GREQAM

Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes Etudes en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail
n°2007-23

RELIABLE INFERENCE FOR THE GINI INDEX

Russel Davidson

November 2007

Reliable Inference for the Gini Index

by

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Québec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

russell.davidson@mcgill.ca

Abstract

Although attention has been given to obtaining reliable standard errors for the plug-in estimator of the Gini index, all standard errors suggested until now are either complicated or quite unreliable. An approximation is derived for the estimator by which it is expressed as a sum of IID random variables. This approximation allows us to develop a reliable standard error that is simple to compute. A simple but effective bias correction is also derived. The quality of inference based on the approximation is checked in a number of simulation experiments, and is found to be very good unless the tail of the underlying distribution is heavy. Bootstrap methods are presented which alleviate this problem except in cases in which the variance is very large or fails to exist. Similar methods can be used to find reliable standard errors of other indices which are not simply linear functionals of the distribution function, such as Sen's poverty index and its modification known as the Sen-Shorrocks-Thon index.

Keywords: Gini index, delta method, asymptotic inference, jackknife, bootstrap

JEL codes: C12, C13, C81, D31, I32

This research was supported by the Canada Research Chair program (Chair in Economics, McGill University) and by grants from the Social Sciences and Humanities Research Council of Canada, and the Fonds Québécois de Recherche sur la Société et la Culture. I thank my colleague Michel Lubrano for useful comments, and for having provoked me to write the paper in the first place.

November 2007

1. Introduction

Some attention has been given recently to the standard error of a Gini index estimated by a plug-in estimator with no distributional assumptions. Quite a number of techniques for computing an asymptotically valid standard error have been proposed, of varying degrees of complexity or computational intensiveness. Sandström, Wretman, and Waldén (1988) discuss estimation of the Gini coefficient with arbitrary probability sampling design, and then propose three ways to compute a standard error. The first is from a complicated analytic formula, the second is based on the jackknife, and the third is discarded as "quite useless".

More recently, Bishop, Formby, and Zheng (1997) have given a discussion of the variance of the Gini index in the context of estimating Sen's index of poverty; their approach is based on U-statistics, as is also that of Xu (2007). Ogwang (2000) provided a method for computing the Gini index by an OLS regression, and discussed how to use this regression to simplify the computation of the jackknife standard error. Then Giles (2004) claimed that the OLS standard error from this regression could be used directly in order to compute the standard error of the Gini index itself. See also the reply by Ogwang (2004).

Subsequently, Modarres and Gastwirth (2006) struck a cautionary note on the use of Giles's approach, showing by simulation that the standard errors it produces are quite inaccurate. They recommended a return to the complex or computationally intensive methods used previously, and, in their replies, Ogwang (2006) and Giles (2006) did not fundamentally disagree with the criticism. More recently still, Bhattacharya (2007) has developed techniques of asymptotic inference for Lorenz curves and the Gini index with stratified and clustered survey data. These techniques are based on sample empirical process theory and the functional delta method, and they lead to a formula for the variance of an estimated Gini index, which is however not at all easy to implement.

This paper shows how to compute an asymptotically correct standard error for an estimated Gini index, based on a reasonably simple formula that is very easy to compute. The proposed standard error is based on the delta method, but makes no use of empirical process theory. The approach also provides a simple and effective bias correction for the estimate of the index. The methods used can be extended to other commonly used indices, including Sen's (1976) poverty index, and the modification of it proposed by Shorrocks (1995), often referred to as the Sen-Shorrocks-Thon (SST) index.

In section 2, we review some well-known properties of the Gini index, and give an expression for the Gini index of a sample. This is then related to the regression proposed by Ogwang (2000). Then, in section 3, an asymptotic approximation for the usual plug-in estimator of the index is derived. This approximation shows that the estimator is asymptotically normal, since it takes the form of a sum of IID random variables. In section 4, inference based on the estimate is investigated. The asymptotic variance is easily found from the approximation, and it is shown how it can easily be estimated from the sample. Bias is studied next, and a simple bias correction proposed.

Section 5 considers the jackknife as an alternative way of doing bias correction and variance estimation. It is found that the jackknife does not give reliable inference. The bootstrap is discussed briefly in section 6. Unlike the jackknife, the bootstrap can yield reasonably reliable inference. Section 7 provides simulation evidence that bears out the main conclusions of the paper, and reveals their limitations when used with heavy-tailed distributions. The empirical study given in Giles (2004) is redone in section 8 so as to make clear how the methods of this paper differ from those used by Giles. In section 9, the methods of the paper are used to find the asymptotic variance of Sen's (1976) poverty index and the SST variant. Section 10 concludes.

2. Properties of the Gini index

The classical definition of the Gini index of inequality is twice the area between the 45°-line and the Lorenz curve. If we denote by F the cumulative distribution function (CDF) of the incomes under study, the Lorenz curve is defined implicitly by the equation

$$L\big(F(x)\big) = \frac{1}{\mu}\int_0^x y\,\mathrm{d}F(y), \tag{1}$$

where $\mu \equiv \int_0^\infty y\,\mathrm{d}F(y)$ is expected income. It is assumed that there are no negative incomes. The function L is increasing and convex, and maps the [0,1] interval into itself. Twice the area between the graph of L and the 45°-line is then

$$G = 1 - 2\int_0^1 L(y)\,\mathrm{d}y. \tag{2}$$

Using the definition (1) in (2), we find that

$$G = 1 - 2\int_0^\infty L\big(F(x)\big)\,\mathrm{d}F(x) = 1 - \frac{2}{\mu}\int_0^\infty\int_0^x y\,\mathrm{d}F(y)\,\mathrm{d}F(x).$$

