Statistica & Applicazioni - Vol. VI, n. 2, 2008, pp. 99-116

MINIMUM SAMPLE SIZES IN ASYMPTOTIC CONFIDENCE INTERVALS FOR GINI'S INEQUALITY MEASURE

Francesca Greselin*
Leo Pasquazzi**

SUMMARY

Statistical inference for inequality measures has been of considerable interest in recent years. Income studies often deal with very large samples, hence precision would not seem a serious issue. Yet, in many empirical studies large standard errors are observed (Maasoumi, 1997). Therefore, it is important to provide methodologies to assess whether differences in estimates are statistically significant. This paper presents an analysis of the performance of asymptotic confidence intervals for Gini's index, virtually the most widely used inequality index. To determine minimum sample sizes assuring a given accuracy in confidence intervals, an extensive simulation study has been carried out. A wide set of underlying distributions has been considered, choosing from specific models for income data. As expected, the minimum sample sizes are seriously affected by some population characteristics as tail heaviness and asymmetry. However, in a wide range of cases, it turns out that they are smaller than sample sizes actually used in social sciences.

Keywords: *Gini concentration index, confidence intervals, estimated coverage, heavy tails.*

1. INTRODUCTION

Statistical inference for inequality measures has been of considerable interest in recent years. Income studies often deal with very large samples, hence precision would not seem a serious issue. Yet, in many empirical studies large standard errors are observed (Maasoumi, 1997). Therefore it is important to provide methodologies to assess whether differences in estimates are statistically significant. This work focuses on Gini's ratio R, one of the most popular and enduring inequality indices. Hoeffding, in his seminal work (Hoeffding, 1948), set the theoretical basis for inference on Gini's index. Applying the delta method to his general theory on U-statistics, he showed that the sample Gini index has a normal asymptotic distribution. Later works extended the asymptotic theory to include more general classes of con-

* Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali - Università degli Studi di Milano Bicocca - Piazza dell'Ateneo Nuovo, 1 - 20126, Milano (e-mail: francesca.greselin@unimib.it).

** Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali – Università degli Studi di Milano Bicocca - Piazza dell'Ateneo Nuovo, 1 - 20126, Milano (e-mail: leo.pasquazzi@unimib.it).

Although the present paper is the result of a close cooperation between the authors, Leo Pasquazzi drew up sections 2, 4, 6, while the other sections are mainly due to Francesca Greselin.

centration measures (see Sendler, 1979; Barret and Pendakur, 1995; Zitikis, 2002).

Giorgi and Provasi (1995) and Palmitesta *et al.* (1999) explored the asymptotic approach. Their results show that this inferential procedure can be affected by the slow speed of convergence of the sample distribution. Further studies revealed a better performance of the t-bootstrap method over the asymptotic one, for sample sizes smaller than 200. Among them, Palmitesta *et al.* (2000), and Giorgi *et al.* (2006), who carried out simulation studies sampling from heavy-tailed parametric distributions. For the best of our knowledge, only Andres and Calonge (2005) compare asymptotic and bootstrap approaches on moderately large samples (of size up to 2000), using Spanish income data as parent distribution. Their results confirm the better performance of the t-bootstrap method.

Nowadays, most authors employ bootstrap methods in their applied research. Mills and Zwandakily (1997) study youth inequality on data provided by the National Longitudinal Survey of Youth (sample size of 4,266 units). Xu (2000) evaluates changes in income inequality in the U.S. over time using data of the Panel Study of Income Dynamics (sample sizes from 2,206 to 6,244), and Moran (2005) compares results obtained from different bootstrap methods on data from income surveys of the LIS (sample sizes from 1,813 to 49,351).

Bootstrap methods are however computationally expensive. Moreover, the difference with respect to the asymptotic approach becomes less significant as the sample size increases. In inference studies involving large samples (i.e. income surveys), it seems therefore reasonable to retain the asymptotic approach. Latorre (1990) showed that sample sizes currently in use are large enough for constructing confidence intervals based on the maximum likelihood estimator for Gini's concentration index. Are they also adequate to assure a good coverage of asymptotic non parametric confidence intervals? This work's aim is to provide an answer to this question by investigating the minimum sample sizes in a non parametric setting. As expected, in accordance with the literature, the threshold size depends on the population distribution, mainly on its tail heaviness and asymmetry. Nevertheless, it appears that, in a wide range of cases, the minimum sample size required for the asymptotic methodology is smaller than sample sizes actually used in social sciences.

This paper is organized and presented as follows. The second section introduces asymptotic confidence intervals for Gini's index. The third section describes the methodology for the empirical verification and assessment of simulation results, which are presented in section four. In section five the length of the confidence intervals is examined, while section six presents some further considerations about the tails of the sample distribution of Gini's concentration measure. Conclusions and final remarks end the paper.

2. CONFIDENCE INTERVALS FOR GINI'S INEQUALITY INDEX

It is well known that the Gini index R may be expressed as ratio of two regular functionals, the Gini mean difference Δ and twice the mean μ. Let $X_1, X_2, \dots, X_n$ be i.i.d. with unknown distribution function F, whose variance σ^2 is assumed to exist. Let $\hat{\Delta}$ be the (unbiased) sample mean difference, $\bar{X}$ the sample mean and $\hat{R} = \hat{\Delta}/(2\bar{X})$ the sample Gini concentration ratio. In an application of his asymptotic distribution theory for U-statistics, Hoeffding (1948) proved that, if

$$\begin{aligned}\sigma_R^2 &= \lim_{n\to\infty} \sigma_{R,n}^2 \\ &= \lim_{n\to\infty} \left\{ \frac{1}{4\mu^2} nVar(\hat{\Delta}) + \frac{\Delta^2}{4\mu^4}\sigma^2 - \frac{\Delta}{2\mu^3} nCov(\bar{X}, \hat{\Delta}) \right\} > 0,\end{aligned}$$

then

$$\sqrt{n}\left(\frac{\hat{R} - R}{\sigma_{R,n}^2}\right) \xrightarrow{d} \mathrm{N}(0,1).$$

It can be shown (Dall'Aglio, 1965) that the condition $\sigma_R^2 > 0$ does not hold if and only if the distribution F is degenerate or F assigns probability $p > 0$ to $x_0 > 0$ and probability $1 - p$ to $x_1 = x_0 p^2/(1-p)^2$.

