Rapporto n. 208

Cross-regional results on income inequality in Italy: issues and evidence from survey data

Francesca Greselin
Leo Pasquazzi

Maggio 2011

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Piazza dell'Ateneo Nuovo, 1 - 20126 Milano - Italia
Tel +39/02/64483103/39 - Fax +39/2/64483105
Segreteria di redazione:
Andrea Bertolini

Noname manuscript No.
(will be inserted by the editor)

Cross-regional results on income inequality in Italy: issues and evidence from survey data

Francesca Greselin · **Leo Pasquazzi**

Received: date / Accepted: date

F. Greselin
Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Università di Milano Bicocca, Milan, Italy, Tel.: +39 02 64483118
Fax: +39 02 64483105
E-mail: francesca.greselin@unimib.it

L. Pasquazzi
Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali, Università di Milano Bicocca, Milan, Italy E-mail: leo.pasquazzi@unimib.it

Keywords Asymptotic confidence interval, Dagum, Gini, Income inequality, Lognormal, Zenga.

Mathematics Subject Classification (2000) 62D05 · 62F12 · 62G20 · 62P25.

Abstract In this paper we analyze the differences in household income inequality among Italian regions. Using data from the 2008 Bank of Italy's *Survey on Household Income and Wealth*, a remarkable gap in inequality among different geographic areas of the country has been observed. Besides, a thorough analysis of the theoretical and practical aspects of obtaining parametric and non parametric confidence intervals for Gini's and Zenga's inequality measures has been provided. The performance of the inferential procedures has been assessed and their effectiveness in developing a cross-regional study is shown.

1 Introdution and motivation

Since more than a century, economists and statisticians have been concerned with the problem of modeling income and wealth distributions and measuring inequality. Such research helps in making well informed decisions for elected representatives, businesses, unions and non-profit organizations, as well as individuals. In particular, the assessment of relationships between inequality and growth is getting an increased attention among scholars and policy makers, in connection with other variables such as human capital, employment, and so on. Furthermore, in recent studies about EU countries the evolution in time of regional disparities appears to be related to the dynamics of economic and monetary integration. Measuring inequality is hence a crucial issue for economics and social sciences.

Unfortunately, we cannot rely only on point estimates of inequality measures, because in many empirical studies large standard errors are observed (see, among many others, Maasoumi 1997). Therefore, it is important to provide methodologies to assess whether differences in estimates are statistically significant. In this work we contribute to the issue by comparing parametric and non parametric estimation techniques. In particular, we will consider two parametric families, the Lognormal and the Dagum models. Our purpose is to obtain confidence intervals for Zenga's new inequality measure (Zenga, 2007) and for Gini's traditional index (Gini, 1914). The aim is to verify what is actually gained when we exploit the information about the underlying model. We will compare the two methodologies (parametric versus non-parametric) firstly in a simulation study, in order to assess coverage accuracy and length. Afterwards, we employ the same techniques on real income data, from Bank of Italy's *Survey on Household Income and Wealth*, to obtain a cross-regional analysis of differences in inequality. Our results show a remarkable gap in inequality among different geographic areas of the country, ranging from the its lowest level observed in Aosta Valley, to its highest level in the Apulia region. Besides Gini's traditional index,

we choose to analyze confidence intervals for Zenga's new inequality index because of its many interesting features: indeed, Zenga's index is a normalized inequality measure; it is the area under a new inequality curve (Zenga, 2007) and it may be decomposed by subgroups as well as by income sources (Radaelli, 2008; Greselin et al. 2009; Radaelli, 2010). For an analysis of the partial order induced by Zenga's inequality curve in some well known models for economic size distributions, we refer to Polisicchio and Porro (2008) and Porro (2008). Properties of the curve and its index have been investigated in Polisicchio (2008) and Maffenini and Polisicchio (2010). For inferential results see Greselin and Pasquazzi (2009) and Greselin et al. (2010).

The rest of the paper is organized as follows. Section 2 provides a brief description of technical details for the computation of the confidence intervals. Simulation results developed to assess the performance of the inferential procedures are presented in section 3. In section 4 we present and discuss our results on inequality in Italy's regions. Conclusions and final remarks end the paper in section 5. A detailed description about the computation of parametric confidence intervals derived from Lognormal and Dagum families has been given in the Appendix.

2 Asymptotic confidence intervals

Let $X_1, X_2, ..., X_n$ be an i.i.d. sample from an unknown distribution F. Gini's index may be defined by

$$G(F) = \int_0^1 2(p - L(p;F))dp, \tag{2.1}$$

where

$$L(p;F) = \frac{\int_0^p F^{-1}(t)dt}{\int_0^1 F^{-1}(t)dt}, \quad 0 < p < 1 \tag{2.2}$$

is the Lorenz curve, while Zenga's new measure is given by

$$Z(F) = \int_0^1 \left(1 - \frac{1-p}{p} \cdot \frac{L(p;F)}{1-L(p;F)}\right) dp. \tag{2.3}$$

As usual, we assume that the support of F is a subset of the non negative real line. Moreover, in order that the two inequality measures be well defined we need to

assume that the first moment of F is finite. The rationale behind definition 2.3 is straightforward. Indeed, $\frac{L(p;F)}{p}$ is the mean income of the poorest p percent of the population, while $\frac{1-L(p;F)}{1-p}$ is the mean income of the remaining part of the population. Notice that the ratio, say $U(p;F)$, between this two means takes on values in $[0,1]$. Small values of this ratio correspond to high inequality between the two groups, while large values correspond to situations close to equity. Thus $Z(F)$ is the mean, i.e. a synthesis, of the point inequality measures given by $I(p;F) = 1 - U(p;F)$.

If $\widehat{F}_n$ is the empirical CDF associated to the observed sample, we may estimate the two inequality measures simply by plugging in $\widehat{F}_n$ instead of F in (2.1), (2.2) and (2.3). Under mild restrictions on F (Hoeffding, 1945; Greselin et al., 2009b) both inequality measures may be represented as

$$T(\widehat{F}_n) = T(F) + \frac{1}{n}\sum_{i=1}^{n} h_T(X_i;F) + o_p(n^{-1/2}) \tag{2.4}$$

where the $h_T(X_i;F)$ is the influence function evaluated at the point X_i, i.e. (as usual δ_X denotes the distribution with unit mass at the point X)

$$h_T(X_i;F) = \lim_{\lambda \downarrow 0} \frac{T(F + \lambda(\delta_{X_i} - F)) - T(F)}{\lambda}.$$

It follows that both inequality measures have normal asymptotic distribution, i.e.

$$\sqrt{n}\left(T(\widehat{F}_n) - T(F)\right) \xrightarrow{L} N(0, \sigma_T^2),$$

where $\sigma_T^2 = Var_F(h_T(X_i;F))$. Knowing a consistent estimor $S_{T;n}^2$ for σ_T^2, we may compute the non-parametric normal $(1-2\alpha)$ confidence interval given by

$$\left(T(\widehat{F}_n) - z_{1-\alpha}\frac{S_{T;n}}{\sqrt{n}}; T(\widehat{F}_n) + z_{1-\alpha}\frac{S_{T;n}}{\sqrt{n}}\right),$$

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

If F is known to belong to a parametric family $\mathscr{F}_\Theta$ indexed by a k-dimensional real parameter vector $\theta \in \Theta \subset \mathbb{R}^k$, then the two functionals in (2.1) and (2.3) are functions of θ and we will simply write $T(\theta)$ instead of $T(F_\theta)$. In this case we may estimate $T(\theta)$ by $T(\widehat{\theta}_n)$, where $\widehat{\theta}_n$ is the maximum likelihood estimate of the

unknown value of θ. If T and $\mathscr{F}_\Theta$ satisfy suitable regularity conditions, then this estimator is asymptotically normal and efficient, i.e.

$$\sqrt{n}\left(T(\widehat{\theta}_n) - T(\theta)\right) \xrightarrow{L} N(0, \sigma_T^2(\theta)), \tag{2.5}$$

where $\sigma_T^2(\theta) = \frac{\partial T}{\partial \theta'} \mathbf{I}_\theta^{-1} \frac{\partial T}{\partial \theta}$. In the variance expression $\frac{\partial T}{\partial \theta}$ and $\mathbf{I}_\theta$ indicate the (column) vector of partial derivatives of T with respect to the components of the parameter vector and the information matrix at the unknown value of θ, respectively. If $\sigma_T^2(\theta)$ is continuous in θ, then $\sigma_T^2(\widehat{\theta}_n)$ is a consistent estimator of $\sigma_T^2(\theta)$.

Besides the normal confidence intervals just described we will also consider different types of bootstrap confidence intervals, i.e. percentile, Bias Corrected Accelerated Bootstrap (Bca) and t-bootstrap confidence intervals.