Then, on interchanging the order of integration and simplifying, we obtain

$$\begin{aligned} G &= 1 - \frac{2}{\mu}\int_0^\infty y\int_y^\infty \mathrm{d}F(x)\,\mathrm{d}F(y) = 1 - \frac{2}{\mu}\int_0^\infty y\big(1-F(y)\big)\,\mathrm{d}F(y) \\ &= \frac{1}{\mu}\int_0^\infty \big(2yF(y) - y\big)\,\mathrm{d}F(y) = \frac{2}{\mu}\int_0^\infty yF(y)\mathrm{d}F(y) - 1. \end{aligned} \tag{3}$$

The last expression above corresponds to a result cited in Modarres and Gastwirth (2004) according to which G is $2/\mu$ times the covariance of Y and $F(Y)$, where Y denotes the random variable "income" of which the CDF is F. There are of course numerous other ways of expressing the index G, but (3) is most convenient for present purposes. See Appendix A for further discussion of this point.

Suppose now that an IID sample of size n is drawn randomly from the population, and let its empirical distribution function (EDF) be denoted as $\hat{F}$. The natural plug-in estimator of G is then $\hat{G}$, defined as

$$\hat{G} = \frac{2}{\hat{\mu}} \int_0^\infty y\hat{F}(y)\,\mathrm{d}\hat{F}(y) - 1. \tag{4}$$

Evaluating $\hat{G}$ using (4) reveals an ambiguity: different answers are obtained if the EDF is defined to be right- or left-continuous. The ambiguity can be resolved by splitting the difference, or by noting that we can write

$$\begin{aligned}\hat{G} &= \frac{1}{\hat{\mu}} \int_0^\infty y\,\mathrm{d}\big(\hat{F}(y)\big)^2 - 1 = \frac{1}{\hat{\mu}} \sum_{i=1}^n y_{(i)} \Big(\Big(\frac{i}{n}\Big)^2 - \Big(\frac{i-1}{n}\Big)^2\Big) - 1 \\ &= \frac{2}{\hat{\mu}n^2} \sum_{i=1}^n y_{(i)}(i - \tfrac{1}{2}) - 1. \end{aligned} \tag{5}$$

Here the $y_{(i)}$, $i = 1, \ldots, n$, are the order statistics. The definition (5) has the advantage over alternative possibilities that, when $y_{(i)} = \hat{\mu}$ for every i, $\hat{G} = 0$.

In order to compute $\hat{G}$ itself, Ogwang (2000) suggested the use of the regression

$$i = \theta + u_i, \qquad i = 1, \ldots, n, \tag{6}$$

estimated by weighted least squares under the assumption that the variance of u_i is proportional to $1/y_{(i)}$. The parameter estimate $\hat{\theta}$ is then

$$\hat{\theta} = \Big(\sum_{i=1}^n y_i\Big)^{-1} \sum_{i=1}^n iy_{(i)}.$$

It is easy to check that $\hat{G}$, as given by (5), is equal to $2\hat{\theta}/n - 1 - 1/n$. Giles (2004) reformulated the weighted regression as

$$i\sqrt{y_{(i)}} = \theta\sqrt{y_{(i)}} + v_i, \qquad i = 1, \ldots, n, \tag{7}$$

now to be estimated by OLS. His proposal was then simply to use the OLS standard error, multiplied by $2/n$, as the standard error of $\hat{G}$. As pointed out by Modarres and Gastwirth (2004), however, the fact that the order statistics are correlated means that the OLS standard error may be unreliable.

3. An asymptotic expression for the Gini index

Standard arguments show that the estimator (4) is consistent under weak regularity conditions. Among these, we require the existence of the second moment of the distribution characterised by F. This is not quite enough, as the class of admissible CDFs F must be further restricted so as to avoid the Bahadur-Savage problem; see Bahadur and Savage (1956). Asymptotic normality calls for a little more regularity, but not a great deal. In this section, we examine the quantity $n^{1/2}(\hat{G}-G)$ that should be asymptotically normal under the required regularity, and derive the variance of its limiting distribution as $n \to \infty$.

Let

$$I \equiv \int_0^\infty yF(y)\,\mathrm{d}F(y) \quad \text{and} \quad \hat{I} \equiv \int_0^\infty y\hat{F}(y)\,\mathrm{d}\hat{F}(y). \tag{8}$$

Notice that the integral defining I exists if we assume that the first moment of F exists, since $F(y)$ is bounded above by 1. Then we have

$$\begin{aligned} n^{1/2}(\hat{G}-G) &= n^{1/2}\Big(\frac{2\hat{I}}{\hat{\mu}} - \frac{2I}{\mu}\Big) = n^{1/2}\frac{2}{\mu\hat{\mu}}(\mu\hat{I} - \hat{\mu}I) \\ &= \frac{2}{\mu\hat{\mu}}\big(\mu n^{1/2}(\hat{I}-I) - In^{1/2}(\hat{\mu}-\mu)\big). \end{aligned} \tag{9}$$

Our assumed regularity ensures that both $n^{1/2}(\hat{\mu}-\mu)$ and $n^{1/2}(\hat{I}-I)$ are of order 1 in probability. To leading order, then, we may approximate (9) by replacing $\mu\hat{\mu}$ in the denominator by μ^2.