Now, $Var(\hat{\Delta})$ and $Cov(\bar{X}, \hat{\Delta})$ may be expressed as (Zenga *et al.,* 2004; Polisicchio, 1997)

$$Var(\hat{\Delta}) = \frac{4}{n(n-1)}\left[\sigma^2 + (n-2)\mathfrak{F} - \left(n - \frac{3}{2}\right)\Delta^2\right]$$

$$Cov(\bar{X}, \hat{\Delta}) = \frac{2}{n}(\mathfrak{D} - \mu\Delta),$$

where

$$\mathfrak{F} = \iiint |x_1 - x_2||x_1 - x_3| dF(x_1)dF(x_2)dF(x_3)$$

$$\mathfrak{D} = \iint x_1|x_1 - x_2| dF(x_1)dF(x_2).$$

Since variances and covariances of U-statistics are regular functionals, the variance $\sigma_{(R)}^2$ can be consistently estimated from the sample data, provided the second moment of the underlying distribution is finite (Lee 1990, p.122, theorem 3). In particular, the unbiased estimators

$$\hat{V}ar\left(\hat{\Delta}\right) \tfrac{4}{(n-2)(n-3)}\left[S^2 + (n-2)\hat{\mathfrak{F}} - \left(n - \tfrac{3}{2}\right)\hat{\Delta}^2\right]$$

$$\hat{C}ov\left(\bar{X}, \hat{\Delta}\right) = \tfrac{2}{(n-2)}\left(\hat{\mathfrak{D}} - \bar{X}\hat{\Delta}\right),$$

may be used. In the above formulae, S^2 is the unbiased variance estimator, while $\hat{\mathfrak{F}}$, $\hat{\mathfrak{D}}$ are the U-statistics corresponding to $\mathfrak{F}$ and $\mathfrak{D}$:

$$\hat{\mathfrak{F}} = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} |X_i - X_j||X_i - X_k|$$
$$\hat{\mathfrak{D}} = \frac{1}{n(n-1)} \sum_{i \neq j} X_i |X_i - X_j|. \quad (1)$$

It follows that

$$\hat{\sigma}^2_{(R)} = \frac{1}{4\bar{X}^2} n\hat{V}ar(\hat{\Delta}) + \frac{\hat{\Delta}^2}{4\bar{X}^2} S^2 - \frac{\hat{\Delta}}{2\bar{X}^3} n\hat{C}ov(\bar{X}, \hat{\Delta})$$

is a strongly consistent estimator for $\sigma^2_{(R)}$. Writing $\Phi^{-1}(p)$ to mean the p-th quantile of the standard normal distribution, it follows that the interval

$$\left(\hat{R} - \frac{\hat{\sigma}_{(R)}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right);\ \hat{R} + \frac{\hat{\sigma}_{(R)}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \quad (2)$$

contains the true value of Gini's concentration index with probability approximately equal to $1 - \alpha$, provided the sample size is large enough.

The confidence interval in (2) is symmetric with respect to the sample Gini concentration index. The bias of $\hat{R}$ is well approximated by

$$C = \frac{1}{n} \frac{R\sigma^2}{\mu^2} - \frac{1}{2\mu^2} Cov(\bar{X}, \hat{\Delta})$$

(Greselin, 1997). Taking into account the bias, the confidence interval (2) becomes

$$\left(\hat{R} - \hat{C} - \frac{\hat{\sigma}_{(R)}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right);\quad \hat{R} - \hat{C} + \frac{\hat{\sigma}_{(R)}}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right), \quad (3)$$

where

$$\hat{C} = \frac{1}{n} \frac{\hat{R}\, S^2}{\bar{X}^2} - \frac{1}{2\bar{X}^2} \hat{C}ov\left(\bar{X}, \hat{\Delta}\right).$$

It is worth noting that direct use of the formulae in (1) implies that computation time is *O(n³)* for $\hat{\mathfrak{F}}$ and *O(n²)* for $\hat{\mathfrak{D}}$ and $\hat{\Delta}$. Average computation time can, however, be drastically reduced to become *O(n ln n)* by the use of a sorting algorithm (i.e. Quicksort) to obtain the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, and exploiting the following formulae (see Zenga *et al.*, 2004; Polisicchio and Zenga, 1995):

$$\hat{\Delta} = \frac{2}{n(n-1)} \sum_{i=1}^{n} X_{(i)}(2i - n - 1)$$

$$\hat{\mathfrak{D}} \frac{1}{n(n-1)} \sum_{i=1}^{n} X^2_{(i)}(2i - n - 1)$$

$$\hat{\mathfrak{F}} = S^2 + \frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} (iX_{(i)} - T_i)^2 +$$

$$+ \frac{4}{(n-1)(n-2)} \sum_{i=1}^{n} (\bar{X} - X_{(i)})(iX_{(i)} - T_i),$$

where $T_i = \sum_{j=1}^{i} X_{(j)}$.

3. METHODOLOGY FOR SIMULATIONS AND ASSESSMENT OF RESULTS

This section determines the minimum sample sizes such that the coverage probability of the confidence intervals in (2) and (3) is reasonably close to the nominal confidence level $1-\alpha$. For this purpose, a simulation study has been performed involving a wide set of distributions, as will be described below. The simulation procedure was implemented using the software package R and may be summarized by the following steps:

a) Fix the sample size n. The sample sizes considered in this study vary between 30 and 3840.
b) Draw $N = 5000$ samples of size n, fixed at the previous step. The choice of N has to take into account two opposite requirements. On one hand, N as large as possible to give a better representation of the simulated sample space, and, on the other hand, N bounded to limit computation time.
c) For each sample compute the confidence intervals (2) and (3), and check whether they contain the true value of Gini's concentration index. The proportion of intervals satisfying this condition provides an estimate for the unknown coverage probability.
d) For comparison, estimate also the coverage probability of the traditional confidence interval for the mean μ.

