulative distribution function

$$F(y) \geq 1 - \left(\frac{\overline{y}}{w}\right)^{\gamma} \ \forall y \geq \overline{y}$$

where $\overline{y} > 0$ *is the scale parameter and* $\gamma > 1$ *is the shape parameter*

HYPOTHESIS (II) *The right tail of* $F(y)$ *is a Pareto* $P(\overline{y}, \gamma)$ *distribution of type I:*

$$F(y) = 1 - \left(\frac{\overline{y}}{w}\right)^{\gamma} \ \forall y \geq \overline{y}$$

where $\overline{y} > 0$ *is the scale parameter and* $\gamma > 1$ *is the shape parameter.*

Let $\hat{W}$ be a random variable with Pareto distribution of type I $P(\varpi_0, \rho)$ where $\varpi_0 > 0$ and $\rho > 1$. The random variable $-\hat{W}$ follows a negative Pareto distribution with distribution $\widehat{P}(-\varpi_0, \rho)$ and mean $-E(\hat{W}) = -\frac{\rho\varpi_0}{\rho-1}$.

HYPOTHESIS (III) *The left tail of* $F(y)$ *is bounded by a negative Pareto distribution of type I* $\widehat{P}(-\underline{y}, \rho)$ *with distribution* $\widehat{G}(y)$ *where* $-\underline{y} > 0$ *and* $\rho > 1$*, i.e.*

$$\widehat{G}(y) \geq F(y) \ \forall y \leq \underline{y}.$$

HYPOTHESIS (IV) *The left tail of* $F(y)$ *is a negative Pareto distribution of type I* $\widehat{P}(-\underline{y}, \rho)$ *with distribution* $\widehat{G}(y)$ *where* $-\underline{y} > 0$ *and* $\rho > 1$*, i.e.*

$$F(y) = \widehat{G}(y) \ \forall y \leq \underline{y}.$$

In the last section, we have proposed nonparametric CIs for I_B using nonparametric CBs for distribution functions and projection techniques. We have also proposed nonparametric CIs for I_{LB} in the case where $F(y)$ satisfies the hypothesis (i) or the hypothesis (ii). To build CIs for μ, we study CIs for I_{UB} based on the same procedure as those used to build CIs for I_{LB}, under the assumption that Y satisfies the hypothesis (iii) or the hypothesis (iv).

Under hypothesis (iii), the mean of the random variable $Y_{UB} = Y \mid Y \leq \overline{y}$ is lower bounded:

$$-E(\hat{W}) \leq E(Y_{UB}).$$

Given that $E(\hat{W}) = \frac{\rho\varpi_0}{\rho-1} = \frac{-\rho\underline{y}}{\rho-1}$, then

$$\frac{\rho\underline{y}}{\rho-1} \leq E(Y_{UB}). \tag{3.16}$$

Let's define the following assumptions:

ASSUMPTION 3.1: *Let $\mathcal{L}$ be a space of functions such that the Stieltjes integral $\Gamma[G] = \int_{-\infty}^{+\infty} y\, dG(y)$ is finite, $\mathcal{F}$ be a space of distribution functions included in $\mathcal{L}$, and Y be a random variable with distribution function $F(y) \in \mathcal{F}$ such that*

$$F(y) \geq 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma} \forall y \geq \overline{y}$$

where $\gamma > 1$ and $\overline{y} > 0$ are known and $F(y) \leq \widehat{G}_{-\underline{y},\rho}(y)\ \forall y \leq \underline{y}$ where $\widehat{G}_{-\underline{y},\rho}(y)$ is the cumulative distribution function of a negative Pareto $\widehat{P}(-\underline{y},\rho)$ distribution where $-\underline{y} > 0$ is the scale parameter and $\rho > 1$ is the shape parameter. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of a sample of n i.i.d. observations on Y.

ASSUMPTION 4: *Let*

$$C_F(\alpha_1) = \left\{F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\}$$

be a confidence band for $F(y)$ with level $1-\alpha_1$ where $G_n^L \in \mathcal{L}$, $G_n^U \in \mathcal{L}$, and $G_n^L(y)$ and $G_n^U(y)$ are step functions with jumps only at $Y_{(1)}, \cdots, Y_{(n)}$ and

$$C_{F_{Y_B}}(\alpha_2) = \left\{F_0 \in \mathcal{L} : \breve{G}_n^L(y) \leq F_0(y) \leq \breve{G}_n^U(y),\ \forall y\right\}$$

a confidence band for $F_{Y|\underline{y}\leq Y\leq\overline{y}}(y)$ with level $1-\alpha_2$ where $\breve{G}_n^L \in \mathcal{L}$, $\breve{G}_n^U \in \mathcal{L}$, and $\breve{G}_n^L(y)$ and $\breve{G}_n^U(y)$ are step functions with jumps only at $Y_{(m_1)}, \ldots, Y_{(m_2)}$ where $m_1 = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)} \leq \underline{y}\right]$ and $m_2 = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)} \leq \overline{y}\right]$.

Using equation *(3.16)* and the former results on I_B and I_{LB}, we can built CIs for μ in the following proposition, under the hypothesis that all parameters are known.

PROPOSITION 6.8. **[Nonparametric CIs for the mean of a random variable with Pareto-bounded tails of distribution when parameters are known]** *Let*

$$\mu_L = \frac{\rho \underline{y}}{\rho - 1}\widetilde{F}_n^L(\underline{y}) + \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=m_1}^{m_2} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right] Y_{(k)} \right) \left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right] + \overline{y}[1 - \widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \underline{y}\widetilde{F}_n^L(\underline{y}) + \left(\left[1 - \widehat{F}_n^L(Y_{(m)})\right] \overline{y} + \sum_{k=m_1}^{m_2} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right] Y_{(k)} \right) \left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right] + \frac{\gamma \overline{y}}{\gamma - 1}[1 - \widetilde{F}_n^L(\overline{y})],$$

$Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ *and* $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$ $\forall y$, *and* $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ *and* $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$ $\forall y$. *Under assumptions 3.1 and 4, the following confidence interval for* μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{3.17}$$

has level greater than or equal to $1 - \alpha$ *where* $\alpha = \alpha_1 + \alpha_2$.

Similar CIs can be derived for μ under hypothesis (iv). Under this hypothesis, the mean of Y_{UB} is:

$$E(Y_{UB}) = -E(\hat{W}) = \frac{\rho \underline{y}}{\rho - 1}.$$

Let's define the following assumption:

ASSUMPTION 3.2: *Let* $\mathcal{L}$ *be a space of functions such that the Stieltjes integral* $\Gamma[G] = \int_{-\infty}^{+\infty} y \, dG(y)$ *is finite,* $\mathcal{F}$ *be a space of distribution functions included in* $\mathcal{L}$, *and* Y *be a*

random variable with distribution function $F(y) \in \mathcal{F}$ *such that*

$$F(y) = 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma}, \forall y \geq \overline{y}$$

where $\gamma > 1$ *and* $\overline{y} > 0$ *are known and* $F(y) = \widehat{G}_{-\underline{y},\rho}(y)\ \forall y \leq \underline{y}$ *where* $\widehat{G}_{-\underline{y},\rho}(y)$ *is the cumulative distribution function of a negative Pareto* $\widehat{P}(-\underline{y}, \rho)$ *distribution where* $-\underline{y} > 0$ *and* $\rho > 1$ *are known. Let* $Y_{(1)} \leq \cdots \leq Y_{(n)}$ *be the order statistics of a sample of* n *i.i.d. observations on* Y.

PROPOSITION 6.9. **[Nonparametric CIs for the mean of a random variable with Pareto tails of distribution when parameters are known]** *Let*

$$\mu_L = \frac{\rho\underline{y}}{\rho - 1}\widetilde{F}_n^L(\underline{y}) + \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=m_1}^{m_2}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right]$$
$$+ \frac{\gamma\overline{y}}{\gamma - 1}[1 - \widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \frac{\rho\underline{y}}{\rho - 1}\widetilde{F}_n^U(\underline{y}) + \left(\left[1 - \widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=m_1}^{m_2}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right]$$
$$+ \frac{\gamma\overline{y}}{\gamma - 1}[1 - \widetilde{F}_n^L(\overline{y})],$$

$Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ *and* $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}\ \forall y$, *and* $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ *and* $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}\ \forall y$. *Under assumptions 3.2 and 4, the following confidence interval for* μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{3.18}$$

has level greater than or equal to $1 - \alpha$ *where* $\alpha = \alpha_1 + \alpha_2$.

When the parameters of the Pareto distributions are known, Proposition 6.10 allows to build nonparametric CIs for the mean of a random variable when the tails of the involved distribution are bounded by Pareto distributions or are Pareto distributions. When

parameters are not fully known, CIs for these parameters can be built using Corollary 6.4. and Corollary 6.5. and used to build CIs for μ. The following proposition provides a CI for μ under hypotheses (ii) and (iv) when $\overline{y}$ and $\underline{y}$ are known but γ and ρ are not.

PROPOSITION 6.10. **[Nonparametric CIs for the mean of a random variable with Pareto tails of distribution when the shape parameters are unknown]** *Let*

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

be a confidence interval for γ with level $1-\alpha_3$, and

$$C_\rho(\alpha_4) = \{\rho_0 \in \mathbb{R} : \rho_l \leq \rho_0 \leq \rho_u\}$$

a confidence interval for ρ with level $1-\alpha_4$. Let

$$\mu_L = \frac{\underline{y}}{1-\frac{1}{\rho_l}}\widetilde{F}_n^L(\underline{y}) + \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=m_1}^{m_2}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_u}}[1-\widetilde{F}_n^U(\overline{y})],$$

$$\mu_U = \frac{\underline{y}}{1-\frac{1}{\rho_u}}\widetilde{F}_n^U(\underline{y}) + \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=m_1}^{m_2}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_l}}[1-\widetilde{F}_n^L(\overline{y})],$$

$Y_{(0)} = \underline{y}$, $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$ and $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$ $\forall y$, and $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$ and $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$ $\forall y$. Under assumptions 3.2 and 4, the following confidence interval for μ

$$\widetilde{C}_\mu(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\} \tag{3.19}$$

has level greater than or equal to $1-\alpha$ where $\alpha = \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4$.

3.7 Application to inequality measures when income is positive

Let Y be a positive random variable with distribution function $F(y)$ which represents the income of a community's households. Let $Z = h(Y)$ with distribution function $H(z)$. Let $\mathcal{L}$ be a space of functions such that the Stieltjes integrals $\Lambda_\delta(F) = \int_{-\infty}^{+\infty} y^\delta \, dF(y)$, $\delta \neq 0$ and $\Upsilon(F) = \int_{-\infty}^{+\infty} z \, dH(z)$ are finite, $\mathcal{F}$ be a space of distribution functions included in $\mathcal{L}$. Let's assume that $F(y)$ satisfies one the two following hypotheses:

HYPOTHESIS (I.I) *The right tail of $F(y)$ is bounded by a Pareto $P(\overline{y}, \gamma)$ distribution of type I:*

$$F(y) \geq 1 - \left(\frac{\overline{y}}{y}\right)^\gamma \ \forall y \geq \overline{y}$$

where $\overline{y} > 0$ and $\gamma > 1$ are known.

HYPOTHESIS (I.II) *The right tail of $F(y)$ is a Pareto $P(\overline{y}, \gamma)$ distribution of type I:*

$$F(y) = 1 - \left(\frac{\overline{y}}{y}\right)^\gamma \ \forall y \geq \overline{y}$$

where $\overline{y} > 0$ and $\gamma > 1$ are known.

Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of a sample of n i.i.d. observations on Y and

$$C_F(\alpha_1) = \left\{F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y), \ \forall y\right\}$$

be a confidence band for $F(y)$ with level $1 - \alpha_1$ where $G_n^L \in \mathcal{L}$, $G_n^U \in \mathcal{L}$, and $G_n^L(y)$ and $G_n^U(y)$ are step functions with jumps only at $Y_{(1)}, \ldots, Y_{(n)}$. Let

$$C_{F_{Y_B}}(\alpha_2) = \left\{F_0 \in \mathcal{L} : \breve{G}_n^L(y) \leq F_0(y) \leq \breve{G}_n^U(y), \ \forall y\right\}$$

be a confidence band for $F_{Y|0 \leq Y \leq \overline{y}}(y)$ with level $1 - \alpha_2$ where $\breve{G}_n^L \in \mathcal{L}$, $\breve{G}_n^U \in \mathcal{L}$, and $\breve{G}_n^L(y)$ and $\breve{G}_n^U(y)$ are step functions with jumps only at $Y_{(1)}, \ldots, Y_{(m)}$ where $m = \sum_{k=1}^{n} \mathbb{1}\left[Y_{(k)} \leq \overline{y}\right]$.

Given that Y is positive, Y^δ is also positive. Following Propositions 6.2. and 6.3., a confidence interval for $\Lambda_\delta(F)$ with level greater than or equal to $1-\alpha$ where $\alpha = \alpha_1 + \alpha_2$ is:

$$\widetilde{C}_{\Lambda_\delta}(\alpha) = \{\Lambda_0 \in \mathbb{R} : \Lambda_{\delta,\min} \leq \Lambda_0 \leq \Lambda_{\delta,\max}\}$$

where

- under hypothesis (i.i):

$$\begin{aligned}\Lambda_{\delta,\min} &= \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y}^\delta + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}^\delta\right)\left[\widetilde{F}_n^L(\overline{y}^\delta) - \widetilde{F}_n^U(\underline{y}^\delta)\right] \\ &\quad + \overline{y}^\delta[1-\widetilde{F}_n^U(\overline{y}^\delta)];\end{aligned}$$

$$\begin{aligned}\Lambda_{\delta,\max} &= \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y}^\delta + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}^\delta\right)\left[\widetilde{F}_n^U(\overline{y}^\delta) - \widetilde{F}_n^L(\underline{y}^\delta)\right] \\ &\quad + \frac{\gamma\overline{y}^\delta}{\gamma-1}[1-\widetilde{F}_n^L(\overline{y}^\delta)];\end{aligned}$$

- under hypothesis (i.ii):

$$\begin{aligned}\Lambda_{\delta,\min} &= \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y}^\delta + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}^\delta\right)\left[\widetilde{F}_n^L(\overline{y}^\delta) - \widetilde{F}_n^U(\underline{y}^\delta)\right] \\ &\quad + \frac{\gamma\overline{y}^\delta}{\gamma-1}[1-\widetilde{F}_n^U(\overline{y}^\delta)];\end{aligned}$$

$$\begin{aligned}\Lambda_{\delta,\max} &= \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y}^\delta + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}^\delta\right)\left[\widetilde{F}_n^U(\overline{y}^\delta) - \widetilde{F}_n^L(\underline{y}^\delta)\right] \\ &\quad + \frac{\gamma\overline{y}^\delta}{\gamma-1}[1-\widetilde{F}_n^L(\overline{y}^\delta)];\end{aligned}$$

$Y_{(0)} = \underline{y} = 0$; and $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$, $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$, $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$, and $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$.

Setting $\delta = 1$ in the equations above provides expressions for CIs for Λ_1. Using these CIs and CIs for $E(Z) = E[h(Y)]$ for adequate choice of $h(Y)$, we can propose CIs for the most popular inequality measures.

3.7.1 Confidence intervals for the class of generalized entropy index when $\delta \neq 0, 1$

The generalized entropy measure is

$$I_E^\delta(y) = \int \frac{1}{\delta(\delta-1)}[(\frac{y}{\mu})^\delta - 1]\ dF(y)$$

where $\delta \neq 0, 1$, and $\mu = \int y\ dF(y)$. We previously showed that this measure can be rewritten as follows:

$$I_E^\delta(y) = \frac{1}{\delta(\delta-1)}\left(\frac{\Lambda_\delta(F)}{\Lambda_1^\delta(F)} - 1\right)$$

where $\Lambda_\delta(F) = \int y^\delta\ dF(y)$ is the non centered moment of order δ of Y and $\Lambda_1(F) = \int y\ dF(y)$ is the mean of Y. Using the previous results, if $F(y)$ satisfies one hypothesis (i.i) or (i.ii), a nonparametric CI for I_E^δ is:

$$C_{I_E^\delta}(\alpha) = \{I_0 \in \mathbb{R} : I_{\delta,\min} \leq I_0 \leq I_{\delta,\max}\}$$

where

$$I_{\delta,\min} = \frac{1}{\delta(\delta-1)} \min_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{\frac{\Lambda_\delta}{\Lambda_1^\delta} - 1\right\},$$

$$I_{\delta,\max} = \frac{1}{\delta(\delta-1)} \max_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{\frac{\Lambda_\delta}{\Lambda_1^\delta} - 1\right\},$$

and $\Lambda_{\delta,\min}$ and $\Lambda_{\delta,\max}$ are defined above for hypotheses (i.i) and (i.ii). $C_{I_E^\delta}(\alpha)$ is of level greater than or equal to $1 - \alpha$ where $\alpha = \alpha_1 + \alpha_2$.

Let's assume now that Y satisfies hypotheses (i.ii) where γ is unknown. Let

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

be a confidence interval for γ with level $1-\alpha_3$. Then, following Proposition 6.7., a nonparametric CI for I_E^δ with level greater than or equal to $1-\alpha = 1-\alpha_1-\alpha_2-\alpha_3$ is

$$C_{I_E^\delta}(\alpha) = \{I_0 \in \mathbb{R} : I_{\delta,\min} \leq I_0 \leq I_{\delta,\max}\}$$

where

$$I_{\delta,\min} = \frac{1}{\delta(\delta-1)} \min_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{ \frac{\Lambda_\delta}{\Lambda_1^\delta} - 1 \right\},$$

$$I_{\delta,\max} = \frac{1}{\delta(\delta-1)} \max_{\Lambda_{\delta,\min} \leq \Lambda_\delta \leq \Lambda_{\delta,\max}} \left\{ \frac{\Lambda_\delta}{\Lambda_1^\delta} - 1 \right\},$$

$$\Lambda_{\delta,\min} = \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y}^\delta + \sum_{k=1}^{m} \left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)}) \right] Y_{(k)}^\delta \right) \left[\widetilde{F}_n^L(\overline{y}^\delta) - \widetilde{F}_n^U(\underline{y}^\delta) \right] + \frac{\overline{y}^\delta}{1-\frac{1}{\gamma_u}} [1-\widetilde{F}_n^U(\overline{y}^\delta)],$$

$$\Lambda_{\delta,\max} = \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y}^\delta + \sum_{k=1}^{m} \left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)}) \right] Y_{(k)}^\delta \right) \left[\widetilde{F}_n^U(\overline{y}^\delta) - \widetilde{F}_n^L(\underline{y}^\delta) \right] + \frac{\overline{y}^\delta}{1-\frac{1}{\gamma_l}} [1-\widetilde{F}_n^L(\overline{y}^\delta)],$$

$Y_{(0)} = \underline{y} = 0$, and $\forall y$ $\widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$, $\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$, $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$, and $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$.

3.7.2 Confidence intervals for the Theil index

The Theil index is:

$$I_E^1 = \int \frac{y}{\mu} \log(\frac{y}{\mu}) \, dF(y)$$

where $\mu = \int y \, dF(y)$. We showed that rewriting the Theil index allows to express it as a function of the means of two random variables:

$$I_E^1 = \frac{\Upsilon(F)}{\Lambda_1(F)} - log(\Lambda_1(F))$$

where $\Lambda_1(F) = \mu = \int y\, dF(y) \neq 0$ is the mean of Y—which is positive— and $\Upsilon(F) = \int y\, log(y)\, dF(y)$ is the mean of $Z = Y\, log(Y)$−which belongs to the interval $\left[-\frac{1}{e}, +\infty\right)$. We have proposed nonparametric CIs for $\Lambda_1(F)$ using inference techniques for lower bounded random variables proposed in subsection 6.1. Using the same techniques, we can build CIs for $\Upsilon(F)$ and derive from these nonparametric CIs for I_E^1. The level of the corresponding CI for I_E^1 is computed using the Bonferroni inequality.

Let $H(z)$ be the distribution function of $Z = Y \log Y$. Let's assume that $H(z)$ satisfies one of the following hypotheses:

RESULT 7.1. *If $F(y)$ satisfies hypothesis (i.i) then*

$$E\left(Z \mid Z \geq \overline{z}\right) \leq \frac{\gamma\overline{y}\log\overline{y}}{\gamma - 1} + \frac{\gamma\overline{y}}{(\gamma-1)^2}$$

where $\overline{z} = \max\left\{0, \overline{y}\log\overline{y}\right\}$.

RESULT 7.2. *If $F(y)$ satisfies hypothesis (i.ii) then*

$$E\left(Z \mid Z \geq \overline{z}\right) = \frac{\gamma\overline{y}\log\overline{y}}{\gamma - 1} + \frac{\gamma\overline{y}}{(\gamma-1)^2}$$

where $\overline{z} = \max\left\{0, \overline{y}\log\overline{y}\right\}$.

Let $Z_{(1)}, \ldots, Z_{(n)}$ be the ordered values of Z corresponding to the sample from Y. Let $C_n^L(z) \in \mathcal{L}$ and $C_n^U(z) \in \mathcal{L}$ be two step functions with jumps only at $Z_{(1)}, \ldots, Z_{(n)}$. Let

$$C_H(\beta_1) = \left\{H_0 \in \mathcal{L} : C_n^L(z) \leq H_0(z) \leq C_n^U(z),\ \forall z\right\}.$$

be a confidence band for $H(z)$ with level $1 - \beta_1$.

Let

$$C_{H_{Z_B}}(\beta_2) = \left\{F_0 \in \mathcal{L} : \check{C}_n^L(y) \leq F_0(y) \leq \check{C}_n^U(y),\ \forall y\right\}$$

be a confidence band for $H_{Z|\underline{z}\leq Z\leq \overline{z}}(z)$ with level $1-\beta_2$ where $\check{C}_n^L \in \mathcal{L}$, $\check{C}_n^U \in \mathcal{L}$, and $\check{C}_n^L(y)$ and $\check{C}_n^U(y)$ are step functions with jumps only at $Z_{(1)}, \ldots, Z_{(\widetilde{m})}$ where $\widetilde{m} = \sum_{k=1}^{n} \mathbb{1}\left[Z_{(k)\leq \overline{z}}\right]$ and $\overline{z} > 1$.

Following Propositions 6.2. and 6.3., a confidence interval for $\Upsilon(F)$ with level greater than or equal to $1-\beta_1-\beta_2$ is:

$$\widetilde{C}_\Upsilon(\beta) = \{\Upsilon_0 \in \mathbb{R} : \Upsilon_{\min} \leq \Upsilon_0 \leq \Upsilon_{\max}\}$$

where

- under hypothesis (i.i):

$$\Upsilon_{\min} = \left([1-\widehat{H}_n^U(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widehat{H}_n^U(Z_{(k)}) - \widehat{H}_n^U(Z_{(k-1)})\right]Z_{(k)}\right)\left[\widetilde{H}_n^L(\overline{z}) - \widetilde{H}_n^U(\underline{z})\right] + \overline{z}[1-\widetilde{H}_n^U(\overline{z})],$$

$$\Upsilon_{\max} = \left([1-\widehat{H}_n^L(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widehat{H}_n^L(Z_{(k)}) - \widehat{H}_n^L(Z_{(k-1)})\right]Z_{(k)}\right)\left[\widetilde{H}_n^U(\overline{z}) - \widetilde{H}_n^L(\underline{z})\right] + \left[\frac{\gamma\overline{y}\log\overline{y}}{\gamma-1} + \frac{\gamma\overline{y}}{(\gamma-1)^2}\right][1-\widetilde{H}_n^L(\overline{z})],$$

- under hypothesis (i.ii):

$$\Upsilon_{\min} = \left([1-\widehat{H}_n^U(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widehat{H}_n^U(Z_{(k)}) - \widehat{H}_n^U(Z_{(k-1)})\right]Z_{(k)}\right)\left[\widetilde{H}_n^L(\overline{z}) - \widetilde{H}_n^U(\underline{z})\right] + \left[\frac{\gamma\overline{y}\log\overline{y}}{\gamma-1} + \frac{\gamma\overline{y}}{(\gamma-1)^2}\right][1-\widetilde{H}_n^U(\overline{z})],$$

$$\Upsilon_{\max} = \left([1 - \widehat{H}_n^L(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}} \left[\widehat{H}_n^L(Z_{(k)}) - \widehat{H}_n^L(Z_{(k-1)}) \right] Z_{(k)} \right) \left[\widetilde{H}_n^U(\overline{z}) - \widetilde{H}_n^L(\underline{z}) \right]$$
$$+ \left[\frac{\gamma\overline{y}\log\overline{y}}{\gamma - 1} + \frac{\gamma\overline{y}}{(\gamma-1)^2} \right] [1 - \widetilde{H}_n^L(\overline{z})],$$

$Z_{(0)} = \underline{z} = -\frac{1}{e}$, and $\forall z$ $\widetilde{H}_n^L(z) = \max\{C_n^L(z), 0\}$, $\widetilde{H}_n^U(z) = \min\{C_n^U(z), 1\}$, $\widehat{H}_n^L(z) = \max\{\check{C}_n^L(z), 0\}$, and $\widehat{H}_n^U(z) = \min\{\check{C}_n^U(z), 1\}$.