Next, we note that

$$n^{1/2}(\hat{\mu}-\mu) = n^{-1/2}\sum_{j=1}^n (y_j - \mu).$$

Clearly this is an asymptotically normal random variable. For $n^{1/2}(\hat{I}-I)$, we calculate as follows.

$$\begin{aligned} n^{1/2}(\hat{I}-I) &= n^{1/2}\Big(\int_0^\infty y\hat{F}(y)\,\mathrm{d}\hat{F}(y) - \int_0^\infty yF(y)\,\mathrm{d}F(y)\Big) \\ &= n^{1/2}\Big(\int_0^\infty yF(y)\,\mathrm{d}(\hat{F}-F)(y) + \int_0^\infty y\big(\hat{F}(y)-F(y)\big)\,\mathrm{d}F(y) \\ &\qquad + \int_0^\infty y\big(\hat{F}(y)-F(y)\big)\,\mathrm{d}(\hat{F}-F)(y)\Big). \end{aligned} \tag{10}$$

The last term above is of order $n^{-1/2}$ as $n \to \infty$, and so will be ignored for the purposes of our asymptotic approximation.

The first term in the rightmost member of (10) is

$$n^{1/2}\int_0^\infty yF(y)\,\mathrm{d}(\hat{F}-F)(y) = n^{-1/2}\sum_{j=1}^n \big(y_jF(y_j) - I\big); \tag{11}$$

note from (8) that $I = \mathrm{E}\big(YF(Y)\big)$. Evidently, this is asymptotically normal, since the terms are IID, the expectation of each term in the sum is 0, and the variance exists. The second term is

$$n^{-1/2}\sum_{j=1}^{n}\Big(\int_0^\infty y\,\mathrm{I}(y_j \le y)\,\mathrm{d}F(y) - I\Big), \tag{12}$$

where $\mathrm{I}(\cdot)$ is an indicator function, equal to 1 if its argument is true, and to 0 if not. Define the deterministic function $m(y) \equiv \int_0^y x\,\mathrm{d}F(x)$. We see that

$$\begin{aligned}\mathrm{E}\big(m(Y)\big) &= \int_0^\infty m(y)\,\mathrm{d}F(y) = \int_0^\infty\int_0^y x\,\mathrm{d}F(x)\,\mathrm{d}F(y)\\ &= \int_0^\infty x\int_x^\infty \mathrm{d}F(y)\,\mathrm{d}F(x) = \int_0^\infty x\big(1-F(x)\big)\,\mathrm{d}F(x)\\ &= \mathrm{E}\big(Y(1-F(Y))\big) = \mu - I.\end{aligned}$$

Consequently,

$$\begin{aligned}\int_0^\infty y\,\mathrm{I}(y_j \le y)\,\mathrm{d}F(y) - I &= \int_{y_j}^\infty y\,\mathrm{d}F(y) - I\\ &= \mu - m(y_j) - I = -\Big(m(y_j) - \mathrm{E}\big(m(Y)\big)\Big).\end{aligned}$$

Thus (12) becomes

$$-n^{-1/2}\sum_{j=1}^{n}\Big(m(y_j) - \mathrm{E}\big(m(Y)\big)\Big), \tag{13}$$

which is again asymptotically normal. It follows that $n^{1/2}(\hat I - I)$ is also asymptotically normal, and, from (10), (11), and (13),

$$\begin{aligned}n^{1/2}(\hat I - I) &= n^{-1/2}\sum_{j=1}^{n}\Big(y_jF(y_j) - m(y_j) - \mathrm{E}\big(YF(Y) - m(Y)\big)\Big)\\ &= n^{-1/2}\sum_{j=1}^{n}\Big(y_jF(y_j) - m(y_j) - (2I - \mu)\Big).\end{aligned} \tag{14}$$

Finally, we obtain from (9) an approximate expression for $n^{1/2}(\hat G - G)$:

$$n^{1/2}(\hat G - G) \approx -\frac{2}{\mu^2}\,I\,n^{1/2}(\hat\mu - \mu) + \frac{2}{\mu}\,n^{1/2}(\hat I - I) \tag{15}$$

This expression can of course be regarded as resulting from the application of the delta method to expression (4). It is useful to express (15) as the sum of contributions from the individual observations, as follows:

$$n^{1/2}(\hat G - G) \approx n^{-1/2}\frac{2}{\mu}\sum_{j=1}^{n}\Big(-\frac{I}{\mu}(y_j - \mu) + y_jF(y_j) - m(y_j) - (2I - \mu)\Big)$$

In this way, $n^{1/2}(\hat{G}-G)$ is expressed approximately as the normalised sum of a set of IID random variables of expectation zero, so that asymptotic normality is an immediate consequence. Since from (3) and (8) we have $G = 2I/\mu - 1$, the variance of the limiting distribution of $n^{1/2}(\hat{G}-G)$ is

$$\frac{1}{n\mu^2}\sum_{j=1}^{n}\mathrm{Var}\Big(-(G+1)y_j + 2\big(y_jF(y_j) - m(y_j)\big)\Big). \tag{16}$$

4. Inference for the Gini index

To estimate the variance (16), one can replace μ by $\hat{\mu}$ and G by $\hat{G}$. But the functions F and m are normally unknown, and so they, too, must be estimated. The value of $F(y_{(i)})$ at the order statistic $y_{(i)}$ is estimated by $\hat{F}(y_{(i)}) = (2i-1)/(2n)$, where we continue to evaluate $\hat{F}$ at its points of discontinuity by the average of the lower and upper limits. Since by definition $m(y) = \mathrm{E}\big(Y\,\mathrm{I}(Y \le y)\big)$, we can estimate $m(y_j)$ by

$$\hat{m}(y_j) = \hat{\mathrm{E}}\big(Y\,\mathrm{I}(Y \le y_j)\big) = \frac{1}{n}\sum_{i=1}^{n} y_i\,\mathrm{I}(y_i \le y_j). \tag{17}$$

If $y_j = y_{(i)}$, then we see that $\hat{m}(y_{(i)}) = (1/n)\sum_{j=1}^{i} y_{(j)}$.