The estimates $\hat{p}_n$ of the unknown coverage probabilities p_n will be used to determine minimum sample sizes as in Greselin and Maffenini (2007). In general, the effective unknown coverage p_n, estimated by $\hat{p}_n$, depends on the chosen sample size n, the $(1-\alpha)$-confidence level and the underlying distribution. In any case, $\lim_{n\to\infty} p_n = (1-\alpha)$. The issue is now to distinguish good values $\hat{p}_n$ from unacceptable ones. As the estimates $\hat{p}_n$ arise from simulations, they vary among different runs. Therefore, a testing procedure is needed, in order to distinguish fluctuations of $\hat{p}_n$ due to sampling variability from substantial differences in estimation. Relying on the assumption that (as in most applications) some coverage error γ with respect to the nominal confidence level $1-\alpha$ is tolerable, let

$$H_0 : p_n \geq 1 - \alpha - \gamma$$

be the hypothesis that the sample size n is sufficiently large, and

$$H_1 : p_n < 1 - \alpha - \gamma$$

be the hypothesis that the sample size is still too small. Since, for large values of N the estimated coverage probability $\hat{p}_n$ is approximately normally distributed, the decision rule is to reject H_0 if

$$\hat{p}_n < 1 - \alpha - \gamma - \Phi^{-1}(1-\delta)\sqrt{\frac{(1-\alpha-\gamma)(\alpha+\gamma)}{N}}$$

for some small positive value of δ. It is easily seen that, in this way, the rejection probability is roughly bounded by δ under H_0. δ is therefore the size of this test. In the following table the critical values are displayed for the case $N = 5000$, $\gamma = 0.1\alpha$ and $\delta = 0.05$.

TABLE 1. - *Critical values*

1-α	***0.90***	***0.95***	***0.975***	***0.99***
Critical values	0.8827	0.9397	0.9687	0.9866

4. SIMULATION RESULTS

Now, the selection of the underlying distributions for the simulations deserves a brief comment. The choice has to meet two different needs: on one hand, the requirement to explore populations with different characteristics, considering symmetric, asymmetric and/or heavy-tailed models; on the other hand, to analyze distributions widely employed in income studies. To this aim the rectangular, the exponential, the log-normal, the Pareto and the Dagum type I (Dagum, 1977) models have been considered. The following subsections present the simulation results.

4.1 *Rectangular distribution*

This distribution was included in this study merely with the aim of considering a symmetric model for comparison with asymmetric ones. Only one choice of the parameters is considered, fixing the support on the interval (0;1). The left half of table 2 displays estimated coverage probabilities of the confidence interval (2), whereas the right half refers to the traditional confidence interval for the mean. Table 3 reports the estimated coverage probabilities of the confidence interval (3). Figures in bold are coverage probabilities which lie in the acceptance region of H_0 (see table 1).

Tables 2 and 3 show that with sample sizes larger than 60 in almost all cases an acceptable coverage is obtained. As expected, the threshold sample sizes tend to be larger for high nominal confidence levels, as convergence in the tails of the sample distribution is more problematic.

TABLE 2. - *Estimated coverage probabilities (Rectangular distribution on (0,1))*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	***$1-\alpha$***	***0.99***	***0.95***	***0.975***	***0.99***
30	**0.8868**	0.9362	0.9610	0.9788		**0.8904**	**0.9400**	**0.9696**	0.9852
60	**0.8976**	**0.9446**	**0.9702**	0.9856		**0.8918**	**0.9410**	0.9678	**0.9868**
120	**0.8990**	**0.9514**	**0.9728**	**0.9866**		**0.8978**	**0.9496**	**0.9722**	**0.9878**
240	**0.8940**	**0.9442**	**0.9702**	**0.9882**		**0.8970**	**0.9450**	**0.9734**	**0.9870**
480	**0.8938**	**0.9452**	**0.9738**	**0.9912**		**0.8982**	**0.9464**	**0.9728**	**0.9900**
960	**0.8996**	**0.9530**	**0.9738**	**0.9886**		**0.9000**	**0.9500**	**0.9726**	**0.9874**
1920	**0.8956**	**0.9462**	**0.9742**	0.9900		**0.8966**	**0.9476**	**0.9730**	**0.9908**
3840	**0.9088**	**0.9530**	**0.9760**	**0.9890**		**0.9038**	**0.9554**	**0.9756**	**0.9892**

TABLE 3. - *Estimated coverage probabilities of c.i. (3) (Rectangular distr. on (0.1))*

	Bias corrected confidence intervals for R			
	$1-\alpha$			
n	***0.9***	***0.95***	***0.975***	***0.99***
30	**0.8908**	0.9366	0.9624	0.9788
60	**0.8994**	**0.9462**	**0.9696**	0.9858
120	**0.9004**	**0.9494**	**0.9724**	0.9862
240	**0.8934**	**0.9444**	**0.9704**	**0.9884**
480	**0.8950**	**0.9448**	**0.9744**	**0.9914**
960	**0.8998**	**0.9534**	**0.9742**	**0.9888**
1920	**0.8950**	**0.9460**	**0.9742**	**0.9902**
3840	**0.9082**	**0.9538**	**0.9754**	**0.9888**

4.2 *Exponential distribution*

The exponential family is acknowledged to provide models for many real fenomena. The density function is given by

$$f(x) = \lambda e^{-\lambda x}, \qquad \text{where} \qquad x \geq 0, \lambda > 0.$$

As λ is a scale parameter, the value of Gini's concentration index is 0.5 regardless of its value.

For this distribution, the minimum sample size for the Gini index confidence interval is of 240 units at all considered nominal confidence levels. On the other hand, it appears that the threshold sample size for the mean confidence interval increases as the nominal confidence level increases. A comparison between the estimated coverage probabilities in table 4 (left part) and table 5 suggests that the bias correction is not very influential.

