ATTI DELLA XLV RIUNIONE SCIENTIFICA (2010)

XLV Riunione Scientifica
(Padova, 16-18 giugno 2010)

ISBN 978-88-6129-566-7
Editore: CLEUP – Padova

Francesca Greselin, Leo Pasquazzi, Ričardas Zitikis

Asymptotic Theory for Zenga's New Index of Economic Inequality

http://old.sis-statistica.org/wp-content/old_uploads/2013/09/RS10-Asymptotic-Theory-for-Zengas-New-Index-of-Economic-Inequality.pdf

Asymptotic Theory for Zenga's New Index of Economic Inequality

Francesca Greselin, Leo Pasquazzi and Ričardas Zitikis

Abstract 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 state 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).

Key words: Zenga index, confidence interval, Lorenz curve, Vervaat process, measuring poverty and inequality.

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 has been arguably the

Francesca Greselin, Leo Pasquazzi
Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Universitá di Milano Bicocca, Milan, Italy, e-mail: francesca.greselin@unimib.it,leo.pasquazzi@unimib.it

Ričardas Zitikis
Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada, e-mail: zitikis@stats.uwo.ca

most popular measure of inequality, with a number of extensions and generalizations available in the literature. Recently, Zenga (2007) suggested a new index for measuring inequality. We shall next recall the Gini and Zenga indices.

Let $F(x)$ denote the cumulative distribution function (cdf) of X, which we assume to be non-negative throughout the paper. We assume that the cdf $F(x)$ is continuous, which is a natural choice when modeling economic size distributions. Let $F^{-1}(p)$ denote the corresponding quantile function. The Lorenz curve is

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

where $\mu_F = \mathbf{E}[X]$ is the true (unknown) mean of X.

The classical Gini index G_F can be written as follows:

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

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

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

is a relative measure of inequality and called the Gini curve. 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 (2007) index Z_F of inequality is defined by the formula

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

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

$$Z_F(p) = 1 - \frac{L_F(p)}{p} \cdot \frac{1-p}{1-L_F(p)}. \tag{2}$$

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 sub-populations.

The Gini and Zenga indices are averages of point inequality measures, that is, of the Gini and Zenga curves, respectively. However, while in the case of the Gini index the weight function $\psi(p) = 2p$ is used, in the case of the Zenga index the uniform weight function 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, with the same sensibility, the index detects 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

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\theta}, \quad x \geq x_0, \tag{3}$$

where $x_0 > 0$ and $\theta > 0$ are parameters. Corresponding to distribution (3), the Lorenz curve is $L_F(p) = 1 - (1-p)^{1-1/\theta}$, and the Gini curve is $G_F(p) = ((1-p)^{1-1/\theta} - (1-p))/p$. In Figure 1 (left-hand panel) 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 (right-hand panel), alongside the Gini curve $G_F(p)$ for an easy comparison. The left-hand panel allows us to appre-

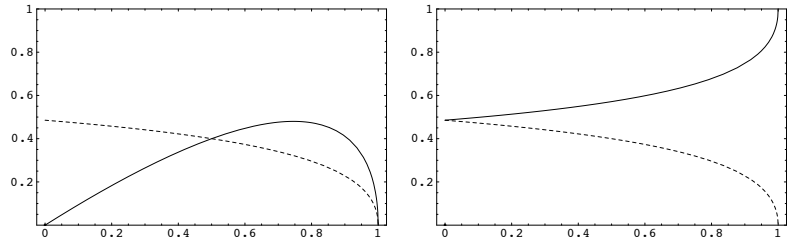

Fig. 1 The Gini curve $G_F(p)$ (dashed; both panels), the weighted Gini curve $g_F(p)$ (solid; left panel), and the Zenga curve $Z_F(p)$ (solid; right panel) in the Pareto case with $x_0 = 1$ and $\theta = 2.06$.

ciate how the 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 identical subgroups. The outcome is that the Gini index underestimates inequality. In the right-hand panel 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 $Z_F(0.8) = 0.678$ and thus 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 and state statistical inferential results. In Section 3 we present results of a simulation study and an analysis of the Bank of Italy SHIW 2006 data.

