Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2010, Article ID 718905, 26 pages
doi:10.1155/2010/718905

Research Article

Zenga's New Index of Economic Inequality, Its Estimation, and an Analysis of Incomes in Italy

Francesca Greselin,[1] Leo Pasquazzi,[1] and Ričardas Zitikis[2]

[1] *Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Università di Milano, Bicocca 20126, Milan, Italy*
[2] *Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON, Canada N6A 5B7*

Correspondence should be addressed to Ričardas Zitikis, zitikis@stats.uwo.ca

Received 2 October 2009; Accepted 28 February 2010

Academic Editor: Madan L. Puri

For at least a century academics and governmental researchers have been developing measures that would aid them in understanding income distributions, their differences with respect to geographic regions, and changes over time periods. It is a fascinating area due to a number of reasons, one of them being the fact that different measures, or indices, are needed to reveal different features of income distributions. Keeping also in mind that the notions of poor and rich are relative to each other, Zenga (2007) proposed a new index of economic inequality. The index is remarkably insightful and useful, but deriving statistical inferential results has been a challenge. For example, unlike many other indices, Zenga's new index does not fall into the classes of L-, U-, and V-statistics. In this paper we derive desired statistical inferential results, explore their performance in a simulation study, and then use the results to analyze data from the Bank of Italy Survey on Household Income and Wealth (SHIW).

1. Introduction

Measuring and analyzing incomes, losses, risks, and other random outcomes, which we denote by X, has been an active and fruitful research area, particularly in the fields of econometrics and actuarial science. The Gini index is arguably the most popular measure of inequality, with a number of extensions and generalizations available in the literature. Keeping in mind that the notions of poor and rich are relative to each other, Zenga [1] constructed an index that reflects this relativity. We will next recall the definitions of the Gini and Zenga indices.

Let $F(x) = \mathbf{P}[X \leq x]$ denote the cumulative distribution function (cdf) of the random variable X, which we assume to be nonnegative throughout the paper. Let $F^{-1}(p) = \inf\{x : F(x) \geq p\}$ denote the corresponding quantile function. The Lorenz curve $L_F(p)$ is given by the formula (see [2])

$$L_F(p) = \frac{1}{\mu_F} \int_0^p F^{-1}(s)ds, \tag{1.1}$$

where $\mu_F = \mathbf{E}[X]$ is the unknown true mean of X. Certainly, from the rigorous mathematical point of view we should call $L_F(p)$ the Lorenz *function*, but this would deviate from the widely accepted usage of the term "Lorenz curve". Hence, curves and functions are viewed as synonyms throughout this paper.

The classical Gini index G_F can now be written as follows:

$$G_F = \int_0^1 \left(1 - \frac{L_F(p)}{p}\right) \varphi(p)dp, \tag{1.2}$$

where $\varphi(p) = 2p$. Note that $\varphi(p)$ is a density function on $[0,1]$. Given the usual econometric interpretation of the Lorenz curve [3], the function

$$G_F(p) = 1 - \frac{L_F(p)}{p}, \tag{1.3}$$

which we call the Gini curve, is a relative measure of inequality (see [4]). Indeed, $L_F(p)/p$ is the ratio between (i) the mean income of the poorest $p \times 100\%$ of the population and (ii) the mean income of the entire population: the closer to each other these two means are, the lower is the inequality.

Zenga's [1] index Z_F of inequality is defined by the formula

$$Z_F = \int_0^1 Z_F(p)dp, \tag{1.4}$$

where the Zenga curve $Z_F(p)$ is given by

$$Z_F(p) = 1 - \frac{L_F(p)}{p} \cdot \frac{1-p}{1-L_F(p)}. \tag{1.5}$$

The Zenga curve measures the inequality between (i) the poorest $p \times 100\%$ of the population and (ii) the richer remaining $(1-p) \times 100\%$ part of the population by comparing the mean incomes of these two disjoint and exhaustive subpopulations. We will elaborate on this interpretation later, in Section 5.

The Gini and Zenga indices G_F and Z_F are (weighted) averages of the Gini and Zenga curves $G_F(p)$ and $Z_F(p)$, respectively. However, while in the case of the Gini index the weight function (i.e., the density) $\varphi(p) = 2p$ is employed, in the case of the Zenga index the uniform weight function $\varphi(p) = 1$ is used. As a consequence, the Gini index underestimates

comparisons between the very poor and the whole population, and emphasizes comparisons which involve almost identical population subgroups. From this point of view, the Zenga index is more impartial: it is based on all comparisons between complementary disjoint population subgroups and gives the same weight to each comparison. Hence, the Zenga index Z_F detects, with the same sensibility, all deviations from equality in any part of the distribution.

To illustrate the Gini curve $G_F(p)$ and its weighted version $g_F(p) = G_F(p)\psi(p)$, and to also facilitate their comparisons with the Zenga curve $Z_F(p)$, we choose the Pareto distribution

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\theta}, \quad x \geq x_0, \tag{1.6}$$

where $x_0 > 0$ and $\theta > 0$ are parameters. Later in this paper, we will use this distribution in a simulation study, setting $x_0 = 1$ and $\theta = 2.06$. Note that when $\theta > 2$, then the second moment of the distribution is finite. The "heavy-tailed" case $1 < \theta < 2$ is also of interest, especially when modeling incomes of countries with very high economic inequality. We will provide additional details on the case in Section 5.

Note 1. Pareto distribution (1.6) is perhaps the oldest model for income distributions. It dates back to Pareto [5], and Pareto [6]. Pareto's original empirical research suggested him that the number of tax payers with income x is roughly proportional to $x^{-(\theta+1)}$, where θ is a parameter that measures inequality. For historical details on the interpretation of this parameter in the context of measuring economic inequality, we refer to Zenga [7]. We can view the parameter $x_0 > 0$ as the lowest taxable income. In addition, besides being the greatest lower bound of the distribution support, x_0 is also the scale parameter of the distribution and thus does not affect our inequality indices and curves, as we will see in formulas below.

Note 2. The Pareto distribution is positively supported, $x \geq x_0 > 0$. In real surveys, however, in addition to many positive incomes we may also observe some zero and negative incomes. This happens when evaluating net household incomes, which are the sums of payroll incomes (net wages, salaries, fringe benefits), pensions and net transfers (pensions, arrears, financial assistance, scholarships, alimony, gifts). Paid alimony and gifts are subtracted in forming the incomes. However, negative incomes usually happen in the case of very few statistical units. For example, in the 2006 Bank of Italy survey we observe only four households with nonpositive incomes, out of the total of 7,766 households. Hence, it is natural to fit the Pareto model to the positive incomes and keep in mind that we are actually dealing with a conditional distribution. If, however, it is desired to deal with negative, null, and positive incomes, then instead of the Pareto distribution we may switch to different ones, such as Dagum distributions with three or four parameters [8–10].

Corresponding to Pareto distribution (1.6), the Lorenz curve is given by the formula $L_F(p) = 1 - (1-p)^{1-1/\theta}$ (see [11]), and thus the Gini curve becomes $G_F(p) = ((1-p)^{1-1/\theta} - (1-p))/p$. In Figure 1(a) we have depicted the Gini and weighted Gini curves. The corresponding Zenga curve is equal to $Z_F(p) = (1-(1-p)^{1/\theta})/p$ and is depicted in Figure 1(b), alongside the Gini curve $G_F(p)$ for an easy comparison. Figure 1(a) allows us to appreciate how the Gini weight function $\psi(p) = 2p$ disguises the high inequality between the mean income of the very poor and that of the whole population, and overemphasizes comparisons between almost

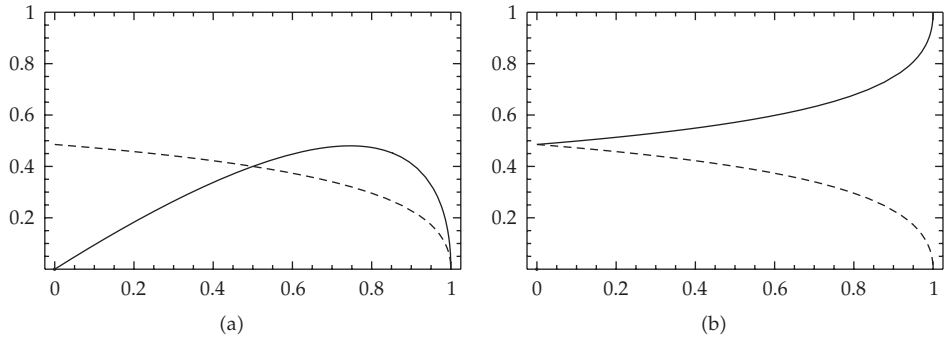

Figure 1: The Gini curve $G_F(p)$(dashed; (a) and (b)), the weighted Gini curve $g_F(p)$(solid; (a)), and the Zenga curve $Z_F(p)$(solid; (b)) in the Pareto case with $x_0 = 1$ and $\theta = 2.06$.

identical subgroups. The outcome is that the Gini index G_F underestimates inequality. In Figure 1(b) we see the difference between the Gini and Zenga inequality curves. For example, $G_F(p)$ for $p = 0.8$ yields 0.296, which tells us that the mean income of the poorest 80% of the population is 29.6% lower than the mean income of the whole population, while the corresponding ordinate of the Zenga curve is $Z_F(0.8) = 0.678$, which tells us that the mean income of the poorest 80% of the population is 67.8% lower than the mean income of the remaining (richer) part of the population.

The rest of this paper is organized as follows. In Section 2 we define two estimators of the Zenga index Z_F and develop statistical inferential results. In Section 3 we present results of a simulation study, which explores the empirical performance of two Zenga estimators, $\widehat{Z}_n$ and $\widetilde{Z}_n$, including coverage accuracy and length of several types of confidence intervals. In Section 4 we present an analysis of the the Bank of Italy Survey on Household Income and Wealth (SHIW) data. In Section 5 we further contribute to the understanding of the Zenga index Z_F by relating it to lower and upper conditional expectations, as well as to the conditional tail expectation (CTE), which has been widely used in insurance. In Section 6 we provide a theoretical background of the aforementioned two empirical Zenga estimators. In Section 7 we justify the definitions of several variance estimators as well as their uses in constructing confidence intervals. In Section 8 we prove Theorem 2.1 of Section 2, which is the main technical result of the present paper. Technical lemmas and their proofs are relegated to Section 9.

2. Estimators and Statistical Inference

Unless explicitly stated otherwise, our statistical inferential results are derived under the assumption that data are outcomes of independent and identically distributed (i.i.d.) random variables.