TABLE 4. - *Estimated coverage probabilities (Exponential distribution, $\lambda = 0.2$)*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	*1−α*	***0.99***	***0.95***	***0.975***	***0.99***
30	0.8608	0.9156	0.9472	0.9724		0.8640	0.9140	0.9406	0.9642
60	0.8810	0.9372	0.9680	0.9820		**0.8852**	0.9312	0.9548	0.9722
120	0.8812	0.9366	**0.9688**	0.9860		**0.8916**	**0.9446**	**0.9694**	0.9842
240	**0.8956**	**0.9484**	**0.9722**	**0.9880**		**0.8952**	**0.9440**	**0.9712**	**0.9882**
480	**0.8940**	**0.9464**	**0.9730**	**0.9908**		**0.8932**	**0.9442**	**0.9736**	**0.9870**
960	**0.8988**	**0.9452**	**0.9702**	**0.9878**		**0.8890**	**0.9418**	**0.9692**	**0.9860**
1920	**0.8986**	**0.9490**	**0.9722**	**0.9892**		**0.8994**	**0.9494**	**0.9724**	**0.9892**
3840	**0.9038**	**0.9486**	**0.9754**	**0.9918**		**0.9018**	**0.9470**	**0.9748**	**0.9876**

TABLE 5. - *Estimated coverage probabilities of c.i.(3) (Exponential distr.,$\lambda = 0.2$)*

	Bias corrected confidence intervals for R			
	1−α			
n	***0.9***	***0.95***	***0.975***	***0.99***
30	0.8528	0.9128	0.9446	0.9716
60	0.8792	0.9354	0.9646	0.9816
120	0.8810	0.9356	0.9678	0.9854
240	**0.8942**	**0.9470**	**0.9722**	**0.9880**
480	**0.8940**	**0.9462**	**0.9732**	**0.9904**
960	**0.8984**	**0.9448**	**0.9702**	**0.9878**
1920	**0.8986**	**0.9490**	**0.9722**	**0.9890**
3840	**0.9038**	**0.9486**	**0.9754**	**0.9918**

4.3 *Lognormal distribution*

From now on, parametric models for income distributions will be considered. The lognormal density function is defined on the non-negative real axis by

$$f(x) = \frac{1}{\delta\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}\left(\frac{\ln x-\gamma}{\delta}\right)^2} \qquad \text{for } \delta > 0.$$

Three different sets of parameters will be chosen, in order to obtain values of R as in real economic distributions. The maximum likelihood estimates of γ and δ obtained on the Italian income and expences data (Banca d'Italia, 2002), are $\gamma = 10$ and $\delta = 0.5$, so that $R = 0.276$. Varying the value of δ, distributions with lower or higher degree of concentration may be obtained. Accordingly, also the choices

$\gamma = 10$, $\delta = 0.35829$ and $\gamma = 10$, $\delta = 0.74161$ will be considered. The simulation results are shown in Tables 6-8.

In particular, tables 6 and 7 show once again that estimated coverage probabilities for the two confidence intervals for R are very similar. Since this is true for all distributions included in this study, from now on, the estimated coverage probabilities for the bias corrected confidence interval will be omitted.

The threshold sample size is 480 for $\delta = 0.35829$ and $\delta = 0.5$ and increases to 1920 for $\delta = 0.74161$, suggesting perhaps a relation with a measure of asymmetry in the underlying distribution. The minimum sample sizes of the confidence interval for the mean are lower in almost all considered cases.

TABLE 6. - *Estimated coverage probabilities (Lognormal distribution, $\gamma = 10$; $\delta = 0.5$)*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	*1−α*	***0.99***	***0.95***	***0.975***	***0.99***
30	0.8488	0.9048	0.9360	0.9602		0.8738	0.9224	0.9498	0.9698
60	0.8674	0.9230	0.9540	0.9742		0.8806	0.9308	0.9582	0.9800
120	0.8826	0.9322	0.9604	0.9808		**0.8922**	**0.9402**	0.9672	0.9844
240	**0.8920**	0.9392	0.9640	0.9816		**0.8866**	**0.9408**	**0.9710**	0.9854
480	**0.8994**	**0.9478**	**0.9724**	**0.9876**		**0.9050**	**0.9492**	**0.9726**	**0.9872**
960	**0.8982**	**0.9482**	**0.9732**	**0.9874**		**0.8962**	**0.9496**	**0.9782**	**0.9918**
1920	**0.8970**	**0.9478**	**0.9726**	**0.9892**		**0.8940**	**0.9478**	**0.9744**	**0.9890**
3840	**0.9044**	**0.9542**	**0.9786**	**0.9898**		**0.8986**	**0.9514**	**0.9754**	**0.9906**

TABLE 7. - *Estimated coverage probabilities of c.i.(3) (Lognormal distr., $\gamma = 10$; $\delta = 0.5$)*

Bias corrected confidence intervals for R				
	1−α			
n	***0.9***	***0.95***	***0.975***	***0.99***
30	0.8464	0.9034	0.9350	0.9590
60	0.8662	0.9216	0.9536	0.9744
120	0.8820	0.9304	0.9604	0.9806
240	**0.8914**	0.9388	0.9644	0.9814
480	**0.8988**	**0.9472**	**0.9718**	**0.9874**
960	**0.8976**	**0.9476**	**0.9730**	**0.9874**
1920	**0.8968**	**0.9480**	**0.9730**	**0.9892**
3840	**0.9048**	**0.9544**	**0.9786**	**0.9898**

