*Statistica, anno LXI, n. 1, 2001*

# DISTRIBUTION-FREE ESTIMATION OF THE GINI INEQUALITY INDEX: THE KERNEL METHOD APPROACH

P.L. Conti, G.M. Giorgi

## 1. INTRODUCTION AND PROBLEM STATEMENT

The measurement of income inequality allows the quantification of its magnitude as well as of the time and/or spatial evolution of the inequality itself. It may be also used to evaluate the efficiency of the economic and financial policy implemented by the governments and, in particular, to analyze the effect of taxation on the redistribution of resources. A large amount of literature on these and related problems is available; see, for instance, the monographs by Piesch (1975), Kakwani (1980), Nygard and Sandström (1981), and Silber (ed.) (1999). From all these references, it is apparent that there are several different inequality measures. Each of them possesses different properties, leading to possible advantages and disadvantages in their use. Therefore, the choice of a specific inequality measure must be based on both its features and the main aspects which characterize the phenomenon under study. The most well-known and most frequently used measure of inequality is the Gini inequality index (in the sequel denoted by $R$), also called Gini concentration ratio. Its fortune and topical interest (Giorgi, 1990, 1993), more than 85 years after its appearance in literature (Gini, 1914), are due to its intuitive appeal and to its close link to the Lorenz (1905) curve, as well as to its extensions and interpretations from various points of view made by several scholars over the last 30 years (Giorgi, 1992).

For a long time, inequality measures in general and the Gini index in particular, have been used from a descriptive point of view. However, data available from statistical agencies frequently come from sample surveys; inequality indices turn out to be computed on the basis of sample data. Therefore, it is necessary to use them not only as descriptive tools, but also as tools for formal statistical inference. The approach to statistical inference can be either nonparamentric (distribution-free) or parametric. A comprehensive survey of the main results in the estimation of $R$ according to these two approaches is in Giorgi (1999). In the latter case the inequality measures are expressed as functions of the unknown parameter of the underlying population distribution function. Estimates of such parameters naturally lead to estimates of inequality measures. More precisely, it is usually assumed that the data come from some heavy tailed distribution with positive asymmetry (e.g.

lognormal, Burr, Weibull; see also Dagum, 1977; Sing and Maddala, 1976). However, the underlying distribution is unknown in practice, and the assumed model may not be suitable for fitting the data or it may not be generally true. In fact, virtually all parametric models commonly used are unimodal. On the other hand, empirical evidence shows that the income distribution is frequently bimodal, or even multimodal; see Park and Marron (1990), where the income distribution in the United States in the Seventy's and Eighty's is analyzed. To make clear the kind of error due to the use of a "wrong" model, Park and Marron (1990) showed that if, for instance, the underlying population density data is assumed lognormal and one wants to study its structure over a time period, it seems that it is unimodal and does not change over the years. On the other hand, if a distribution-free approach based on a kernel estimate of the income density function is used, then completely different conclusions are drawn. In fact, data show that the population density is at least bimodal, and that its structure significantly changes over time. The population distribution can be actually considered as a mixture of a poor subpopulation and an average income receivers population. In particular, in the case studied by Park and Marron (1990) there was an increase in the poor over the years.

Because of the good flexibility of the distribution-free approach and the ever increasing use of the Gini index $R$ not only for measuring inequality but also poverty (see, for example, Foster and Sen, 1997), inferential problems for $R$ should be studied more deeply, following a distribution-free approach. In fact, the problems of estimating the income distribution and related inequality measures are strictly related. As a first step, one could be interested in estimating the underlying population distribution, in order to get at least some qualitative ideas about its characteristics (unimodality vs. multimodality, or other). As stressed by Park and Marron (1990), the kernel method proves to be a "very useful tools for exploring the distribution structure of unknown populations". As a second step, the density estimate obtained at the first step could be used to study the inequality in income distribution, and in particular to produce an estimate of the Gini inequality index.