For the non-parametric versions of these confidence intervals, we proceed as in Greselin and Pasquazzi (2009) and estimate the bootstrap distributions by taking $R = 9,999$ resamples from the original sample (i.e. from $\widehat{F}_n$). As variance estimator for σ_T^2 we use

$$S_{T;n}^2 = \frac{1}{n} \sum_{i=1}^{n} h_T(X_i; \widehat{F}_n)^2,$$

and, following Efron (1987), we estimate the acceleration constant for the Bca confidence intervals by

$$\widehat{a} = \frac{1}{6} \frac{\sum_{i=1}^{n} h_T(X_i; \widehat{F}_n)^3}{\left(\sum_{i=1}^{n} h_T(X_i; \widehat{F}_n)^2\right)^{3/2}}. \tag{2.6}$$

Heuristically, we may say that in the parametric versions of the confidence intervals nothing changes with respect to the non-parametric setting, except that $F_{\widehat{\theta}_n}$ plays the role of $\widehat{F}_n$. Indeed, the expansion corresponding to (2.4) in the non-parametric setting, may now be replaced by

$$T(\widehat{\theta}_n) = T(\theta) + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f_\theta(X_i)}{\partial \theta'} \mathbf{I}_\theta^{-1} \frac{\partial T}{\partial \theta} + o_p(n^{-1/2}), \tag{2.7}$$

where $\frac{\partial \ln f_\theta(X_i)}{\partial \theta}$ is the score vector (a column vector) of the i-th sample component X_i at the unknown true value of θ. Thus, we may use

$$h_T(X_i; \widehat{\theta}_n) = \left.\frac{\partial \ln f_\theta(X_i)}{\partial \theta'}\right|_{\theta=\widehat{\theta}_n} \mathbf{I}_{\widehat{\theta}_n}^{-1} \left.\frac{\partial T}{\partial \theta}\right|_{\theta=\widehat{\theta}_n} \tag{2.8}$$

instead of $h_T(X_i;\widehat{F}_n)$ for estimating the variance $\sigma_T^2(\theta)$, which results in

$$\begin{aligned} V_{T;n}^2 &= \frac{1}{n}\sum_{i=1}^{n} h_T(X_i;F_{\widehat{\theta}_n})^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial T}{\partial \theta'}\right|_{\theta=\widehat{\theta}_n} \mathbf{I}_{\widehat{\theta}_n}^{-1} \left.\frac{\partial \ln f_\theta(X_i)}{\partial \theta}\right|_{\theta=\widehat{\theta}_n} \left.\frac{\partial \ln f_\theta(X_i)}{\partial \theta'}\right|_{\theta=\widehat{\theta}_n} \mathbf{I}_{\widehat{\theta}_n}^{-1} \left.\frac{\partial T}{\partial \theta}\right|_{\theta=\widehat{\theta}_n}. \end{aligned} \tag{2.9}$$

In the same way we just substitute $h_T(X_i;\widehat{F}_n)$ by $h_T(X_i;\widehat{\theta}_n)$ in (2.6) to get the estimate of the acceleration constant for the parametric Bca confidence interval.

A detailed discussion about the computation of parametric confidence intervals derived from Lognormal and Dagum families is given in the Appendix.

3 Simulation results

In an effort to gauge the actual performance of the inferential procedures, we developed a simulation study to assess coverage and mean length of the confidence intervals. As mentioned above, we considered two parent distributions, the Lognormal and the Dagum distribution, with parameter values given by the maximum likelihood estimates obtained on the Italian equivalent income distribution in 2006 ($\gamma = 9.6823$ and $\delta = 0.6093$ for the Lognormal; $a = 3.6781$, $b = 19261.86$ and $p = 0.6875$ for the Dagum model). We drew $10,000$ samples from each parent distribution and computed the parametric and non-parametric confidence intervals for Gini's and Zenga's new index. The coverage accuracies in Table 3.1 and Table 3.2 are the proportion of confidence intervals that contain the true value of the inequality measure, while the mean lengths are the average of the $10,000$ lengths of the 95% confidence intervals. From theory we expect that, because of the asymptotic efficiency of the ML estimators, the parametric confidence intervals should be shorter than the non-parametric ones. Indeed, in the light tailed Lognormal case (Table 3.1) we notice that the parametric confidence intervals clearly outperform the non-parametric ones. The former are both shorter and have larger coverage probability than the corresponding non-parametric ones. In estimating Gini's index with samples of size 100, *on average*, the

					Lognormal parent distribution					
$1-2\alpha$	0.9	0.95	0.975	0.99	mean length	0.9	0.95	0.975	0.99	mean length
sample size	**Normal confidence intervals**					**Percentile confidence intervals**				
					Gini - non-parametric					
100	0.8624	0.9200	0.9531	0.9769	0.0918	0.8452	0.8989	0.9327	0.9583	0.0913
200	0.8773	0.9363	0.9671	0.9828	0.0669	0.8716	0.9269	0.9553	0.9760	0.0668
400	0.8839	0.9347	0.9640	0.9844	0.0481	0.8759	0.9302	0.9590	0.9793	0.0480
					Gini - parametric					
100	0.8932	0.9461	0.9720	0.9886	0.0861	0.8847	0.9362	0.9649	0.9827	0.0861
200	0.8977	0.9493	0.9734	0.9877	0.0612	0.8976	0.9454	0.9693	0.9863	0.0612
400	0.8909	0.9452	0.9731	0.9888	0.0433	0.8891	0.9423	0.9703	0.9873	0.0433
					Zenga - non-parametric					
100	0.8671	0.9255	0.9593	0.9813	0.1040	0.8594	0.9104	0.9408	0.9657	0.1037
200	0.8784	0.9372	0.9658	0.9843	0.0745	0.8737	0.9296	0.9567	0.9766	0.0744
400	0.8781	0.9368	0.9671	0.9855	0.0532	0.8742	0.9287	0.9581	0.9790	0.0531
					Zenga - parametric					
100	0.8929	0.9469	0.9737	0.9895	0.1033	0.8847	0.9362	0.9649	0.9827	0.1033
200	0.8973	0.9496	0.9742	0.9882	0.0728	0.8976	0.9454	0.9693	0.9863	0.0728
400	0.8915	0.9449	0.9733	0.9890	0.0514	0.8891	0.9423	0.9703	0.9873	0.0514
	Bca confidence intervals					**t-bootstrap confidence intervals**				
					Gini - non-parametric					
100	0.8691	0.9226	0.9543	0.9783	0.0940	0.8881	0.9403	0.9672	0.9864	0.1054
200	0.8807	0.9366	0.9657	0.9828	0.0685	0.8928	0.9461	0.9733	0.9880	0.0727
400	0.8837	0.9350	0.9644	0.9847	0.0488	0.8906	0.9413	0.9697	0.9876	0.0503
					Gini - parametric					
100	0.9018	0.9513	0.9768	0.9903	0.0880	0.9015	0.9487	0.9695	0.9832	0.0929
200	0.9002	0.9521	0.9761	0.9886	0.0619	0.9003	0.9492	0.9725	0.9866	0.0635
400	0.8928	0.9462	0.9744	0.9891	0.0436	0.8922	0.9477	0.9730	0.9894	0.0441
					Zenga - non-parametric					
100	0.8766	0.9286	0.9599	0.9824	0.1031	0.8865	0.9377	0.9687	0.9866	0.1117
200	0.8845	0.9377	0.9642	0.9837	0.0744	0.8955	0.9469	0.9725	0.9880	0.0783
400	0.8805	0.9358	0.9660	0.9850	0.0533	0.8873	0.9426	0.9711	0.9873	0.0549
					Zenga - parametric					
100	0.9018	0.9513	0.9768	0.9903	0.1029	0.9095	0.9593	0.9808	0.9925	0.1062
200	0.9002	0.9521	0.9761	0.9886	0.0727	0.9045	0.9562	0.9778	0.9905	0.0738
400	0.8928	0.9462	0.9744	0.9891	0.0514	0.8950	0.9490	0.9758	0.9911	0.0518

Table 3.1 *Simulation results: coverages and mean lengths of confidence intervals for Gini's and Zenga's inequality measures. Samples are drawn from the Lognormal distribution with parameters* $\gamma = 9.682300$ *and* $\delta = 0.609262$

length of the interval decreases by 0.0074 and the coverage improves by 2.11% (the corresponding figures are 0.0068 and $+1.08\%$ for samples of size 200; 0.0052 and $+0.77$, respectively, for samples of size 400). Further, choosing a nominal coverage of 0.95, the coverage improves from having a *maximum difference from the nominal* of 5.11% to a value of 1.38% with samples of size 100 for the Lognormal model (analogous results for the other cases). As for Zenga's index, the length of the interval decreases by 0.0031 and coverage increases by 1.86%, on average, when the sample size is 100 (respectively 0.0033 and 1.12% for samples of size 200; 0.0037 and 0.86% for samples of size 400).