TABLE 8. - *Estimated coverage probabilities (Lognormal distribution)*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	***1-α***	***0.99***	***0.95***	***0.975***	***0.99***
					Lognormal distribution $\gamma = 10, \delta = 0.35829$				
30	0.8638	0.9104	0.9418	0.9650		0.8810	0.9302	0.9570	0.9752
60	0.8774	0.9298	0.9568	0.9738		**0.8910**	**0.9424**	0.9652	0.9848
120	**0.8882**	0.9380	0.9668	0.9842		**0.9032**	**0.9502**	**0.9734**	**0.9900**
240	**0.8902**	**0.9422**	**0.9706**	0.9864		**0.8928**	**0.9468**	**0.9702**	**0.9880**
480	**0.8966**	**0.9486**	**0.9728**	**0.9902**		**0.8960**	**0.9502**	**0.9760**	**0.9908**
960	**0.8988**	**0.9484**	**0.9754**	**0.9898**		**0.8986**	**0.9508**	**0.9750**	**0.9906**
1920	**0.9018**	**0.9480**	**0.9710**	**0.9870**		**0.9024**	**0.9526**	**0.9752**	**0.9902**
3840	**0.8978**	**0.9476**	**0.9714**	**0.9864**		**0.9038**	**0.9528**	**0.9730**	**0.9886**
					Lognormal distribution $\gamma = 10, \ \delta = 0.74161$				
30	0.8236	0.8858	0.9192	0.9498		0.8530	0.9056	0.9314	0.9530
60	0.8510	0.9068	0.9380	0.9604		0.8710	0.9206	0.9500	0.9676
120	0.8730	0.9244	0.9558	0.9744		**0.8878**	**0.9400**	0.9634	0.9814
240	0.8790	0.9322	0.9616	0.9818		**0.9006**	**0.9468**	0.9686	0.9852
480	**0.8960**	**0.9440**	0.9684	0.9848		**0.8970**	**0.9424**	0.9678	0.9846
960	**0.8952**	**0.9402**	0.9662	0.9854		**0.8998**	**0.9448**	**0.9702**	0.9858
1920	**0.8950**	**0.9466**	**0.9720**	**0.9886**		**0.8988**	**0.9476**	**0.9734**	**0.9892**
3840	**0.8986**	**0.9518**	**0.9772**	**0.9912**		**0.8960**	**0.9498**	**0.9732**	**0.9900**

4.4 *Pareto distribution*

It is well known that the Pareto family provides an accurate description of many income distributions, especially of the right tail. The density function is given by

$$f(x) = \theta x_0 x^{-(\theta+1)} \qquad \text{for} \quad x > x_0 \qquad \text{where} \quad x_0 > 0 \quad \text{and} \quad \theta > 0.$$

Let us notice that only moments of order less than θ exist. Since the theory of asymptotic confidence intervals for R relies on the existence of the second moment, only cases with $\theta > 2$ will be considered. The parameter x_0 is the greatest lower bound of the support and will be fixed at 1.

Table 9 displays the simulation results. The minimum sample size drastically depends on the value of θ: while for $\theta = 5$ sample sizes as large as 1920 ensure accurate confidence intervals at all nominal confidence levels for both R and the mean, for $\theta = 2.5$ sample sizes as large as 3840 still seem inadequate for confidence intervals for Gini's index.

TABLE 9. - *Estimated coverage probabilities (Pareto distribution)*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	*1−α*	***0.99***	***0.95***	***0.975***	***0.99***
					Pareto distribution, $\theta = 5$				
30	0.7810	0.8362	0.8664	0.8968		0.8524	0.9008	0.9246	0.9442
60	0.8178	0.8746	0.9036	0.9306		0.8592	0.9144	0.9430	0.9636
120	0.8522	0.8974	0.9268	0.9516		0.8800	0.9300	0.9530	0.9740
240	0.8646	0.9158	0.9412	0.9672		**0.8864**	0.9320	0.9580	0.9774
480	**0.8828**	0.9324	0.9572	0.9772		**0.8908**	**0.9422**	0.9670	0.9862
960	**0.8854**	0.9374	0.9644	0.9806		**0.8944**	**0.9466**	**0.9716**	0.9856
1920	**0.8896**	**0.9444**	**0.9710**	**0.9882**		**0.8998**	**0.9446**	**0.9712**	**0.9882**
3840	**0.9048**	**0.9460**	**0.9720**	**0.9876**		**0.9024**	**0.9492**	**0.9742**	**0.9858**
					Pareto distribution, $\theta = 3$				
30	0.7206	0.7738	0.8138	0.8484		0.8030	0.8490	0.8794	0.9080
60	0.7696	0.8256	0.8588	0.8944		0.8340	0.8856	0.9152	0.9402
120	0.8172	0.8722	0.9038	0.9292		0.8610	0.9108	0.9354	0.9562
240	0.8426	0.8920	0.9202	0.9486		0.8720	0.9206	0.9488	0.9712
480	0.8638	0.9154	0.9428	0.9658		0.8810	0.9318	0.9550	0.9776
960	**0.8836**	0.9330	0.9574	0.9738		**0.8904**	**0.9436**	**0.9688**	0.9832
1920	**0.8938**	0.9350	0.9576	0.9776		**0.8942**	**0.9448**	0.9672	0.9830
3840	**0.8928**	**0.9426**	0.9686	0.9832		**0.9048**	**0.9500**	**0.9706**	0.9856
					Pareto distribution, $\theta = 2.5$				
30	0.6986	0.7572	0.7926	0.8254		0.7878	0.8272	0.8598	0.8900
60	0.7478	0.8034	0.8388	0.8744		0.8106	0.8568	0.8908	0.9186
120	0.7950	0.8456	0.8794	0.9120		0.8400	0.8884	0.9186	0.9450
240	0.8064	0.8610	0.8960	0.9294		0.8404	0.8928	0.9276	0.9532
480	0.8428	0.8942	0.9258	0.9508		0.8670	0.9160	0.9434	0.9666
960	0.8594	0.9106	0.9406	0.9642		0.8804	0.9294	0.9570	0.9738
1920	0.8692	0.9208	0.9468	0.9728		**0.8864**	**0.9354**	0.9614	0.9800
3840	0.8794	0.9288	0.9568	0.9756		**0.8924**	**0.9428**	**0.9698**	0.9824

4.5 *Dagum distribution*

A flexible three parameter family for the description of income distributions is given by the Dagum (type I) model. This distribution "enjoys the good features of both the Pareto and the Lognormal models, without having their drawbacks" (Latorre, 1990). In fact, the density function may be unimodal resembling the low incomes part of the Lognormal density function, while the right tail decays slowly as in the Pareto case. Its analytic expression is given by

$$f(x) = \lambda\beta\theta\, x^{-(\theta+1)}\left(1 + \lambda x^{-\theta}\right)^{-(\beta+1)} \quad \text{, where} \quad \lambda, \beta, \theta > 0 \quad \text{and} \quad x > 0.$$

Notice that λ is merely a scale parameter, whereas β is a shape parameter which affects asymmetry. Also in this case the existence of moments is determined by θ (only moments of order less than θ exist). In the simulations the parameters β and θ were chosen to lie in a neighbourhood of their maximum likelihood estimates ($\beta = 1.055$ and $\theta = 3.095$) obtained on the Italian expenses distribution as given by the Banca d'Italia survey on Household Income and Wealth in 2002 (8001 households).