Hence, let $X_1, \ldots, X_n$ be independent copies of X. We use two nonparametric estimators for the Zenga index Z_F. The first one [12] is given by the formula

$$\widehat{Z}_n = 1 - \frac{1}{n}\sum_{i=1}^{n-1} \frac{i^{-1}\sum_{k=1}^{i} X_{k:n}}{(n-i)^{-1}\sum_{k=i+1}^{n} X_{k:n}}, \tag{2.1}$$

where $X_{1:n} \le \cdots \le X_{n:n}$ are the order statistics of $X_1, \ldots, X_n$. With $\overline{X}$ denoting the sample mean of $X_1, \ldots, X_n$, the second estimator of the Zenga index Z_F is given by the formula

$$\begin{aligned}\widetilde{Z}_n = &-\sum_{i=2}^{n} \frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}} \log\left(\frac{i}{i-1}\right) \\ &+ \sum_{i=1}^{n-1} \left(\frac{\overline{X}}{X_{i:n}} - 1 - \frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}}\right) \log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right).\end{aligned} \tag{2.2}$$

The two estimators $\widehat{Z}_n$ and $\widetilde{Z}_n$ are asymptotically equivalent. However, despite the fact that the estimator $\widetilde{Z}_n$ is more complex, it will nevertheless be more convenient to work with when establishing asymptotic results later in this paper.

Unless explicitly stated otherwise, we assume throughout that the cdf $F(x)$ of X is a continuous function. We note that continuous cdf's are natural choices when modeling income distributions, insurance risks, and losses (see, e.g., [13]).

Theorem 2.1. *If the moment* $\mathbf{E}[X^{2+\alpha}]$ *is finite for some* $\alpha > 0$*, then one has the asymptotic representation*

$$\sqrt{n}\left(\widetilde{Z}_n - Z_F\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) + o_{\mathbf{P}}(1), \tag{2.3}$$

where $o_{\mathbf{P}}(1)$ *denotes a random variable that converges to* 0 *in probability when* $n \to \infty$*, and*

$$h(X_i) = \int_0^{\infty} (\mathbf{1}\{X_i \le x\} - F(x)) w_F(F(x)) dx \tag{2.4}$$

with the weight function

$$w_F(t) = -\frac{1}{\mu_F} \int_0^t \left(\frac{1}{p} - 1\right) \frac{L_F(p)}{(1 - L_F(p))^2} dp + \frac{1}{\mu_F} \int_t^1 \left(\frac{1}{p} - 1\right) \frac{1}{1 - L_F(p)} dp. \tag{2.5}$$

In view of Theorem 2.1, the asymptotic distribution of $\sqrt{n}\,(\widetilde{Z}_n - Z_F)$ is centered normal with the variance $\sigma_F^2 = \mathbf{E}[h^2(X)]$, which is finite (see Theorem 7.1) and can be written as follows:

$$\sigma_F^2 = \int_0^{\infty} \int_0^{\infty} (\min\{F(x), F(y)\} - F(x)F(y)) w_F(F(x)) w_F(F(y)) dx\, dy. \tag{2.6}$$

Alternatively,

$$\sigma_F^2 = \int_0^1 \left(\int_{[0,u)} t w_F(t) dF^{-1}(t) - \int_{[u,1)} (1-t) w_F(t) dF^{-1}(t)\right)^2 du. \tag{2.7}$$

The latter expression of σ_F^2 is particularly convenient when working with distributions for which the first derivative (when it exists) of the quantile $F^{-1}(t)$ is a relatively simple function, as is the case for a large class of distributions (see, e.g., [14]). However, irrespectively of what expression for the variance σ_F^2 we use, the variance is unknown since the cdf $F(x)$ is unknown, and thus σ_F^2 needs to be estimated empirically.

2.1. One Sample Case

Replacing the population cdf everywhere on the right-hand side of (2.6) by the empirical cdf $F_n(x) = n^{-1}\sum_{i=1}^{n}\mathbf{1}\{X_i \le x\}$, where $\mathbf{1}$ denotes the indicator function, we obtain (Theorem 7.2) the following estimator of the variance σ_F^2:

$$\begin{aligned} S_{X,n}^2 = &\sum_{k=1}^{n-1}\sum_{l=1}^{n-1}\left(\frac{\min\{k,l\}}{n} - \frac{k}{n}\frac{l}{n}\right) \\ &\times w_{X,n}\left(\frac{k}{n}\right)w_{X,n}\left(\frac{l}{n}\right)(X_{k+1:n} - X_{k:n})(X_{l+1:n} - X_{l:n}), \end{aligned} \tag{2.8}$$

where

$$w_{X,n}\left(\frac{k}{n}\right) = -\sum_{i=1}^{k} I_{X,n}(i) + \sum_{i=k+1}^{n} J_{X,n}(i) \tag{2.9}$$

with the following expressions for the summands $I_{X,n}(i)$ and $J_{X,n}(i)$: first,

$$I_{X,n}(1) = -\frac{\sum_{k=2}^{n} X_{k:n} - (n-1)X_{1:n}}{\left(\sum_{k=1}^{n} X_{k:n}\right)\left(\sum_{k=2}^{n} X_{k:n}\right)} + \frac{1}{X_{1,n}}\log\left(1 + \frac{X_{1:n}}{\sum_{k=2}^{n} X_{k:n}}\right). \tag{2.10}$$

Furthermore, for every $i = 2, \ldots, n-1$,

$$\begin{aligned} I_{X,n}(i) = &\, n\frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)^2}\log\left(\frac{i}{i-1}\right) \\ &- \frac{\left(\sum_{k=i+1}^{n} X_{k:n} - (n-i)X_{i:n}\right)\left(\sum_{k=1}^{n} X_{k:n}\right)}{\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)\left(\sum_{k=i+1}^{n} X_{k:n}\right)\left(\sum_{k=i}^{n} X_{k:n}\right)} \\ &+ \left(\frac{1}{X_{i:n}} + n\,\frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)^2}\right)\log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right), \end{aligned} \tag{2.11}$$

$$\begin{aligned} J_{X,n}(i) = &\, \frac{n}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}}\log\left(\frac{i}{i-1}\right) \\ &- \frac{\sum_{k=i+1}^{n} X_{k:n} - (n-i)X_{i:n}}{X_{i:n}\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)}\log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right). \end{aligned} \tag{2.12}$$

Finally,

$$J_{X,n}(n) = \frac{1}{X_{n,n}} \log\left(\frac{n}{n-1}\right). \tag{2.13}$$

With the just defined estimator $S^2_{X,n}$ of the variance σ^2_F, we have the asymptotic result:

$$\frac{\sqrt{n}\left(\widetilde{Z}_n - Z_F\right)}{S_{X,n}} \longrightarrow_d \mathcal{N}(0,1), \tag{2.14}$$

where $\rightarrow_d$ denotes convergence in distribution.

2.2. Two Independent Samples

We now discuss a variant of statement (2.14) in the case of two populations when samples are independent. Namely, let the random variables $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim H$ be independent within and between the two samples. Just like in the case of the cdf $F(x)$, here we also assume that the cdf $H(x)$ is continuous and $\mathbf{E}[Y^{2+\alpha}] < \infty$ for some $\alpha > 0$. Furthermore, we assume that the sample sizes n and m are comparable, which means that there exists $\eta \in (0,1)$ such that

$$\frac{m}{n+m} \longrightarrow \eta \in (0,1) \tag{2.15}$$

when both n and m tend to infinity. From statement (2.3) and its counterpart for $Y_i \sim H$ we then have that the quantity $\sqrt{nm/(n+m)}\ ((\widetilde{Z}_{X,n} - \widetilde{Z}_{Y,m}) - (Z_F - Z_H))$ is asymptotically normal with mean zero and the variance $\eta\sigma^2_F + (1-\eta)\sigma^2_H$. To estimate the variances σ^2_F and σ^2_H, we use $S^2_{X,n}$ and $S^2_{Y,n}$, respectively, and obtain the following result:

$$\frac{\left(\widetilde{Z}_{X,n} - \widetilde{Z}_{Y,m}\right) - (Z_F - Z_H)}{\sqrt{(1/n)S^2_{X,n} + (1/m)S^2_{Y,m}}} \longrightarrow_d \mathcal{N}(0,1). \tag{2.16}$$

2.3. Paired Samples

Consider now the case when the two samples $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim H$ are paired. Thus, we have that $m = n$, and we also have that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed. Nothing is assumed about the joint distribution of (X, Y). As before, the cdf's $F(x)$ and $H(y)$ are continuous and both have finite moments of order $2 + \alpha$, for some $\alpha > 0$. From statement (2.3) and its analog for Y we have that $\sqrt{n}\ ((\widetilde{Z}_{X,n} - \widetilde{Z}_{Y,n}) - (Z_F - Z_H))$ is asymptotically normal with mean zero and the variance $\sigma^2_{F,H} = \mathbf{E}[(h(X) - h(Y))^2]$. The latter variance can of course be written as $\sigma^2_F - 2\mathbf{E}[h(X)h(Y)] + \sigma^2_H$. Having already constructed estimators $S^2_{X,n}$ and $S^2_{Y,n}$, we are only left to construct an estimator

for $\mathbf{E}[h(X)h(Y)]$. (Note that when X and Y are independent, then $\mathbf{P}[X \le x, Y \le y] = F(x)H(y)$ and thus the expectation $\mathbf{E}[h(X)h(Y)]$ vanishes.) To this end, we write the equation

$$\mathbf{E}[h(X)h(Y)] = \int_0^\infty \int_0^\infty \big(\mathbf{P}[X \le x, Y \le y] - F(x)H(y)\big) w_F(F(x)) w_H(H(y))\, dx\, dy. \tag{2.17}$$

Replacing the cdf's $F(x)$ and $H(y)$ everywhere on the right-hand side of the above equation by their respective empirical estimators $F_n(x)$ and $H_n(y)$, we have (Theorem 7.3)

$$\begin{aligned} S_{X,Y,n} = \sum_{k=1}^{n-1}\sum_{l=1}^{n-1} &\left(\frac{1}{n}\sum_{i=1}^{k} \mathbf{1}\{Y_{(i,n)} \le Y_{l:n}\} - \frac{k}{n}\frac{l}{n} \right) \\ &\times w_{X,n}\left(\frac{k}{n}\right) w_{Y,n}\left(\frac{l}{n}\right) (X_{k+1:n} - X_{k:n})(Y_{l+1:n} - Y_{l:n}), \end{aligned} \tag{2.18}$$

where $Y_{(1,n)}, \ldots, Y_{(n,n)}$ are the induced (by $X_1, \ldots, X_n$) order statistics of $Y_1, \ldots, Y_n$. (Note that when $Y \equiv X$, then $Y_{(i,n)} = Y_{i:n}$ and so the sum $\sum_{i=1}^{k} \mathbf{1}\{Y_{(i,n)} \le Y_{l:n}\}$ is equal to $\min\{k, l\}$; hence, estimator (2.18) coincides with estimator (2.8), as expected.) Consequently, $S^2_{X,n} - 2S_{X,Y,n} + S^2_{Y,n}$ is an empirical estimator of $\sigma^2_{F,H}$, and so we have that

$$\frac{\sqrt{n}\,\left(\widetilde{Z}_{X,n} - \widetilde{Z}_{Y,n}\right) - (Z_F - Z_H)}{\sqrt{S^2_{X,n} - 2S_{X,Y,n} + S^2_{Y,n}}} \longrightarrow_d \mathcal{N}(0,1). \tag{2.19}$$

We conclude this section with a note that the above established asymptotic results (2.14), (2.16), and (2.19) are what we typically need when dealing with two populations, or two time periods, but extensions to more populations and/or time periods would be a worthwhile contribution. For hints and references on the topic, we refer to Jones et al. [15] and Brazauskas et al. [16].