Using these CIs and equations, we can propose the following nonparametric CI for I_E^1 :

$$C_{I_E^1}(\eta) = \{I_0 \in \mathbb{R} : I_{1,\min} \leq I_0 \leq I_{1,\max}\}$$

where

$$I_{1,\min} = \min \left\{ \frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max} \right\},$$

$$I_{1,\max} = \max \left\{ \frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max} \right\},$$

where $\Lambda_{1,\min}$; $\Lambda_{1,\max}$; $\Upsilon_{\min}$; and $\Upsilon_{\max}$ are defined above for hypotheses (i.i, i.ii, ii.i, and ii.ii). The corresponding CI is of level greater than or equal to $1 - \eta = 1 - \alpha - \beta$ where $\alpha = \alpha_1 + \alpha_2$ and $\beta = \beta_1 + \beta_2$.

Let's assume now that Y satisfies hypotheses (i.ii) where the parameter γ is not known. Let

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

be a confidence interval for γ with level $1-\alpha_3$. Following Proposition 6.7., a nonparametric CI for I_E^1 with level $1 - \eta = 1 - \alpha_1 - \alpha_2 - \alpha_3 - \beta_1 - \beta_2$ is

$$C_{I_E^1}(\eta) = \{I_0 \in \mathbb{R} : I_{1,\min} \leq I_0 \leq I_{1,\max}\}$$

where

$$I_{1,\min} = \min\left\{\frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \le \Upsilon \le \Upsilon_{\max},\ \Lambda_{1,\min} \le \Lambda_1 \le \Lambda_{1,\max}\right\},$$

$$I_{1,\max} = \max\left\{\frac{\Upsilon}{\Lambda_1} - log(\Lambda_1) : \Upsilon_{\min} \le \Upsilon \le \Upsilon_{\max},\ \Lambda_{1,\min} \le \Lambda_1 \le \Lambda_{1,\max}\right\},$$

$$\Lambda_{1,\min} = \left([1 - \widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_u}}[1 - \widetilde{F}_n^U(\overline{y})],$$

$$\Lambda_{1,\max} = \left(\left[1 - \widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_l}}[1 - \widetilde{F}_n^L(\overline{y})],$$

$$\Upsilon_{\min} = \left([1 - \widehat{H}_n^U(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widehat{H}_n^U(Z_{(k)}) - \widehat{H}_n^U(Z_{(k-1)})\right]Z_{(k)}\right)\left[\widetilde{H}_n^L(\overline{z}) - \widetilde{H}_n^U(\underline{z})\right] + \left[\frac{\overline{y}\log\overline{y}}{1-\frac{1}{\gamma_u}} + \frac{\overline{y}}{\left(1-\frac{1}{\gamma_u}\right)^2}\right][1 - \widetilde{H}_n^U(\overline{z})],$$

$$\Upsilon_{\max} = \left([1 - \widehat{H}_n^L(Z_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widehat{H}_n^L(Z_{(k)}) - \widehat{H}_n^L(Z_{(k-1)})\right]Z_{(k)}\right)\left[\widetilde{H}_n^U(\overline{z}) - \widetilde{H}_n^L(\underline{z})\right] + \left[\frac{\overline{y}\log\overline{y}}{1-\frac{1}{\gamma_l}} + \frac{\overline{y}}{\left(1-\frac{1}{\gamma_l}\right)^2}\right][1 - \widetilde{H}_n^L(\overline{z})],$$

$Z_{(0)} = \underline{z} = -\frac{1}{e}$, $Y_{(0)} = \underline{y} = 0$, $\forall z\ \widetilde{H}_n^L(z) = \max\{C_n^L(z), 0\}$, $\widetilde{H}_n^U(z) = \min\{C_n^U(z), 1\}$, $\widehat{H}_n^L(z) = \max\{\check{C}_n^L(z), 0\}$, and $\widehat{H}_n^U(z) = \min\{\check{C}_n^U(z), 1\}$; and $\forall y\ \widetilde{F}_n^L(y) = \max\{G_n^L(y), 0\}$,

$\widetilde{F}_n^U(y) = \min\{G_n^U(y), 1\}$, $\widehat{F}_n^L(y) = \max\{\breve{G}_n^L(y), 0\}$, and $\widehat{F}_n^U(y) = \min\{\breve{G}_n^U(y), 1\}$.

3.7.3 Confidence intervals for the mean logarithmic deviation, the logarithmic deviation, and the Atkinson inequality measures

In this subsection we propose nonparametric CIs for the Mean Logarithmic Deviation, the Logarithmic Deviation, and the Atkinson inequality measures using similar techniques as those for the generalized entropy class of measures.

The Mean Logarithmic Deviation index is:

$$I_E^0 = -\int \log(\frac{y}{\mu})\ dF(y) = log(\mu)\ -\int log(y)\ dF(y)$$

where $\mu = \int y\ dF(y)$. Rewriting it:

$$I_E^0 = log(\Lambda_1(F)) - \Omega(F)$$

where $\Lambda_1(F) = \mu = \int y\ dF(y)$ is the mean of Y—which is positive— and $\Omega(F) = \int log(y)\ dF(y)$ is the mean of $Z = log(Y)$−which belongs to $(-\infty, +\infty)$. In the last subsection, we proposed nonparametric CIs for $\Lambda_1(F)$ under hypotheses (i.i) and (i.ii). Similar techniques can be used to build CIs for $\Omega(F)$ using the same type of regularity conditions for the distribution function of Z and Propositions 6.8., 6.9., and 6.10.

The Logarithmic Variation index is:

$$I_{LV} = \int [\log(\frac{y}{\mu})]^2\ dF(y) = E\left([\log Y - \log(\Lambda_1(F))]^2\right) = E(Z)$$

where $\mu = \int y\ dF(y)$ and $Z = [\log Y - \log(\Lambda_1(F))]^2 \left[\log\left(\frac{Y}{E(Y)}\right)\right]^2$. Hence, I_{LV} is the mean of a positive random variable. By imposing regularity conditions to the tail of the distribution function of Z of the same type as hypotheses (i.i) and (i.ii), nonparametric CIs can be easily built using Propositions 6.2. and 6.3.

The Atkinson class of inequality measures is:

$$I_A^{\varepsilon} = \begin{cases} 1 - [\int (\frac{y}{\mu})^{1-\varepsilon} \, dF(y)]^{\frac{1}{1-\varepsilon}} & \text{if } \varepsilon > 0 \text{ and } \varepsilon \neq 1 \\ 1 - e^{-I_E^0} & \text{if } \varepsilon = 1 \end{cases}$$

When $\varepsilon \neq 1$ and $\varepsilon > 0$, the Atkinson inequality measure is:

$$I_A^{\varepsilon} = 1 - \left[\delta(\delta+1)(I_E^{\delta}+1)\right]^{1/\delta}$$

where $\delta = 1 - \varepsilon$ and I_E^{δ} is the generalized entropy measure of order δ. Hence, if

$$C_{I_E^{\delta}}(\alpha) = \{I_0 \in \mathbb{R} : I_{\delta,l} \leq I_0 \leq I_{\delta,u}\}$$

is a CI for I_E^{δ} with level $1 - \alpha$, then

$$C_{I_A^{\varepsilon}}(\alpha) = \left\{I_0 \in \mathbb{R} : 1 - [\delta(\delta+1)(I_{\delta,u}+1)]^{1/\delta} \leq I_0 \leq 1 - [\delta(\delta+1)(I_{\delta,l}+1)]^{1/\delta}\right\}$$

is a CI for I_A^{ε} with level $1 - \alpha$ too. Therefore, the nonparametric CIs we proposed in the last subsection for I_E^{δ} can be used to build CIs for I_A^{ε} with the same level of confidence.

When $\varepsilon = 1$, the Atkinson measure is $I_A^1 = 1 - e^{-I_E^0}$. Hence, nonparametric CIs for the mean logarithmic deviation can be used to build CIs for I_A^1 with the same level of confidence. Moreover, rewriting I_A^1 as follows:

$$I_A^1 = 1 - \frac{\exp[E(\log(Y))]}{E(Y)}$$

Propositions 6.2., 6.3., 6.7., 6.8., 6.9., and 6.10. can be used to build CIs under some regularity conditions.

3.7.4 Confidence intervals for the Lorenz curve

The Lorenz curve is:

$$L(p) = \frac{1}{E(Y)} \int_0^{F_Y^{-1}(p)} y \; dF(y)$$

where $p \in (0,1)$. The Lorenz curve is an illustration of the distribution of resources in a community. For each p, $L(p)$ represents the proportion of the households' community which owns 100 p percent of the total income of the community. When there is perfect equity, income is equally distributed among households. In this case, the Lorenz curve is the straight line $L(p) = p$.

In this section, we propose nonparametric CIs for $L(p)$ using CBs for $F(y)$ and projection techniques. Rewriting the Lorenz curve, we can express the Lorenz curve as a function of the means of two variables, as follows:

$$L(p) = \frac{1}{E(Y)} \int_0^{+\infty} y \; \mathbb{1}[y \leq F_Y^{-1}(p)] \; dF(y) = \frac{E\left[Y \; \mathbb{1}[y \leq F_Y^{-1}(p)]\right]}{E(Y)} = \frac{E\left[Y \mid Y \leq F^{-1}(p)\right]}{E(Y)}$$

where Y is positive and $Y \mid Y \leq F^{-1}(p)$ is bounded over $[0, F^{-1}(p)]$. Let $\Lambda_1(F) = E(Y) \neq 0$ and $\Upsilon(F) = E\left[Y \mid Y \leq F^{-1}(p)\right]$. Then,

$$L(p) = \frac{\Upsilon(F)}{\Lambda_1(F)}.$$

We have proposed CIs for $\Lambda_1(F)$ under hypotheses (i.i) and (i.ii). Similarly, we can build nonparametric CIs for $\Upsilon(F)$ using the general expression of CIs for the mean of a bounded random variable. Let $H(z)$ be the distribution function of $Z = Y \mid Y \leq F^{-1}(p)$ and $\widetilde{m} = \sum_{k=1}^{n} \mathbb{1}[Y_{(k)} \leq F^{-1}(p)]$. Let

$$C_H(\beta_1) = \left\{ H_0 \in \mathcal{L} : C_n^L(z) \leq H_0(z) \leq C_n^U(z), \; \forall z \right\}$$

be a confidence band for $H(z)$ with level $1 - \beta_1$ where $C_n^L(z) \in \mathcal{L}$ and $C_n^U(z) \in \mathcal{L}$ are two step functions with jumps only at $Y_{(1)}, \ldots, Y_{(\widetilde{m})}$. A confidence interval for $\Upsilon(F)$ with

level greater than or equal to $1-\beta_1$ is:

$$\widetilde{C}_{\Upsilon}(\beta_1) = \{\Upsilon_0 \in \mathbb{R} : \Upsilon_{\min} \leq \Upsilon_0 \leq \Upsilon_{\max}\}$$

where

$$\Upsilon_{\min} = [1 - \widetilde{H}_n^U(Y_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}} \left[\widetilde{H}_n^U(Y_{(k)}) - \widetilde{H}_n^U(Y_{(k-1)})\right] Y_{(k)}$$

and

$$\Upsilon_{\max} = [1 - \widetilde{H}_n^L(Y_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}} \left[\widetilde{H}_n^L(Y_{(k)}) - \widetilde{H}_n^L(Y_{(k-1)})\right] Y_{(k)}$$

where $\overline{z} = F^{-1}(p)$ and $\forall z\ \widetilde{H}_n^L(z) = \max\{C_n^L(z), 0\},\ \widetilde{H}_n^U(z) = \min\{C_n^U(z), 1\}$.