TABLE 10. - *Estimated coverage probabilities (Dagum distribution)*

	Confidence intervals for R					*Confidence intervals for* μ			
n	***0.9***	***0.95***	***0.975***	***0.99***	***1−α***	***0.99***	***0.95***	***0.975***	***0.99***
				Dagum distribution, $\lambda = 1,\ \beta = 1,\ \theta = 5$					
30	0.8410	0.8964	0.9264	0.9516		0.8796	0.9252	0.9574	0.9740
60	0.8684	0.9192	0.9458	0.9674		**0.8876**	0.9370	0.9634	0.9818
120	0.8772	0.9306	0.9584	0.9744		**0.8982**	**0.9440**	**0.9692**	0.9862
240	**0.8870**	**0.9416**	0.9642	0.9788		**0.9016**	**0.9470**	**0.9700**	0.9848
480	**0.8986**	**0.9436**	0.9672	0.9836		**0.8910**	**0.9480**	**0.9718**	**0.9886**
960	**0.8898**	**0.9392**	**0.9696**	0.9852		**0.8942**	**0.9448**	**0.9742**	**0.9912**
1920	**0.8930**	**0.9474**	**0.9738**	**0.9878**		**0.8940**	**0.9488**	**0.9738**	**0.9892**
3840	**0.8972**	**0.9474**	**0.9730**	**0.9876**		**0.8986**	**0.9484**	**0.9738**	**0.9904**
				Dagum distribution, $\lambda = 1,\ \beta = 1,\ \theta = 3$					
30	0.7958	0.8556	0.8936	0.9256		0.8440	0.8926	0.9236	0.9488
60	0.8144	0.8718	0.9126	0.9438		0.8628	0.9090	0.9400	0.9638
120	0.8430	0.8936	0.9266	0.9550		0.8738	0.9168	0.9466	0.9698
240	0.8596	0.9146	0.9386	0.9600		0.8800	0.9308	0.9580	0.9770
480	0.8724	0.9222	0.9496	0.9698		**0.8910**	0.9382	0.9660	0.9840
960	0.8784	0.9284	0.9544	0.9730		**0.8928**	**0.9440**	**0.9706**	0.9848
1920	0.8826	0.9356	0.9640	0.9826		**0.8930**	**0.9484**	**0.9716**	**0.9866**
3840	**0.8896**	0.9390	0.9654	0.9830		**0.8888**	**0.9444**	**0.9720**	**0.9854**
				Dagum distribution, $\lambda = 1,\ \beta = 1,\ \theta = 2.5$					
30	0.7486	0.8084	0.8536	0.8978		0.8188	0.8710	0.8996	0.9258
60	0.7766	0.8406	0.8834	0.9208		0.8426	0.8884	0.9210	0.9488
120	0.8102	0.8722	0.9054	0.9356		0.8576	0.9108	0.9388	0.9628
240	0.8282	0.8894	0.9246	0.9482		0.8700	0.9194	0.9470	0.9664
480	0.8492	0.9028	0.9336	0.9568		0.8756	0.9280	0.9534	0.9756
960	0.8558	0.9132	0.9468	0.9678		**0.8860**	0.9366	0.9576	0.9772
1920	0.8720	0.9226	0.9476	0.9700		**0.8834**	0.9378	0.9642	0.9840
3840	0.8792	0.9356	0.9626	0.9774		**0.8934**	**0.9444**	0.9684	0.9858

The three distributions presented in Table 10 differ only for the value of the parameter θ. Hence one can appreciate how the coverage depends on tail heaviness: all coverage estimates decrease dramatically and go far from acceptable values as θ approaches 2. Table 11 has to be compared with the first part of table 10: the parent distributions differ only for the value of the shape parameter β, assessing how the sole asymmetry affects the coverage. As in the lognormal case, it appears that increasing asymmetry has a negative influence on the coverage probability, and that the impact is more severe on the confidence intervals for Gini's index than on the confidence intervals for the mean.

TABLE 11. - *Estimated coverage probabilities (Dagum distribution)*

	Confidence intervals for R					*Confidence intervals for μ*			
n	***0.9***	***0.95***	***0.975***	***0.99***	***1−α***	***0.99***	***0.95***	***0.975***	***0.99***
				Dagum distribution, $\lambda = 1,\ \beta = 0.5,\ \theta = 5$					
30	0.8598	0.9128	0.9426	0.9646		0.8782	0.9322	0.9586	0.9772
60	0.8822	0.9302	0.9576	0.9740		0.8906	0.9378	0.9646	0.9820
120	0.8778	0.9326	0.9596	0.9788		0.8818	0.9392	0.9676	0.9850
240	**0.8944**	**0.9408**	0.9684	0.9824		**0.8972**	**0.9542**	**0.9762**	**0.9880**
480	**0.8914**	**0.9422**	**0.9702**	**0.9882**		**0.8856**	**0.9406**	**0.9694**	**0.9866**
960	**0.8860**	**0.9408**	**0.9692**	**0.9852**		**0.8988**	**0.9482**	**0.9714**	**0.9876**
1920	**0.8916**	**0.9446**	**0.9734**	**0.9894**		**0.9054**	**0.9496**	**0.9720**	**0.9860**
3840	**0.8976**	**0.9482**	**0.9724**	**0.9894**		**0.9042**	**0.9552**	**0.9750**	**0.9908**
				Dagum distribution, $\lambda = 1,\ \beta = 1.5,\ \theta = 5$					
30	0.8320	0.8882	0.9180	0.9410		0.8752	0.9260	0.9526	0.9740
60	0.8512	0.9034	0.9326	0.9584		**0.8876**	0.9366	0.9624	0.9786
120	0.8822	0.9270	0.9528	0.9718		**0.8964**	**0.9468**	**0.9694**	0.9838
240	0.8812	0.9348	0.9598	0.9784		**0.8932**	**0.9442**	**0.9720**	**0.9870**
480	**0.8856**	0.9378	0.9648	0.9834		**0.8936**	**0.9448**	**0.9728**	**0.9878**
960	**0.8862**	**0.9416**	0.9658	0.9810		**0.9002**	**0.9508**	**0.9744**	**0.9894**
1920	**0.8970**	**0.9444**	**0.9712**	0.9850		**0.9016**	**0.9478**	**0.9746**	**0.9898**
3840	**0.8944**	**0.9428**	**0.9730**	**0.9886**		**0.9042**	**0.9526**	**0.9766**	**0.9896**

5. SIZE OF CONFIDENCE INTERVALS

Figure 1 shows how the mean length of the $N = 5000$ confidence intervals for Gini's index approaches zero as the sample size increases. Since the length of the confidence intervals is proportional to the quantiles of the standard normal distribution, only the mean length of confidence intervals with nominal coverage 0.95 is depicted.