3. A Simulation Study

Here we investigate the numerical performance of the estimators $\widehat{Z}_n$ and $\widetilde{Z}_n$ by simulating data from Pareto distribution (1.6) with $x_0 = 1$ and $\theta = 2.06$. These choices give the value $Z_F = 0.6$, which is approximately seen in real income distributions. As to the (artificial) choice $x_0 = 1$, we note that since x_0 is the scale parameter in the Pareto model, the inequality indices and curves are invariant to it. Hence, all results to be reported in this section concerning the coverage accuracy and size of confidence intervals will not be affected by the choice $x_0 = 1$.

Following Davison and Hinkley [17, Chapter 5], we compute four types of confidence intervals: normal, percentile, BCa, and t-bootstrap. For normal and studentized bootstrap confidence intervals we estimate the variance using empirical influence values. For the estimator $\widetilde{Z}_n$, the influence values $h(X_i)$ are obtained from Theorem 2.1, and those for the estimator $\widehat{Z}_n$ using numerical differentiation as in Greselin and Pasquazzi [12].

In Table 1 we report coverage percentages of 10,000 confidence intervals, for each of the four types: normal, percentile, BCa, and t-bootstrap. Bootstrap-based approximations

Table 1: Coverage proportions of confidence intervals from the Pareto parent distribution with $x_0 = 1$ and $\theta = 2.06$ ($Z_F = 0.6$).

	$\widehat{Z}_n$				$\widetilde{Z}_n$			
	0.9000	0.9500	0.9750	0.9900	0.9000	0.9500	0.9750	0.9900
n	Normal confidence intervals							
200	0.7915	0.8560	0.8954	0.9281	0.7881	0.8527	0.8926	0.9266
400	0.8059	0.8705	0.9083	0.9409	0.8047	0.8693	0.9078	0.9396
800	0.8256	0.8889	0.9245	0.9514	0.8246	0.8882	0.9237	0.9503
n	Percentile confidence intervals							
200	0.7763	0.8326	0.8684	0.9002	0.7629	0.8190	0.8567	0.8892
400	0.8004	0.8543	0.8919	0.9218	0.7934	0.8487	0.8864	0.9179
800	0.8210	0.8777	0.9138	0.9415	0.8168	0.8751	0.9119	0.9393
n	BCa confidence intervals							
200	0.8082	0.8684	0.9077	0.9383	0.8054	0.867	0.9047	0.9374
400	0.8205	0.8863	0.9226	0.9531	0.8204	0.886	0.9212	0.9523
800	0.8343	0.8987	0.9331	0.9634	0.8338	0.8983	0.9323	0.9634
n	t-bootstrap confidence intervals							
200	0.8475	0.9041	0.9385	0.9658	0.8485	0.9049	0.9400	0.9675
400	0.8535	0.9124	0.9462	0.9708	0.8534	0.9120	0.9463	0.9709
800	0.8580	0.9168	0.9507	0.9758	0.8572	0.9169	0.9504	0.9754

have been obtained from 9,999 resamples of the original samples. As suggested by Efron [18], we have approximated the acceleration constant for the BCa confidence intervals by one-sixth times the standardized third moment of the influence values. In Table 2 we report summary statistics concerning the size of the 10,000 confidence intervals. As expected, the confidence intervals based on $\widehat{Z}_n$ and $\widetilde{Z}_n$ exhibit similar characteristics. We observe from Table 1 that all confidence intervals suffer from some undercoverage. For example, with sample size 800, about 97.5% of the studentized bootstrap confidence intervals with 0.99 nominal confidence level contain the true value of the Zenga index. It should be noted that the higher coverage accuracy of the studentized bootstrap confidence intervals (when compared to the other ones) comes at the cost of their larger sizes, as seen in Table 2. Some of the studentized bootstrap confidence intervals extend beyond the range $[0, 1]$ of the Zenga index Z_F, but this can easily be fixed by taking the minimum between the currently recorded upper bounds and 1, which is the upper bound of the Zenga index Z_F for every cdf F. We note that for the BCa confidence intervals, the number of bootstrap replications of the original sample has to be increased beyond 9,999 if the nominal confidence level is high. Indeed, for samples of size 800, it turns out that the upper bound of 1,598 (out of 10,000) of the BCa confidence intervals based on $\widehat{Z}_n$ and with 0.99 nominal confidence level is given by the largest order statistics of the bootstrap distribution. For the confidence intervals based on $\widetilde{Z}_n$, the corresponding figure is 1,641.

4. An Analysis of Italian Income Data

In this section we use the Zenga index Z_F to analyze data from the Bank of Italy Survey on Household Income and Wealth (SHIW). The sample of the 2006 wave of this survey contains 7,768 households, with 3,957 of them being panel households. For detailed information on

Table 2: Size of the 95% asymptotic confidence intervals from the Pareto parent distribution with $x_0 = 1$ and $\theta = 2.06$ ($Z_F = 0.6$).

	$\widehat{Z}_n$			$\widetilde{Z}_n$		
	min	mean	max	min	mean	max
n	Normal confidence intervals					
200	0.0680	0.1493	0.7263	0.0674	0.1500	0.7300
400	0.0564	0.1164	0.7446	0.0563	0.1167	0.7465
800	0.0462	0.0899	0.6528	0.0462	0.0900	0.6535
n	Percentile confidence intervals					
200	0.0673	0.1456	0.4751	0.0667	0.1462	0.4782
400	0.0561	0.1140	0.4712	0.0561	0.1143	0.4721
800	0.0467	0.0883	0.4110	0.0468	0.0884	0.4117
n	BCa confidence intervals					
200	0.0668	0.1491	0.4632	0.0661	0.1497	0.4652
400	0.0561	0.1183	0.4625	0.0558	0.1186	0.4629
800	0.0465	0.0925	0.4083	0.0467	0.0927	0.4085
n	t-bootstrap confidence intervals					
200	0.0677	0.2068	2.4307	0.0680	0.2099	2.5148
400	0.0572	0.1550	2.0851	0.0573	0.1559	2.1009
800	0.0473	0.1159	2.2015	0.0474	0.1162	2.2051

the survey, we refer to the Bank of Italy [19] publication. In order to treat data correctly in the case of different household sizes, we work with equivalent incomes, which we have obtained by dividing the total household income by an equivalence coefficient, which is the sum of weights assigned to each household member. Following the modified Organization for Economic Cooperation and Development (OECD) equivalence scale, we give weight 1 to the household head, 0.5 to the other adult members of the household, and 0.3 to the members under 14 years of age. It should be noted, however, that—as is the case in many surveys concerning income analysis—households are selected using complex sampling designs. In such cases, statistical inferential results are quite complex. To alleviate the difficulties, in the present paper we follow the commonly accepted practice and treat income data as if they were i.i.d.

In Table 3 we report the values of $\widehat{Z}_n$ and $\widetilde{Z}_n$ according to the geographic area of the households, and we also report confidence intervals for Z_F based on the two estimators. We note that two households in the sample had negative incomes in 2006, and so we have not included them in our computations.

Note 3. Removing the negative incomes from our current analysis is important as otherwise we would need to develop a much more complex methodology than the one offered in this paper. To give a flavour of technical challenges, we note that the Gini index may overestimate the economic inequality when negative, zero, and positive incomes are considered. In this case the Gini index needs to be renormalized as demonstrated by, for example, Chen et al. [20]. Another way to deal with the issue would be to analyze the negative incomes and their concentration separately from the zero and positive incomes and their concentration.

Consequently, the point estimates of Z_F are based on 7,766 equivalent incomes with $\widehat{Z}_n = 0.6470$ and $\widetilde{Z}_n = 0.6464$. As pointed out by Maasoumi [21], however, good care is

Table 3: Confidence intervals for Z_F in the 2006 Italian income distribution.

	$\widehat{Z}_n$ estimator				$\widetilde{Z}_n$ estimator			
	95%		99%		95%		99%	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Northwest: $n = 1988$, $\widehat{Z}_n = 0.5953$, $\widetilde{Z}_n = 0.5948$								
Normal	0.5775	0.6144	0.5717	0.6202	0.5771	0.6138	0.5713	0.6196
Student	0.5786	0.6168	0.5737	0.6240	0.5791	0.6172	0.5748	0.6243
Percent	0.5763	0.6132	0.5710	0.6193	0.5758	0.6124	0.5706	0.6185
BCa	0.5789	0.6160	0.5741	0.6234	0.5785	0.6156	0.5738	0.6226
Northeast: $n = 1723$, $\widehat{Z}_n = 0.6108$, $\widetilde{Z}_n = 0.6108$								
Normal	0.5849	0.6393	0.5764	0.6478	0.5849	0.6393	0.5764	0.6479
Student	0.5874	0.6526	0.5796	0.6669	0.5897	0.6538	0.5836	0.6685
Percent	0.5840	0.6379	0.5773	0.6476	0.5839	0.6379	0.5772	0.6475
BCa	0.5894	0.6478	0.5841	0.6616	0.5894	0.6479	0.5842	0.6615
Center: $n = 1574$, $\widehat{Z}_n = 0.6316$, $\widetilde{Z}_n = 0.6316$								
Normal	0.5957	0.6708	0.5839	0.6826	0.5956	0.6708	0.5838	0.6827
Student	0.5991	0.6991	0.5897	0.7284	0.6036	0.7016	0.5977	0.7311
Percent	0.5948	0.6689	0.5864	0.6818	0.5948	0.6688	0.5863	0.6818
BCa	0.6024	0.6850	0.5963	0.7021	0.6024	0.6850	0.5963	0.7020
South: $n = 1620$, $\widehat{Z}_n = 0.6557$, $\widetilde{Z}_n = 0.6543$								
Normal	0.6358	0.6770	0.6293	0.6834	0.6346	0.6756	0.6282	0.6820
Student	0.6371	0.6805	0.6313	0.6902	0.6371	0.6796	0.6320	0.6900
Percent	0.6351	0.6757	0.6286	0.6828	0.6337	0.6742	0.6276	0.6812
BCa	0.6375	0.6793	0.6325	0.6888	0.6363	0.6778	0.6315	0.6873
Islands: $n = 861$, $\widehat{Z}_n = 0.6109$, $\widetilde{Z}_n = 0.6095$								
Normal	0.5918	0.6317	0.5856	0.6380	0.5910	0.6302	0.5848	0.6364
Student	0.5927	0.6339	0.5864	0.6405	0.5928	0.6330	0.5874	0.6401
Percent	0.5897	0.6297	0.5839	0.6360	0.5885	0.6275	0.5831	0.6340
BCa	0.5923	0.6324	0.5868	0.6414	0.5914	0.6307	0.5860	0.6394
Italy (entire population): $n = 7766$, $\widehat{Z}_n = 0.6470$, $\widetilde{Z}_n = 0.6464$								
Normal	0.6346	0.6596	0.6307	0.6636	0.6341	0.6591	0.6302	0.6630
Student	0.6359	0.6629	0.6327	0.6686	0.6358	0.6627	0.6331	0.6683
Percent	0.6348	0.6597	0.6314	0.6640	0.6343	0.6592	0.6309	0.6635
BCa	0.6363	0.6619	0.6334	0.6676	0.6358	0.6613	0.6330	0.6669

needed when comparing point estimates of inequality measures. Indeed, direct comparison of the point estimates corresponding to the five geographic areas of Italy would lead us to the conclusion that the inequality is higher in the central and southern areas when compared to the northern area and the islands. But as we glean from pairwise comparisons of the confidence intervals, only the differences between the estimates corresponding to the northwestern and southern areas and perhaps to the islands and the southern area may be deemed statistically significant.