Let $Z_i \equiv -(G+1)y_{(i)} + 2\big(y_{(i)}F(y_{(i)}) - m(y_{(i)})\big)$. Clearly, we can estimate Z_i by

$$\hat{Z}_i \equiv -(\hat{G}+1)y_{(i)} + \frac{2i-1}{n}y_{(i)} - \frac{2}{n}\sum_{j=1}^{i} y_{(j)}. \tag{18}$$

Then $\bar{Z} \equiv n^{-1}\sum_{i=1}^{n}\hat{Z}_i$ is an estimate of $\mathrm{E}(Z_i)$, and $n^{-1}\sum_{i=1}^{n}(\hat{Z}_i - \bar{Z})^2$ is an estimate of $\mathrm{Var}(Z_i)$. Since the sum in (16) can be rewritten as the sum of the variances of the Z_i, $i = 1,\ldots,n$, the variance of $\hat{G}$ can be estimated by

$$\widehat{\mathrm{Var}}(\hat{G}) = \frac{1}{(n\hat{\mu})^2}\sum_{i=1}^{n}(\hat{Z}_i - \bar{Z})^2. \tag{19}$$

Having a reasonable estimate of the variance of $\hat{G}$ is only one part of getting reasonable inference, since $\hat{G}$ can be quite severely biased. First, we note that, since $\mathrm{E}(\hat{\mu}) = \mu$, the expectation of the first term on the right-hand side of (15) vanishes. Therefore we need consider only $\mathrm{E}(\hat{I} - I)$ in order to approximate $\mathrm{E}(\hat{G} - G)$.

Replacing population values by estimates, we see that $\hat{I} = \hat{\mu}(\hat{G}+1)/2$, and, from the expression (5) for $\hat{G}$, it follows that

$$\hat{I} = \frac{1}{n^2}\sum_{i=1}^{n} y_{(i)}(i - \tfrac{1}{2}).$$

In order to compute the expectation of $\hat{I}$, we need the expectations of the order statistics $y_{(i)}$. It is known that, if the order statistics are those of an IID sample of size n drawn from the continuous distribution F with density $f \equiv F'$, then the density of $y_{(i)}$ is

$$f_{(i)}(x) = i\binom{n}{i}\big(F(x)\big)^{i-1}\big(1-F(x)\big)^{n-i}f(x).$$

Thus

$$\mathrm{E}(\hat{I}) = \frac{1}{n^2}\sum_{i=1}^{n} i(i-\tfrac{1}{2})\binom{n}{i}\int_0^\infty x\,\big(F(x)\big)^{i-1}\big(1-F(x)\big)^{n-i}\mathrm{d}F(x). \tag{20}$$

Now it is easy to check that

$$i\binom{n}{i} = n\binom{n-1}{i-1} \quad \text{and} \quad i^2\binom{n}{i} = n(n-1)\binom{n-2}{i-2} + n\binom{n-1}{i-1}. \tag{21}$$

If $i = 1$, the first term on the right-hand side of the second equation above is replaced by 0. We see that

$$\begin{aligned}\sum_{i=1}^{n} i\binom{n}{i}F^{i-1}(1-F)^{n-i} &= n\sum_{i=1}^{n}\binom{n-1}{i-1}F^{i-1}(1-F)^{n-i}\\ &= n\sum_{i=0}^{n-1}\binom{n-1}{i}F^{i}(1-F)^{n-1-i} = n,\end{aligned} \tag{22}$$

where the last step follows from the binomial theorem. Similarly, from (21) along with (22), we have

$$\begin{aligned}\sum_{i=1}^{n} i^2\binom{n}{i}F^{i-1}(1-F)^{n-i} &= n(n-1)\sum_{i=1}^{n}\binom{n-2}{i-2}F^{i-1}(1-F)^{n-i} + n\\ &= n(n-1)\sum_{i=0}^{n-2}\binom{n-2}{i}F^{i+1}(1-F)^{n-2-i} + n\\ &= n(n-1)F + n.\end{aligned} \tag{23}$$

Thus, with (22) and (23), (20) becomes

$$\begin{aligned}\mathrm{E}(\hat{I}) &= \frac{1}{n^2}\int_0^\infty \big(n(n-1)F(x) + n - \tfrac{1}{2}n\big)\,x\,\mathrm{d}F(x)\\ &= \int_0^\infty xF(x)\,\mathrm{d}F(x) - \frac{1}{n}\int_0^\infty x\big(F(x)-\tfrac{1}{2}\big)\mathrm{d}F(x)\\ &= I - \frac{1}{n}(I - \frac{\mu}{2}).\end{aligned}$$

From (15) we can now obtain an approximate expression for the bias of $\hat{G}$:

$$\mathrm{E}(\hat{G} - G) \approx \frac{2}{\mu}\mathrm{E}(\hat{I} - I) = -\frac{1}{n\mu}(2I - \mu) = -G/n. \tag{24}$$

If follows from this that $n\hat{G}/(n-1)$ is a bias-corrected estimator of G. Although still biased, its bias is of order smaller than n^{-1}.

It may be helpful here to summarise the steps needed in the computation of (19) and (24). After computing $\hat{\mu}$ as the sample mean, the steps are as follows.