According to the ideas exposed above, in this paper we study an estimator of the Gini inequality index $R$ obtained *via* a preliminary estimate of the population density based on the kernel approach. We mainly concentrate on its asymptotic properties, that provide useful large sample approximations. In particular, (strong) consistency and asymptotic normality are proved in Propositions 1 and 2, respectively. Such results are used to obtain confidence intervals for $R$ (Proposition 5).

## 2. ESTIMATION OF GINI'S INDEX: BASICS AND CONSISTENCY RESULTS

Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a population $X$ with density function $f(\cdot)$. Furthermore, let $\hat{f}_b(\cdot)$ be an estimate of $f(\cdot)$ based on the kernel method:

$$\hat{f}_b(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right) \qquad x \in \mathbb{R} \tag{1}$$

where $K(\cdot)$ is the kernel function, and $h > 0$ is the bandwidth. As usual, $K(\cdot)$ is assumed to be a continuous density function, symmetric with respect to (w.r.t.) 0.

From the estimate (1) it is fairly natural to draw the following estimator of the Gini's index:

$$\hat{R} = \frac{\hat{\Delta}}{2\hat{\mu}} \tag{2}$$

where

$$\hat{\Delta} = \int_{\mathbb{R}^2} |x - y| \hat{f}_h(x) \hat{f}_h(y) dx\, dy \tag{3}$$

$$\hat{\mu} = \int_{\mathbb{R}} x \hat{f}_h(x) dx = \frac{1}{n} \sum_{i=1}^{n} X_i$$

The first result we obtain is the strong consistency of the estimator (2).

*Proposition 1.* Suppose that $h$ tends to 0 as the sample size goes to infinity. Then

$$\hat{R} \xrightarrow{a.s.} R \qquad \text{as } n \to \infty.$$

*Proof.* Since $\hat{\mu}$ converges a.s. to $\mu$ by the strong law of large numbers, it is enough to prove that

$$\hat{\Delta} \xrightarrow{a.s.} \Delta \qquad \text{as } n \to \infty. \tag{4}$$

Let $\Delta_K$ be equal to

$$\Delta_K = \int_{\mathbb{R}^2} |x - y| K(x) K(y) dx\, dy$$

By a little algebra, it is seen that

$$\hat{\Delta} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \int_{\mathbb{R}^2} |x-y| \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \frac{1}{h} K\left(\frac{y - X_j}{h}\right) dx\, dy \right\}$$

$$= \frac{h}{n^2} \sum_{i=1}^{n} \int_{\mathbb{R}^2} |u - v| K(u) K(v) du\, dv + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} |h(u - v) + (X_i - X_j)| K(u) K(v) du\, dv$$

$$= \frac{h}{n} \Delta_K + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} |h(u - v) + (X_i - X_j)| K(u) K(v) du\, dv$$

Consider now the $V$-statistic of degree 2 (see Serfling, 1980, pp. 174-175)

$$V_\Delta = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |X_i - X_j|$$

Using the inequality $||a| - |b|| \leq |a - b|$, it is not difficult to obtain the following chain of inequalities

$$|\hat{\Delta} - V_{\Delta}| \leq \frac{h}{n} \Delta_K$$

$$+ \left| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \int_{\mathbb{R}^2} \{ |h(u - v) + (X_i - X_j)| - |X_i - X_j| \} K(u) K(v) du\, dv \right|$$

$$\leq \frac{h}{n} \Delta_K + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} \int_{\mathbb{R}^2} |h(u - v)| K(u) K(v) du\, dv = h\Delta_K$$

Taking into account that, from the strong consistency of $V$-statistics (see, for instance, Serfling, 1980, p. 174 and p. 206)

$$V_{\Delta} \xrightarrow{a.s.} \Delta \qquad \text{as } n \to \infty$$

we finally obtain

$$|\hat{\Delta} - \Delta| \leq |\hat{\Delta} - V_{\Delta}| + |V_{\Delta} - \Delta| \xrightarrow{a.s.} 0$$

with probability 1, as $n$ goes to infinity, from which (4) follows.