Moreover, we have used the paired samples of the 2004 and 2006 incomes of the 3,957 panel households in order to check whether during this time period there was a change in inequality among households. In Table 4 we report the values of $\widetilde{Z}_n$ based on the panel households for these two years, and the 95% confidence intervals for the difference between the values of the Zenga index for the years 2006 and 2004. These computations have been

Table 4: 95% confidence intervals for the difference of the Zenga indices between 2006 and 2004 in the Italian income distribution.

Northwest (926 pairs)			Northeast (841 pairs)		Center (831 pairs)	
$\widetilde{Z}_n^{(2006)}$		0.5797	$\widetilde{Z}_n^{(2006)}$	0.6199	$\widetilde{Z}_n^{(2006)}$	0.5921
$\widetilde{Z}_n^{(2004)}$		0.5955	$\widetilde{Z}_n^{(2004)}$	0.6474	$\widetilde{Z}_n^{(2004)}$	0.5766
Difference		−0.0158	Difference	−0.0275	Difference	0.0155
	Lower	Upper	Lower	Upper	Lower	Upper
Normal	−0.0426	0.0102	−0.0573	0.0003	−0.0183	0.0514
Student	−0.0463	0.0103	−0.0591	0.0017	−0.0156	0.0644
Percent	−0.0421	0.0108	−0.0537	0.0040	−0.0183	0.0505
BCa	−0.0440	0.0087	−0.0551	0.0022	−0.0130	0.0593
South (843 pairs)			Islands (512 pairs)		Italy (3953 pairs)	
$\widetilde{Z}_n^{(2006)}$		0.6200	$\widetilde{Z}_n^{(2006)}$	0.6179	$\widetilde{Z}_n^{(2006)}$	0.6362
$\widetilde{Z}_n^{(2004)}$		0.6325	$\widetilde{Z}_n^{(2004)}$	0.6239	$\widetilde{Z}_n^{(2004)}$	0.6485
Difference		−0.0125	Difference	−0.0060	Difference	−0.0123
	Lower	Upper	Lower	Upper	Lower	Upper
Normal	−0.0372	0.0129	−0.0333	0.0213	−0.0259	0.0007
Student	−0.0365	0.0166	−0.0351	0.0222	−0.0264	0.0013
Percent	−0.0372	0.0131	−0.0333	0.0214	−0.0253	0.0016
BCa	−0.0351	0.0162	−0.0331	0.0216	−0.0255	0.0013

based on formula (2.19). Having removed the four households with at least one negative income in the paired sample, we were left with a total of 3,953 observations. We see that even though we deal with large sample sizes, the point estimates alone are not reliable. Indeed, for Italy as the whole and for all geographic areas except the center, the point estimates suggest that the Zenga index decreased from the year 2004 to 2006. However, the 95% confidence intervals in Table 4 suggest that this change is not significant.

5. An Alternative Look at the Zenga Index

In various contexts we have notions of rich and poor, large and small, risky and secure. They divide the underlying populations into two parts, which we view as subpopulations. The quantile $F^{-1}(p)$, for some $p \in (0,1)$, usually serves as a boundary separating the two subpopulations. For example, we may define rich if $X > F^{-1}(p)$ and poor if $X \leq F^{-1}(p)$. Calculating the mean value of the former subpopulation gives rise to the upper conditional expectation $\mathbf{E}[X \mid X > F^{-1}(p)]$, which is known in the actuarial risk theory as the conditional tail expectation (CTE). Calculating the mean value of the latter subpopulation gives rise to the lower conditional expectation $\mathbf{E}[X \mid X \leq F^{-1}(p)]$, which is known in the econometric literature as the absolute Bonferroni curve, as a function of p.

Clearly, the ratio

$$R_F(p) = \frac{\mathbf{E}\left[X \mid X \leq F^{-1}(p)\right]}{\mathbf{E}\left[X \mid X > F^{-1}(p)\right]} \tag{5.1}$$

of the lower and upper conditional expectations takes on values in the interval $[0,1]$. When X is equal to any constant, which can be interpreted as the egalitarian case, then $R_F(p)$ is equal

to 1. The ratio $R_F(p)$ is equal to 0 for all $p \in (0,1)$ when the lower conditional expectation is equal to 0 for all $p \in (0,1)$. This means extreme inequality in the sense that, loosely speaking, there is only one individual who possesses the entire wealth. Our wish to associate the egalitarian case with 0 and the extreme inequality with 1 leads to function $1-R_F(p)$, which coincides with the Zenga curve (1.5) when the cdf $F(x)$ is continuous. The area

$$1 - \int_0^1 \frac{\mathbf{E}\left[X \mid X \le F^{-1}(p)\right]}{\mathbf{E}\left[X \mid X > F^{-1}(p)\right]} dp \tag{5.2}$$

beneath the function $1 - R_F(p)$ is always in the interval $[0,1]$. Quantity (5.2) is a measure of inequality and coincides with the earlier defined Zenga index Z_F when the cdf $F(x)$ is continuous, which we assume throughout the paper.

Note that under the continuity of $F(x)$, the lower and upper conditional expectations are equal to the absolute Bonferroni curve $p^{-1}AL_F(p)$ and the dual absolute Bonferroni curve $(1-p)^{-1}(\mu_F - AL_F(p))$, respectively, where

$$AL_F(p) = \int_0^p F^{-1}(t)dt \tag{5.3}$$

is the absolute Lorenz curve. This leads us to the expression of the Zenga index Z_F given by (1.4), which we now rewrite in terms of the absolute Lorenz curve as follows:

$$Z_F = 1 - \int_0^1 \left(\frac{1}{p} - 1\right) \frac{AL_F(p)}{\mu_F - AL_F(p)} dp. \tag{5.4}$$

We will extensively use expression (5.4) in the proofs below. In particular, we will see in the next section that the empirical Zenga index $\widetilde{Z}_n$ is equal to Z_F with the population cdf $F(x)$ replaced by the empirical cdf $F_n(x)$.

We are now in the position to provide additional details on the earlier noted Pareto case $1 < \theta < 2$, when the Pareto distribution has finite $\mathbf{E}[X]$ but infinite $\mathbf{E}[X^2]$. The above derived asymptotic results and thus the statistical inferential theory fail in this case. The required adjustments are serious and rely on the use of the extreme value theory, instead of the classical central limit theorem (CLT). Specifically, the task can be achieved by first expressing the absolute Lorenz curve $AL_F(p)$ in terms of the conditional tail expectation (CTE):

$$\mathrm{CTE}_F(p) = \frac{1}{1-p} \int_p^1 F^{-1}(t)dt \tag{5.5}$$

using the equation $AL_F(p) = \mu_F - (1-p)\mathrm{CTE}_F(p)$. Hence, (5.4) becomes

$$Z_F = 1 - \int_0^1 \frac{1}{p} \left(\frac{\mathrm{CTE}_F(0)}{\mathrm{CTE}_F(p)} - (1-p)\right) dp, \tag{5.6}$$

where $\mathrm{CTE}_F(0)$ is of course the mean μ_F. Note that replacing the population cdf $F(x)$ by its empirical counterpart $F_n(x)$ on the right-hand side of (5.6) would not lead to an estimator

that would work when $\mathbf{E}[X^2] = \infty$, and thus when the Pareto parameter $1 < \theta < 2$. A solution to this problem is provided by Necir et al. [22], who have suggested a new estimator of the conditional tail expectation $\mathrm{CTE}_F(p)$ for heavy-tailed distributions. Plugging in that estimator instead of the CTE on the right-hand side of (5.6) produces an estimator of the Zenga index when $\mathbf{E}[X^2] = \infty$. Establishing asymptotic results for the new "heavy-tailed" Zenga estimator would, however, be a complex technical task, well beyond the scope of the present paper, as can be seen from the proofs of Necir et al. [22].

6. A Closer Look at the Two Zenga Estimators

Since samples are "discrete populations", (5.2) and (5.4) lead to slightly different empirical estimators of Z_F. If we choose (5.2) and replace all population-related quantities by their empirical counterparts, then we will arrive at the estimator $\widehat{Z}_n$, as seen from the proof of the following theorem.

Theorem 6.1. *The empirical Zenga index $\widehat{Z}_n$ is an empirical estimator of Z_F.*

Proof. Let U be a uniform on $[0,1]$ random variable independent of X. The cdf of $F^{-1}(U)$ is F. Hence, we have the following equations:

$$\begin{aligned} Z_F &= 1 - \mathbf{E}_U\left(\frac{\mathbf{E}_X\left[X \mid X \le F^{-1}(U)\right]}{\mathbf{E}_X\left[X \mid X > F^{-1}(U)\right]}\right) \\ &= 1 - \int_{(0,\infty)} \frac{1-F(x)}{F(x)} \frac{\mathbf{E}[X\ \mathbf{1}\{X \le x\}]}{\mathbf{E}[X\ \mathbf{1}\{X > x\}]} dF(x) \\ &= 1 - \int_{(0,\infty)} \frac{1-F(x)}{F(x)} \frac{\int_{(0,x]} y\, dF(y)}{\int_{(x,\infty)} y\, dF(y)} dF(x). \end{aligned} \tag{6.1}$$

Replacing every F on the right-hand side of (6.1) by F_n, we obtain

$$1 - \frac{1}{n}\sum_{i=1}^{n-1} \frac{1-F_n(X_{i:n})}{F_n(X_{i:n})} \frac{\sum_{k=1}^{n} X_{k:n}\mathbf{1}\{X_{k:n} \le X_{i:n}\}}{\sum_{k=1}^{n} X_{k:n}\mathbf{1}\{X_{k:n} > X_{i:n}\}}, \tag{6.2}$$

which simplifies to

$$1 - \frac{1}{n}\sum_{i=1}^{n-1} \frac{1-i/n}{i/n} \frac{\sum_{k=1}^{i} X_{k:n}}{\sum_{k=i+1}^{n} X_{k:n}}. \tag{6.3}$$

This is the estimator $\widehat{Z}_n$ [12]. □

If, on the other hand, we choose (5.4) as the starting point for constructing an empirical estimator for Z_F, then we first replace the quantile $F^{-1}(p)$ by its empirical counterpart

$$\begin{aligned} F_n^{-1}(p) &= \inf\{x : F_n(x) \geq p\} \\ &= X_{i:n} \quad \text{when } p \in \left(\frac{(i-1)}{n}, \frac{i}{n}\right] \end{aligned} \tag{6.4}$$

in the definition of $AL_F(p)$, which leads to the empirical absolute Lorenz curve $AL_n(p)$, and then we replace each $AL_F(p)$ on the right-hand side of (5.4) by the just constructed $AL_n(p)$. (Note that $\mu_F = AL_F(1) \approx AL_n(1) = \overline{X}$.) These considerations produce the empirical Zenga index $\widetilde{Z}_n$, as seen from the proof of the following theorem.