Hence, a nonparametric CI for $L(p)$ is:

$$C_{L(p)}(\eta) = \{L_0 \in \mathbb{R} : L_{1,\min} \leq L_0 \leq L_{1,\max}\}$$

where

$$L_{1,\min} = \min\left\{\frac{\Upsilon}{\Lambda_1} : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

$$L_{1,\max} = \max\left\{\frac{\Upsilon}{\Lambda_1} : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

where $\Lambda_{1,\min}$; $\Lambda_{1,\max}$; $\Upsilon_{\min}$; and $\Upsilon_{\max}$ are defined above under the adequate hypothesis (i.i and i.ii). The level of the corresponding CI is greater than or equal to $1-\alpha-\beta_1$ where $\alpha = \alpha_1 + \alpha_2$.

Let's assume now that Y satisfies hypotheses (i.ii) where the parameter γ is unknown. Let

$$C_{\gamma}(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

be a confidence interval for γ with level $1-\alpha_3$. A nonparametric CI for $L(p)$ with level

greater than or equal to $1-\eta$ where $\eta = \alpha_1 + \alpha_2 + \alpha_3 + \beta_1$ is:

$$C_{L(p)}(\eta) = \{L_0 \in \mathbb{R} : L_{1,\min} \leq L_0 \leq L_{1,\max}\}$$

where

$$L_{1,\min} = \min\left\{\frac{\Upsilon}{\Lambda_1} : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

$$L_{1,\max} = \max\left\{\frac{\Upsilon}{\Lambda_1} : \Upsilon_{\min} \leq \Upsilon \leq \Upsilon_{\max},\ \Lambda_{1,\min} \leq \Lambda_1 \leq \Lambda_{1,\max}\right\},$$

$$\Lambda_{1,\min} = \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_u}}[1-\widetilde{F}_n^U(\overline{y})],$$

$$\Lambda_{1,\max} = \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right]Y_{(k)}\right)\left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right] + \frac{\overline{y}}{1-\frac{1}{\gamma_l}}[1-\widetilde{F}_n^L(\overline{y})];$$

$$\Upsilon_{\min} = [1-\widetilde{H}_n^U(Y_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widetilde{H}_n^U(Y_{(k)}) - \widetilde{H}_n^U(Y_{(k-1)})\right]Y_{(k)};$$

and

$$\Upsilon_{\max} = [1-\widetilde{H}_n^L(Y_{(\widetilde{m})})]\overline{z} + \sum_{k=1}^{\widetilde{m}}\left[\widetilde{H}_n^L(Y_{(k)}) - \widetilde{H}_n^L(Y_{(k-1)})\right]Y_{(k)}.$$

3.7.5 Confidence intervals for the Gini index

The literature presents several expressions for the Gini index. One of the most popular of these is:

$$I_{Gini} = 1 - 2R(F)$$

where $\forall 0 \leq q \leq 1$

$$R(F) = \frac{1}{\mu} \int_0^1 C(F;q)\ dq,$$

$$C(F;q) = \int_0^{Q(F;q)} y\ dF(y),$$

$$Q(F;Q) = \inf\{y \mid F(y) \geq q\},$$

and $\mu = \int y\ dF(y)$. $C(F;q)$ is the cumulative income function and $Q(F;q)$ is the quantile function of $F(y)$.

The Gini index can be expressed as a function of the Lorenz curve:

$$I_{Gini} = 1 - 2\int_0^1 L(p)\ dp = 2\int_0^1 [p - L(p)]\ \ dp] \tag{3.21}$$

For each $F(y)$, the Gini index represents twice the area between the Lorenz curve and the perfect equity line. It measures how far the distribution of income of a households' community is from perfect equity. The values of the Gini index lie between 0 and 1.

Applying the projection techniques we have been using all along the paper, nonparametric CIs for the Gini index can be derived from CIs for the Lorenz curve. In particular, the CIs for the Lorenz curve we proposed earlier can be projected to build nonparametric CIs for the Gini index when income is positive, under regularity conditions.

3.8 Monte Carlo study

In this section, we study the properties of the asymptotic and exact CIs for the Theil index I_E^1 using Monte Carlo techniques. We compare the performance of the proposed nonparametric CIs for the Theil index with the performance of the asymptotic CI and the bootstrap CI. Two bootstrap procedures are considered: the percentile-t (i.i.d.) bootstrap and the semi-parametric bootstrap. The latter was proposed by Davidson and Flachaire (2007). This bootstrap is like the standard bootstrap procedure for all but the right-hand tail of the distribution. At each step of this procedure, each observation is a drawing, with probability $1 - p_{tail}$, from the empirical distribution of the sample of

the smallest $n(1-p_{tail})$ order statistics and , with probability p_{tail}, a drawing from the $Pareto(y_0, \rho)$ distribution function with cumulative distribution function

$$F(y) = 1 - (y/y_0)^{-\rho}\,,\ y > y_0$$

where $y_0 = Y_{(n(1-p_{tail})}$. ρ and p_{tail} are estimated from the sample $Y_1, \ldots, Y_n$, of observations as follows:

$$\widehat{\rho} = \left[k^{-1} \sum_{i=0}^{k-1} \log Y_{(n-i)} - \log Y_{(n-k+1)} \right]^{-1}$$

and

$$p_{tail} = \frac{hk}{n}$$

where $k = \sqrt{n}$ and h is to be choosen. In their simulations, Davidson and Flachaire (2007) used several values of h: $h = 0.3, 0.4, 0.6, 0.8, 1$. In our simulations, we set $h = 0.4$.

We suppose in our simulations that the distribution of the income of households is the following mixture:

$$Y = \begin{cases} Z \text{ with probability } P_0 \\ \overline{y} \text{ with probability } 1 - P_0 \end{cases}$$

where Z follows a Singh-Maddala distribution $SM(a, b, c)$ with cumulative distribution function $F(y) = 1 - [1 + ay^b]^{-c}$ and $\overline{y}$ is some positive number. The Singh-Maddala distribution has been proven by Brachman, Stich, and Trede (1996) to mimic the income of several developed countries, such as Germany, well. Davidson and Flachaire (2004) used this distribution to explain the failure of the asymptotic and the bootstrap inference methods to perform well when applied to the Theil index with small and fairly large sample. Following Davidson and Flachaire (2007), we set $a = 100$, $b = 2.8$, $c = 1.7$, $\overline{y} = F^{-1}(0.99961) = 1.00078$, and $P_0 = 0.9$. We suppose that the right-hand tail of the distribution function of Y is bounded by a $Pareto(\overline{y}, \delta)$ where $\delta = 2$. CIs with level 95% are simulated for sample sizes $n = 50, 100, 200$, $n = 500$ and $n = 1000$ using $N = 500$, $N = 250$ and $N = 150$ replications, respectively. For the CIs based on the regularized

statistics, we use $\zeta_E = \zeta_{AD} = 0.07$. We showed in a former paper that these values deliver CIs of minimal width for the poverty measure P_2 with a distribution of income slightly different from the distribution proposed above. With our choice of parameters, the true value of the Theil index is $I_0 = 0.3907$.

Table 3.1 shows the coverage probability and the average width of the simulated CIs for I_E^1 using both continuous conservative critical points (corresponding to the case $P_0 = 1$) and adequate noncontinuous critical points.

The results confirm that asymptotic and bootstrap CIs for the Theil index are not reliable. Like the asymptotic CI, both the standard bootstrap and the semiparametric bootstrap proposed by Davidson and Flachaire (2007) deliver coverage probability far below the theoretical level of 95%. With our setting, the estimation value of ρ is lower than 2. Moreover, with our choice of k, $\widehat{\rho}$ is infinite for some sample size and some samples, in particular when all observations in the tail of the sample are equal to $\overline{y}$. In this case, the mean of the Pareto law is $y_0 = \overline{y}$. In small samples, both bootstrap methods experience problems with the distribution function under study. So does the asymptotic CI. In contrast, nonparametric methods perform well. They are reliable and conservative for all sample sizes. Among the nonparametric methods, the regularized Eicker-type and Anderson Darling-type CIs provide the smaller widths. The Berk-Jones type CI performs better than the KS-type CI but less than the methods based on the regularized statistics. The regularized Eicker-type method provides the shortest CI for $n = 50$ while the regularized Anderson Darling-type CI is the shortest for larger sample sizes.

Table 3.1: Simulated confidence intervals for the Theil index I_E^1

with $Y = \begin{cases} \overline{y} \text{ with probability } 1 - P_0 = 0.1 \\ SM(100, 2.8, 1.7) \text{ with probability } P_0 = 0.9 \end{cases}$, $\zeta_E = \zeta_{AD} = 0.07$,

$N = 500$ replications for $n = 50, 100, 200$, and

$N = 250$ for $n = 500$, and $N = 150$ for 1000

		Coverage probability (in %)				
	n	50	100	200	500	1000
Asymptotic	$P0 = 1$	92.60	95.20	95.60	93.60	95.6
	$P0 = 0.9$	-	-	-	-	-
t-Bootstrap	$P0 = 1$	93.40	95.80	95.40	95.60	95.6
	$P0 = 0.9$	-	-	-	-	-
Bootstrap DF	$P0 = 1$	78.00	84.80	56.40	56.00	59.00
	$P0 = 0.9$	-	-	-	-	-
KS	$P0 = 1$	100.00	100.00	100.00	100.00	100.00
	$P0 = 0.9$	100.00	100.00	100.00	100.00	100.00
E_ζ	$P0 = 1$	100.00	100.00	100.00	100.00	100.00
	$P0 = 0.9$	100.00	100.00	100.00	100.00	100.00
AD_ζ	$P0 = 1$	100.00	100.00	100.00	100.00	100.00
	$P0 = 0.9$	100.00	100.00	100.00	100.00	100.00
BJ	$P0 = 1$	100.00	100.00	100.00	100.00	100.00
	$P0 = 0.9$	100.00	100.00	100.00	100.00	100.00

		Width				
	n	50	100	200	500	1000
Asymptotic	$P0 = 1$	0.2234	0.1592	0.1130	0.0705	0.0503
	$P0 = 0.9$	-	-	-	-	-
t-Bootstrap	$P0 = 1$	0.4086	0.1740	0.1185	0.0724	0.0516
	$P0 = 0.9$	-	-	-	-	-
Bootstrap DF	$P0 = 1$	0.4508	0.2027	0.1429	0.0823	0.0564
	$P0 = 0.9$	-	-	-	-	-
KS	$P0 = 1$	1.1461	0.7359	0.4666	0.2535	0.1573
	$P0 = 0.9$	1.1481	0.7364	0.4668	0.2539	0.1574
E_ζ	$P0 = 1$	0.6918	0.4352	0.2765	0.1499	0.0940
	$P0 = 0.9$	0.6945	0.4353	0.2766	0.1501	0.0939
AD_ζ	$P0 = 1$	0.6995	0.4304	0.2710	0.1461	0.0912
	$P0 = 0.9$	0.6997	0.4305	0.2711	0.1461	0.0913
BJ	$P0 = 1$	0.7105	0.4459	0.2940	0.1858	0.1331
	$P0 = 0.9$	0.7327	0.4460	0.2941	0.1864	0.1336

3.9 Empirical illustration

In this section, we analyze the level of inequality of rural Mexican households using the proposed inference methods for the Gini inequality index. We employ data that have been collected as part of the targeting and evaluation program: PROGRESA.[1] A census of households in a set of 506 rural communities has been conducted in 1997, 1998, and 1999 and the data processed to insure comparability. Data about households' characteristics are extracted from the November 1997 survey and expenditure aggregate is constructed using the March 1998 survey.[2]

[1]PROGRESA is a health, education, and nutrition program of the Mexican government aimed to reduce poverty in targeted rural communities.