Commenting the results of the Dagum case (Table 3.2), it is important to keep in mind the problem of non existence of the ML estimates discussed in the Appendix. Indeed, the difference in coverage accuracy between the parametric and non-parametric

	Dagum parent distribution									
$1-2\alpha$	0.9	0.95	0.975	0.99	mean length	0.9	0.95	0.975	0.99	mean length
sample size	**Normal confidence intervals**					**Percentile confidence intervals**				
	Gini - non-parametric									
100	0.8469	0.9088	0.9430	0.9691	0.0977	0.8759	0.9346	0.9633	0.9819	0.0968
200	0.8666	0.9229	0.9551	0.9752	0.0727	0.8802	0.9357	0.9644	0.9841	0.0722
400	0.8835	0.9384	0.9653	0.9832	0.0533	0.8897	0.9442	0.9721	0.9876	0.0532
	Gini - parametric									
100	0.7839	0.8439	0.8810	0.9042	0.0993	0.8092	0.8631	0.8915	0.9106	0.0994
200	0.8620	0.9175	0.9485	0.9678	0.0723	0.8730	0.9254	0.9539	0.9698	0.0723
400	0.8932	0.9459	0.9718	0.9862	0.0520	0.8938	0.9464	0.9726	0.9883	0.0520
	Zenga - non-parametric									
100	0.8486	0.9143	0.9506	0.9748	0.1218	0.8757	0.9336	0.9643	0.9833	0.1210
200	0.8644	0.9240	0.9596	0.9818	0.0890	0.8815	0.9359	0.9672	0.9870	0.0886
400	0.8781	0.9375	0.9676	0.9857	0.0647	0.8899	0.9456	0.9719	0.9883	0.0646
	Zenga - parametric									
100	0.8081	0.8645	0.8933	0.9140	0.1234	0.8263	0.8745	0.9029	0.9181	0.1234
200	0.8786	0.9276	0.9561	0.9727	0.0880	0.8841	0.9324	0.9587	0.9747	0.0881
400	0.8962	0.9499	0.9737	0.9891	0.0625	0.8983	0.9509	0.9742	0.9895	0.0625
	Bca confidence intervals					**t-bootstrap confidence intervals**				
	Gini - non-parametric									
100	0.8279	0.8836	0.9213	0.9504	0.1016	0.8530	0.9135	0.9483	0.9725	0.1196
200	0.8514	0.9063	0.9401	0.9661	0.0758	0.8644	0.9206	0.9534	0.9778	0.0837
400	0.8757	0.9309	0.9581	0.9778	0.0553	0.8783	0.9339	0.9657	0.9833	0.0586
	Gini - parametric									
100	0.8182	0.8672	0.8939	0.9103	0.1045	0.8120	0.8678	0.8944	0.9141	0.1046
200	0.8806	0.9319	0.9584	0.9737	0.0742	0.8751	0.9290	0.9557	0.9714	0.0740
400	0.9004	0.9503	0.9735	0.9885	0.0526	0.8963	0.9497	0.9742	0.9889	0.0523
	Zenga - non-parametric									
100	0.8363	0.8939	0.9275	0.9564	0.1213	0.8603	0.9215	0.9539	0.9770	0.1359
200	0.8495	0.9064	0.9425	0.9680	0.0895	0.8678	0.9242	0.9562	0.9816	0.0973
400	0.8744	0.9263	0.9567	0.9773	0.0654	0.8793	0.9352	0.9640	0.9843	0.0692
	Zenga - parametric									
100	0.8158	0.8677	0.8942	0.9116	0.1240	0.8251	0.8735	0.9014	0.9173	0.1252
200	0.8781	0.9304	0.9574	0.9744	0.0883	0.8831	0.9322	0.9589	0.9753	0.0886
400	0.8967	0.9505	0.9746	0.9874	0.0626	0.8994	0.9506	0.9750	0.9898	0.0626

Table 3.2 *Simulation results: coverages and mean lengths of confidence intervals for Gini's and Zenga's inequality measures. Samples are drawn from the Dagum distribution with parameters* $a = 3.678111$, $b = 19261.86$ *and* $p = 0.687456$

confidence intervals for samples of size $n = 100$ is approximately equal to the proportion of samples for which our root search algorithm did not come up with a *good* solution of the likelihood equations (see section 2.2 for the figures). For $n = 200$ and $n = 400$, when the problem of non-existence of the solutions of the likelihood equations becomes negligible, the parametric Dagum confidence intervals perform clearly better than the corresponding non-parametric ones. For Zenga's index, the length of the interval decreases by 0.0029 and coverage increases by 0.48%, on average, when the sample size is 200 (0.0034 and 1.15%, respectively, for samples of size 400). Similar results have been obtained for Gini's measure.

Summarizing, we may say that switching to parametric confidence intervals leads to a considerable improvement in terms of coverage accuracy for both inequality measures.

4 Cross-regional results of income inequality in Italy

In this section we compute the above confidence intervals for Gini's and Zenga's inequality measures on income samples from the 20 Italian regions. Our aim is to assess differences in inequality among regions and to show how the methodology developed above can be effective to this aim. The data come from the 2008 wave of the Survey on Household Income and Wealth conducted by the Bank of Italy. Along with income, this survey reports different characteristics of $7,977$ households and their members such as geographic location, age, employment status etcetera. For detailed information on the survey, we refer to the Bank of Italy (2008) publication. Since the theory in the preceding section applies only to positive incomes, we deleted the non-positive incomes we found in the sample.[1]

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 OECD (Organization for Economic Co-operation and Developement) 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 (see Bank of Italy, 2008).

The confidence intervals we present in this section should, however, be interpreted with some care. Indeed, as in virtually every large official survey, the SHIW data were collected according to a complex sample design, but the methods presented in the preceding sections apply only to simple random sampling. There is growing attention

[1] Overall, there are 19 non positive incomes in the sample: 1 in each of the samples from Trentino, Veneto, Liguria, Abruzzo, Molise and Sardinia, 2 in the samples from Emilia Romagna and Campania, 3 in the sample from Sicily and 6 in the sample from Apulia.

in the literature to methods for variance and interval estimation for complex designs. A correct use of these methods is however far from trivial, and many researchers therefore stick with simple random sampling methods to get at least approximations to the results they would have obtained if they employed more involved procedures. In this work we will follow this custom.

Let us now analyze the content of Tables 5.3, 5.4 and 5.5. The first group of columns reports for each region the non-parametric point estimates of the inequality measures and the corresponding 95% confidence intervals. The second and third groups of columns refer to the parametric Lognormal and Dagum confidence intervals. Besides the maximum likelihood estimates of the inequality measures, the values of the Supremum Class Anderson Darling test statistic, i.e.

$$\begin{aligned} AD &= \sup_x \frac{|F_n(x) - F_{\widehat{\theta}_n}(x)|}{F_{\widehat{\theta}_n}(x)(1 - F_{\widehat{\theta}_n}(x))} \\ &= \sup_{i=1,2,\ldots,n} \frac{|i/n - F_{\widehat{\theta}_n}(x_i)|}{F_{\widehat{\theta}_n}(x_i)(1 - F_{\widehat{\theta}_n}(x_i))} \end{aligned}$$

and the p-values of the goodness of fit test based on this statistic are reported. The p-value is the fraction of bootstrap replicates of the AD statistic which is larger than the value of the AD statistic observed on the original sample. While in the Lognormal case the p-value is always based on $9,999$ bootstrap replicates, in the Dagum case the number of bootstrap replicates for the computation of the p-value may be larger, since in case our root search algorithm does not find a local maximum for some bootstrap resample we put the value of the bootstrap replicate of the AD statistic equal to infinity and take an additional resample in order to get anyway $9,999$ bootstrap replicates of the inequality measures and their studentized versions for the computation of the confidence intervals. The figures in brackets to the right of the p-values for the Dagum model are the number of resamples we needed to take. To put the value of the AD statistic equal to infinity when the ML estimates of the parameters do not exist

seems a natural choice, as the probability of this event tends to zero if the Dagum model is indeed the *true* model.

The p-values for the Lognormal model reveal the rather poor fit provided by this model. Of the 20 regional income distributions considered in this work, the Lognormal model seems to fit only those from Aosta Valley, Trentino, Veneto, Liguria, Umbria, Abruzzo, Basilicata, Calabria and Sardinia. The Dagum model performs better, having a p-value larger than 0.025 in 18 out of the 20 regional income distributions.[2]

Figures 4.1 and 4.2 graphically show the normal, Bca and t-bootstrap confidence intervals, for ease of comparison among regions and between the parametric and non parametric types of estimation. The parametric Lognormal and/or Dagum confidence intervals are reported only if the p-value for the model is greater than 0.01. The regions on the abscissa are ordered according to the value of the non-parametric point estimate of Gini's or Zenga's measure. It is worth noting that the orderings induced by the non-parametric point estimates of Gini's and Zenga's index are different, because different measures reveal different features of the income distribution (for a deeper insight on this issue, see Greselin et al., 2010).

The non-parametric point estimates for the Gini (Zenga) index range from 0.1621 (0.4137) in Aosta Valley to 0.3400 (0.6665) in Apulia. As perhaps expected, the difference between the non-parametric and the ML-estimates of the inequality measures is quite small (usually less than 0.02, with few exceptions given by Molise, Abruzzo and Apulia) whenever, according to the p-value, the model fits well the data.

Finally, we will now comment on the confidence intervals. In general, we observe that the different methods to construct confidence intervals seem to have little impact on interval locations (whenever the parametric model has a good fit) but a slight improvement arises in establishing interval boundaries, when we compare results from the parametric (giving shorter intervals) and the non parametric approach. Besides this, we notice that even though moderately large sample sizes, the confidence inter-

[2] The only exceptions are Lazio and Campania regions.