*Remark 1.* Proposition 1 states that the estimator (2) is (strongly) consistent under the only assumption that $h$ tends to zero as the sample size goes to infinity. At a first glance, this result could appear far from intuition, since the condition $h \to 0$ as $n \to \infty$ does not ensure that the kernel estimator (1) is consistent. In fact, $\hat{f}_h(\cdot)$ is consistent iff both the conditions $h \to 0$ and $nh \to \infty$ are fulfilled (Parzen, 1962). In other words, (2) can be consistent even when (1) is not.

## 3. ASYMPTOTIC DISTRIBUTION AND CONFIDENCE INTERVALS

The main goal of the present section is to study confidence intervals for the Gini's index $R$. Now, unless to make specific parametric assumptions on the population, it is virtually impossible to derive the exact sampling distribution of the estimator (2). For this reason, we resort to the asymptotic distribution, as the sample size goes to infinity. The following proposition plays a key role in all subsequent developments.

*Proposition 2.* Suppose that $h$ tends to zero as $n$ goes to infinity. Then, the following result holds true:

$$\sqrt{n}(\hat{\Delta} - \mathrm{E}[\hat{\Delta}]) \xrightarrow{d} N(0, 4\sigma_1^2) \tag{5}$$

as $n$ goes to infinity, where

$$\sigma_1^2 = V(\mathrm{E}[|X_1 - X_2| \,|\, X_1])$$

*Proof.* Let $W_h(X_i, X_j)$ be equal to

$$W_h(X_i, X_j) = \frac{1}{h^2} \int_{\mathbb{R}^2} |x - y| K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - X_j}{h}\right) dx\, dy\,, \qquad j \neq i$$

Then, it is easy to see that the equality

$$\hat{\Delta} = U_h + S_h \tag{6}$$

holds, where

$$U_h = \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j>i} W_h(X_i, X_j) + S_h$$

and

$$S_h = \frac{h}{n} \Delta_K - \frac{2}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j>i} W_h(X_i, X_j)$$

Now, it is not difficult to see that $S_h = O_p(n^{-1})$. Furthermore, from Fubini's theorem, we can also write

$$\begin{aligned} \mathrm{E}[W_h(X_i, X_j)] &= \mathrm{E}\left[\int_{\mathbb{R}^2} |x - y| K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - X_j}{h}\right) dx\, dy\right] \\ &= \mathrm{E}\left[\int_{\mathbb{R}^2} |h(u - v) + (X_i - X_j)| K(u) K(v)\, du\, dv\right] \\ &= \int_{\mathbb{R}^2} \mathrm{E}[|h(u - v) + (X_i - X_j)| K(u) K(v)\, du\, dv] \end{aligned}$$

from which $\mathrm{E}[S_h] = O(n^{-1})$ follows. Hence, we have essentially proved that

$$\hat{\Delta} = U_h + O_p(n^{-1}); \qquad \mathrm{E}[\hat{\Delta}] = \mathrm{E}[U_h] + O(n^{-1}) \tag{7}$$

From relationships (7) it follows that $\sqrt{n}(\hat{\Delta} - \mathrm{E}[\hat{\Delta}])$ possesses the same asymptotic distribution as $\sqrt{n}(U_h - \mathrm{E}[U_h])$, provided that it exists.

The statistic $U_h$ is a U-statistic of degree 2 with kernel depending on $h$, and hence on the sample size $n$. In can be studied in the same way as in the standard case where the kernel does not vary with $n$ (see, for instance, Lee, 1990, p. 12).

Let $g_h(X_i)$ be equal to

$$\begin{aligned} g_h(X_i) &= W_h(X_i, X_j) - \mathrm{E}[W_h(X_i, X_j)] \\ &= W_h(X_i, X_j) - \mathrm{E}[U_h] \end{aligned}$$

and

$$G_h = \frac{2}{n} \sum_{i=1}^{n} g_h(X_i)$$

Furthermore, define the variances

$$\sigma_{1h}^2 = V[g_h(X_1)]; \qquad \sigma_{2h}^2 = V[W_h(X_i, X_j)]$$

Now, it is easy to see that

$$V[U_h] = \binom{n}{2}^{-1} \{2(n-2)\sigma_{1h}^2 + \sigma_{2h}^2\}$$

$$V[G_h] = \frac{4}{n}\sigma_{1h}^2 ; \qquad Cov[U_h, G_h] = \frac{4}{n}\sigma_{1h}^2$$