- Sort the sample in increasing order, so as to obtain the series of order statistics $y_{(i)}$.
- Form the two series $w_i \equiv (2i-1)y_{(i)}/(2n)$ and $v_i \equiv n^{-1}\sum_{j=1}^{i} y_{(j)}$. Then $\hat{I} = \bar{w}$, the mean of the w_i.
- Compute the bias-corrected estimate of the Gini index, $\hat{G} = n(2\hat{I}/\hat{\mu} - 1)/(n-1)$.
- Form the series $\hat{Z}_i = -(\hat{G}+1)y_{(i)} + 2(w_i - v_i)$, and compute the mean $\bar{Z}$. The estimated variance of $\hat{G}$ is the sum of the squares of the $\hat{Z}_i - \bar{Z}$, divided by $(n\hat{\mu})^2$, as in (19). The standard error is the square root of the estimated variance.

It is often of considerable interest to test whether the Gini indices for two populations are the same. If independent samples are drawn from both populations, one can compute the two estimated indices, $\hat{G}_1$ and $\hat{G}_2$ say, along with two standard errors $\hat{\sigma}_{G1}$ and $\hat{\sigma}_{G2}$. A suitable test statistic is then $\tau \equiv (\hat{G}_1 - \hat{G}_2)/\sqrt{\hat{\sigma}^2_{G1} + \hat{\sigma}^2_{G2}}$. If correlated samples are available, the covariance of the two estimated indices should be taken into account. In order to do so, two series, with elements $\hat{Z}_{1i}$ and $\hat{Z}_{2i}$ say, should be formed, using (18), for each sample. Then, after making sure that the elements of the two series are ordered *in the same way*, the covariance of $\hat{G}_1$ and $\hat{G}_2$ is estimated by

$$\widehat{\text{cov}}(\hat{G}_1, \hat{G}_2) = \frac{1}{n^2\hat{\mu}_1\hat{\mu}_2}\sum_{i=1}^{n}(\hat{Z}_{1i} - \bar{Z}_1)(\hat{Z}_{2i} - \bar{Z}_2), \tag{25}$$

where n is the size of each sample, $\hat{\mu}_k$, $k = 1, 2$, are the sample means, and $\bar{Z}_k$, $k = 1, 2$, the means of the $\hat{Z}_{ki}$. The same technique can be used to estimate covariances of a set of more than two estimated Gini indices.

5. The jackknife

Among the various "computationally intensive" suggestions for obtaining a standard error for the Gini index is the jackknife; it is proposed by Modarres and Gastwirth (2006) among others. However, we will see in this section that the jackknife does *not* yield a reliable estimate of the standard error, and further that it is not even appropriate for its usual main function, namely bias correction.

A first remark is called for here. Given the regression (6) proposed by Ogwang (2000) as modified by Giles (2004) to take the form (7) that can be estimated by OLS, implementation of the jackknife is by no means computationally intensive. Consider Giles's regression modified further as follows:

$$\left(\frac{2i-1}{n}-1\right)\sqrt{y_{(i)}} = \theta\sqrt{y_{(i)}} + \text{residual}, \tag{26}$$

where the term "residual" is used to emphasise the fact that this regression is a computational tool, and has no direct statistical interpretation. It is straightforward to check that the OLS estimate $\hat{\theta}$ from this regression is equal to the (biased) estimator $\hat{G}$ given by (5).

If we denote by $\hat{\theta}^{(i)}$ the estimate obtained by leaving out observation i, then the jackknife estimate of G is

$$\hat{G}_J \equiv \hat{\theta} + \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}-\hat{\theta}^{(i)}). \tag{27}$$

The jackknife bias estimator is thus n^{-1} times the negative of

$$b_J \equiv (n-1)\sum_{i=1}^{n}(\hat{\theta}-\hat{\theta}^{(i)}). \tag{28}$$

From the result of the previous section, this should be an estimate of G for the jackknife to correct properly for bias.

For the general linear regression

$$y_i = \boldsymbol{X}_i\boldsymbol{\theta} + \text{residual}, \quad i = 1,\ldots,n,$$

estimated by OLS, with $\boldsymbol{X}_i$ a $1\times k$ vector of regressors and $\boldsymbol{\theta}$ a $k\times 1$ vector of parameters, the vector $\boldsymbol{\theta}^{(i)}$ of OLS estimates found by omitting observation i is related to the full-sample estimate $\hat{\boldsymbol{\theta}}$ by the equation

$$\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{(i)} = \frac{1}{1-h_i}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_i{}^\top\hat{u}_i, \tag{29}$$

where $\boldsymbol{X}$ is the $n\times k$ matrix with i^{th} row $\boldsymbol{X}_i$, $\hat{u}_i$ is the OLS residual for observation i, and $h_i = (\boldsymbol{P}_{\boldsymbol{X}})_{ii}$, the i^{th} diagonal element of the orthogonal projection matrix $\boldsymbol{P}_{\boldsymbol{X}} \equiv \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. See Davidson and MacKinnon (2004), section 2.6, for details.

In order to specialise (29) for use with regression (26), we note that $\boldsymbol{X}$ becomes a vector with typical component $\sqrt{y_{(i)}}$, and $h_i = y_{(i)}/(n\hat{\mu})$. The residual is

$$\hat{u}_i = \sqrt{y_{(i)}}\left(\frac{2i-1}{n}-1-\hat{\theta}\right).$$

Thus, noting that $\hat{\theta} = \hat{G}$, we see that

$$\hat{\theta}-\hat{\theta}^{(i)} = \frac{y_{(i)}\big(2i-1-n(1+\hat{G})\big)}{n(n\hat{\mu}-y_{(i)})}. \tag{30}$$

Since this is trivially easy to compute after computing $\hat{G}$, by running regression (26) or otherwise, implementing the formula (27) is also very simple.