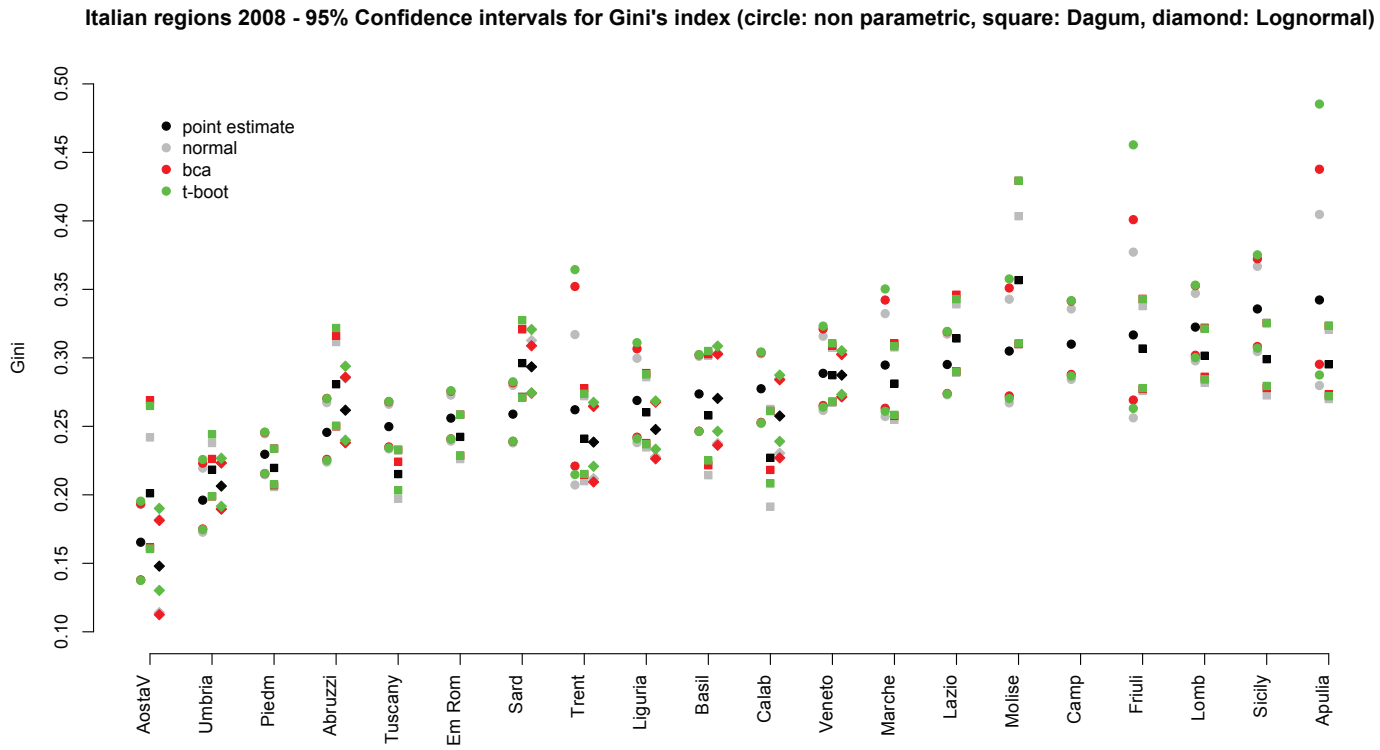

Fig. 4.1 *Gini's index: Confidence intervals based in the non-parametric, the Lognormal and the Dagum parametric approaches for the twenty Italian regions*

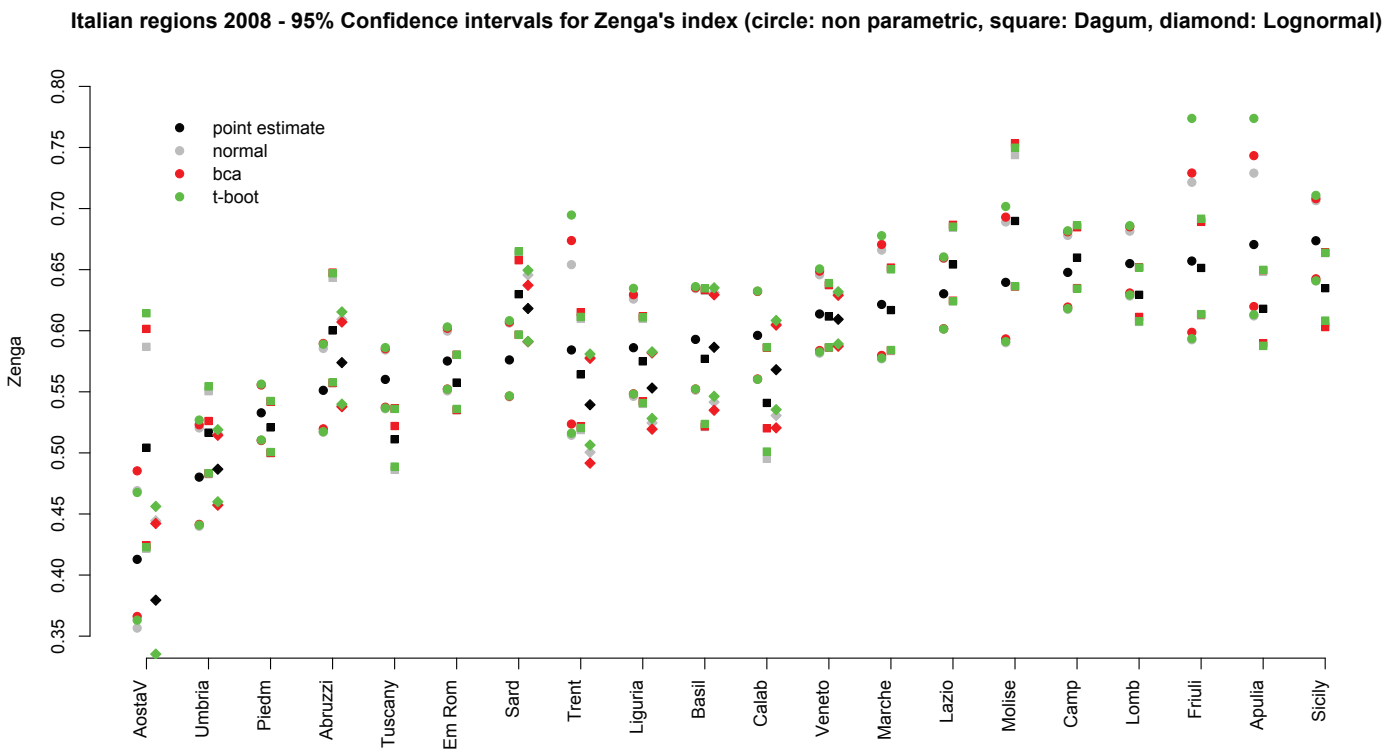

Fig. 4.2 *Zenga's index: Confidence intervals based in the non-parametric, the Lognormal and the Dagum parametric approaches for the twenty Italian regions*

vals for the inequality measures are occasionally very large. For example, on the 445 observations from the Apulia region, the non-parametric normal confidence interval for Gini's (Zenga's) index stretches from 0.2798 to 0.4047 (0.6121 to 0.7290). The corresponding Lognormal confidence interval is much shorter, but the p-value for the Lognormal model is virtually zero. The ML-estimates for the parameters of the Dagum model do exist in this region giving raise to a good model fit to data, assessed by a p-value of 0.1426, so that we obtain the shorter Dagum confidence intervals stretching from 0.2699 to 0.3207 (0.5875 to 0.6484). On the other side, the shortest nonparametric interval arises for Piedmont, with a width of 0.0298 for the Gini's index (respectively 0.0456 for the Zenga's index). For Piedmont the estimation is based on a sample of 789 statistical units, and we can rely on the non-parametric and the parametric Dagum approach.

The t-bootstrap confidence intervals in figures 4.1 and 4.2 are in some instances markedly right skewed, suggesting that convergence to the normal limit distribution has not yet occurred for the considered sample sizes.

From Liguria data, the estimation of the Gini index gives a non-parametric confidence interval of about 6 percentage points, while the parametric interval estimate is 4 percentage points wide for the Lognormal and 5 percentage points wide for the Dagum model. In the case of the Zenga index, those figures are respectively 8 percentage points, 6 for the Lognormal and 7 for the Dagum. Those figure are more interesting if we rephrase them in terms of the point estimate of the index, giving a relative variation of about 23% for the Gini and of only 14% for the Zenga index.

When confidence intervals do not overlap, as in Aosta Valley and Abruzzi (non parametric and Dagum confidence intervals), for example, we can unambiguously assess this difference to be statistically significant. So we can say that Aosta Valley has the lowest degree of inequality with respect to all other regions (with the only exception of Umbria). We can also infer that Umbria has a lower inequality in household income than Emilia Romagna, Sardinia, Liguria, Basilicata, and so on.

5 Conclusions

The results reported and discussed in this paper concern a cross-regional analysis of household income inequality in Italy, based on real income data coming from Bank of Italy's 2008 *Survey on Household Income and Wealth*. We observe considerable differences among geographic areas as the estimates for the Gini (Zenga) index range from a value of 0.1621 (0.4137) in Aosta Valley to 0.3400 (0.6665) in Apulia region, denoting a remarkable gap. Further, looking for statistical significance of comparisons among regions, we explored parametric inference for Gini's and Zenga's new index of inequality and compared its performance with respect to the non-parametric approach. We considered two common models in income analysis: the Lognormal and the Dagum families. Recalling their popularity with income distributions, we saw that in our case the Lognormal model hardly fits the real data, because of its light tail, while the Dagum model gives a further confirmation of its ability to describe income distributions, in almost all cases. The first results, based on simulations, confirm that there is some advantage in terms of length and coverage accuracy with parametric confidence intervals. However, as our application to Italy's regional income distributions shows, comparisons based on non-parametric and parametric inferential methods (when they may be applied) lead to almost the same conclusions.