from which the relationship

$$\mathrm{E}[(U_h - G_h)^2] = O(n^{-2}) \tag{8}$$

follows, uniformly for $0 \leq h \leq h_0$, whatever $h_0 > 0$ may be. Relationship (8) implies that

$$\sqrt{n}(U_h - G_h) = o_p(1)$$

On the other hand, the central limit theorem for triangular arrays allows us to write

$$\sqrt{n}G_h \xrightarrow{d} N(0, 4\sigma_1^2)$$

as $n$ goes to infinity, from which (5) follows.

*Remark 2.* Proposition 2 is proved under the only condition that $h$ tends to zero as $n$ goes to infinity. It is important to notice that the condition $h \to 0$ does not ensure that $\sqrt{n}(\mathrm{E}[\hat{\Delta}] - \Delta)$ tends to zero as $n$ increases. Hence, the asymptotic normality of $\sqrt{n}(\hat{\Delta} - \delta)$ does not follows from Proposition 2, unless further condition on $h$ are considered. This problem is dealt with in Proposition 3.

*Proposition 3.* If $h = o(n^{-1/2})$, then

$$\sqrt{n}(\hat{\Delta} - \delta) \xrightarrow{d} N(0, 4\sigma_1^2) \tag{9}$$

as the sample size $n$ tends to infinity.

*Proof.* Using the inequality $||a| - |b|| \leq |a - b|$ we first obtain

$$|\mathrm{E}|h(u - v) + (X_i - X_j)| - \mathrm{E}|X_i - X_j|| \leq h|u - v|$$

Taking into account that $\mathrm{E}|X_i - X_j| = \Delta$, we then have

$$\mathrm{E}[W_h(X_i, X_j)] = \Delta + O(h) \tag{10}$$

The conclusion (9) then follows from (10).

We are now in a position to obtain the main result of the present section, *i.e.* the asymptotic distribution of the estimator (2).

*Proposition 4.* Let us denote by $\sigma^2$ the population variance, which is assumed to exist. Then, the following two statements hold true.

(I) If the bandwidth $h$ tends to zero as the sample size goes to infinity, then:

$$\sqrt{n}\left(\hat{R} - \frac{\mathrm{E}[\hat{\Delta}]}{\mu}\right) \xrightarrow{d} N(0, \sigma_R^2) \qquad \text{as } n \to \infty \tag{11}$$

where

$$\sigma_R^2 = \frac{1}{4}\left\{4\frac{\sigma_1^2}{\mu^2} + \frac{\Delta^2}{\mu^4} - 2\frac{\Delta}{\mu^3}(\mathrm{E}[X_1 | X_1 - X_2|] - \mu\Delta)\right\} \tag{12}$$

(II) Under the additional assumption $h = o(n^{-1/2})$, we have further

$$\sqrt{n}\,(\hat{R} - R) \xrightarrow{d} N(0, \sigma_R^2) \qquad \text{as } n \to \infty \tag{13}$$

*Proof.* (I) Using the *delta*-method, it is seen that the large sample distribution of $\sqrt{n}\,(\hat{R} - \mathrm{E}[\hat{\delta}]/\mu)$ coincides with the large sample distribution of

$$\frac{1}{2}\left\{\frac{1}{\hat{\mu}}\sqrt{n}\,(\hat{\Delta} - \mathrm{E}[\hat{\Delta}]) - \frac{\mathrm{E}[\hat{\Delta}]}{\mu^2}\sqrt{n}\,(\hat{\mu} - \mu)\right\}$$

From Proposition 2, conclusion (11) easily follows.

(II) To prove (13), it is enough to take into account that the assumption $h = o(n^{-1/2})$ implies that $\sqrt{n}(\mathrm{E}[\hat{\Delta}] - \Delta)$ tends to zero as $n$ goes to infinity.

Statement (II) of Proposition 4 is useful to construct large sample confidence intervals for $R$. To this purpose, we have to estimate the asymptotic variance (12).