Let us take a closer look at expression (28). From (30), we find that

$$b_J = \frac{n-1}{n^2\hat\mu}\sum_{i=1}^{n}\Bigl(y_{(i)}\bigl(2i-1-n(1+\hat G)\bigr)\bigl(1-\frac{y_{(i)}}{n\hat\mu}\bigr)^{-1}\Bigr). \tag{31}$$

Since $n^{-2}\sum_i y_{(i)}(2i-1) = \hat\mu(1+\hat G)$ (equation (5)), and $n^{-1}\sum_i y_{(i)} = \hat\mu$, it follows that

$$\sum_{i=1}^{n} y_{(i)}\bigl(2i-1-n(1+\hat G)\bigr) = 0.$$

Thus we have

$$b_J = \frac{1}{n^2\hat\mu^2}\sum_{i=1}^{n} y_{(i)}^2\bigl(2i-1-n(1+\hat G)\bigr) + O_p(n^{-1}). \tag{32}$$

Now $n^{-1}\sum_i y_{(i)}^2 = \hat\sigma^2 + \hat\mu^2$, where $\hat\sigma^2$ is the sample variance, while

$$\frac{1}{n^2}\sum_{i=1}^{n} y_{(i)}^2(2i-1) = 2\int_0^\infty y^2\hat F(y)\,\mathrm{d}\hat F(y) \equiv 2\hat e_2,$$

where we define $e_2 = \mathrm{E}\bigl(Y^2F(Y)\bigr)$. Substituting these results into (32), we find that

$$b_J = \frac{2\hat e_2}{\hat\mu^2} - (1+\hat G)\Bigl(1+\frac{\hat\sigma^2}{\hat\mu^2}\Bigr), \tag{33}$$

which is a consistent estimator of the rather complicated functional defined by the same expression without the hats. In general, b_J is not, therefore, a consistent estimator of G,[1] as would be needed if the jackknife estimator $\hat G_J$ were to be unbiased to an order smaller than n^{-1}. It may well be that $\hat G_J$ is *less* biased than $\hat G$, but its bias converges to 0 as $n \to \infty$ no faster. Since the properly bias-corrected estimator $(n+1)\hat G/n$ is even easier to compute than the jackknife estimator $\hat G_J$, there is no need to bother with the latter.

The jackknife estimator of the variance of $\hat G$ is

$$\widehat{\mathrm{Var}}_J(\hat G) = \frac{n-1}{n}\sum_{i=1}^{n}\Bigl(\hat\theta^{(i)} - \frac{1}{n}\sum_{j=1}^{n}\hat\theta^{(j)}\Bigr)^2, \tag{34}$$

with the $\hat\theta^{(i)}$ given by (30). The calculations needed to analyse (34) are similar in spirit to those above for the jackknife estimator itself, but a good deal more complicated, and so we omit them here. They show that it is not a consistent estimator of the asymptotic variance of $\hat G$. This fact also emerges very clearly from some of the simulations reported in section 7.

[1] Exceptionally, b_J is consistent for G if the underlying distribution is the exponential distribution. This is noted here because many of the simulations reported in section 7 use the exponential distribution to generate simulated samples.

6. The bootstrap

Unlike the jackknife, the bootstrap can reasonably be expected to yield fairly reliable inference about the Gini index. Indeed, if used in combination with the asymptotic standard error derived from (19), it should give rise to asymptotic refinements relative to inference based on the variance estimate (19); see Beran (1988).

Specifically, in order to test the hypothesis that the population value of the Gini index is G_0, one first computes the statistic $\tau \equiv (\hat{G} - G_0)/\hat{\sigma}_G$, where here $\hat{G}$ is the almost unbiased estimate $n(2\hat{I}/\hat{\mu} - 1)/(n-1)$, and the standard error $\hat{\sigma}_G$ is the square root of the variance estimate (19). Then one generates B bootstrap samples of size n by resampling with replacement from the observed sample (assumed to be also of size n). For bootstrap sample j, one computes a bootstrap statistic τ_j^*, in exactly the same way as τ was computed from the original data, but with G_0 replaced by $\hat{G}$, in order that the hypothesis tested should be true of the bootstrap data-generating process. The bootstrap P value is then the proportion of the τ_j^* that are more extreme than τ. For a test at significance level α, rejection occurs if the bootstrap P value is less than α. For such a test, it is also desirable to choose B such that $\alpha(B+1)$ is an integer; see, among other references, Davidson and MacKinnon (2000).

Bootstrap confidence intervals can also be based on the empirical distribution of the bootstrap statistics τ_j^*. For an interval at nominal confidence level $1-\alpha$, one estimates the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical distribution, normally as the $\lceil \alpha B/2 \rceil$ and $\lceil (1-\alpha/2)B \rceil$ order statistics of the τ_j^*. Here $\lceil \cdot \rceil$ denotes the ceiling function: $\lceil x \rceil$ is the smallest integer not smaller than x. Let these estimated quantiles be denoted as $q_{\alpha/2}$ and $q_{1-\alpha/2}$ respectively. Then the bootstrap confidence interval is constructed as $[\hat{G} - \hat{\sigma}_G q_{1-\alpha/2}, \hat{G} - \sigma_G q_{\alpha/2}]$. It is of the sort referred to as a percentile-t, or bootstrap-t, confidence interval; see for instance Hall (1992).