6 Appendix

6.1 Lognormal confidence intervals

Recall first that a distribution belongs to the Lognormal family if its density function is given by

$$f(x;\gamma,\delta) = \frac{1}{\delta\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}\left(\frac{\ln x-\gamma}{\delta}\right)^2}, \quad x>0$$

for some $-\infty<\gamma<\infty$ and $\delta>0$. In our simulation study we used a Lognormal parent distribution with $\gamma=0.6823$ and $\delta=0.6093$, i.e. the maximum likelihood

	non-parametric		Lognormal		Dagum	
Abruzzo (Sample size $n = 200$, mean income 18,103.1887)						
AD	-		3.4861		2.0651	
p-value	-		0.0826		0.4548	(14,225)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2443	0.5491	0.2487	0.5541	0.2676	0.5834
normal	0.2239÷0.2673	0.5169÷0.5855	0.2383÷0.2853	0.5385÷0.6093	0.2500÷0.3115	0.5571÷0.6434
perc	0.2219÷0.2654	0.5134÷0.5817	0.2246÷0.2719	0.5165÷0.5879	0.2249÷0.2857	0.5239÷0.6093
Bca	0.2259÷0.2704	0.5197÷0.5896	0.2380÷0.2859	0.5378÷0.6071	0.2500÷0.3162	0.5572÷0.6476
t-boot	0.2251÷0.2702	0.5174÷0.5890	0.2399÷0.2939	0.5399÷0.6154	0.2505÷0.3218	0.5578÷0.6470
Aosta Valley (Sample size $n = 45$, mean income 18,743.6443)						
AD			1.8304		1.3251	
p-value			0.3005		0.9223	(35,729)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.1621	0.4137	0.1690	0.4192	0.1822	0.4635
normal	0.1378÷0.1930	0.3566÷0.4691	0.1140÷0.1819	0.3145÷0.4444	0.1604÷0.2419	0.4217÷0.5868
perc	0.1298÷0.1854	0.3586÷0.4723	0.1332÷0.2007	0.3477÷0.4767	0.1244÷0.2055	0.3401÷0.5049
Bca	0.1379÷0.1934	0.3661÷0.4853	0.1125÷0.1813	0.3027÷0.4422	0.1618÷0.2691	0.1608÷0.2649
t-boot	0.1375÷0.1954	0.3631÷0.4676	0.1302÷0.1901	0.3353÷0.4562	0.4244÷0.6015	0.4227÷0.6143
Apulia (Sample size $n = 445$, mean income 14,770.4706)						
AD	-		70785.5222		63.7935	
p-value	-		0.0000		0.1426	(11,662)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3400	0.6665	0.3323	0.6660	0.2976	0.6230
normal	0.2798÷0.4047	0.6121÷0.7290	0.3097÷0.3510	0.6393÷0.6884	0.2699÷0.3207	0.5875÷0.6484
perc	0.2821÷0.4055	0.6057÷0.7207	0.3113÷0.3523	0.6403÷0.6891	0.2749÷0.3260	0.5971÷0.6586
Bca	0.2953÷0.4377	0.6198÷0.7433	0.3095÷0.3506	0.6381÷0.6871	0.2733÷0.3234	0.5898÷0.6496
t-boot	0.2875÷0.4853	0.6131÷0.7738	0.3122÷0.3534	0.6404÷0.6894	0.2720÷0.3234	0.5877÷0.6498
Basilicata (Sample size $n = 128$, mean income 14,261.9247)						
AD	-		6.3792		2.1434	
p-value	-		0.0169		0.3388	(12,527)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2716	0.5907	0.2857	0.6069	0.2709	0.5932
normal	0.2462÷0.3011	0.5513÷0.6345	0.2375÷0.3033	0.5416÷0.6313	0.2143 ÷0.3017	0.5212 ÷0.6328
perc	0.2422÷0.2966	0.5460÷0.6296	0.2509÷0.3166	0.5574÷0.6470	0.2419 ÷0.3301	0.5530 ÷0.6656
Bca	0.2465÷0.3022	0.5523÷0.6352	0.2362÷0.3027	0.5349÷0.6294	0.2216 ÷0.3027	0.5217 ÷0.6334
t-boot	0.2464÷0.3024	0.5523÷0.6361	0.2464÷0.3086	0.5463÷0.6351	0.2251 ÷0.3050	0.5237 ÷0.6347
Calabria (Sample size $n = 190$, mean income 14,087.3514)						
AD	-		4.0953		1.9182	
p-value	-		0.0561		0.5553	(16,409)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2759	0.5941	0.2797	0.5987	old	old
normal	0.2520÷0.3029	0.5603÷0.6321	0.2304÷0.2847	0.5306÷0.6056	0.1914 ÷0.2627	0.4952 ÷0.5866
perc	0.2492÷0.2996	0.5555÷0.6273	0.2509÷0.3054	0.5574÷0.6329	0.2485 ÷0.3202	0.5601 ÷0.6515
Bca	0.2529÷0.3038	0.5605÷0.6322	0.2270÷0.2840	0.5205÷0.6046	0.2182 ÷0.2612	0.5201 ÷0.5860
t-boot	0.2525÷0.3043	0.5602÷0.6325	0.2390÷0.2874	0.5354÷0.6084	0.2083 ÷0.2613	0.5011 ÷0.5864
Campania (Sample size $n = 625$, mean income 12,523.8394)						
AD	-		3,857,032.5014		35.3209	
p-value	-		0.0000		0.0005	(10,000)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3093	0.6462	0.3466	0.6826	0.3156	0.6509
normal	0.2843÷0.3357	0.6175÷0.6780	0.3232÷0.3591	0.6560÷0.6970	0.3038 ÷0.3466	0.6347 ÷0.6850
perc	0.2842÷0.3353	0.6149÷0.6752	0.3283÷0.3641	0.6613÷0.7021	0.2853 ÷0.3276	0.6167 ÷0.6672
Bca	0.2879÷0.3414	0.6194÷0.6810	0.3233÷0.3592	0.6551÷0.6968	0.3039 ÷0.3492	0.6349 ÷0.6847
t-boot	0.2867÷0.3418	0.6181÷0.6818	0.3254÷0.3604	0.6567÷0.6975	0.3038 ÷0.3513	0.6345 ÷0.6864
Emilia Romagna (Sample size $n = 718$, mean income 23,532.0146)						
AD	-		33.3814		2.4770	
p-value	-		0.0009		0.1895	(10,000)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2556	0.5741	0.2715	0.5873	0.2495	0.5664
normal	0.2392÷0.2728	0.5507÷0.5997	0.2556÷0.2826	0.5650÷0.6031	0.2261 ÷0.2585	0.5349 ÷0.5802
perc	0.2389÷0.2725	0.5486÷0.5975	0.2579÷0.2849	0.5678÷0.6059	0.2407 ÷0.2732	0.5524 ÷0.5976
Bca	0.2409÷0.2756	0.5523÷0.6020	0.2561÷0.2828	0.5651÷0.6030	0.2286 ÷0.2585	0.5351 ÷0.5803
t-boot	0.2408÷0.2760	0.5521÷0.6031	0.2568÷0.2835	0.5656÷0.6035	0.2287 ÷0.2586	0.5360 ÷0.5804

Table 5.1 *Cross-regional levels of income inequality expressed by the Gini and the Zenga indexes, followed by their 95% confidence intervals, in non-parametric (columns 1-2) and parametric setting (columns 3-6), for Italian regions from Abruzzo to Emilia Romagna (segue).*