The main result is that the mean length of the confidence intervals is smaller than 0.05 if the sample size is about 2000.

As expected, the confidence interval tends to be larger if the parent distribution has a heavier tail (see Figure 1, right part). Trying to understand the effect of asymmetry, evidence from the log-normal distribution suggests that the mean length of the confidence intervals increases if the parent distribution becomes more asymmetric. This result is consistent with what happens in the Dagum case if asymmetry is augmented by

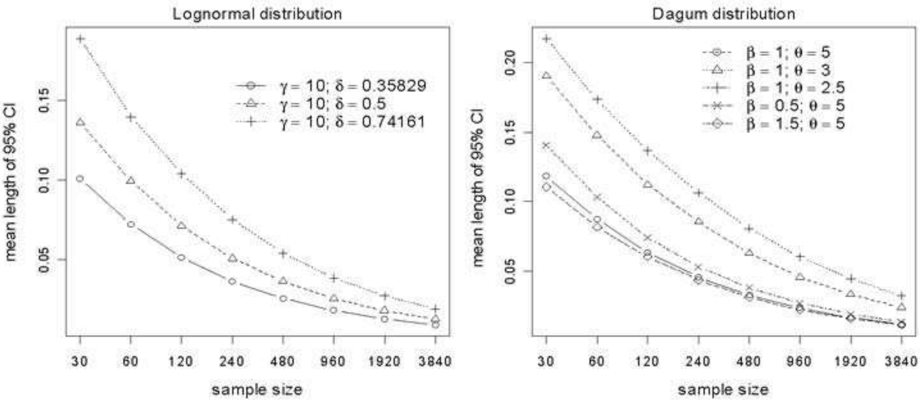

FIGURE 1. - *Mean length of 95% confidence intervals*

decreasing the value of θ (holding β fixed). On the other hand, when asymmetry is augmented by increasing the value of β, variability (measured by R or the coefficient of variation) decreases. The outcome is that the variance of the asymptotic distribution of R decreases, so that the confidence intervals tend to be slightly shorter.

6. LEFT AND RIGHT TAILS OF THE SAMPLE GINI INDEX DISTRIBUTION

The purpose of this section is to investigate left and right rejection probabilities, as evaluated from simulations. For each confidence interval which fails to contain the true value R, it was checked whether R lies on its left (right) side. The proportion of intervals satisfying this condition provides an estimate of the right (left) rejection probability (denoted by RRP, and, respectively, LRP). This terminology is taken from the theory of statistical hypothesis testing: the true value of R would be rejected in favour of a larger one, if it lied on the left of the confidence interval. Therefore, the RRP describes the right tail of the sample distribution, whereas the LRP refers to the left tail. Figure 2 exhibits the behaviour of the estimates of the LRP and RRP in case of sampling from Dagum distributions. The nominal confidence level is fixed at 0.95, so that the estimates should be close to 0.025 when convergence in the tails of the sample distribution has occurred. For nominal confidence levels different from 0.95 similar patterns are observed.

The LRP decreases as sample size increases, reaching values close to 0.025 if the underlying Dagum distribution has finite fourth moment. In the other cases, samples of size up to 3840 appear to be too small for convergence.

The right tail of the sample distribution behaves quite differently. Figure 2 (right part) suggests that the RRP approaches its asymptotic value 0.025 from below. As opposed to the LRP, the estimated RRP stays within two percentage points from its asymptotic value (note that the scales on the two ordinate axes in Figure 2 are different). The irregular behavior of the RRP may deserve further research.

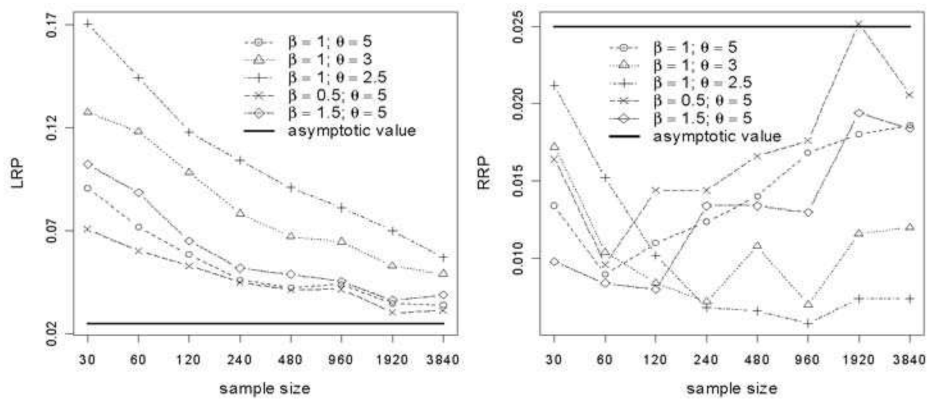

FIGURE 2. - *Estimated LRP and RRP sampling from Dagum distribution (nominal confidence level 0.95)*

In any case, the results presented in this section show that the sample distribution has a heavy left tail, which approaches the asymptotic normal distribution only if the sample size is very large, whereas its right tail is close to the asymptotic distribution.