Theorem 6.2. *The empirical Zenga index $\widetilde{Z}_n$ is an estimator of Z_F.*

Proof. By construction, the estimator $\widetilde{Z}_n$ is given by the equation:

$$\widetilde{Z}_n = 1 - \int_0^1 \left(\frac{1}{p} - 1\right) \frac{AL_n(p)}{\overline{X} - AL_n(p)} dp. \tag{6.5}$$

Hence, the proof of the lemma reduces to verifying that the right-hand sides of (2.2) and (6.5) coincide. For this, we split the integral in (6.5) into the sum of integrals over the intervals $((i-1),\ i/n)$ for $i = 1, \ldots, n$. For every $p \in ((i-1)/n, i/n)$, we have $AL_n(p) = C_{i,n} + pX_{i:n}$, where

$$C_{i,n} = \frac{1}{n} \sum_{k=1}^{i-1} X_{k:n} - \frac{i-1}{n} X_{i:n}. \tag{6.6}$$

Hence, (6.5) can be rewritten as $\widetilde{Z}_n = \sum_{i=1}^n \zeta_{i,n}$, where

$$\zeta_{i,n} = \frac{1}{n} - \int_{(i-1)/n}^{i/n} \left(\frac{1}{p} - 1\right) \frac{\Lambda_{i,n} + p}{\Psi_{i,n} - p} dp \tag{6.7}$$

with

$$\Lambda_{i,n} = \frac{C_{i,n}}{X_{i:n}}, \qquad \Psi_{i,n} = \frac{\overline{X} - C_{i,n}}{X_{i:n}}. \tag{6.8}$$

Consider first the case $i = 1$. We have $C_{1,n} = 0$ and thus $\Lambda_{1,n} = 0$, which implies

$$\zeta_{1,n} = \left(\frac{\overline{X}}{X_{1:n}} - 1\right) \log\left(1 + \frac{X_{1:n}}{\sum_{k=2}^n X_{k:n}}\right). \tag{6.9}$$

Next, we consider the case $i = n$. We have $C_{n,n} = \overline{X} - X_{n:n}$ and thus $\Psi_{n,n} = 1$, which implies

$$\zeta_{n,n} = \left(1 - \frac{\overline{X}}{X_{n:n}}\right) \log\left(\frac{n}{n-1}\right). \tag{6.10}$$

When $2 \le i \le n-1$, then the integrand in the definition of $\zeta_{i,n}$ does not have any singularity, since $\Psi_{i,n} > i/n$ due to $\sum_{k=i+1}^{n} X_{k:n} > 0$ almost surely. Hence, after simple integration we have that, for $i = 2, \ldots, n-1$,

$$\begin{aligned}\zeta_{i,n} = {} & \frac{(i-1)X_{i:n} - \sum_{k=1}^{i-1} X_{k:n}}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}} \log\left(\frac{i}{i-1}\right) \\ & + \left(\frac{\overline{X}}{X_{i:n}} - 1 + \frac{(i-1)X_{i:n} - \sum_{k=1}^{i-1} X_{k:n}}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}}\right) \log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right).\end{aligned} \tag{6.11}$$

With the above formulas for $\zeta_{i,n}$ we easily check that the sum $\sum_{i=1}^{n} \zeta_{i,n}$ is equal to the right-hand side of (2.2). This completes the proof of Theorem 6.2. □

7. A Closer Look at the Variances

Following the formulation of Theorem 2.1 we claimed that the asymptotic distribution of $\sqrt{n}\ (\widetilde{Z}_n - Z_F)$ is centered normal with the *finite* variance $\sigma_F^2 = \mathbf{E}[h^2(X)]$. The following theorem proves this claim.

Theorem 7.1. *When* $\mathbf{E}[X^{2+\alpha}] < \infty$ *for some* $\alpha > 0$, *then* $n^{-1/2} \sum_{i=1}^{n} h(X_i)$ *converges in distribution to the centered normal random variable*

$$\Gamma = \int_0^\infty \mathcal{B}(F(x)) w_F(F(x)) dx, \tag{7.1}$$

where $\mathcal{B}(p)$ *is the Brownian bridge on the interval* $[0,1]$. *The variance of* Γ *is finite and equal to* σ_F^2.

Proof. Note that $n^{-1/2} \sum_{i=1}^{n} h(X_i)$ can be written as $\int_0^\infty e_n(F(x)) w_F(F(x)) dx$, where $e_n(p) = \sqrt{n}(E_n(p) - p)$ is the empirical process based on the uniform on $[0,1]$ random variables $U_i = F(X_i)$, $i = 1, \ldots, n$. We will next show that

$$\int_0^\infty e_n(F(x)) w_F(F(x)) dx \longrightarrow_d \int_0^\infty \mathcal{B}(F(x)) w_F(F(x)) dx. \tag{7.2}$$

The proof is based on the well-known fact that, for every $\varepsilon > 0$, the following weak convergence of stochastic processes takes place:

$$\left\{\frac{e_n(p)}{p^{1/2-\varepsilon}(1-p)^{1/2-\varepsilon}},\ 0 \le p \le 1\right\} \Longrightarrow \left\{\frac{\mathcal{B}(p)}{p^{1/2-\varepsilon}(1-p)^{1/2-\varepsilon}},\ 0 \le p \le 1\right\}. \tag{7.3}$$

Hence, in order to prove statement (7.2), we only need to check that the integral

$$\int_0^\infty F(x)^{1/2-\varepsilon}(1-F(x))^{1/2-\varepsilon} w_F(F(x))dx \tag{7.4}$$

is finite. For this, by considering, for example, the two cases $p \le 1/2$ and $p > 1/2$ separately, we first easily verify the bound $|w_F(p)| \le c + c\log(1/p) + c\log(1/(1-p))$. Hence, for every $\varepsilon > 0$, there exists a constant $c < \infty$ such that, for all $p \in (0,1)$,

$$\left|w_F(p)\right| \le \frac{c}{p^\varepsilon (1-p)^\varepsilon}. \tag{7.5}$$

Bound (7.5) implies that integral (7.4) is finite when $\int_0^\infty (1-F(x))^{1/2-2\varepsilon}dx < \infty$, which is true since the moment $\mathbf{E}[X^{2+\alpha}]$ is finite for some $\alpha > 0$ and the parameter $\varepsilon > 0$ can be chosen as small as desired. Hence, $n^{-1/2}\sum_{i=1}^n h(X_i) \to_d \Gamma$ with Γ denoting the integral on the right-hand side of statement (7.2). The random variable Γ is normal because the Brownian bridge $\mathcal{B}(p)$ is a Gaussian process. Furthermore, Γ has mean zero because $\mathcal{B}(p)$ has mean zero for every $p \in [0,1]$. The variance of Γ is equal to σ_F^2 because $\mathbf{E}[\mathcal{B}(p)\mathcal{B}(q)] = \min\{p,q\} - pq$ for all $p, q \in [0,1]$. We are left to show that $\mathbf{E}[\Gamma^2] < \infty$. For this, we write the bound:

$$\begin{aligned}\mathbf{E}\left[\Gamma^2\right] &= \int_0^\infty \int_0^\infty \mathbf{E}\left[\mathcal{B}(F(x))\mathcal{B}(F(y))\right] w_F(F(x)) w_F(F(y))dx\,dy \\ &\le \left(\int_0^\infty \sqrt{\mathbf{E}[\mathcal{B}^2(F(x))]}\; w_F(F(x))dx\right)^2.\end{aligned} \tag{7.6}$$

Since $\mathbf{E}[\mathcal{B}^2(F(x))] = F(x)(1-F(x))$, the finiteness of the integral on the right-hand side of bound (7.6) follows from the earlier proved statement that integral (7.4) is finite. Hence, $\mathbf{E}[\Gamma^2] < \infty$ as claimed, which concludes the proof of Theorem 7.1. □

Theorem 7.2. *The empirical variance $S_{X,n}^2$ is an estimator of σ_F^2.*

Proof. We construct an empirical estimator for σ_F^2 by replacing every F on the right-hand side of (2.6) by the empirical F_n. Consequently, we replace the function $w_F(t)$ by its empirical version

$$w_{X,n}(t) = -\int_0^t \left(\frac{1}{p} - 1\right) \frac{AL_n(p)}{\left(\overline{X} - AL_n(p)\right)^2} dp + \int_t^1 \left(\frac{1}{p} - 1\right) \frac{1}{\overline{X} - AL_n(p)} dp. \tag{7.7}$$

We denote the resulting estimator of σ_F^2 by $S_{X,n}^2$. The rest of the proof consists of verifying that this estimator coincides with the one defined by (2.8). Note that $\min\{F_n(x), F_n(y)\} - F_n(x)F_n(y) = 0$ when $x \in [0, X_{1:n}) \cup [X_{n:n}, \infty)$ and/or $y \in [0, X_{1:n}) \cup [X_{n:n}, \infty)$. Hence, the just defined $S_{X,n}^2$ is equal to

$$\int_{X_{1:n}}^{X_{n:n}} \int_{X_{1:n}}^{X_{n:n}} \left(\min\{F_n(x), F_n(y)\} - F_n(x)F_n(y)\right) w_{X,n}(F_n(x)) w_{X,n}(F_n(y))dx\,dy. \tag{7.8}$$

Since $F_n(x) = k/n$ when $x \in [X_{k:n}, X_{k+1:n})$, we therefore have that

$$S^2_{X,n} = \sum_{k=1}^{n-1}\sum_{l=1}^{n-1}\left(\frac{\min\{k,l\}}{n} - \frac{k}{n}\frac{l}{n}\right) \times w_{X,n}\left(\frac{k}{n}\right)w_{X,n}\left(\frac{l}{n}\right)(X_{k+1:n} - X_{k:n})(X_{l+1:n} - X_{l:n}). \tag{7.9}$$