[2]The data set excludes households in the expenditure survey that had not been interviewed in November 1997 and 10 communities with fewer than 10 households with expenditure information, leaving 20544 households in 496 communities (see Demombynes, Elbers, Lanjouw and Lanjouw, 2007)

In a former paper, we used these data to analyze poverty in Mexico both at the national and regional levels. First, we used the census as a whole to build CIs for the level of poverty P_2 of rural households in Mexico. Then, drawing samples randomly from the census, we studied the poverty profile of PROGRESA-targeted communities and analyzed the determinants of poverty in rural areas in Mexico for the involved communities using two characteristics of the heads of households: the gender and the level of education.

In this section, we study the profile of inequality of PROGRESA targeted communities using samples of size $n = 500$ and 1000 drawn from the census. We employ the same samples as those used in our former paper and the same values for the regularization parameters—the latter were derived by applying a split sample procedure. For $n = 500$, we found that the smallest widths for the CIs for the poverty measure P_2 were achieved by $\zeta_{AD} = 0.45$ and $\zeta_E = 0.039$ while those values were $\zeta_{AD} = 0.5$ and $\zeta_E = 0.05$ for $n = 1000$. We compare the inequality profile of households with a female head to those of households which head is a male, and the profile of households with an educated head to those with a non-educated head.

Tables 3.2 and 3.3 show the estimated CIs for the Gini index corresponding to $n = 500$ and $n = 1000$, respectively. Asymptotic and bootstrap CIs are estimated using the whole sample, including the auxiliairy sample on which the optimal value of the regularization parameters are computed. The Berk Jones-type CI uses simulated critical points.

Results obtained with a relatively small sample of $n = 500$ are consistent with those obtained with $n = 1000$. As someone would expect, results delivered by the smaller sample are less accurate than those obtained for $n = 1000$ but do not contradict the latter.

According to CIs using the regularized statistics—which we proved were the best performing CIs, the level of inequality among rural households targeted by PROGRESA is relatively low. For $n = 1000$, these CIs show that the highest level of inequality among those households, as provided by the Gini index, is about 22%. This results is in line with the objectives of the program which targets fairly homogenous rural households and provide them help to improve their standards of living. Against this bright global picture,

inequality seems to be unevenly spread among types of households. In fact, inequality among households with a female head can be as high as 67% while inequality among households with a male head still lies in the average 22%. This reflects atypical problems faced by female households' heads in providing resources to their dependants compared to uniform shocks faced by male heads. Likewise, households with a non-educated head register more inequality (44%) than households with an educated head (24%), even if the level of inequality among the latter is slightly higher than the total average.

This picture of the distribution of inequality among rural households targeted by PROGRESA completes the poverty profile we derived in our former paper. In addition to implementing policies that would help reduce poverty among households with a female head or a non-educated head, authorities policies targeted to the most vulnerable among those households to help them catch up with other households and get insured against negative shocks would decrease inequality further.

Table 3.2: Mexican households in PROGRESA: Confidence intervals for I_{Gini} for different types of households' heads

$n = 500$, $\zeta_{AD} = 0.45$, $\zeta_E = 0.039$

Table 3.2a: All households

	Confidence Intervals		
	min	max	width
Asymp	-0.337	0.895	1.232
Bootstrap	-0.121	0.842	0.963
KS	-0.660	0.216	0.877
E_ζ	-0.511	0.286	0.797
AD_ζ	-0.674	0.231	0.906
BJ	-0.650	0.240	0.890

Table 3.2b: Households with a female head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.289	0.864	1.153
Bootstrap	-0.079	0.856	0.934
KS	-0.900	0.625	1.525
E_ζ	-0.900	0.760	1.660
AD_ζ	-0.900	0.653	1.553
BJ	-0.881	0.678	1.560

Table 3.2c: Households with a male head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.336	0.898	1.234
Bootstrap	-0.124	0.842	0.965
KS	-0.685	0.224	0.909
E_ζ	-0.544	0.290	0.835
AD_ζ	-0.700	0.244	0.944
BJ	-0.676	0.254	0.930

Table 3.2d: Households with a non educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.266	0.783	1.050
Bootstrap	-0.109	0.870	0.980
KS	-0.877	0.444	1.321
E_ζ	-0.851	0.529	1.380
AD_ζ	-0.861	0.463	1.323
BJ	-0.826	0.500	1.326

Table 3.2e: Households with an educated head

.

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.346	0.900	1.246
Bootstrap	-0.125	0.836	0.962
KS	-0.734	0.248	0.982
E_ζ	-0.613	0.323	0.935
AD_ζ	-0.747	0.270	1.016
BJ	-0.723	0.296	1.019

3.10 Conclusion

Inference studies for inequality measures show that asymptotic and bootstrap methods do not perform well when applied to these measures. Davidson and Flachaire (2007) showed that asymptotic approximations provide a poor approximation to the real distributions of statistics, for small and even fairly large samples while the standard bootstrap deliver a poor performance when applied to distributions with heavy tails, as it is often the case for income distributions. They proposed a semiparametric bootstrap procedure for the Theil inequality index to improve the performance of the bootstrap when distribution functions have heavy tails. However, the performance of bootstrap inference is also known to be sensitive to distribution functions with probability masses. In each case, the origin of the failure of the bootstrap must be identified to correct the drawback, which is not obvious when data comes from an unknown distribution function.

Table 3.3: Mexican households in PROGRESA: Confidence intervals for I_{Gini} for different types of households' heads
$n = 1000$, $\zeta_{AD} = 0.5$, and $\zeta_E = 0.05$

Table 3.3a: All households

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.342	0.897	1.238
Bootstrap	-0.093	0.839	0.932
KS	-0.455	0.152	0.606
E_ζ	-0.275	0.218	0.494
AD_ζ	-0.462	0.166	0.628
BJ	-0.501	0.177	0.678

Table 3.3b: Households with a female head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.320	0.900	1.220
Bootstrap	0.007	0.836	0.829
KS	-0.879	0.566	1.445
E_ζ	-0.851	0.675	1.527
AD_ζ	-0.838	0.622	1.461
BJ	-0.820	0.607	1.427

Table 3.3c: Households with a male head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.344	0.897	1.240
Bootstrap	-0.106	0.839	0.946
KS	-0.518	0.161	0.680
E_ζ	-0.356	0.219	0.575
AD_ζ	-0.532	0.172	0.703
BJ	-0.561	0.187	0.748

Table 3.3d: Households with a non educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.329	0.900	1.229
Bootstrap	-0.053	0.839	0.892
KS	-0.668	0.369	1.037
E_ζ	-0.567	0.438	1.005
AD_ζ	-0.660	0.405	1.065
BJ	-0.662	0.423	1.085

Table 3.3e: Households with an educated head

	Confidence Intervals (in %)		
	min	max	width
Asymp	-0.340	0.899	1.239
Bootstrap	-0.111	0.840	0.951
KS	-0.565	0.180	0.745
E_ζ	-0.397	0.239	0.636
AD_ζ	-0.580	0.187	0.767
BJ	-0.608	0.220	0.827

In this paper, we propose nonparametric CIs for the most popular inequality measures in the literature. We show that inequality measures can be reexpressed as a function of the mean of two random variables. When the involved random variables are bounded, we employ inference methods for the mean of a bounded random variable we derived in a former paper (Diouf and Dufour (2005b) to build CIs from confidence bands for the underlying distribution using projection techniques. When the involved variables are unbounded, we generalize these projection techniques to random variables with distribution functions which tails are bounded by a Pareto distribution or follow a Pareto distribution. Under these regularity conditions, we propose nonparametric CIs for the mean of a lower bounded random variable and for the mean of an unbounded random variable using confidence bands for the distribution functions.

We apply these CIs to build CIs for the means involved in the inequality measures and derive CIs for the latter. The levels of the corresponding CIs are computed using the Bonferroni inequality.

Owing to the CBs for distribution functions they are derived from, the CIs for inequality measures need a single set of critical points to be built, when applied to continuous distribution functions. This property makes them convenient to implement. When the income distribution function is noncontinuous, adequate critical points for continuous dis-

tribution functions provide CIs for inequality measures with level greater than or equal to the nominal level. Moreover, exploiting embeddedness of the image sets of distribution functions allows to improve the performance of the inference methods.

Monte Carlo simulations are performed to study the performance of these methods for the Theil index. The results show that the standard bootstrap procedure and the alternative proposed by Davidson and Flachaire (2007), as well as the asymptotic method can fail in providing reliable CIs for the Theil index while nonparametric inference methods are strongly reliable and provide informative CIs. The regularized statistics deliver the best width among the latter.

Last, the profile of inequality of Mexican households involved in PROGRESA as assessed by the Gini index is analyzed. Results show that there are more inequalities among households with a female head or a non-educated head. Hence, in addition to implementing policies that would help reduce poverty among households with a female head or a non-educated head, authorities plan policies targeted to the most vulnerable among those households to help them catch up with other households and get insured against negative shocks that would increase inequality further.

3.11 Appendix: Proof of theorems and propositions

PROOF OF RESULT 5.1. The Generalized Entropy measure is

$$I_E^{\delta}(y) = \frac{1}{\delta(\delta-1)}\left(\frac{\Lambda_{\delta}(F)}{\Lambda_1^{\delta}(F)} - 1\right)$$

where $\Lambda_{\delta}(F) = \int y^{\delta}\ dF(y)$ is the mean of Y^{δ}, which is bounded over $[0, \overline{y}^{\delta}]$ when Y is bounded over $[0, \overline{y}]$. Following Diouf and Dufour (2005b), we can build CIs for $\Lambda_{\delta}(F)$ using confidence bands for the distribution functions of Y^{δ}. Let

$$C_F(\alpha) = \left\{F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y\right\}.$$

define a confidence band for $F(y)$ with level $1-\alpha$ where $G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Y_{(1)}, \cdots, Y_{(n)}$. A nonparametric CI for $\Lambda_1[F]$ with level $1-\alpha$ is:

$$\widetilde{C}_{\Lambda_1}(\alpha) = \{\mu_0 \in \mathbb{R} : \Lambda_{1,\min} \leq \mu_0 \leq \Lambda_{1,\max}\ \}$$

and a nonparametric CI for $\Lambda_{\delta}[F]$ with level $1-\alpha$ is:

$$\widetilde{C}_{\Lambda_{\delta}}(\alpha) = \{\lambda_0 \in \mathbb{R} : \Lambda_{\delta,\min} \leq \lambda_0 \leq \Lambda_{\delta,\max}\ \}$$

where

$$\Lambda_{\delta,\min} = [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)}^{\delta} + \sum_{k=1}^{n}\left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)})\right]Y_{(k)}^{\delta},$$

$$\Lambda_{\delta,\max} = [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)}^{\delta} + \sum_{k=1}^{n}\left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)})\right]Y_{(k)}^{\delta},$$

$Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, and $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ and $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$. Hence, the corresponding CI for $I_E^{\delta}(y)$ is:

$$C_{I_E^{\delta}}(\alpha) = \left\{I_0 \in \mathbb{R} : \frac{1}{\delta(\delta-1)}[\frac{\Lambda_{\delta,\min}}{\Gamma_{\max}^{\delta}} - 1] \leq I_0 \leq \frac{1}{\delta(\delta-1)}[\frac{\Lambda_{\delta,\max}}{\Gamma_{\min}^{\delta}} - 1]\right\}.$$

The level of $C_{I_E^\delta}(\alpha)$ can be computed using the levels of $\widetilde{C}_{\Lambda_\delta}$ and $\widetilde{C}_{\Lambda_1}$ and the Bonferroni inequality. The latter states that for two events E_1 and E_2:

$$\Pr(E_1 \cap E_2) \geq 1 - \Pr(\overline{E_1}) - \Pr(\overline{E_2}).$$

The CIs of $\Lambda_1(F)$ and $\Lambda_\delta(F)$ are such that

$$\Pr[\Lambda_0 \in C_{\Lambda_1}(\alpha)] = 1 - \alpha \text{ and } \Pr[\Lambda_0 \in C_{\Lambda_\delta}(\alpha)] = 1 - \alpha.$$

Then, the inequality of Bonferroni yields that

$$\Pr[\Gamma_0 \in C_\Gamma(\alpha), \Lambda_0 \in C_{\Lambda_\delta}(\alpha)] \geq 1 - \alpha - \alpha$$

The level of the simultaneous confidence set of $(\Gamma[F], \Lambda_\delta[F])$ is also the level of the CI of any function of the vector $(\Lambda_1[F], \Lambda_\delta[F])$. Thus, the level of $C_{I_E^\delta}(\alpha)$ is $1 - 2\alpha$.