In order to test a hypothesis that the Gini indices are the same for two populations from which two independent samples have been observed, a suitable test statistic is $(\hat{G}_1 - \hat{G}_2)/\sqrt{\hat{\sigma}_{G1}^2 + \hat{\sigma}_{G2}^2}$. For each bootstrap repetition, a bootstrap sample is generated by resampling with replacement from each of the two samples, and then the bootstrap statistic is computed as $(G_1^* - G_2^* - \hat{G}_1 + \hat{G}_2)/\sqrt{(\sigma_{G1}^*)^2 + (\hat{\sigma}_{G2})^2}$ in what should be obvious notation. If the samples are correlated, the denominator of the statistic should take account of the covariance, which can be estimated using the formula (25). Bootstrap samples are then generated by resampling pairs of observations.

7. Simulation evidence

In this section, we study by simulation to what extent the methods proposed here give reliable inference, and we compare them with methods previously proposed.

First, in order to see whether the asymptotic normality assumption yields a good approximation, simulations were undertaken with drawings from the exponential distribution, with CDF $F(x) = 1 - \mathrm{e}^{-x}$, $x \geq 0$. The true value G_0 of the Gini index for this distribution is easily shown to be one half. In Figure 1, graphs are shown of the

EDF of 10,000 realisations of the statistic $\tau = (\hat{G} - G_0)/\hat{\sigma}_G$, using the bias-corrected version of $\hat{G}$ and the standard error $\hat{\sigma}_G$ derived from (19), for sample sizes $n = 10$ and 100. The graph of the standard normal CDF is also given as a benchmark.

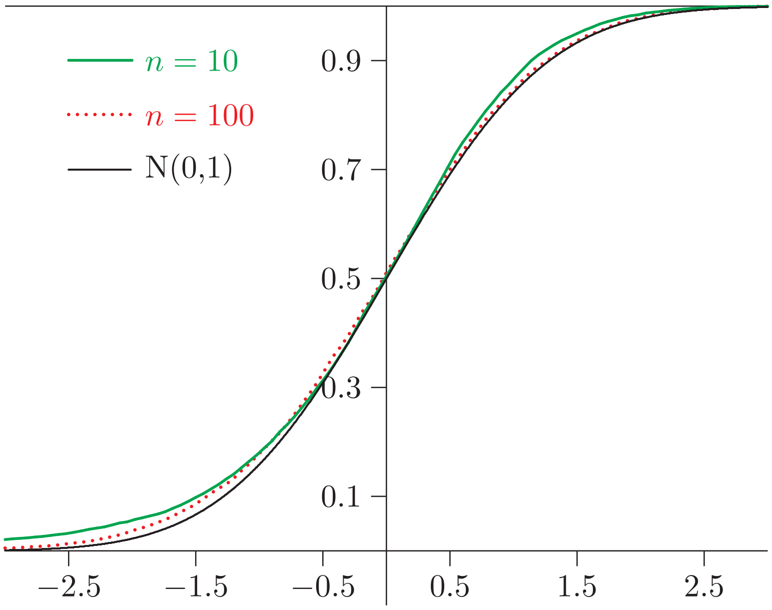

Figure 1. Distribution of standardised statistic as a function of sample size

It can be seen that, even for a very small sample size, the asymptotic standard normal approximation is good. The estimated bias of the of $\hat{G}$ was -0.000444 for $n = 10$, and -0.000717 for $n = 100$. Thus the promise of an approximately unbiased estimator is borne out in these examples. The means of the realised τ were -0.1262 and -0.0478 for $n = 10$ and $n = 100$, and the variances were 1.3709 and 1.0879. The greatest absolute differences between the empirical distributions of the τ and the standard normal CDF were 0.0331 and 0.0208.

In contrast, Figure 2 shows the empirical distributions for $n = 100$ of the statistics $\tau_G \equiv (\hat{G} - G_0)/\hat{\sigma}_{\text{OLS}}$ and $\tau_J \equiv (\hat{G}_J - G_0)/\hat{\sigma}_J$. Here $\hat{\sigma}_{\text{OLS}}$ is the standard error from regression (7), $\hat{\sigma}_J$ is the square root of the variance (34), and $\hat{G}_J$ is given by (27).

It is clear that both of these statistics have distributions that are far from the standard normal distribution. The jackknife estimator does a good job of removing bias, but this is an accidental property of the exponential distribution, whereby the jackknife bias estimator (33) happens to be consistent for G. The mean of the jackknife estimates $\hat{G}_J$ was -0.0057, not quite so good as with true bias correction. The mean of the τ_J was -0.0308, that of τ_G was -0.0012, and the variances of τ_J and τ_G were 0.2693 and 0.4275 respectively.

The exponential distribution may well be fairly characteristic of distributions encountered in practice, but its tail is not heavy. Heavy-tailed distributions are notorious for causing problems for both asymptotic and bootstrap inference, and so in Figure 3 we show empirical distributions for our preferred statistic τ with data generated by the