First of all, it is immediate to see that

$$\sigma_h^2 = \int_{\mathbb{R}} (x - \hat{\mu})^2 \hat{f}_h(x)\,dx$$

and

$$\hat{M} = \int_{\mathbb{R}^2} x\,|x - y|\,\hat{f}_h(x)\,\hat{f}_h(y)\,dx\,dy$$

are (strongly) consistent estimators of $\sigma^2$ and $\mathrm{E}[X_1 | X_1 - X_2|]$, respectively. Furthermore, a little re-elaboration of P.K. Sen's technique (1960) (the only difference is that we have a variable-kernel $U$-statistic) it is not difficult to see that the estimator:

$$\hat{\sigma}_{1,h}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - U_h)^2$$

where

$$u_i = \frac{1}{n-1}\sum_{j \neq i} W_h(X_i, X_j) \qquad i = 1, \ldots, n$$

is a (weakly) consistent estimator of $\sigma_1^2$. Hence, the quantity

$$\hat{\sigma}_R^2 = \frac{1}{4}\left\{4\frac{\hat{\sigma}_1^2}{\hat{\mu}^2} + \frac{\hat{\Delta}^2}{\hat{\mu}^4} - 2\frac{\hat{\Delta}}{\hat{\mu}^3}(\hat{M} - \hat{\mu}\hat{\Delta})\right\}$$

is a (weakly) consistent estimator of $\sigma_R^2$. From this last result, the following proposition is obtained. Proof is immediate.

*Proposition 5.* Suppose that the bandwidth $h$ tends to zero as the sample size $n$ goes to infinity. Then, we have:

$$\sqrt{n}\,\frac{\hat{R} - \mathrm{E}[\hat{\Delta}]/\mu}{\hat{\sigma}_R} \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty. \tag{14}$$

Furthermore, under the additional assumption $h = o(n^{-1/2})$, we can also write:

$$\sqrt{n}\,\frac{\hat{R} - R}{\hat{\sigma}_R} \xrightarrow{d} N(0,1) \tag{15}$$

as $n$ goes to infinity.

Using a technique similar to that of Proposition 1, it is not difficult to see that $\mathrm{E}[\hat{\Delta}]$ tends to $\Delta$ as $h$ tends to zero, *i.e.* as $n$ goes to infinity. Roughly speaking, this means that we can approximately write $\mathrm{E}[\hat{\Delta}]/\mu \approx R$ as $n$ is large and $h$ is small. Hence, from Proposition 5 it follows that the interval $\hat{R} \mp z_{\alpha/2}\,\hat{\sigma}_R/\sqrt{n}$ is an approximated confidence interval of size $1 - \alpha$, where $z_\gamma$ is the $(1 - \gamma)$-th quantile of the normal standard distribution.

*Remark 3.* Equation (15) seems to provide a more intuitive support to the approximated confidence interval $\hat{R} \mp z_{\alpha/2}\,\hat{\sigma}_R/\sqrt{n}$, under the additional assumption $h = o(n^{-1/2})$. However, the kernel density estimator $\hat{f}_h(\cdot)$ achieves its maximum efficiency if $h \sim n^{-1/5}$, and this suggest not to use a bandwidth $h = o(n^{-1/2})$. At any rate, the approximated confidence interval obtained from Proposition 5 essentially rests on the approximation $\mathrm{E}[\hat{\Delta}] \approx \Delta$ when $h$ is "small", and hence it can be used without any restriction on the speed at which $h$ tends to zero as the sample size $n$ increases.

*Dipartimento di Scienze statistiche* PIER LUIGI CONTI
*Università degli Studi di Bologna*

*Dipartimento di Statistica, Probabilità e Statistiche applicate* GIOVANNI MARIA GIORGI
*Università degli Studi di Roma "La Sapienza"*

## REFERENCES

P.L. CONTI (1995), *Un'introduzione alla stima di funzioni di densità*, "Rivista Italiana di Economia, Demografia e Statistica", 49, pp. 93-113.

C. DAGUM (1977), *A new model of personal income distribution: specification and estimation*, "Economie Appliquée", 30, pp. 413-436.