7. CONCLUSIONS

In this paper minimum sample sizes for asymptotic confidence intervals for Gini's inequality ratio are determined. As expected, the threshold is affected by some characteristics of the underlying distribution, as heavy tails and asymmetry. The minimum sample sizes tends to be larger for high nominal confidence levels, reflecting the fact that convergence in the tails of the sample distribution is more awkward. As long as simulations were done, the coverage of confidence intervals for R was compared to the corresponding coverage of confidence intervals for the mean as a "milestone". Although the minimum sample sizes for the confidence intervals for R are usually larger than those for the confidence intervals for the mean, the estimated coverage probabilities tend to be close to each other whenever the mean confidence interval performs well. Simulations on an underlying rectangular distribution suggest samples of size 60 suffice to ensure satisfactory coverage of the confidence intervals, both for R and for the mean. Further, considering asymmetric and/or heavy tailed distribution families, the coverage probabilities decrease and the minimum sample sizes increase. For Pareto and Dagum distributions with moments up to order five, the minimum sample sizes vary between 240 and 1920 for the 95% confidence interval. If only moments of order less than 2.5 exist, none of the considered sample sizes suffices.

A final remark concerning income studies can be made: the simulation results show that the approximations provided by the asymptotic theory are generally accurate. Failure occurs only in extreme cases, when the underlying distribution is very heavy tailed and/or asymmetric. However, even in these cases the coverage of the asymptotic confidence intervals is fairly good with samples of size 3840. This insight seems to be of interest from an operational point of view, since samples of several thousands statistical units are customary in income studies.

ACKNOWLEDGEMENT

This research was partially supported by F.A.R. 2005 - Università degli studi Milano Bicocca.

RIASSUNTO

Negli ultimi anni si è osservato un crescente interesse verso l'inferenza per misure di concentrazione. Infatti, anche se le analisi sul reddito si basano spesso su grandi campioni, così che la precisione nella stima non sembrerebbe un problema, in molti studi empirici è stato osservato uno scarto quadratico medio elevato (Maasoumi, 1997). In questo lavoro si analizza il comportamento degli intervalli di confidenza asintotici per l'indice di concentrazione di Gini. In particolare, si presentano i risultati di un vasto studio simulativo volto ad indicare le ampiezze campionarie minime che assicurano una data precisione nella copertura degli intervalli di confidenza. Un'ampia gamma di distribuzioni è stata considerata, con particolare attenzione per modelli adeguati a descrivere i redditi. Come ci si attendeva, le soglie campionarie sono influenzate da alcune caratteristiche della distribuzione da cui derivano i campioni, in particolare dalla pesantezza delle code e dall'asimmetria. In molti casi, tuttavia, si osserva che tali ampiezze campionarie sono inferiori a quelle generalmente in uso nelle scienze sociali.

REFERENCES

Andres R., Calonge S. (2005). Inference on Income Inequality and Tax Progressivity Indices: U-Statistics and Bootstrap Methods. *ECINEQ* w.p. 2005 – 09.

Banca d'Italia (2004). I bilanci delle famiglie italiane nel 2002. Supplement to *Bollettino Statistico*, XIV, n. 1-2.

Barret G.F., Pendakur K. (1995). The asymptotic distribution of the generalized Gini index of inequality. *Candian Journal of Economics*, **28**, 1042-1055.

Dall'Aglio G. (1965). Comportamento asintotico delle stime della differenza media e del rapporto di concentrazione di Gini. *Metron*, 24, 379-414.

Dagum C. (1977). A New Model of Personal Income Distribution: Specification and Estimation. *Economie Appliquée*, **26**, 1, 843-876.

Giorgi G.M., Palmitesta P., Provasi C. (2006). Asymptotic and bootstrap inference for the generalized Gini indices. *Metron*, **64**, 1, 107-124.

Giorgi G.M., Provasi C. (1995). An experimental analysis to verify some properties of inequality indices of the Gini family. In Barbini P., Dagum C., Lemmi A., Provasi C. (eds.), *Quantitative Methods for Applied Sciences*, Nuova Immagine, Siena, 254-285.

Greselin F. (1997). Valutazione approssimata della distorsione dello stimatore $\hat{R}$ nella distribuzione di Pareto *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali,* **19**, 3, 269-279.

Greselin F., Maffenini W. (2007). On sample sizes for confidence intervals for Gini's mean difference: a new approach for their determination. *Statistica & Applicazioni*, **5**, 1, 103-122.

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

Latorre G. (1990). Asymptotic Distributions of Indices of Concentration: Empirical Verification and Application, *Studies in Contemporary Economics. Income and Wealth Distribution, Inequality and Poverty*, C. Dagum, M. Zenga (Eds), Springer-Verlag, Berlin.

Lee A.J. (1990). U-Statistics, Theory and Practice. Marcel Dekker, Inc. New York and Basel.

Maasoumi E. (1997). *Empirical analysis of inequality and welfare.* Handbook of Applied Microeconomics. P.Schmidt and H.Pesaran (eds).

Mills J.A., Zvandakily S. (1997). Statistical Inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics*, **12**, 133-150.

Moran P. (2005). Bootstrapping the LIS: statistical inference with the Gini index and patterns of inequality in the Global North. *Luxembourg Income Study Working Paper*, 378.

Palmitesta P., Provasi C., Spera C. (1999). Approximated distributions of Sampling Inequality Indices. *Computational Economics*, **13**, 211-226.

Palmitesta P., Provasi C., Spera C. (2000). Confidence Interval Estimation for Inequality Indices of the Gini Family. *Computational Economics*, **16**, 137-147.

Polisicchio M., Zenga M. (1995). La covarianza fra la media campionaria e la differenza media nel campionamento da popolazioni finite. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali,* **17**, 4, 23-30.

Polisicchio M. (1997). Stimatore corretto della covarianza fra media e differenza media campionaria. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico Sociali.*

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sendler W. (1979). On Statistical Inference in Concentration Measurement. *Metrika*, **26**, 109-122.

Xu K. (2000). Inference for Generalized Gini indices using the iterated bootstrap method. *Journal of Business & Economic Statistics,* **18**, 2, 223-227.

Zenga M., Polisicchio M., Greselin F. (2004). The variance of Gini's mean difference and its estimators. *Statistica*, **3**, 455-475.

Zitikis R. (2002). Large sample estimation of a family of economic inequality indices. *Pakistan Journal of Statistics* (special Issue in Honour of Dr. S. Ejaz Ahmad), 18, 225-248.