2 Estimators and statistical inference

Let $X_1,\dots,X_n$ be independent copies of X. We use two non-parametric estimators of the Zenga index. The first one (Greselin and Pasquazzi, 2009) is

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

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

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

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

Theorem 1. *If the moment* $\mathbf{E}[X^{2+\alpha}]$ *is finite for some* $\alpha > 0$*, then we have the asymptotic representation*

$$\sqrt{n}\,(\widetilde{Z}_n - Z_F) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(X_i) + o_{\mathbf{P}}(1), \tag{3}$$

where

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

with the weight function

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

In view of Theorem 1, the asymptotic distribution of $\sqrt{n}\,(\widetilde{Z}_n - Z_F)$ is centered normal with variance $\sigma_F^2 = \mathbf{E}[h^2(X)]$, which is finite (see Theorem 7.1 in Greselin *et al.*, 2009a) and can be rewritten as follows:

$$\sigma_F^2 = \int_0^\infty \int_0^\infty \big(\min\{F(x),F(y)\} - F(x)F(y)\big) w_F(F(x)) w_F(F(y))\,dx\,dy. \tag{4}$$

Replacing the cdf $F(x)$ on the right-hand side of equation (4) 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 the following variance estimator:

$$S^2_{X,n} = \sum_{k=1}^{n-1}\sum_{l=1}^{n-1}\left(\frac{\min\{k,l\}}{n} - \frac{k}{n}\frac{l}{n}\right) \times w_{X,n}\left(\frac{k}{n}\right) w_{X,n}\left(\frac{l}{n}\right)(X_{k+1:n} - X_{k:n})(X_{l+1:n} - X_{l:n}), \quad (5)$$

where

$$w_{X,n}(k/n) = -\sum_{i=1}^{k} I_{X,n}(i) + \sum_{i=k+1}^{n} J_{X,n}(i)$$

with the following expressions for the summands $I_{X,n}(i)$ and $J_{X,n}(i)$. First, we have

$$I_{X,n}(1) = -\frac{\sum_{k=2}^{n} X_{k:n} - (n-1)X_{1:n}}{(\sum_{k=1}^{n} X_{k:n})(\sum_{k=2}^{n} X_{k:n})} + \frac{1}{X_{1,n}}\log\left(1 + \frac{X_{1:n}}{\sum_{k=2}^{n} X_{k:n}}\right). \quad (6)$$

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

$$\begin{aligned} I_{X,n}(i) =& n\frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)^2}\log\left(\frac{i}{i-1}\right) \\ &- \frac{(\sum_{k=i+1}^{n} X_{k:n} - (n-i)X_{i:n})(\sum_{k=1}^{n} X_{k:n})}{(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n})(\sum_{k=i+1}^{n} X_{k:n})(\sum_{k=i}^{n} X_{k:n})} \\ &+ \left(\frac{1}{X_{i:n}} + n\frac{\sum_{k=1}^{i-1} X_{k:n} - (i-1)X_{i:n}}{\left(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}\right)^2}\right)\log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right) \end{aligned} \quad (7)$$

and

$$\begin{aligned} J_{X,n}(i) =& \frac{n}{\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n}}\log\left(\frac{i}{i-1}\right) \\ &- \frac{\sum_{k=i+1}^{n} X_{k:n} - (n-i)X_{i:n}}{X_{i:n}(\sum_{k=i+1}^{n} X_{k:n} + iX_{i:n})}\log\left(1 + \frac{X_{i:n}}{\sum_{k=i+1}^{n} X_{k:n}}\right). \end{aligned} \quad (8)$$

Finally,

$$J_{X,n}(n) = \frac{1}{X_{n,n}}\log\left(\frac{n}{n-1}\right). \quad (9)$$

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

$$\frac{\sqrt{n}(\widetilde{Z}_n - Z_F)}{S_{X,n}} \to_d \mathcal{N}(0,1). \quad (10)$$

Proofs can be found in Greselin *et al.* (2009a), which in some cases crucially rely on the Vervaat process (see Zitikis, 1998, Greselin *et al.*, 2009b, and references therein).

3 A simulation study and an analysis of Italian income data

In this section, we first investigate the numerical performance of the estimators $\widehat{Z}_n$ and $\widetilde{Z}_n$ in a simulation study, and then evaluate confidence intervals for the Zenga index on real income data.

We begin by simulating data from Pareto distribution (3) with $x_0 = 1$ and $\theta = 2.06$. These choices give the value $Z_F = 0.6$, which we approximately see 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, and thus our results concerning the coverage accuracy and size of confidence intervals will not be affected by the choice.