Furthermore,

$$\begin{aligned} w_{X,n}\left(\frac{k}{n}\right) &= -\int_0^{k/n}\left(\frac{1}{p} - 1\right)\frac{AL_n(p)}{\left(\overline{X} - AL_n(p)\right)^2}dp + \int_{k/n}^1\left(\frac{1}{p} - 1\right)\frac{1}{\overline{X} - AL_n(p)}dp \\ &= -\sum_{i=1}^{k} I_{X,n}(i) + \sum_{i=k+1}^{n} J_{X,n}(i), \end{aligned} \tag{7.10}$$

where, using notations (6.6) and (6.8), the summands on the right-hand side of (7.10) are

$$I_{X,n}(i) = \frac{1}{X_{i:n}}\int_{(i-1)/n}^{i/n}\left(\frac{1}{p} - 1\right)\frac{\Lambda_{i,n} + p}{\left(\Psi_{i,n} - p\right)^2}dp \tag{7.11}$$

for all $i = 1, \ldots, n-1$, and

$$J_{X,n}(i) = \frac{1}{X_{i:n}}\int_{(i-1)/n}^{i/n}\left(\frac{1}{p} - 1\right)\frac{1}{\Psi_{i,n} - p}dp \tag{7.12}$$

for all $i = 2, \ldots, n$. When $i = 1$, then $\Lambda_{i,n} = 0$. Hence, we immediately arrive at the expression for $I_{X,n}(1)$ given by (2.10). When $2 \le i \le n-1$, then

$$\begin{aligned} I_{X,n}(i) = {} & \frac{\Lambda_{i,n}}{X_{i:n}\Psi^2_{i,n}}\log\left(\frac{i}{i-1}\right) - \frac{(\Lambda_{i,n} + \Psi_{i,n})(\Psi_{i,n} - 1)}{nX_{i:n}\Psi_{i,n}(\Psi_{i,n} - (i-1)/n)(\Psi_{i,n} - i/n)} \\ & + \frac{1}{X_{i:n}}\left(1 + \frac{\Lambda_{i,n}}{\Psi^2_{i,n}}\right)\log\left(\frac{\Psi_{i,n} - (i-1)/n}{\Psi_{i,n} - i/n}\right), \end{aligned} \tag{7.13}$$

and, after some algebra, we arrive at the right-hand side of (2.11). When $2 \le i \le n-1$, then we have the expression

$$J_{X,n}(i) = \frac{1}{X_{i:n}\Psi_{i,n}}\log\left(\frac{i}{i-1}\right) - \frac{1}{X_{i:n}}\left(1 - \frac{1}{\Psi_{i,n}}\right)\log\left(\frac{\Psi_{i,n} - (i-1)/n}{\Psi_{i,n} - i/n}\right), \tag{7.14}$$

which, after some algebra, becomes the expression recorded in (2.12). When $i = n$, then $\Psi_{i,n} = 1$, and so we see that $J_{X,n}(n)$ is given by (2.13). This completes the proof of Theorem 7.2. □

Theorem 7.3. *The empirical mixed moment $S_{X,Y,n}$ is an estimator of* $\mathbf{E}[h(X)h(Y)]$.

Proof. We proceed similarly to the proof of Theorem 7.2. We estimate the integrand $\mathbf{P}[X \leq x, Y \leq y] - F(x)H(y)$ using

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x, Y_i \leq y\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x\}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{Y_i \leq y\}. \tag{7.15}$$

After some rearrangement of terms, estimator (7.15) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_{i:n} \leq x, Y_{(i,n)} \leq y\} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_{i:n} \leq x\}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{Y_{i:n} \leq y\}. \tag{7.16}$$

When $x \in [X_{k:n}, X_{k+1:n})$ and $y \in [Y_{l:n}, Y_{l+1:n})$, then estimator (7.16) is equal to $n^{-1}\sum_{i=1}^{k}\mathbf{1}\{Y_{(i,n)} \leq Y_{l:n}\} - (k/n)(l/n)$, which leads us to the estimator $S_{X,Y,n}$. This completes the proof of Theorem 7.3. □

8. Proof of Theorem 2.1

Throughout the proof we use the notation $AL_F^*(p)$ for the dual absolute Lorenz curve $\int_p^1 F^{-1}(t)dt$, which is equal to $\mu_F - AL_F(p)$. Likewise, we use the notation $AL_n^*(p)$ for the *empirical* dual absolute Lorenz curve.

Proof. Simple algebra gives the equations

$$\begin{aligned}
\sqrt{n}\left(\widetilde{Z}_n - Z_F\right) &= -\sqrt{n}\int_0^1\left(\frac{1}{p}-1\right)\left(\frac{AL_n(p)}{AL_n^*(p)} - \frac{AL_F(p)}{AL_F^*(p)}\right)dp \\
&= -\sqrt{n}\int_0^1\left(\frac{1}{p}-1\right)\frac{AL_n(p) - AL_F(p)}{AL_F^*(p)}dp \\
&\quad + \sqrt{n}\int_0^1\left(\frac{1}{p}-1\right)\frac{AL_F(p)}{AL_F^{*2}(p)}(AL_n^*(p) - AL_F^*(p))dp \\
&\quad + O_{\mathbf{P}}(r_{n,1}) + O_{\mathbf{P}}(r_{n,2})
\end{aligned} \tag{8.1}$$

with the remainder terms

$$\begin{aligned}
r_{n,1} &= \sqrt{n}\int_0^1\left(\frac{1}{p}-1\right)(AL_n(p) - AL_F(p))\left(\frac{1}{AL_n^*(p)} - \frac{1}{AL_F^*(p)}\right)dp, \\
r_{n,2} &= \sqrt{n}\int_0^1\left(\frac{1}{p}-1\right)\frac{AL_F(p)}{AL_F^*(p)}(AL_n^*(p) - AL_F^*(p))\left(\frac{1}{AL_n^*(p)} - \frac{1}{AL_F^*(p)}\right)dp.
\end{aligned} \tag{8.2}$$

We will later show (Lemmas 9.1 and 9.2) that the remainder terms $r_{n,1}$ and $r_{n,2}$ are of the order $o_{\mathbf{P}}(1)$. Hence, we now proceed with our analysis of the first two terms on the right-hand side of (8.1), for which we use the (general) Vervaat process

$$V_n(p) = \int_0^p \left(F_n^{-1}(t) - F^{-1}(t)\right)dt + \int_0^{F^{-1}(p)} (F_n(x) - F(x))dx \tag{8.3}$$

and its dual version

$$V_n^*(p) = \int_p^1 \left(F_n^{-1}(t) - F^{-1}(t)\right)dt + \int_{F^{-1}(p)}^{\infty} (F_n(x) - F(x))dx. \tag{8.4}$$

For mathematical and historical details on the Vervaat process, see Zitikis [23], Davydov and Zitikis [24], Greselin et al. [25], and references therein. Since $\int_0^1 (F_n^{-1}(t) - F^{-1}(t))dt = \overline{X} - \mu_F$ and $\int_0^{\infty} (F_n(x) - F(x))dx = -(\overline{X} - \mu_F)$, adding the right-hand sides of (8.3) and (8.4) gives the equation $V_n^*(p) = -V_n(p)$. Hence, whatever upper bound we have for $|V_n(p)|$, the same bound holds for $|V_n^*(p)|$. In fact, the absolute value can be dropped from $|V_n(p)|$ since $V_n(p)$ is always nonnegative. Furthermore, we know that $V_n(p)$ does not exceed $(p - F_n(F^{-1}(p)))(F_n^{-1}(p) - F^{-1}(p))$. Hence, with the notation $e_n(p) = \sqrt{n}(F_n(F^{-1}(p)) - p)$, which is the uniform on $[0,1]$ empirical process, we have that

$$\sqrt{n}\, V_n(p) \le \left|e_n(p)\right|\left|F_n^{-1}(p) - F^{-1}(p)\right|. \tag{8.5}$$

Bound (8.5) implies the following asymptotic representation for the first term on the right-hand side of (8.1):

$$\begin{aligned} &-\sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right)\frac{AL_n(p) - AL_F(p)}{AL_F^*(p)}dp \\ &\quad = \sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right)\frac{1}{AL_F^*(p)}\left(\int_0^{F^{-1}(p)} (F_n(x) - F(x))dx\right)dp + O_{\mathbf{P}}(r_{n,3}), \end{aligned} \tag{8.6}$$

where

$$r_{n,3} = \int_0^1 \left(\frac{1}{p} - 1\right)\frac{1}{AL_F^*(p)}\left|e_n(p)\right|\left|F_n^{-1}(p) - F^{-1}(p)\right|dp. \tag{8.7}$$

We will later show (Lemma 9.3) that $r_{n,3} = o_{\mathbf{P}}(1)$. Furthermore, we have the following asymptotic representation for the second term on the right-hand side of (8.1):

$$\begin{aligned} &\sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right)\frac{AL_F(p)}{AL_F^{*2}(p)}(AL_n^*(p) - AL_F^*(p))dp \\ &\quad = -\sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right)\frac{AL_F(p)}{AL_F^{*2}(p)}\left(\int_{F^{-1}(p)}^{\infty} (F_n(x) - F(x))dx\right)dp + O_{\mathbf{P}}(r_{n,4}), \end{aligned} \tag{8.8}$$

where

$$r_{n,4} = \int_0^1 \left(\frac{1}{p} - 1\right) \frac{AL_F(p)}{AL_F^{*2}(p)} \left|e_n(p)\right| \left|F_n^{-1}(p) - F^{-1}(p)\right| dp. \tag{8.9}$$

We will later show (Lemma 9.4) that $r_{n,4} = o_{\mathbf{P}}(1)$. Hence, (8.1), (8.6) and (8.8) together with the aforementioned statements that $r_{n,1}, \ldots, r_{n,4}$ are of the order $o_{\mathbf{P}}(1)$ imply that

$$\begin{aligned}
\sqrt{n}\left(\widetilde{Z}_n - Z_F\right) &= \sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right) \frac{1}{AL_F^*(p)} \left(\int_0^{F^{-1}(p)} (F_n(x) - F(x))dx\right) dp \\
&\quad - \sqrt{n}\int_0^1 \left(\frac{1}{p} - 1\right) \frac{AL_F(p)}{AL_F^{*2}(p)} \left(\int_{F^{-1}(p)}^{\infty} (F_n(x) - F(x))dx\right) dp + o_{\mathbf{P}}(1) \qquad (8.10) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) + o_{\mathbf{P}}(1).
\end{aligned}$$

This completes the proof of Theorem 2.1. □

9. Negligibility of Remainder Terms

The following four lemmas establish the above noted statements that the remainder terms $r_{n,1}, \ldots, r_{n,4}$ are of the order $o_{\mathbf{P}}(1)$. In the proofs of the lemmas we will use a parameter $\delta \in (0, 1/2]$, possibly different from line to line but never depending on n. Furthermore, we will frequently use the fact that

$$\mathbf{E}[X^q] < \infty \quad \text{implies} \quad \int_0^1 \left|F_n^{-1}(t) - F^{-1}(t)\right|^q dt = o_{\mathbf{P}}(1). \tag{9.1}$$

Another technical result that we will frequently use is the fact that, for any $\varepsilon > 0$ as small as desired,

$$\sup_{x \in \mathbf{R}} \frac{\sqrt{n}|F_n(x) - F(x)|}{F(x)^{1/2-\varepsilon}(1 - F(x))^{1/2-\varepsilon}} = O_{\mathbf{P}}(1) \tag{9.2}$$

when $n \to \infty$.