PROOF OF RESULT 5.2. The tail index is

$$I_E^1 = \frac{\Upsilon(F)}{\Lambda_1(F)} - log(\Lambda_1(F))$$

where $\Lambda_1(F) = \mu = \int y \; dF(y) \neq 0$ is the mean of Y and $\Upsilon(F) = \int y \; log(y) \; dF(y)$ is the mean of $Y \; log(Y)$. When Y is bounded over $[0, \overline{y}]$, $YlogY$ is also bounded. Let υ_1 and υ_2 be respectively the lower and the upper bounds of $YlogY$ where $\upsilon_1 < \upsilon_2$. If

$$C_F(\alpha_1) = \left\{ F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y), \; \forall y \right\}.$$

define a confidence band for $F(y)$ with level $1 - \alpha_1$ where$G_n^L(y) \in \mathcal{L}$ and $G_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Y_{(1)}, \cdots, Y_{(n)}$ then a nonparametric CI for $\Lambda_1[F]$ with level $1 - \alpha_1$ is:

$$\widetilde{C}_{\Lambda_1}(\alpha_1) = \{\mu_0 \in \mathbb{R} : \Lambda_{1,\min} \leq \mu_0 \leq \Lambda_{1,\max} \}$$

where

$$\Lambda_{1,\min} = [1 - \widetilde{F}_n^U(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^U(Y_{(k)}) - \widetilde{F}_n^U(Y_{(k-1)})\right] Y_{(k)},$$

$$\Lambda_{1,max} = [1 - \widetilde{F}_n^L(Y_{(n)})]Y_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{F}_n^L(Y_{(k)}) - \widetilde{F}_n^L(Y_{(k-1)})\right] Y_{(k)},$$

$Y_{(0)} = 0$, $Y_{(n+1)} = \overline{y}$, and $\forall y$, $\widetilde{F}_n^L(y) = \max\left\{G_n^L(y), 0\right\}$ and $\widetilde{F}_n^U(y) = \min\left\{G_n^U(y), 1\right\}$. Likewise, if

$$C_H(\alpha_2) = \left\{H_0 \in \mathcal{L} : \breve{G}_n^L(z) \leq H_0(z) \leq \breve{G}_n^U(z),\ \forall z\right\}.$$

is a confidence band for $H(z)$ with level $1 - \alpha_2$ where $Z = YlogY$ and $H(z)$ is the distribution function of Z, $\breve{G}_n^L(y) \in \mathcal{L}$ and $\breve{G}_n^U(y) \in \mathcal{L}$ are two step functions with jumps only at $Z_{(1)}, \cdots, Z_{(n)}$ then a nonparametric CI for $\Upsilon[F]$ with level $1 - \alpha_2$ is

$$C_\Upsilon(\alpha_2) = \{\Upsilon_0 \in \mathbb{R} : \Upsilon_{\min} \leq \Upsilon_0 \leq \Upsilon_{\max}\ \}$$

where

$$\Upsilon_{\min} = [1 - \widetilde{H}_n^U(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{H}_n^U(Z_{(k)}) - \widetilde{H}_n^U(Z_{(k-1)})\right] Z_{(k)},$$

$$\Upsilon_{\max} = [1 - \widetilde{H}_n^L(Z_{(n)})]Z_{(n+1)} + \sum_{k=1}^{n} \left[\widetilde{H}_n^L(Z_{(k)}) - \widetilde{H}_n^L(Z_{(k-1)})\right] Z_{(k)},$$

$Z_{(0)} = \upsilon_1$, $Z_{(n+1)} = \upsilon_2$, and $\forall z$, $\widetilde{H}_n^L(z) = \max\left\{\breve{G}_n^L(z), 0\right\}$ and $\widetilde{H}_n^U(y) = \min\left\{\breve{G}_n^U(z), 1\right\}$. The corresponding CI for the Theil index is:

$$C_{I_E^1}(\alpha) = \left\{I_0 \in \mathbb{R} : \frac{\Upsilon_{\min}}{\Lambda_{1,\max}} - log(\Lambda_{1,\max}) \leq I_0 \leq \frac{\Upsilon_{\max}}{\Lambda_{1,\min}} - log(\Lambda_{1,\min})\right\}.$$

Following the Bonferroni inequality, the level of $C_{I_E^1}(\alpha)$ is greater than or equal to $1 - \alpha_1 - \alpha_1 - \alpha_2 = 1 - 2\alpha_1 - \alpha_2 = 1 - \alpha$.

PROOF OF RESULT 6.1. Given that $\gamma > 1$, the mean of W exists and is:

$$E(W) = \int_{\overline{y}}^{+\infty} w \; dG(w) < \infty$$

where $\lim_{\overline{y} \to +\infty} \int_{\overline{y}}^{+\infty} w \; dG(w) = 0$. Developing $d\left[wG(w)\right]$:

$$\begin{aligned} d\left[wG(w)\right] &= (dw)\, G(w) + w dG(w) \\ \Rightarrow d\left[w\left(1 - G(w)\right)\right] &= (dw)\left[1 - G(w)\right] + w \; d\left[1 - G(w)\right] \\ \Rightarrow d\left[w\left(1 - G(w)\right)\right] &= (dw)\left[1 - G(w)\right] - w \; dG(w) \end{aligned}$$

Hence,

$$\int_{\overline{y}}^{z} d\left[w\left(1 - G(w)\right)\right] = \int_{\overline{y}}^{z} (dw)\left[1 - G(w)\right] - \int_{\overline{y}}^{z} w \; dG(w),$$

where

$$\lim_{z \to +\infty} \int_{\overline{y}}^{z} d\left[w\left(1 - G(w)\right)\right] = \lim_{z \to +\infty} z\left(1 - G(z)\right) = 0,$$

and

$$\begin{aligned} \int_{\overline{y}}^{z} w \; dG(w) &= \int_{\overline{y}}^{z} (dw)\left[1 - G(w)\right] - \int_{\overline{y}}^{z} d\left[w\left(1 - G(w)\right)\right] \\ &= \int_{\overline{y}}^{z} (dw)\left[1 - G(w)\right] - z\left(1 - G(z)\right), \end{aligned}$$

$$E(W) = \int_{\overline{y}}^{+\infty} w \; dG(w) = \int_{\overline{y}}^{+\infty} \left[1 - G(w)\right] \; dw.$$

Likewise,

$$E(Y_{LB}) = \int_{\overline{y}}^{+\infty} w \; dF_{Y_{LB}}(w) = \int_{\overline{y}}^{+\infty} \left[1 - F_{Y_{LB}}(w)\right] \; dw.$$

By assumption:

$$\begin{aligned} G(w) &\leq F_{Y_{LB}}(w)\ \forall w \\ &\Rightarrow\ 1 - F_{Y_{LB}}(w) \leq 1 - G(w) \\ &\Rightarrow\ \int_{\overline{y}}^{+\infty} [1 - F_{Y_{LB}}(w)]\ dw \leq \int_{\overline{y}}^{+\infty} [1 - G(w)]\ dw \end{aligned}$$

i.e.

$$E(Y_{LB}) \leq E(W) = \frac{\gamma \overline{y}}{\gamma - 1}$$

where the last equality follows from equation (*3.10*).

PROOF OF PROPOSITION 6.2. Following equation (*3.11*), the mean of Y is

$$\mu = I_B\ \Pr(\underline{y} \leq Y \leq \overline{y}) + I_{LB}\ \Pr(Y \geq \overline{y}).$$

If

$$C_F(\alpha_1) = \left\{ F_0 \in \mathcal{L} : G_n^L(y) \leq F_0(y) \leq G_n^U(y),\ \forall y \right\}.$$

is a CB for $F(y)$ with level $1 - \alpha_1$ then ,

$$C_{\Pr(Y \geq \overline{y})}(\alpha_1) = \left\{ F_0 \in \mathcal{L} : 1 - \widetilde{F}_n^U(\overline{y}) \leq F_0(y) \leq 1 - \widetilde{F}_n^L(\overline{y}),\ \forall y \right\}$$

and

$$C_{\Pr(\underline{y} \leq Y \leq \overline{y})}(\alpha_1) = \left\{ F_0 \in \mathcal{L} : \widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y}) \leq F_0(y) \leq \widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y}),\ \forall y \right\}$$

are respectively CIs for $\Pr(Y \geq \overline{y})$ and $C_{\Pr(\underline{y} \leq Y \leq \overline{y})}(\alpha)$ with level $1 - \alpha_1$.

Moreover, equation *(3.12)* provides the following CI for I_B with level $1-\alpha_2$:

$$\begin{aligned}\widetilde{C}_{I_B}(\alpha_2) &= \left\{\mu_0 \in \mathbb{R} : [1-\widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right] Y_{(k)} \leq \mu_0 \right. \\ &\left. \leq \left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right] Y_{(k)}\right\}.\end{aligned}$$

To end,

$$\begin{aligned} I_{LB} &= E(Y_{LB}) = E(Y \mid Y \geq \overline{y}) \\ &\geq \overline{y} \end{aligned} \tag{3.22}$$

and result 6.1 shows that under hypothesis (i)

$$I_{LB} = E(Y_{LB}) \leq \frac{\gamma\overline{y}}{\gamma-1}. \tag{3.23}$$

The CI for μ which corresponds to these CIs is:

$$\widetilde{C}_{\mu}(\alpha) = \{\mu_0 \in \mathbb{R} : \mu_L \leq \mu_0 \leq \mu_U\}$$

where

$$\mu_L = \left([1-\widehat{F}_n^U(Y_{(m)})]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^U(Y_{(k)}) - \widehat{F}_n^U(Y_{(k-1)})\right] Y_{(k)}\right)\left[\widetilde{F}_n^L(\overline{y}) - \widetilde{F}_n^U(\underline{y})\right] + \overline{y}[1-\widetilde{F}_n^U(\overline{y})]$$

and

$$\mu_U = \left(\left[1-\widehat{F}_n^L(Y_{(m)})\right]\overline{y} + \sum_{k=1}^{m}\left[\widehat{F}_n^L(Y_{(k)}) - \widehat{F}_n^L(Y_{(k-1)})\right] Y_{(k)}\right)\left[\widetilde{F}_n^U(\overline{y}) - \widetilde{F}_n^L(\underline{y})\right] + \frac{\gamma\overline{y}}{\gamma-1}[1-\widetilde{F}_n^L(\overline{y})].$$

Following the Bonferroni inequality, the level of this CI is greater than or equal to $1-2\alpha_1-\alpha_2$.

PROOF OF PROPOSITION 6.3. Under under hypothesis (ii),

$$E(Y_{LB}) = \frac{\gamma \overline{y}}{\gamma - 1}$$

Hence, following the proof of Proposition 6.2. and replacing the lower bound of $E(Y_{LB})$ in equation *(3.22)* by $\frac{\gamma \overline{y}}{\gamma - 1}$ yields the corresponding CI for μ.

PROOF OF COROLLARY 6.4. Let W be a random variable with a Pareto(w_0, δ) distribution. The density function of W is

$$g(w) = \begin{cases} \frac{\delta w_0^{\delta}}{w^{\delta+1}} \text{ for } w \geq w_0 > 0 \\ \qquad 0 \text{ otherwise} \end{cases}$$

and its cumulative distribution function

$$G(w) = 1 - \left(\frac{w_0}{w}\right)^{\delta} \text{ for } w \geq w_0 > 0$$

where $\delta > 0$ and $w_0 > 0$ are respectively the shape and the scale parameters.