	non-parametric		Lognormal		Dagum	
Friuli (Sample size $n = 253$, mean income $23,412.4072$)						
AD	-		175.2041		3.4627	
p-value	-		0.0000		0.0704	(10,024)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3143	0.6519	0.3269	0.6595	0.2985	0.6373
normal	0.2562÷0.3772	0.5926÷0.7215	0.3159÷0.3694	0.6464÷0.7110	0.2758÷0.3378	0.6125÷0.6903
perc	0.2579÷0.3769	0.5827÷0.7094	0.2992÷0.3531	0.6249÷0.6900	0.2604÷0.3221	0.5836÷0.6616
Bca	0.2692÷0.4009	0.5987÷0.7291	0.3156÷0.3675	0.6457÷0.7058	0.2773÷0.3431	0.6134÷0.6891
t-boot	0.2631÷0.4555	0.5935÷0.7738	0.3172÷0.3776	0.6469÷0.7136	0.2779÷0.3429	0.6136÷0.6917
Lazio (Sample size $n = 413$, mean income $19,815.4741$)						
AD	-		10,328,146.4376		31.6265	
p-value	-		0.0000		0.0006	(9,999)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2945	0.6291	0.3372	0.6718	0.3067	0.6448
normal	0.2730÷0.3173	0.6013÷0.6592	0.3393÷0.3822	0.6739÷0.7242	0.2893 ÷0.3391	0.6240 ÷0.6845
perc	0.2721÷0.3162	0.5983÷0.6563	0.3153÷0.3580	0.6454÷0.6954	0.2755 ÷0.3252	0.6055 ÷0.6656
Bca	0.2740÷0.3190	0.6017÷0.6596	0.3390÷0.3804	0.6739÷0.7194	0.2901 ÷0.3461	0.6246 ÷0.6867
t-boot	0.2737÷0.3193	0.6014÷0.6604	0.3398÷0.3892	0.6746÷0.7259	0.2900 ÷0.3429	0.6242 ÷0.6852
Liguria (Sample size $n = 314$, mean income 22154.2403)						
AD	-		4.2168		1.3990	
p-value	-		0.0531		0.8331	(13,711)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2677	0.5839	0.2648	0.5777	0.2617	0.5769
normal	0.2381÷0.2997	0.5463÷0.6259	0.2278÷0.2676	0.5245÷0.5817	0.2346 ÷0.2859	0.5401 ÷0.6098
perc	0.2370÷0.2984	0.5412÷0.6203	0.2448÷0.2841	0.5482÷0.6047	0.2385 ÷0.2899	0.5438 ÷0.6142
Bca	0.2421÷0.3066	0.5484÷0.6295	0.2263÷0.2677	0.5194÷0.5820	0.2378 ÷0.2889	0.5423 ÷0.6117
t-boot	0.2409÷0.3111	0.5482÷0.6347	0.2333÷0.2685	0.5282÷0.5827	0.2369 ÷0.2879	0.5408 ÷0.6109
Lombardy (Sample size $n = 844$, mean income $25,119.3734$)						
AD	-		1,182,207,030.5158		226.0723	
p-value	-		0.0000		0.0258	(10,264)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3220	0.6539	0.3315	0.6650	0.3081	0.6382
normal	0.2978÷0.3471	0.6284÷0.6815	0.3111÷0.3409	0.6408÷0.6763	0.2819÷0.3212	0.6072÷0.6516
perc	0.2983÷0.3475	0.6267÷0.6797	0.3166÷0.3459	0.6469÷0.6818	0.2957÷0.3348	0.6249÷0.6691
Bca	0.3019÷0.3527	0.6309÷0.6851	0.3108÷0.3410	0.6397÷0.6762	0.2862÷0.3220	0.6113÷0.6519
t-boot	0.3002÷0.3531	0.6293÷0.6860	0.3130÷0.3416	0.6417÷0.6766	0.2842÷0.3214	0.6077÷0.6517
Marche (Sample size $n = 354$, mean income $20,695.0153$)						
AD	-		471,034,784.5457		347.5015	
p-value	-		0.0000		0.0388	(10,403)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2936	0.6189	0.3266	0.6591	0.2815	0.6118
normal	0.2572÷0.3323	0.5770÷0.6660	0.2792÷0.3247	0.6016÷0.6566	0.2548÷0.3076	0.5834÷0.6504
perc	0.2567÷0.3316	0.5713÷0.6607	0.3029÷0.3484	0.6297÷0.6846	0.2565÷0.3092	0.5736÷0.6408
Bca	0.2632÷0.3422	0.5798÷0.6706	0.2815÷0.3245	0.6012÷0.6567	0.2576÷0.3105	0.5841÷0.6517
t-boot	0.2610÷0.3503	0.5776÷0.6778	0.2857÷0.3256	0.6038÷0.6577	0.2584÷0.3084	0.5842÷0.6503
Molise (Sample size $n = 137$, mean income $17,100.6103$)						
AD			16.7838		1.9577	
p-value			0.0026		0.3262	(11,210)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3026	0.6356	0.3183	0.6491	0.3348	0.6688
normal	0.2671÷0.3428	0.5902÷0.6890	0.2609÷0.3318	0.5781÷0.6661	0.3101÷0.4034	0.6359÷0.7438
perc	0.2635÷0.3395	0.5807÷0.6804	0.2815÷0.3522	0.6012÷0.6890	0.2688÷0.3611	0.5938÷0.7005
Bca	0.2722÷0.3510	0.5933÷0.6930	0.2584÷0.3313	0.5684÷0.6648	0.3100÷0.4292	0.6363÷0.7534
t-boot	0.2703÷0.3576	0.5911÷0.7017	0.2715÷0.3359	0.5819÷0.6687	0.3105÷0.4294	0.6365÷0.7498
Piedmont (Sample size $n = 789$, mean income $19,844.2667$)						
AD	-		49.3686		2.1106	
p-value	-		0.0004		0.3041	(10,027)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2292	0.5319	0.2379	0.5376	0.2234	0.5241
normal	0.2147÷0.2445	0.5100÷0.5556	0.2180÷0.2408	0.5067÷0.5421	0.2057÷0.2336	0.4997÷0.5421
perc	0.2142÷0.2438	0.5082÷0.5531	0.2263÷0.2490	0.5193÷0.5546	0.2134÷0.2411	0.5059÷0.5483
Bca	0.2155÷0.2455	0.5102÷0.5556	0.2167÷0.2407	0.5037÷0.5419	0.2070÷0.2338	0.5000÷0.5419
t-boot	0.2156÷0.2458	0.5106÷0.5563	0.2199÷0.2412	0.5082÷0.5425	0.2076÷0.2337	0.5007÷0.5425

Table 5.2 *Cross-regional levels of income inequality expressed by the Gini and the Zenga indexes, followed by their 95% confidence intervals, in non-parametric (columns 1-2) and parametric setting (columns 3-6), for Italian regions from Friuli to Piedmont (continuation).*

	non-parametric		Lognormal		Dagum	
Sardinia (Sample size $n = 334$, mean income $15,826.8052$)						
AD			8.5697		2.1011	
p-value			0.0125		0.2506	(10,084)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2581	0.5746	0.2683	0.5827	0.2798	0.6068
normal	0.2379÷0.2799	0.5459÷0.6063	0.2742÷0.3127	0.5908÷0.6456	0.2718÷0.3205	0.5968÷0.6630
perc	0.2362÷0.2785	0.5422÷0.6028	0.2489÷0.2871	0.5543÷0.6088	0.2402÷0.2891	0.5510÷0.6169
Bca	0.2390÷0.2818	0.5464÷0.6070	0.2742÷0.3088	0.5911÷0.6372	0.2711÷0.3211	0.5966÷0.6577
t-boot	0.2392÷0.2825	0.5467÷0.6082	0.2744÷0.3207	0.5913÷0.6495	0.2710÷0.3275	0.5968÷0.6649
Sicily (Sample size $n = 526$, mean income $13,364.4196$)						
AD			27.0621		3.5450	
p-value			0.0012		0.0855	(10,170)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.3349	0.6721	0.3441	0.6797	0.3133	0.6494
normal	0.3046÷0.3668	0.6408÷0.7065	0.3263÷0.3653	0.6594÷0.7043	0.2726÷0.3255	0.6062÷0.6635
perc	0.3036÷0.3665	0.6371÷0.7031	0.3239÷0.3634	0.6559÷0.7013	0.3017÷0.3544	0.6350÷0.6920
Bca	0.3083÷0.3723	0.6425÷0.7081	0.3260÷0.3653	0.6585÷0.7034	0.2780÷0.3254	0.6030÷0.6641
t-boot	0.3071÷0.3752	0.6409÷0.7108	0.3278÷0.3681	0.6597÷0.7054	0.2795÷0.3254	0.6083÷0.6638
Trentino (Sample size $n = 173$, mean income $21,645.1988$)						
AD	-		2.6089		1.7095	
p-value	-		0.1568		0.4141	(10,334)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2599	0.5787	0.2618	0.5734	0.2419	0.5591
normal	0.2072÷0.3170	0.5144÷0.6541	0.2118÷0.2652	0.5005÷0.5783	0.2100÷0.2720	0.5187÷0.6099
perc	0.2103÷0.3179	0.5063÷0.6431	0.2337÷0.2873	0.5311÷0.6091	0.2128÷0.2751	0.5078÷0.5998
Bca	0.2210÷0.3521	0.5236÷0.6738	0.2094÷0.2645	0.4916÷0.5774	0.2148÷0.2778	0.5217÷0.6152
t-boot	0.2148÷0.3644	0.5161÷0.6947	0.2208÷0.2675	0.5064÷0.5809	0.2152÷0.2738	0.5204÷0.6113
Tuscany (Sample size $n = 619$, mean income $23,840.4409$)						
AD	-		23.6587		2.7008	
p-value	-		0.0013		0.1764	(10,437)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2492	0.5590	0.2545	0.5628	0.2340	0.5380
normal	0.2335÷0.2661	0.5361÷0.5842	0.2411÷0.2689	0.5431÷0.5842	0.1971÷0.2332	0.4861÷0.5362
perc	0.2326÷0.2655	0.5336÷0.5818	0.2401÷0.2680	0.5411÷0.5823	0.2353÷0.2711	0.5400÷0.5898
Bca	0.2350÷0.2679	0.5374÷0.5853	0.2407÷0.2685	0.5420÷0.5831	0.2242÷0.2327	0.5220÷0.5364
t-boot	0.2342÷0.2682	0.5367÷0.5862	0.2425÷0.2708	0.5442÷0.5857	0.2033÷0.2326	0.4888÷0.5360
Umbria (Sample size $n = 267$, mean income $18,926.5312$)						
AD	-		4.2205		2.3714	
p-value	-		0.0520		0.1829	(10,138)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.1952	0.4782	0.2037	0.4819	0.2051	0.4934
normal	0.1727÷0.2195	0.4399÷0.5204	0.1895÷0.2233	0.4577÷0.5156	0.1989÷0.2378	0.4826÷0.5505
perc	0.1721÷0.2189	0.4358÷0.5161	0.1867÷0.2201	0.4519÷0.5093	0.1729÷0.2115	0.4363÷0.5034
Bca	0.1751÷0.2231	0.4415÷0.5230	0.1896÷0.2233	0.4572÷0.5144	0.1989÷0.2262	0.4833÷0.5261
t-boot	0.1746÷0.2258	0.4409÷0.5269	0.1916÷0.2267	0.4600÷0.5190	0.1990÷0.2441	0.4834÷0.5544
Veneto (Sample size $n = 596$, mean income $20,567.9419$)						
AD			7.8753		2.2449	
p-value			0.0166		0.2945	(10,815)
	Gini	Zenga	Gini	Zenga	Gini	Zenga
point est.	0.2880	0.6122	0.2862	0.6076	0.2823	0.6052
normal	0.2616÷0.3158	0.5816÷0.6458	0.2718÷0.3030	0.5883÷0.6304	0.2673÷0.3074	0.5860÷0.6373
perc	0.2613÷0.3155	0.5789÷0.6426	0.2700÷0.3013	0.5852÷0.6275	0.2575÷0.2977	0.5723÷0.6243
Bca	0.2652÷0.3211	0.5838÷0.6488	0.2715÷0.3024	0.5873÷0.6290	0.2680÷0.3089	0.5863÷0.6374
t-boot	0.2640÷0.3232	0.5826÷0.6506	0.2733÷0.3052	0.5892÷0.6319	0.2680÷0.3106	0.5863÷0.6390