Lemma 9.1. *Under the conditions of Theorem 2.1, $r_{n,1} = o_{\mathbf{P}}(1)$.*

Proof. We split the remainder term $r_{n,1} = \sqrt{n}\int_0^1 \cdots dp$ into the sum of $r_{n,1}^*(\delta) = \sqrt{n}\int_0^{1-\delta} \cdots dp$ and $r_{n,1}^{**}(\delta) = \sqrt{n}\int_{1-\delta}^1 \cdots dp$. The lemma follows if

(1) for every $\delta > 0$, the statement $r_{n,1}^*(\delta) = o_{\mathbf{P}}(1)$ holds when $n \to \infty$,

(2) $r_{n,1}^{**}(\delta) = h(\delta)O_{\mathbf{P}}(1)$ for a deterministic $h(\delta) \downarrow 0$ when $\delta \downarrow 0$, where $O_{\mathbf{P}}(1)$ does not depend on δ.

To prove part (1), we first note that when $0 < p < 1 - \delta$, then $AL_F^*(p) \geq \int_{1-\delta}^1 F^{-1}(t)dt$, which is positive, and $AL_n^*(p) \geq \int_{1-\delta}^1 F^{-1}(t)dt + o_{\mathbf{P}}(1)$ due to statement (9.1) with $q = 1$. Hence, we are left to show that, when $n \to \infty$,

$$\sqrt{n}\int_0^{1-\delta} \frac{1}{p}\,\left|AL_n(p) - AL_F(p)\right|\left|AL_n^*(p) - AL_F^*(p)\right|dp = o_{\mathbf{P}}(1). \tag{9.3}$$

Since $AL_n^*(p) - AL_F^*(p) = (\overline{X} - \mu_F) - (AL_n(p) - AL_F(p))$, statement (9.3) follows if

$$\sqrt{n}\left|\overline{X} - \mu_F\right| \int_0^{1-\delta} \frac{1}{p}\left|AL_n(p) - AL_F(p)\right|dp = o_{\mathbf{P}}(1), \tag{9.4}$$

$$\sqrt{n}\int_0^{1-\delta} \frac{1}{p}\left|AL_n(p) - AL_F(p)\right|^2 dp = o_{\mathbf{P}}(1). \tag{9.5}$$

We have $\sqrt{n}\,|\overline{X} - \mu_F| = O_{\mathbf{P}}(1)$ and $|AL_n(p) - AL_F(p)| \leq \sqrt{p}\,(\int_0^1 |F_n^{-1}(p) - F^{-1}(p)|^2 dp)^{1/2}$. Since $\int_0^1 |F_n^{-1}(p) - F^{-1}(p)|^2 dp = o_{\mathbf{P}}(1)$ and $\int_0^{1-\delta} p^{-1}\sqrt{p}\,dp < \infty$, we have statement (9.4). To prove statement (9.5), we use bound (8.5) and reduce the proof to showing that

$$\frac{1}{\sqrt{n}}\int_0^{1-\delta} \frac{1}{p}\left|\int_0^{F^{-1}(p)} \sqrt{n}\,(F_n(x) - F(x))dx\right|^2 dp = o_{\mathbf{P}}(1), \tag{9.6}$$

$$\frac{1}{\sqrt{n}}\int_0^{1-\delta} \frac{1}{p}\,\left|e_n(p)\right|^2\left|F_n^{-1}(p) - F^{-1}(p)\right|^2 dp = o_{\mathbf{P}}(1). \tag{9.7}$$

To prove statement (9.6), we use statement (9.2) and observe that

$$\int_0^{1-\delta} \frac{1}{p}\left(\int_0^{F^{-1}(p)} F(x)^{1/2-\varepsilon}dx\right)^2 dp \leq c(F,\delta)\int_0^{1-\delta} \frac{1}{p}\,p^{1-2\varepsilon}dp < \infty. \tag{9.8}$$

To prove statement (9.7), we use the uniform on $[0,1]$ version of statement (9.2) and Hölder's inequality, and in this way reduce the proof to showing that

$$\frac{1}{\sqrt{n}}\left(\int_0^{1-\delta} \frac{1}{p^{2\varepsilon a}}dp\right)^{1/a}\left(\int_0^{1-\delta} \left|F_n^{-1}(p) - F^{-1}(p)\right|^{2b} dp\right)^{1/b} = o_{\mathbf{P}}(1) \tag{9.9}$$

for some $a, b > 1$ such that $a^{-1} + b^{-1} = 1$. We choose the parameters a and b as follows. First, since $\mathbf{E}[X^{2+\alpha}] < \infty$, we set $b = (2+\alpha)/2$. Next, we choose $\varepsilon > 0$ on the left-hand side of statement (9.9) so that $2\varepsilon a < 1$, which holds when $\varepsilon < \alpha/(4+2\alpha)$ in view of the equation $a^{-1} + b^{-1} = 1$. Hence, statement (9.9) holds and thus statement (9.7) follows. This completes the proof of part (1).

To establish part (2), we first estimate $|r_{n,1}^{**}(\delta)|$ from above using the bounds $AL_F^*(p) \geq (1-p)F^{-1}(1/2)$ and $AL_n^*(p) \geq (1-p)F_n^{-1}(1/2)$, which hold since $\delta \leq 1/2$. Hence, we have

reduced our task to verifying the statement $\sqrt{n}\int_{1-\delta}^{1}|AL_n(p)-AL_F(p)|dp = h(\delta)O_{\mathbf{P}}(1)$. Using the Vervaat process $V_n(p)$ and bound (8.5), we reduce the proof of the statement to showing that the integrals

$$\int_{1-\delta}^{1}\left(\int_{0}^{F^{-1}(p)}\sqrt{n}\,|F_n(x)-F(x)|dx\right)dp, \tag{9.10}$$

$$\int_{1-\delta}^{1}\left|e_n(p)\right|\left|F_n^{-1}(p)-F^{-1}(p)\right|dp \tag{9.11}$$

are of the order $h(\delta)O_{\mathbf{P}}(1)$ with possibly different $h(\delta)\downarrow 0$ in each case. In view of statement (9.2), we have the desired statement for integral (9.10) if the quantity

$$\int_{1-\delta}^{1}\left(\int_{0}^{F^{-1}(p)}(1-F(x))^{1/2-\varepsilon}dx\right)dp \tag{9.12}$$

converges to 0 when $\delta\downarrow 0$, in which case we use it as $h(\delta)$. The inner integral of (9.12) does not exceed $\int_0^{\infty}(1-F(x))^{1/2-\varepsilon}dx$, which is finite for all sufficiently small $\varepsilon>0$ since $\mathbf{E}[X^{2+\alpha}]<\infty$ for some $\alpha>0$. This completes the proof that quantity (9.10) is of the order $h(\delta)O_{\mathbf{P}}(1)$. To show that quantity (9.11) is of a similar order, we use the uniform on $[0,1]$ version of statement (9.2) and reduce the task to showing that $\int_{1-\delta}^{1}|F_n^{-1}(p)-F^{-1}(p)|dp$ is of the order $h(\delta)O_{\mathbf{P}}(1)$. By the Cauchy-Bunyakowski-Schwarz inequality, we have that

$$\int_{1-\delta}^{1}\left|F_n^{-1}(p)-F^{-1}(p)\right|dp \le \sqrt{\delta}\left(\int_{0}^{1}\left|F_n^{-1}(p)-F^{-1}(p)\right|^2dp\right)^{1/2}. \tag{9.13}$$

Since $\mathbf{E}[X^2]<\infty$, we have $\int_0^1|F_n^{-1}(p)-F^{-1}(p)|^2dp = o_{\mathbf{P}}(1)$, and so setting $h(\delta)=\sqrt{\delta}$ establishes the desired asymptotic result for integral (9.11). This also completes the proof of part (2), and also of Lemma 9.1. □

Lemma 9.2. *Under the conditions of Theorem 2.1,* $r_{n,2} = o_{\mathbf{P}}(1)$.

Proof. Like in the proof of Lemma 9.1, we split the remainder term $r_{n,2}=\sqrt{n}\int_0^1\cdots dp$ into the sum of $r_{n,2}^{*}(\delta)=\sqrt{n}\int_0^{1-\delta}\cdots dp$ and $r_{n,2}^{**}(\delta)=\sqrt{n}\int_{1-\delta}^{1}\cdots dp$. To prove the lemma, we need to show the following.

(1) For every $\delta>0$, the statement $r_{n,2}^{*}(\delta)=o_{\mathbf{P}}(1)$ holds when $n\to\infty$.

(2) $r_{n,2}^{**}(\delta)=h(\delta)O_{\mathbf{P}}(1)$ for a deterministic $h(\delta)\downarrow 0$ when $\delta\downarrow 0$, where $O_{\mathbf{P}}(1)$ does not depend on δ.

To prove part (1), we first estimate $|r_{n,2}^{*}(\delta)|$ from above using the bounds $p^{-1}AL_F(p)\le F^{-1}(1-\delta)<\infty$, $AL_F^{*}(p)\ge\int_{1-\delta}^{1}F^{-1}(t)dt>0$, and $AL_n^{*}(p)\ge\int_{1-\delta}^{1}F^{-1}(t)dt+o_{\mathbf{P}}(1)$. This reduces our task to showing that, for every $\delta>0$,

$$\sqrt{n}\int_{0}^{1-\delta}\left|AL_n^{*}(p)-AL_F^{*}(p)\right|^2dp = o_{\mathbf{P}}(1). \tag{9.14}$$

Since $AL_n^*(p) - AL_F^*(p) = (\overline{X} - \mu_F) - (AL_n(p) - AL_F(p))$ and $\sqrt{n}\ (\overline{X} - \mu_F)^2 = o_{\mathbf{P}}(1)$, statement (9.14) follows from

$$\sqrt{n} \int_0^{1-\delta} \left| AL_n(p) - AL_F(p) \right|^2 dp = o_{\mathbf{P}}(1), \tag{9.15}$$

which is an elementary consequence of statement (9.5). This establishes part (1).