If a n-sample from $G(w)$ is available, then the random variables $T_i = -ln\left(\frac{w_0}{W_i}\right)^{\delta}$ with $i = 1, ..., n$ are i.i.d with each, exponential distribution of parameter $\theta = 1$. Let $W_{(1)}, .., W_{(n)}$ be the increasingly ordered sample corresponding to the n-sample from $G(w)$. As the function $-ln\left(\frac{w_0}{x}\right)^{\delta}$ is increasing in x, the increasingly ordered sample which corresponds to $W_{(1)}, .., W_{(n)}$ is $T_{(1)}, .., T_{(n)}$ such that $T_{(i)} = -ln\left(\frac{w_0}{W_{(i)}}\right)^{\delta}$. Following Lawless (1982, pp101-103), if we define the random variable $T = \sum_{i=1}^{r} T_{(i)} + (n-r)T_{(r)}$ then $\frac{2T}{\theta} \sim \chi^2(2r)$ $\forall 3 \leq r \leq n$.

As a consequence, fixing $r = k\epsilon[3, n]$ yields that

$$\begin{aligned}\frac{2T}{\theta} &= 2[\sum_{i=1}^{k} T_{(i)} + (n-k)T_{(k)}] = 2[\sum_{i=1}^{k-1} T_{(i)} + (n-k+1)T_{(k)}] \\ &= 2[\sum_{i=1}^{k-1} - ln\left(\frac{w_0}{W_{(i)}}\right)^{\delta} - (n-k+1)ln\left(\frac{w_0}{W_{(k)}}\right)^{\delta}] \\ &= -2\delta[\sum_{i=1}^{k-1} ln\left(\frac{w_0}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{w_0}{W_{(k)}}\right)]\end{aligned}$$

follows a $\chi^2(2k)$. If w_0 is known, then

$$\Pr[\chi^2_\alpha(2k) \leq -2\delta\left(\sum_{i=1}^{k-1} ln\left(\frac{w_0}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{w_0}{W_{(k)}}\right)\right)] = 1-\alpha$$

where $\chi^2_\alpha(2k)$ is the α^{th} quantile of a χ^2 distribution with $2k$ degrees of freedom. As $\frac{w_0}{W_{(i)}} \leq 1\ \forall i$,

$$\Pr[\frac{\chi^2_\alpha(2k)}{-\sum_{i=1}^{k-1} ln\left(\frac{w_0}{W_{(i)}}\right) - (n-k+1)ln\left(\frac{w_0}{W_{(k)}}\right)} \leq 2\delta] = 1-\alpha$$

and

$$\Pr\left(\frac{-\chi^2_\alpha(2k)}{2\left[\sum_{i=1}^{k-1} ln\left(\frac{w_0}{W_{(i)}}\right) + (n-k+1)ln\left(\frac{w_0}{W_{(k)}}\right)\right]} \leq \delta]\right) = 1-\alpha$$

or

$$\Pr\left(\frac{-\chi^2_\alpha(2k)}{2\left[n\ \ln(w_0) - \sum_{i=1}^{k-1}\ln(W_{(i)}) - (n-k+1)\ln(W_{(k)})\right]} \leq \delta]\right) = 1-\alpha$$

Proof of Proposition 6.5. The proof of this corollary is similar to the proof of the previous one.

Given that $2T \sim \chi^2(2k)$

$$\Pr[\chi^2_{\alpha/2}(2k) \leq -2\delta \left(\sum_{i=1}^{k-1} ln \left(\frac{w_0}{W_{(i)}} \right) + (n-k+1) ln \left(\frac{w_0}{W_{(k)}} \right) \right) \leq \chi^2_{1-\alpha/2}(2k)] = 1-\alpha$$

where $\chi^2_\nu(2k)$ is the ν^{th} quantile of the chi2 distribution of 2k degrees of freedom. The bounds of the CI can be re-expressed as follows:

$$\begin{aligned} \delta^L &= \frac{-\chi^2_{\alpha/2}(2k)}{2\left[n\ln(w_0) - \sum_{i=1}^{k-1} ln\left(W_{(i)}\right) - (n-k+1) ln\left(W_{(k)}\right) \right]} \\ \delta^U &= \frac{-\chi^2_{1-\alpha/2}(2k)}{2\left[n\ln(w_0) - \sum_{i=1}^{k-1} ln\left(W_{(i)}\right) - (n-k+1) ln\left(W_{(k)}\right) \right]} \end{aligned}$$

PROOF OF PROPOSITION 6.6. Under hypothesis (ii), the mean of Y_{LB} is

$$E(Y_{LB}) = \frac{\gamma \overline{y}}{\gamma - 1} = \frac{\overline{y}}{1 - \frac{1}{\gamma}}.$$

Hence, if

$$C_{\overline{y},\gamma}(\alpha_3) = \left\{ (\overline{y}_0, \gamma_0) \in \mathbb{R}^2 : \overline{y}_l \leq \overline{y}_0 \leq \overline{y}_u \text{ and } \gamma_l \leq \gamma_0 \leq \gamma_u \right\}$$

is a confidence region for $\overline{y}$ and γ with level $1-\alpha_3$, then

$$C_{I_{LB}}(\alpha_3) = \left\{ \mu_0 \in \mathbb{R} : \frac{\overline{y}_l}{1 - \frac{1}{\gamma_u}} \leq \mu_0 \leq \frac{\overline{y}_u}{1 - \frac{1}{\gamma_l}} \right\}$$

is a CI for I_{LB} with level $1-\alpha_3$.

Using $C_{I_{LB}}(\alpha_3)$, $C_F(\alpha_1)$, $C_{F_{Y_B}}(\alpha_2)$, and the CI for I_B that has been proposed in equation *(3.12)* provide the result. The Bonferroni inequality gives that the level of this CI is greater than or equal to $1 - 2\alpha_1 - \alpha_2 - \alpha_3$.

PROOF OF PROPOSITION 6.7. Under hypothesis (ii), the mean of Y_{LB} is

$$E(Y_{LB}) = \frac{\gamma\overline{y}}{\gamma - 1} = \frac{\overline{y}}{1 - \frac{1}{\gamma}}.$$

Hence, if

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

is a confidence interval for γ with level $1 - \alpha_3$, then

$$C_{I_{LB}}(\alpha_3) = \left\{\mu_0 \in \mathbb{R} : \frac{\overline{y}}{1 - \frac{1}{\gamma_u}} \leq \mu_0 \leq \frac{\overline{y}}{1 - \frac{1}{\gamma_l}}\right\}$$

is a CI for I_{LB} with level $1 - \alpha_3$. Using $C_{I_{LB}}(\alpha_3)$ and setting $\overline{y}_l = \overline{y}_u = \overline{y}$ in Proposition 6.6. yields the result in Proposition 6.6. The Bonferroni inequality gives that the level of this CI is greater than or equal to $1 - 2\alpha_1 - \alpha_2 - \alpha_3$.

PROOF OF PROPOSITION 6.8. Proposition 6.2 proposes nonparametric CIs for $I_B \Pr(\underline{y} \leq Y \leq \overline{y}) + I_{LB} \Pr(Y \geq \overline{y})$ under hypotheses similar to those assumed in Proposition 7.1. Moreover, under hypothesis (iii):

$$\frac{\rho\underline{y}}{\rho - 1} \leq E(Y_{UB}) \leq \underline{y}$$
$$\Leftrightarrow \frac{\rho\underline{y}}{\rho - 1} \leq I_{UB} \leq \underline{y}.$$

Hence using these bounds and Proposition 6.2. provides the result. The Bonferroni inequality provides that the level of this CI is greater than or equal to $1 - 2\alpha_1 - \alpha_2$.

PROOF OF PROPOSITION 6.9. Under hypothesis (iv), $I_{UB} = \frac{\rho\underline{y}}{\rho - 1}$. Hence, using the proof of Proposition 6.8. and the expression of I_{UB} provide the result.

PROOF OF PROPOSITION 6.10. Proposition 6.7. provides nonparametric CIs for $I_B \Pr(\underline{y} \leq Y \leq \overline{y}) + I_{LB} \Pr(Y \geq \overline{y})$ under hypotheses similar to those assumed in

Proposition 7.3. Moreover, under hypothesis (iv) $I_{UB} = \frac{\rho \underline{y}}{\rho - 1}$ and, hence, if

$$C_\gamma(\alpha_3) = \{\gamma_0 \in \mathbb{R} : \gamma_l \leq \gamma_0 \leq \gamma_u\}$$

is a confidence interval for γ with level $1 - \alpha_3$, then

$$C_{I_{UB}}(\alpha_3) = \left\{ \mu_0 \in \mathbb{R} : \frac{\underline{y}}{1 - \frac{1}{\rho_l}} \leq \mu_0 \leq \frac{\underline{y}}{1 - \frac{1}{\rho_u}} \right\}$$

is a CI for I_{UB} with level $1 - \alpha_3$. Using $C_{I_{UB}}(\alpha_3)$ and Proposition 6.7. yields the result in Proposition 6.10. The Bonferroni inequality gives that the level of this CI is greater than or equal to $1 - 2\alpha_1 - \alpha_2 - \alpha_3$.

PROOF OF RESULT 7.1. If $Y \sim Pareto(\overline{y}, \delta) \mid Y \geq \overline{y}$, $\delta > 1$ and $\overline{y} > 0$ then

$$E\left(Y \log Y \mid Y \geq \overline{y}\right) = \int_{\overline{y}}^{+\infty} y \log(y) dF_{Y|Y \geq \overline{y}}(y)$$

where for $y \geq \overline{y}$

$$\begin{aligned} F_{Y|Y \geq \overline{y}}(y) &= 1 - \left(\frac{\overline{y}}{y}\right)^\gamma \\ &\Rightarrow \quad dF_{Y|Y \geq \overline{y}}(y) = -\gamma \left(\frac{\overline{y}}{y}\right)^{\gamma - 1} \overline{y} \left(-\frac{1}{y^2}\right) dy = \gamma \left(\frac{\overline{y}}{y}\right)^\gamma \left(\frac{1}{y}\right) dy \end{aligned}$$

and

$$\begin{aligned} E\left(Y \log Y \mid Y \geq \overline{y}\right) &= \int_{\overline{y}}^{+\infty} y \log(y) \gamma \left(\frac{\overline{y}}{y}\right)^\gamma \left(\frac{1}{y}\right) dy \\ &= \gamma \overline{y}^\gamma \int_{\overline{y}}^{+\infty} \frac{\log y}{y^\gamma} dy \\ &= \gamma \overline{y}^\gamma \left\{ \left[-\frac{\log y}{(\gamma - 1) y^{\gamma - 1}} \right]_{\overline{y}}^{+\infty} + \int_{\overline{y}}^{+\infty} \frac{1}{(\gamma - 1) y^\gamma} dy \right\} \end{aligned}$$

where the last equality is derived by integration by parts. Then,

$$\begin{aligned} E\left(Y \log Y \mid Y \geq \overline{y}\right) &= \gamma \overline{y}^{\gamma} \left\{ -\frac{1}{(\gamma-1)} \lim_{y \to +\infty} \frac{\log y}{y^{\gamma-1}} + \frac{\log \overline{y}}{(\gamma-1)\overline{y}^{\gamma-1}} + \frac{1}{(\gamma-1)^2 \overline{y}^{\gamma-1}} \right\} \\ &= \frac{\gamma \overline{y} \log \overline{y}}{\gamma-1} + \frac{\gamma \overline{y}}{(\gamma-1)^2}. \end{aligned}$$

Under hypothesis (i.i)

$$F_{Y|Y \geq \overline{y}}(y) \geq 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma}.$$

It follows that,

$$E\left(Y \log Y \mid Y \geq \overline{y}\right) \leq \frac{\gamma \overline{y} \log \overline{y}}{\gamma-1} + \frac{\gamma \overline{y}}{(\gamma-1)^2}.$$