Table 5.3 *Cross-regional levels of income inequality expressed by the Gini and the Zenga indexes, followed by their 95% confidence intervals, in non-parametric (columns 1-2) and parametric setting (columns 3-6), for Italian regions from Sardinia to Veneto (end).*

estimates obtained on the Italian equivalent income distribution in 2006, from which we generated $10,000$ simple random samples of size $n = 100$, 200 and 400.

In this model the likelihood equations have a solution at every point of the sample space (i.e. points of $\mathbb{R}^n$ with strictly positive coordinates), and the maximum likeli-

hood estimators of the parameters are given by

$$\widehat{\gamma}_n = \frac{1}{n}\sum_{i=1}^{n} \ln X_i, \qquad \widehat{\delta}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \ln^2 X_i - \widehat{\gamma}_n^2}. \tag{6.1}$$

The Lorenz curve and Gini's index are respectively given by (Zenga, 1984)

$$L(p;\gamma,\delta) = \Phi\left(\Phi^{-1}(p) - \delta\right) \qquad 0 < p < 1$$

and

$$G(\gamma,\delta) = 2\Phi(\delta/\sqrt{2}) - 1.$$

The value of Zenga's index may be obtained by substituting $L_F(p)$ with $L(p;\gamma,\delta)$ in (2.3) so that, by simple algebra

$$Z(\gamma,\delta) = \int_0^1 \frac{p - \Phi(\Phi^{-1}(p) - \delta)}{p\{1 - \Phi(\Phi^{-1}(p) - \delta)\}} dp, \tag{6.2}$$

where the integral has to be computed by numerical methods. In order to get the influence values $h_T(X_i,\theta)$ we need the gradient vectors of the two inequality functionals with respect to the parameters. Since γ is a scale parameter, we just need to compute the derivatives of the two functionals with respect to δ. For Gini's index we easily get

$$\frac{\partial G}{\partial \delta} = \sqrt{2}\phi(\delta/\sqrt{2}),$$

where ϕ is the standard normal density function. For Zenga's index in (6.2) we may take the derivative under the integral sign. Thus, we obtain

$$\frac{\partial Z}{\partial \delta} = \int_0^1 \frac{1-p}{p} \frac{\phi(\Phi^{-1}(p) - \delta)}{[1 - \Phi(\Phi^{-1}(p) - \delta)]^2} dp.$$

Finally, recalling that the Lognormal information matrix is given by

$$\mathbf{I} = \begin{bmatrix} \frac{1}{\delta^2} & 0 \\ 0 & \frac{1}{2\delta^4} \end{bmatrix}$$

and that the score function is given by

$$\frac{\partial \ln f(x;\gamma,\delta)}{\partial \delta} = \frac{1}{\delta}\left[\left(\frac{\ln x - \gamma}{\delta}\right)^2 - 1\right], \tag{6.3}$$

we see that

$$h_G(x;\gamma,\delta) = 2^{3/2}\delta^3\left[\left(\frac{\ln x-\gamma}{\delta}\right)^2 - 1\right]\phi(\delta/\sqrt{2})$$

when dealing with Gini's index, whereas

$$h_Z(x;\gamma,\delta) = 2\delta^3\left[\left(\frac{\ln x-\gamma}{\delta}\right)^2 - 1\right]\int_0^1 \frac{1-p}{p}\frac{\phi(\Phi^{-1}(p)-\delta)}{[1-\Phi(\Phi^{-1}(p)-\delta)]^2}dp$$

when Zenga's index is considered.

6.2 Dagum confidence intervals

F belongs to the Dagum family if its density function is given by

$$f(x) = \frac{apx^{ap-1}}{b^{ap}\left[1+\left(\frac{x}{b}\right)^a\right]^{p+1}}, \quad x>0$$

for some $a,b,p>0$ (Kleiber and Kotz, 2003). Notice that the first moment of a Dagum distribution is finite if and only if $a>1$, and therefore the inequality measures we consider in this paper are only defined for the subfamily of Dagum distributions with $a>1$.

As in the Lognormal case, we used the maximum likelihood estimates from the Italian equivalent income distribution as parameter values for the parent distribution. Thus we simulated $10,000$ samples from the Dagum distribution with $a=3.6781$, $b=19,262$ and $p=0.6875$ in order to analyze coverage accuracy and length of confidence intervals for Gini's and Zenga's new index.

Given an i.i.d. sample $x_1,x_2,...,x_n$, the likelihood equations for the Dagum family are given by

$$\begin{cases} \dfrac{n}{a} + p\displaystyle\sum_{i=1}^n \ln\left(\frac{x_i}{b}\right) - (p+1)\sum_{i=1}^n \frac{\ln\left(\frac{x_i}{b}\right)}{1+\left(\frac{b}{x_i}\right)^a} = 0 \\ np - (p+1)\displaystyle\sum_{i=1}^n \frac{1}{1+\left(\frac{b}{x_i}\right)^a} = 0 \\ \dfrac{n}{p} + a\displaystyle\sum_{i=1}^n \ln\left(\frac{x_i}{b}\right) - \sum_{i=1}^n \ln\left[1+\left(\frac{x_i}{b}\right)^a\right] = 0 \end{cases} \tag{6.4}$$

However, no explicit solution of this system is known. The ML estimation problem is easier to handle if we observe that the natural logarithm of a Dagum random variable follows a generalized logistic distribution with density function given by

$$f(y) = \frac{\alpha}{\sigma} \frac{e^{-\frac{x-\theta}{\sigma}}}{\left(1+e^{-\frac{x-\theta}{\sigma}}\right)^{\alpha+1}}, \quad -\infty < y < \infty,$$

where $-\infty < \theta < \infty$ and $\alpha, \sigma > 0$. Notice that θ and σ are the location and scale parameter, respectively, while α is a shape parameter that affects asymmetry. The parameters of the generalized logistic distribution are related to those of the Dagum distribution by the relations

$$a = \frac{1}{\sigma}, \quad b = e^{\theta}, \quad p = \alpha.$$

Thus, the problem of solving the system in (6.4) is equivalent to the problem of finding a solution of the likelihood equations of the generalized logistic distribution, which are given by

$$\begin{cases} \dfrac{n}{\alpha} - \displaystyle\sum_{i=1}^{n} \ln\left(1+e^{-\frac{y_i-\theta}{\sigma}}\right) = 0 \\ -\dfrac{n}{\sigma} + \dfrac{1}{\sigma}\displaystyle\sum_{i=1}^{n} \dfrac{y_i-\theta}{\sigma} - \dfrac{\alpha+1}{\sigma}\displaystyle\sum_{i=1}^{n} \dfrac{\frac{y_i-\theta}{\sigma}}{1+e^{\frac{y_i-\theta}{\sigma}}} = 0 \\ \dfrac{n}{\sigma} - \dfrac{\alpha+1}{\sigma}\displaystyle\sum_{i=1}^{n} \dfrac{1}{1+e^{\frac{y_i-\theta}{\sigma}}} = 0 \end{cases} \tag{6.5}$$

In our simulation study we employed an iterative two step procedure for solving this system. At step i we first find an update θ_i and σ_i of the location and scale parameters through a single Newton-Raphson step applied to the last two equations of the system (6.5). Then we use θ_i and σ_i in the first equation in order to get an update α_i of the shape parameter. If the likelihood function has a local maximum, this two step procedure will converge to it provided that the starting point $(\alpha_0, \theta_0, \sigma_0)$ is not too far from the solution. The initial values for this algorithm are thus of crucial importance. We select them by least squares fitting the three quartiles of the generalized logistic distribution with shape parameter $\alpha = 1$ to the corresponding quartiles of the natural

logarithm of the sample observations. Since the quantile function of the generalized logistic distribution, given by

$$y(t) = \theta + \sigma \ln\left(\frac{t^{\alpha}}{1-t^{\alpha}}\right), \qquad 0 < t < 1,$$

is linear in θ and σ, we find a closed form solution for the initial values. Indeed, putting $\alpha_0 = 1$ and denoting by Q_1, Q_2 and Q_3 the quartiles of the natural logarithm of the sample observations x_i, we see that the initial values θ_0 and σ_0 are given by the least squares solution of the linear system

$$\begin{cases} Q_1 = \theta - \sigma \ln 3 \\ Q_2 = \theta \\ Q_3 = \theta + \sigma \ln 3, \end{cases}$$

which yields $\theta_0 = (Q_1 + Q_2 + Q_3)/3$ and $\sigma_0 = (-Q_1 \ln 3 + Q_3 \ln 3)/(2\ln^2 3)$. In our simulations we allow for each sample a maximum number of $1,000$ iterations of the two step procedure above (making an exception for the sample coming from Aosta Valley, for which we needed $5,000$ iterations to reach convergence, due to the very small sample size of 45 statistical units). If the algorithm reduces the gradient of the likelihood function to a value smaller than 10^{-12} within the iterations, we test the hessian matrix for negative definiteness at the solution. If this test is positive we conclude that the solution is a local maximum of the likelihood equation. Notice that beyond a bad choice of the initial values, there may be another simple reason why this procedure does not deliver a local maximum. Indeed, as Shao (2002) points out, there exist points in the sample space such that a solution of the likelihood equations in (6.5), and therefore also of the system in (6.4), does not exist. Nevertheless, with probability tending to 1 as the sample size increases, there exists a sequence of solutions of the likelihood equations of the generalized logistic distribution that is consistent and asymptotically normally distributed (Abberger and Heiler, 2000).