To prove part (2), we first estimate $|r_{n,2}^{**}(\delta)|$ from above using the bounds $AL_F^*(p) \geq (1-p)F^{-1}(1/2)$ and $AL_n^*(p) \geq (1-p)F_n^{-1}(1/2)$, and in this way reduce the task to showing that

$$\sqrt{n} \int_{1-\delta}^{1} \frac{1}{1-p} \left| AL_n^*(p) - AL_F^*(p) \right| dp = h(\delta) O_{\mathbf{P}}(1). \tag{9.16}$$

Using the Vervaat process, statement (9.16) follows if

$$\int_{1-\delta}^{1} \frac{1}{1-p} \left(\int_{F^{-1}(p)}^{\infty} \sqrt{n}\ |F_n(x) - F(x)| dx \right) dp = h(\delta) O_{\mathbf{P}}(1), \tag{9.17}$$

$$\int_{1-\delta}^{1} \frac{1}{1-p}\ |e_n(p)| \left| F_n^{-1}(p) - F^{-1}(p) \right| dp = h(\delta) O_{\mathbf{P}}(1) \tag{9.18}$$

with possibly different $h(\delta) \downarrow 0$ in each case. Using statement (9.2), we have that statement (9.17) holds with $h(\delta)$ defined as the integral

$$\int_{1-\delta}^{1} \frac{1}{1-p} \left(\int_{F^{-1}(p)}^{\infty} (1 - F(x))^{1/2-\varepsilon} dx \right) dp, \tag{9.19}$$

which converges to 0 when $\delta \downarrow 0$ as the following argument shows. First, we write the integrand as the product of $(1-F(x))^{\varepsilon}$ and $(1-F(x))^{1/2-2\varepsilon}$. Then we estimate the first factor by $(1-p)^{\varepsilon}$. The integral $\int_0^{\infty} (1-F(x))^{1/2-2\varepsilon} dx$ is finite for all sufficiently small $\varepsilon > 0$ since $\mathbf{E}[X^{2+\alpha}] < \infty$ for some $\alpha > 0$. Since $\int_{1-\delta}^{1} (1-p)^{-1+\varepsilon} dp \downarrow 0$ when $\delta \downarrow 0$, integral (9.19) converges to 0 when $\delta \downarrow 0$. The proof of statement (9.17) is finished.

We are left to prove statement (9.18). Using the uniform on $[0,1]$ version of statement (9.2), we reduce the task to showing that

$$\int_{1-\delta}^{1} \frac{1}{(1-p)^{1/2+\varepsilon}} \left| F_n^{-1}(p) - F^{-1}(p) \right| dp = h(\delta) O_{\mathbf{P}}(1). \tag{9.20}$$

In fact, we will see below that $O_{\mathbf{P}}(1)$ can be replaced by $o_{\mathbf{P}}(1)$. Using Hölder's inequality, we have that the right-hand side of (9.20) does not exceed

$$\left(\int_{1-\delta}^{1} \frac{1}{(1-p)^{(1/2+\varepsilon)a}} dp \right)^{1/a} \left(\int_{1-\delta}^{1} \left| F_n^{-1}(p) - F^{-1}(p) \right|^b dp \right)^{1/b} \tag{9.21}$$

for some $a, b > 1$ such that $a^{-1} + b^{-1} = 1$. We choose the parameters a and b as follows. Since $\mathbf{E}[X^{2+\alpha}] < \infty$, we set $b = 2 + \alpha$, and so the right-most integral of (9.21) is of the order $o_{\mathbf{P}}(1)$. Furthermore, $a = (2 + \alpha)/(1 + \alpha) < 2$, which can be made arbitrarily close to 2 by choosing sufficiently small $\alpha > 0$. Choosing $\varepsilon > 0$ so small that $(1/2+\varepsilon)a < 1$, we have that the left-most integral in (9.21) converges to 0 when $\delta \downarrow 0$. This establishes statement (9.18) and completes the proof of Lemma 9.2. □

Lemma 9.3. *Under the conditions of Theorem 2.1,* $r_{n,3} = o_{\mathbf{P}}(1)$.

Proof. We split the remainder term $r_{n,3} = \int_0^1 \cdots dp$ into the sum of $r^*_{n,3} = \int_0^{1/2} \cdots dp$ and $r^{**}_{n,3} = \int_{1/2}^1 \cdots dp$. The lemma follows if the two summands are of the order $o_{\mathbf{P}}(1)$.

To prove $r^*_{n,3} = o_{\mathbf{P}}(1)$, we use the bound $AL^*_F(p) \geq \int_{1/2}^1 F^{-1}(p)dp$ and the uniform on $[0,1]$ version of statement (9.2), and in this way reduce our task to showing that

$$\int_0^{1/2} \frac{1}{p^{1/2+\varepsilon}} \left| F_n^{-1}(p) - F^{-1}(p) \right| dp = o_{\mathbf{P}}(1). \tag{9.22}$$

This statement can be established following the proof of statement (9.20), with minor modifications.

To prove $r^{**}_{n,3} = o_{\mathbf{P}}(1)$, we use the bound $AL^*_F(p) \geq (1-p)F^{-1}(1/2)$, the fact that $\sup_t |e_n(t)| = O_{\mathbf{P}}(1)$, and statement (9.1) with $q = 1$. The desired result for $r^{**}_{n,3}$ follows, which finishes the proof of Lemma 9.3. □

Lemma 9.4. *Under the conditions of Theorem 2.1,* $r_{n,4} = o_{\mathbf{P}}(1)$.

Proof. We split $r_{n,4} = \int_0^1 \cdots dp$ into the sum of $r^*_{n,4} = \int_0^{1/2} \cdots dp$ and $r^{**}_{n,4} = \int_{1/2}^1 \cdots dp$, and then show that the two summands are of the order $o_{\mathbf{P}}(1)$.

To prove $r^*_{n,4} = o_{\mathbf{P}}(1)$, we use the bounds $p^{-1}AL_F(p) \leq F^{-1}(1/2) < \infty$ and $AL^*_F(p) \geq \int_{1/2}^1 F^{-1}(p)dp > 0$ together with the uniform on $[0,1]$ version of statement (9.2). This reduces our task to showing that $\int_0^{1/2} |F_n^{-1}(p) - F^{-1}(p)|dp = o_{\mathbf{P}}(1)$, which holds due to statement (9.1) with $q = 1$.

To prove $r^{**}_{n,4} = o_{\mathbf{P}}(1)$, we use the bound $AL^*_F(p) \geq (1-p)F^{-1}(1/2)$ and the uniform on $[0,1]$ version of statement (9.2), and in this way reduce the proof to showing that

$$\int_{1/2}^1 \frac{1}{(1-p)^{1/2+\varepsilon}} \left| F_n^{-1}(p) - F^{-1}(p) \right| dp = o_{\mathbf{P}}(1). \tag{9.23}$$

This statement can be established following the proof of statement (9.20). The proof of Lemma 9.4 is finished. □

Acknowledgments

The authors are indebted to two anonymous referees and the editor in charge of the manuscript, Madan L. Puri, for their constructive criticism and suggestions that helped them to improve the paper. The research has been partially supported by the 2009 F.A.R. (Fondo

di Ateneo per la Ricerca) at the University of Milan Bicocca, and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] M. Zenga, "Inequality curve and inequality index based on the ratios between lower and upper arithmetic means," *Statistica & Applicazioni*, vol. 5, pp. 3–27, 2007.
[2] G. Pietra, "Delle relazioni fra indici di variabilitá, note I e II," *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, vol. 74, pp. 775–804, 1915.
[3] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Journal of the American Statistical Association*, vol. 9, pp. 209–219, 1905.
[4] C. Gini, "Sulla misura della concentrazione e della variabilità dei caratteri," in *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. Anno Accademico*, vol. 48, part 2, pp. 1201–1248, Premiate Officine Grafiche Carlo Ferrari, Venezia, Italy, 1914.
[5] V. Pareto, "La legge della domanda," *Giornale degli Economisti*, vol. 10, pp. 59–68, 1895.
[6] V. Pareto, "Ecrits sur la courve de la répartition de la richesse," in *Complete Works of V. Pareto*, G. Busino, Ed., Librairie Droz, Genève, Switzerland, 1965.
[7] M. Zenga, "Il contributo degli italiani allo studio della concentrazione," in *La Distribuzione Personale del Reddito: Problemi di Formazione, di Ripartizione e di Misurazione*, M. Zenga, Ed., Vita e Pensiero, Milano, Italy, 1987.
[8] C. Dagum, "A new model of personal distribution: specification and estimation," *Economie Appliquée*, vol. 30, pp. 413–437, 1977.
[9] C. Dagum, "The generation and distribution of income. The Lorenz curve and the Gini ratio," *Economie Appliquée*, vol. 33, pp. 327–367, 1980.
[10] C. Dagum, "A model of net wealth distribution specified for negative, null and positive wealth. A case of study: Italy," in *Income and Wealth Distribution, Inequality and Poverty*, C. Dagum and M. Zenga, Eds., pp. 42–56, Springer, Berlin, Germany, 1990.
[11] J. L. Gastwirth, "A general definition of the Lorenz curve," *Econometrica*, vol. 39, pp. 1037–1039, 1971.
[12] F. Greselin and L. Pasquazzi, "Asymptotic confidence intervals for a new inequality measure," *Communications in Statistics: Simulation and Computation*, vol. 38, no. 8, pp. 1742–1756, 2009.
[13] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley Series in Probability and Statistics, Wiley-Interscience, Hoboken, NJ, USA, 2003.
[14] Z. A. Karian and E. J. Dudewicz, *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Method*, CRC Press, Boca Raton, Fla, USA, 2000.
[15] B. L. Jones, M. L. Puri, and R. Zitikis, "Testing hypotheses about the equality of several risk measure values with applications in insurance," *Insurance: Mathematics & Economics*, vol. 38, no. 2, pp. 253–270, 2006.
[16] V. Brazauskas, B. L. Jones, M. L. Puri, and R. Zitikis, "Nested *L*-statistics and their use in comparing the riskiness of portfolios," *Scandinavian Actuarial Journal*, no. 3, pp. 162–179, 2007.
[17] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge University Press, Cambridge, UK, 1997.
[18] B. Efron, "Better bootstrap confidence intervals," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–200, 1987.
[19] Bank of Italy, "Household income and wealth in 2004," *Supplements to the Statistical Bulletin, Sample Surveys*, vol. 16, no. 7, 2006.
[20] C. N. Chen, T. W. Tsaur, and T. S. Rhai, "The Gini coefficient and negative income," *Oxford Economic Papers*, vol. 34, pp. 473–478, 1982.
[21] E. Maasoumi, "Empirical analysis of welfare and inequality," in *Handbook of Applied Econometrics, Volume II: Microeconomics*, M. H. Pesaran and P. Schmidt, Eds., Blackwell, Oxford, UK, 1994.
[22] A. Necir, A. Rassoul, and R. Zitikis, "Estimating the conditional tail expectation in the case of heavy-tailed losses," *Journal of Probability and Statistics*. In press.
[23] R. Zitikis, "The Vervaat process," in *Asymptotic Methods in Probability and Statistics (Ottawa, ON, 1997)*, B. Szyszkowicz, Ed., pp. 667–694, North-Holland, Amsterdam, The Netherlands, 1998.
[24] Y. Davydov and R. Zitikis, "Convex rearrangements of random elements," in *Asymptotic Methods in Stochastics*, vol. 44 of *Fields Institute Communications*, pp. 141–171, American Mathematical Society, Providence, RI, USA, 2004.
[25] F. Greselin, M. L. Puri, and R. Zitikis, "*L*-functions, processes, and statistics in measuring economic inequality and actuarial risks," *Statistics and Its Interface*, vol. 2, no. 2, pp. 227–245, 2009.