Proof of result 7.2. Following the proof of result 7.1., if

$$F_{Y|Y \geq \overline{y}}(y) = 1 - \left(\frac{\overline{y}}{y}\right)^{\gamma}$$

then

$$E\left(Y \log Y \mid Y \geq \overline{y}\right) = \frac{\gamma \overline{y} \log \overline{y}}{\gamma-1} + \frac{\gamma \overline{y}}{(\gamma-1)^2}.$$

Conclusion générale

Cette thèse offre deux types de contributions à la littérature. La première contribution est purement statistique. Elle consiste à proposer des intervalles de confiance non paramétriques exacts pour la moyenne d'une variable aléatoire bornée, que la variable étudiée soit bornée ou pas. Dans le cas ou la variable est bornée, nous montrons que ces intervalles de confiance peuvent être déduites de bandes de confiance pour la fonction de distribution sous-jacente en utilisant des techniques de projection. Lorsque la variable aléatoire n'est pas bornée, nous proposons un principe de projection généralisé qui s'applique aux fonctions de distributions dont les queues sont bornées par des lois de Pareto. A première vue, les méthodes d'inférence proposées concernent uniquement la construction d'intervalles de confiance pour la moyenne. Toutefois, l'approche utilisée est loin d'être aussi restrictive qu'elle paraît. Résoudre le problème pour la moyenne d'une variable Y permet de le résoudre pour tous les moments de Y. Pour cela, il suffit de remplacer la série de données de Y par une fonction de cette dernière. Par exemple, si on s'intéresse à construire des intervalles de confiance pour le moment d'ordre 2 de Y, il suffit de remplacer les observations de Y par le carré de ces observations, de construire la fonction de répartition empirique qui correspond à la nouvelle série de données et d'appliquer les méthodes d'inférence proposées sur cette dernière. Les intervalles de confiance ainsi obtenus pour la moyenne des données transformées constituent des intervalles de confiance pour le moment d'ordre 2 de Y. Utilisant ce schéma, toutes les transformations de variables aléatoires peuvent être envisagées. Les transformations continues sont construites en utilisant les méthodes présentées dans cette thèse alors que pour les transformations non continues, d'intéressantes propriétés de monotonicité fournies dans chacun des trois articles permettent de les étudier.

Le deuxième type de contribution est économétrique. Il consiste à proposer des intervalles de confiance exacts pour les mesures de pauvreté de Foster, Greer et Thorbecke (1984) et les mesures d'inégalités les plus populaires: les mesures d'entropie généralisée, de déviation logarithmique et d'Atkinson et les indices de Theil, de Lorenz, de Gini et de variation logarithmique. Nous proposons des expressions explicites et faciles à calculer

pour ces intervalles et montrons par une étude Monte Carlo que ces intervalles sont fiables et robustes, à l'inverse des intervalles asymptotique et bootstrap. Pour illustration, nous analysons dans les articles 2 et 3 les profils de pauvreté et d'inégalités des ménages ruraux au Mexique en 1998 en utilisant des données du programme PROGRESA. Les résultats montrent que les intervalles asymptotiques sont souvent trop petits pour être réalistes alors que l'intervalle bootstrap peut exploser. L'analyse montre que le profil de pauvreté des ménages Mexicains dépend grandement du type de chef de ménage: les niveaux de pauvreté et d'inégalités des ménages dont le chef est un homme ou est éduqué sont moins élevés que ceux des autres ménages. Par conséquent, les mesures destinées à réduire le taux d'illettrisme et à sécuriser le revenu des ménages dont le chef est une femme pourraient aider à réduire la pauvreté et les inégalités dans le Mexique rural.

Références

Abdelkhalek, T. & Dufour, J.-M. (1998), 'Statistical inference for computable general equilibrium models, with application to a model of the Moroccan economy', **LXXX**, 520–534.

Anderson, T. W. (1969), '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.

Anderson, T. W. & Darling, D. A. (1952), 'Asymptotic theory of certain goodness-of-fit criteria based on stochastic process', *Annals of Mathematical Statistics* **23**, 193–212.

Bahadur, R. R. & Savage, L. R. (1956), 'The nonexistence of certain statistical procedures in nonparametric problems', *Annals of Mathematical Statistics* **27**(4), 1115–1122.

Beran, R. (1988), 'Prepivoting test statistics : a bootstrap view of asymptotic refinements', *Journal of the American Statistical Association* **83**(403), 687–697.

Berk, R. H. & Jones, D. H. (1977), 'Relatively optimal combinations of test statistics', *Scan J Statist.* .

Berk, R. H. & Jones, D. H. (1979), 'Goodnees-of-fit test statistics that dominate the Kolmogorov statistics', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* .

Biewen, M. (2002), 'Bootstrap inference for inequality, mobility and poverty measurement', *Journal of Econometrics* **108**, 317–342.

Brachman, K., Stich, A. & Trede, M. (1996), 'Evaluating parametric income distribution models', *Allgemeines Statistisches Archiv* **80**, 285–298.

Breth, M. (1976), 'Nonparametric confidence intervals for a mean using censored data', *Journal of Royal Statistical Society serie B* **38**, 251–254.

Breth, M., Maritz, J. S. & Williams, E. J. (1978), 'On distribution-free lower confidence limits for the mean of a nonnegative random variable', *Biometrika* **65**, 529–534.

Chambers, J., William, C., Beat, K. & Paul, T. (1983), 'Graphical methods for data analysis,'.

Chen, Z. (1996), 'Joint confidence region for the parameters of pareto distribution', *Metrika* **44**, 191–197.

Cheng, R. C. H. & Iles, T. C. (1988), 'One-sided confidence bands for cumulative distribution function', *The Annals of Statistics* **30**(2), 155–159.

Cochran, W. G. & Snedecor, G. W. (1989), 'Statistical methods'.

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

Cowell, F. A. (1989), 'Sampling variance and decomposable inequality measures', *Journal of Econometrics* **42**, 27–41.

Cowell, F. A. (2003), Theil, inequality and the structure of income distribution. working paper, STICERD - Distributional Analysis Research Programme Papers 67, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Cowell, F. A. & Flachaire, E. (2002), 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.

Cowell, F. A. & Victoria-Feser, M. P. (1996), 'Poverty measurement with contaminated data : a robust approach', *European Economic Review* **40**, 1761–1771.

Cramér, H. (1928), 'On the composition of elementary errors, second paper, statistical applications', *Skand. Actuar* **11**, 141–180.

Dardanoni, V. & Forcina, A. (1999), 'Inference for Lorenz curve orderings', *Econometrics Journal* **2**, 49–75.

Davidson, R. & Duclos, J.-Y. (2000), 'Statistical inference for stochastic dominance and for the measurement of poverty and inequality', *Econometrica* **68**, 1435–1464.

Davidson, R. & Flachaire, E. (2007), 'Asymptotic and bootstrap inference for inequality and poverty measure', *Journal of Econometrics* **141**(1), 141–166.

Demombynes, G., Elbers, C., Lanjouw, J. O. & Lanjouw, P. (2007), How good a map? putting small area estimation to the test. World Bank Policy Research Working Paper No. 4155.

DiCiccio, T. J. & Efron, B. (1996), 'Bootstrap confidence intervals', *Statistical Science* **11**, 189–228.

Diouf, M. A. & Dufour, J. M. (2005*a*), Improved exact nonparametric confidence bands and tests for distribution functions based on standardized empirical distribution functions. Working paper, Département de Sciences Economiques, Université de Montréal.

Diouf, M. A. & Dufour, J. M. (2005*b*), Improved nonparametric inference for the mean of a bounded random variable with application to poverty measures. Working paper, Département de Sciences Economiques, Université de Montréal.

Dufour, J. M. (1990), 'Exact tests and confidence sets in linear regressions with autocorrelated errors', *Econometrica* **58**(2), 475–494.

Dufour, J. M. (1995), Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics. Technical report, C.R.D.E., Université de Montréal.

Dufour, J.-M. (2003), 'Identification, weak instruments and statistical inference in econometrics'.

Dufour, J. M. & Khalaf, L. (2001), *Monte Carlo Tests in Econometrics*, Companion to Theoritical Econometrics,Badi Baltagi, Blockwell, Oxford, U. K., 2001, Chapter 23, 494-519.

Dufour, J.-M. & Kiviet, J. F. (19984), 'Exact inference methods for first-order autoregressive distributed lag models', **66**, 79–104.

Dufour, J.-M. & Neifar, M. (2004), 'Méthodes d'inférence exactes pour un modèle de régression avec erreurs AR(2) gaussiennes'.

Dufour, J. M. & Taamouti, M. (1999), 'Projection-based statistical inference in linear structural models with possibly weak instruments', *Econometrica* .

Dufour, J.-M. & Taamouti, M. (2005), 'Projection-based statistical inference in linear structural models with possibly weak instruments', **73**(4), 1351–1365.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses'.

Eicker, F. (1979), 'The asymptotic distribution of the suprema of the standardized empirical process', *The Annals of Statistics* **7**(1), 116–138.

Fieller, E. C. (1940), 'The biological standardization of insulin', *Journal of the Royal Statistical Society (Supplement)* **7**, 1–64.

Fieller, E. C. (1954), 'Some problems in interval estimation', *Journal of the Royal Statistical Society Series B* **16**, 175–185.

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

Foster, J. E., Greer, J. & Thorbecke, E. (1984), 'A class of decomposable poverty measures', *Econometrica* **52**, 761–765.

Gleser, L. J. (1985), 'Exact power of goodness-of-fit tests of komogorov type for discontinuous distributions', *Journal of the American Statistical Association* **80**(392).

Hoeffding, W. (1963), 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association* **58**, 13–29.

Hora, J. A. & Hora, S. C. (1990), 'Nonparametric bounds for errors in dollar unit sampling based on the cumulative distribution function', *Proceedings of the Decision Sciences Institute* pp. 115–117.

Horowitz, J. L. (2001), 'The bootstrap', *Handbook of Econometrics* **5**, 3159–3228. J. J. HECKMAN and E. E. LEAMER, eds., Elsevier Science.

Jaeschke, D. (1979), 'The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals', *The Annals of Statistics* **7**(1).

Jager, L. & Wellner, J. A. (2004), 'A new goodness of fit test: the reversed Berk-Jones statistic'. Technical Report 443, Department of Statistics, University of Washington.

Kakwani, N. (1993), 'Statistical inference in the measurement of poverty', *Review of Economics and Statistics* **75**, 632–639.

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

Kolmogorov, A. N. (1941), 'Confidence limits for an unknown distribution function', *Annals of Mathematical Statistics* **12**, 461–463.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Lilliefors, H. (1967), 'On the kolmogorov-smirnov test for normality with mean and variance unknown', *Journal of the American Statistical Association* .

Mills, J. & Zandvakili, S. (1997), 'Statistical inference via bootstrapping for measures of inequality', *Journal of Applied Econometrics* **12**, 103–150.

Mises, V. R. (1931), 'F. deuticke leipzig und wein', *Wahrscheinlichkeitsrechnung* pp. 316–335.

Noether, G. E. (1963), 'Note on kolmogorov statistic in the discrete case', *Metrika* **7**, 115–116.

Owen, A. B. (1995), 'Nonparametric likelihood confidence bands for a distribution function', *American Statistical Association* **90**, 516–521.

Rongve, I. (1997), 'Statistical inference for poverty indices with fixed poverty lines', *Applied Economics* **29**.

Shapiro, S. S. & Wilk, M. B. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika* **52**.

Smirnov, N. V. (1944), 'Approximate laws of distribution of random variables from empirical data', *Uspehi MatematiÏceskih Nauk* **10**, 179–206.

Sutton, V. K. & Young, D. M. (1997), 'A comparison of distribution-free confidence interval methods for estimating the mean of a random variable with bounded support', *ASA Proceedings of the Business and Economic Statistics Section* pp. 81–84. American Statistical Association.

Theil, H. (1967), 'Economics and information theory poverty', *Amsterdam : North Holland* .

Van-Garderen, K. J. & Schluter, C. (2003), Improving finite sample confidence intervals for inequality and poverty measures. Discussion paper 2003/02,University Van Amsterdam Econometrics.

Zheng, B. (2001), 'Statistical inference for poverty indices with relative poverty lines', *Journal of Econometrics* **101**, 337–356.