So how do we handle samples on which the algorithm does not deliver a local maximum? And what happens if it finds a local maximum, but the inequality measures are not defined at that point (i.e. the ML estimate of the parameter a is not larger than 1)? Our answers to these questions depend on whether we are dealing with a bootstrap resample or not. In the former case we simply discard the sample and take another bootstrap resample until we reach a total of $9,999$ bootstrap resamples such that the algorithm converges to a local maximum at which the inequality measures are defined. Otherwise, if the sample we are dealing with is one of the *original* samples from the Dagum parent distribution of the simulation study, we use it for estimating the probability of the subset of the sample space on which the ML estimates of the inequality measures do not exist. For the Dagum parent distribution in our simulation study corresponding to the sample sizes $n = 100$, 200 and 400 these estimates are given by 0.0712, 0.0141 and 0.0010, respectively. It is worth noting that, among the samples drawn from the Dagum parent distribution, we never observed one giving rise to a solution provided by the algorithm outside the domain of the inequality measures (i.e. a solution with $a \leq 1$).

In any case, for the 20 *real* samples from regional income distributions we deal with in the application shown in section 4, the algorithm was *always* able to calculate the ML estimates. Let us now turn to the expressions of the inequality measures in the Dagum model. The Lorenz curve and Gini's index are respectively given by (Dagum, 1977)

$$L(t;a,b,p) = B\left(t^{1/p}; p+\frac{1}{a}; 1-\frac{1}{a}\right), \quad 0 < t < 1 \tag{6.6}$$

and

$$G(p,a,b) = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1. \tag{6.7}$$

In (6.6) we used $B(t;a;b)$ to indicate the beta cdf, while $\Gamma(x)$ indicates the Gamma function in (6.7). Substituting the Lorenz curve in the formula of Zenga's index, we

get

$$Z(p,a,b) = \int_0^1 \frac{t - B\left(t^{1/p}; p+\frac{1}{a}; 1-\frac{1}{a}\right)}{t[1 - B\left(t^{1/p}; p+\frac{1}{a}; 1-\frac{1}{a}\right)]} dt. \tag{6.8}$$

As noticed at the beginning of this section, the Lorenz curve, and thus the two inequality measures, are defined if and only if $a > 1$.

In order to get the influence values $h_T(X_i, \theta)$ in the Dagum case we need the gradient vectors of the two inequality functionals with respect to the parameters. Since b is a scale parameter, we just need to compute the derivatives of the two functionals with respect to a and p. For Gini's index in (6.7) it is easily checked that

$$\frac{\partial G}{\partial a} = \frac{G+1}{a^2}\left[\psi(p+1/a) - \psi(2p+1/a)\right],$$
$$\frac{\partial G}{\partial p} = (G+1)\left[\psi(p) + 2\psi(2p+1/a) + 2\psi(2p) - \psi(p+1/a)\right],$$

where $\psi(x)$ is the digamma function, i.e the derivative of $\ln\Gamma(x)$.

The partial derivatives of Zenga's new index with respect to the parameters of the Dagum family are rather cumbersome expressions. We do not report them here. For computational convenience, we suggest not to use the analytic expressions, but to approximate the partial derivatives of Zenga's index by Newton's difference quotient.

Finally, we need the score function (a vector valued function) $\partial \ln f(x;\theta)/\partial\theta$ and the information matrix $\mathbf{I}_\theta$ to compute the parametric influence values in (2.8) for the Dagum family. If we put $n = 1$ in the system of likelihood equations in (6.4) and multiply the second equation by a/b, we get the components of the score function on the LHS, while the components of the information matrix can be found, for example,

in Kleiber and Kotz (2003). We report them here for completeness:

$$I_{11} = \frac{1}{a^2(2+p)}\left\{p[(\psi(p)-\psi(1)-1)^2+\psi'(p)+\psi'(1)]+2[\psi(p)-\psi(1)]\right\};$$
$$I_{12} = \frac{p-1-p[\psi(p)-\psi(1)]}{b(2+p)};$$
$$I_{22} = \frac{a^2 p}{b^2(2+p)};$$
$$I_{13} = \frac{\psi(2)-\psi(p)}{a(1+p)};$$
$$I_{23} = \frac{a}{b(1+p)};$$
$$I_{33} = \frac{1}{p^2}.$$

References

Abberger K. and Heiler S.: Simultaneous estimation of parameters for a generalized logistic distribution and application to time series models. Allgemeines Statistisches Archiv **84(1)**, 41–49 (2000)

Banca d'Italia: Household Income and Wealth. Supplements to the Statistical Bulletin - Sample Surveys, **XX(8)** (2010)

Dagum C.: A new model of personal distribution: specification and estimation. Economie Appliquée **30**, 413–437 (1977)

Davison A.C. and Hinkley D.V.: Bootstrap Methods and their Application Cambridge Series in Statistical and Probabilistic Mathematics (1997)

Efron B.: Better Bootstrap confidence intervals (with discussion), Journal of the American Statistical Association **82**, 171-200 (1987)

Gini C.: Sulla misura della concentrazione e della variabilità dei caratteri. In: Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. Anno Accademico 1913–1914, **LXXII(2)** Premiate Officine Grafiche C. Ferrari, Venezia, 1201–1248 (1914)

Greselin F. and Pasquazzi L.: Asymptotic confidence intervals for a new inequality measure. Communications in Statistics: Computation and Simulation **38(8)**, 17-42 (2009)

Greselin F., Puri M.L., Zitikis R.: *L*-functions, processes, and statistics in measuring economic inequality and actuarial risks. Statistics and Its Interface, **2**, 227–245 (2009)

Greselin F., Pasquazzi L., Zitikis R.: Zenga's New Index of Economic Inequality, Its Estimation, and an Analysis of Incomes in Italy. Journal of Probability and Statistics DOI 10.1155/2010/718905 (2010)

Hoeffding W.: A Class of Statistics with Asymptotically Normal Distribution. The Annals of Mathematical Statistics, **19(3)**, 293–325 (1948)

Kleiber C. and Kotz S.: Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, Hoboken, NJ (2003)

Maasoumi E.: Empirical analysis of Welfare and Inequality. In: M.H. Pesaran and P.Schmidt (eds.) Handbook of Applied Econometrics. Blackwell (1994)

Maffenini W., Polisicchio M.: How potential is the I(p) inequality curve in the analysis of empirical distributions. Technical report No. 186, Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali, Università degli Studi di Milano - Bicocca, available via: http://www.dimequant.unimib.it/_ricerca/ricerca_lista.jsp (2010)

Moran T.P.: Statistical Inference for Measures of Inequality With a Cross-National Bootstrap Application. Sociological Methods & Research, **34(3)**, 296-333 (2006)

Polisicchio M.: The continuous random variable with uniform point inequality measure I(p). Statistica & Applicazioni **6**, 136–151 (2008)

Polisicchio M., Porro F.: The I(p) Curve for Some Classical Income Models. Technical report No. 159, Dipartimento di Metodi Quantitativi per le Scienze Economiche Aziendali, Università degli Studi di Milano - Bicocca, available via: http://www.dimequant.unimib.it/_ricerca/pubblicazione.jsp?id=169 (2008)

Porro F.: Equivalence between partial order based on curve $L(p)$ and partial order based on curve I(p). Proceedings of the XLIV Meeting of the Italian Statistics Society, available via: http://www.sis-statistica.it/files/pdf/atti/rs08_spontanee_a_3_4.pdf (2008)

Radaelli P.: A subgroup decomposition of Zenga's Uniformity and Inequality indexes. Statistica & Applicazioni **6**, 117–136 (2008)

Radaelli P.: (2010) On the Decomposition by Subgroups of the Gini Index and Zenga's Uniformity and Inequality Indexes. International Statistical Review **78**, 81–101 (2010)

Shao Q.: Maximum likelihood estimation for generalized logistic distribution. Communication in Statistics - Theory and Methods, **31(10)**, 1687–1700 (2002)

Zenga M.: Tendenza alla massima ed alla minima concentrazione per variabili casuali continue. Statistica **44**, 619–640 (1984)

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