J. FOSTER, A. SEN (1997), *On economic inequality after a quarter century*, in A. SEN (ed.) "On economic inequality", Clarendon Press, Oxford.

C. GINI (1914), *Sulla misura della concentrazione e della variabilità dei caratteri*, "Atti del R. Istituto Veneto di Scienze, Lettere e Arti", 73, pp. 1203-1248.

G.M. GIORGI (1990), *Bibliographic portrait of the Gini concentration ratio*, "Metron", 48, pp. 183-221.

G.M. GIORGI (1992), *Il Rapporto di concentrazione di Gini: genesi, evoluzione ed una bibliografia commentata*, Editrice Ticci, Siena.

G.M. GIORGI (1993), *A fresh look at the topical interest of the Gini concentration ratio*, "Metron", 51, pp. 83-98.

G.M. GIORGI (1999), *Income inequality measurement: the statistical approach*, in J. SILBER (ed.) "Handbook of income inequality measurement", Kluwer Academic Publishers, Boston, pp. 245-260.

G.M. GIORGI, A. PALLINI (1990), *Inequality indices: theoretical and empirical aspects of their asymptotic behaviour*, "Statistical Papers", 31, pp. 65-76.

N.C. KAKWANI (1980), *Income inequality ad poverty. Methods of estimation and policy applications*, Oxford University Press, New York.

A.J. LEE (1990), *U-Statistics: theory and practice*, Dekker, New York.

M.O. LORENZ (1905), *Methods of measuring the concentration of wealth*, "Quarterly Publications Am. Statist. Ass.", 9, pp. 209-219.

F. NYGARD, A. SANDSTRÖM (1981), *Measuring income inequality*, Almqvist & Wiksell International, Stockholm.

B.Y. PARK, J.S. MARRON (1990), *Comparison of data-driven bandwidth selectors*, "Journal of the American Statistical Association", 85, pp. 66-72.

E. PARZEN (1962), *On estimation of a probability density and mode*, "The Annals of Mathematical Statistics", 33, pp. 1065-1076.

W. PIESCH (1975), *Statistische Konzentrationsmasse*, J.C.B. Mohr (Paul Siebeck), Tübingen.

B. SCHOLKOPF, C.J.C. BURGES, A.J. SMOLA (eds.) (1999), *Advances in Kernel methods*, MIT Press, Cambridge (Mass.).

A. SEN (ed.) (1997), *On economic inequality*, Clarendon Press, Oxford.

P.K. SEN (1960), *On some convergence properties of U-statistics*, "Calcutta Statistical Association Bullettin", 10, pp. 1-18.

R.J. SERFLING (1980), *Approximation theorems of mathematical statistics*, Wiley, New York.

J. SILBER (ed.) (1999), *Handbook of income inequality measurement*, Kluwer Academic Publishers, Boston.

S.K. SING, G.S. MADDALA (1976), *A function for size distribution of incomes*, "Econometrica", 44, pp. 963-970.

## RIASSUNTO

*Stima non parametrica dell'indice di disuguaglianza di Gini: un approccio basato su stimatori nucleo*

In questo lavoro si considera un approccio non parametrico alla stima dell'indice di concentrazione di Gini. L'idea è quella di considerare una stima preliminare, basata sul metodo del nucleo, della funzione di densità della popolazione, e poi nel calcolare il corrispondente indice di concentrazione. Si studiano le proprietà statistiche dello stimatore introdotto, con particolare riferimento al caso di grandi campioni (di solito disponibile dalle rilevazioni di enti statistici). Come sottoprodotto, si ottiene un intervallo di confidenza per l'indice di concentrazione.

SUMMARY

*Distribution-free estimation of the Gini inequality index: the kernel method approach*

In this paper a non-parametric approach to the estimation of the Gini inequality index is introduced. The basic idea consists in taking first a preliminary estimation of the population density function, and then in computing the corresponding inequality index. Statistical properties of the estimator introduced are studied, with particulal reference to the case of large samples (usually available from statistical agencies). As a by-product, approximated confidence intervals are obtained.