Following Davison and Hinkley (1997, 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 1, and those for the estimator $\widehat{Z}_n$ using numerical differentiation as in Greselin and Pasquazzi (2009). In Table 1 we report coverage percentages and

Table 1 Coverage proportions and mean size 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	mean size	0.9000	0.9500	0.9750	0.9900	mean size
n	Normal confidence intervals									
200	0.7915	0.8560	0.8954	0.9281	0.1493	0.7881	0.8527	0.8926	0.9266	0.1500
400	0.8059	0.8705	0.9083	0.9409	0.1164	0.8047	0.8693	0.9078	0.9396	0.1167
800	0.8256	0.8889	0.9245	0.9514	0.0899	0.8246	0.8882	0.9237	0.9503	0.0900
n	Percentile confidence intervals									
200	0.7763	0.8326	0.8684	0.9002	0.1456	0.7629	0.8190	0.8567	0.8892	0.1462
400	0.8004	0.8543	0.8919	0.9218	0.1140	0.7934	0.8487	0.8864	0.9179	0.1143
800	0.8210	0.8777	0.9138	0.9415	0.0883	0.8168	0.8751	0.9119	0.9393	0.0884
n	BCa confidence intervals									
200	0.8082	0.8684	0.9077	0.9383	0.1491	0.8054	0.867	0.9047	0.9374	0.1497
400	0.8205	0.8863	0.9226	0.9531	0.1183	0.8204	0.886	0.9212	0.9523	0.1186
800	0.8343	0.8987	0.9331	0.9634	0.0925	0.8338	0.8983	0.9323	0.9634	0.0927
n	t-bootstrap confidence intervals									
200	0.8475	0.9041	0.9385	0.9658	0.2068	0.8485	0.9049	0.9400	0.9675	0.2099
400	0.8535	0.9124	0.9462	0.9708	0.1550	0.8534	0.9120	0.9463	0.9709	0.1559
800	0.8580	0.9168	0.9507	0.9758	0.1159	0.8572	0.9169	0.9504	0.9754	0.1162

mean sizes of 10,000 confidence intervals, for each of the four types: normal, percentile, BCa, and t-bootstrap. Bootstrap based approximations have been obtained from 9999 resamples of the original samples. We have approximated the acceleration constant for the BCa confidence intervals by one-sixth times the standardized third moment of the influence values. As expected, the confidence intervals based on

$\widehat{Z}_n$ and $\widetilde{Z}_n$ exhibit similar characteristics. We observe from Table 1 that all the 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. 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 $\widehat{Z}_n$ with 0.99 nominal confidence level is given by the largest order statistics of the bootstrap distribution (and $1,641$ respectively for $\widetilde{Z}_n$).

The second set of results we present here arises from using the Zenga index to analyze income data from the Bank of Italy 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 the survey, we refer to the Bank of Italy (2006) publication. In Table 2 we report the values of $\widehat{Z}_n$ and $\widetilde{Z}_n$ according to the geographic area of 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. 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 (1994), however, good care is 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.

Acknowledgements The research has been partially supported by the 2009 FAR, University of Milano Bicocca, and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

BANK OF ITALY (2006). *Household Income and Wealth in 2004. Supplements to the Statistical Bulletin–Sample Surveys*, 16 (7).

DAVISON A.C. AND HINKLEY D.V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.

GRESELIN, F. AND PASQUAZZI, L. (2009). Asymptotic confidence intervals for a new inequality measure. *Communications in Statistics: Computation and Simulation,* 38 (8), 17–42.

Table 2 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

GRESELIN, F., PASQUAZZI, L. AND ZITIKIS, R. (2009a). Zenga's new index of economic inequality, its estimation, and an analysis of incomes in Italy. *MPRA Paper 17147* available at http://ideas.repec.org/p/pra/mprapa/17147.html

GRESELIN, F., PURI, M.L. AND ZITIKIS, R. (2009b). *L*-functions, processes, and statistics in measuring economic inequality and actuarial risks. *Statistics and Its Interface,* 2, 227–245.

MAASOUMI E. (1994). Empirical analysis of welfare and inequality. In: *Handbook of Applied Econometrics, Volume II: Microeconomics*. (Eds.: M.H. Pesaran and P. Schmidt). Blackwell, Oxford.

ZENGA, M.(2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica & Applicazioni*, 5, 3–27.

ZITIKIS, R.(1998). The Vervaat process. In: *Asymptotic Methods in Probability and Statistics*, pp. 667–694. North-Holland, Amsterdam.