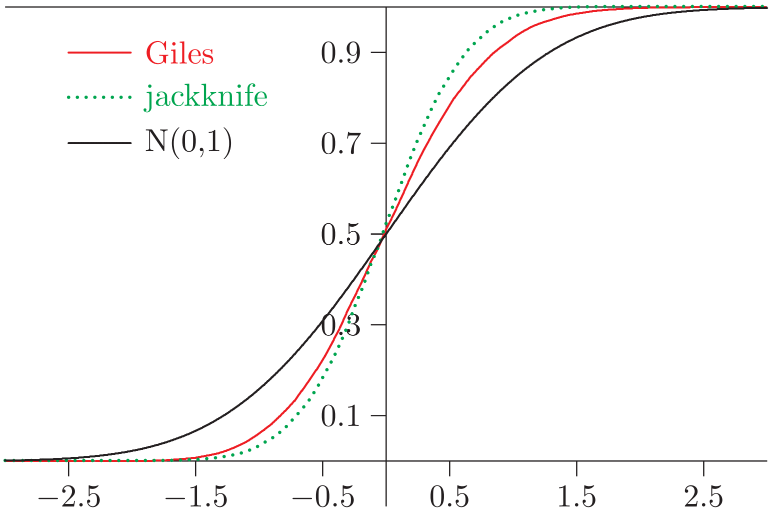

Figure 2. Distribution of Giles's statistic and jackknife statistic

Pareto distribution, of which the CDF is $F_{\text{Pareto}}(x) = 1 - x^{-\lambda}$, $x \geq 1$, $\lambda > 1$. The second moment of the distribution is $\lambda/(\lambda - 2)$, provided that $\lambda > 2$, so that, if $\lambda \leq 2$, no reasonable inference about the Gini index is possible. If $\lambda > 1$, the true Gini index is $1/(2\lambda - 1)$. Plots of the distribution of τ are shown for $n = 100$ and $\lambda = 100, 5, 3, 2$. For values of λ greater than about 50, the distribution does not change much, which implies that there is a distortion of the standard error with the heavy tail even if the tail index is large. The actual index estimate $\hat{G}$, however, is not significantly biased for any value of λ considered.

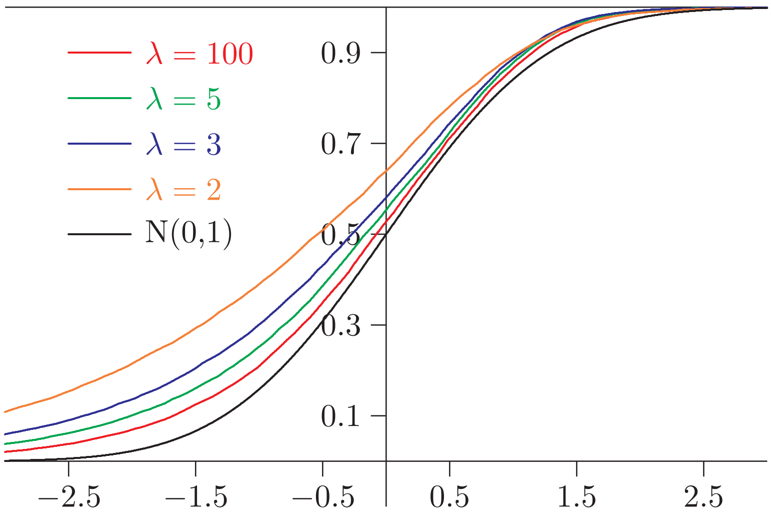

Figure 3. Distribution of τ for the Pareto distribution

Table 1 shows how the bias of τ, its variance, and the greatest absolute deviation of its distribution from standard normal vary with λ.

λ	Bias	Variance	Divergence from N(0,1)
100	-0.1940	1.3579	0.0586
20	-0.2170	1.4067	0.0647
10	-0.2503	1.4798	0.0742
5	-0.3362	1.6777	0.0965
4	-0.3910	1.8104	0.1121
3	-0.5046	2.1011	0.1435
2	-0.8477	3.1216	0.2345

Table 1. Summary statistics for Pareto distribution

It is plain that the usual difficulties with heavy-tailed distributions are just as present here as in other circumstances.

We end this section with some evidence about the behaviour of the bootstrap. In Table 2, coverage rates of percentile-t bootstrap confidence intervals are given for $n = 100$ and for nominal confidence levels from 90% to 99%. The successive rows of the table correspond, first, to the exponential distribution, and then to the Pareto distribution for $\lambda = 10, 5, 2$. The numbers are based on 10,000 replications with 399 bootstrap repetitions each.

Level	90%	92%	95%	97%	99%
Exponential	0.889	0.912	0.943	0.965	0.989
$\lambda = 10$	0.890	0.910	0.942	0.964	0.984
$\lambda = 5$	0.880	0.905	0.937	0.957	0.982
$\lambda = 2$	0.831	0.855	0.891	0.918	0.954

Table 2. Coverage of percentile-t confidence intervals

Apart from the expected serious distortions when $\lambda = 2$, the coverage rate of these confidence intervals is remarkably close to nominal. It seems that, unless the tails are very heavy indeed, the bootstrap can yield acceptably reliable inference in circumstances in which the asymptotic distribution does not.

8. Empirical illustration

Giles (2004) uses his proposed methods, including the jackknife, in order to estimate the Gini index and associated standard error for consumption data for 133 countries, data extracted from the Penn World Tables; see Summers and Heston (1995). The tables in this section compare Giles's results with those obtained using the methods of this paper.

The data are measures of real consumption per capita, in constant dollars, expressed in international prices, base year 1985, for four years, 1970, 1975, 1980, and 1985. The Gini index is therefore a measure of dispersion of consumption across the 133 countries for which data are available. Table 3 shows the estimated index for each of the four years, first as computed using formula (5) (or equivalently, by Giles's regression (7) or (26)), second with the simple bias correction proposed here, and third using the jackknife formula (27). The standard errors associated with these estimates are, respectively, the standard error from regression (26), the square root of the variance estimate (19), and the jackknife standard error, given by the square root of (34).

Year	$\hat{G}$ from (5)	Bias corrected	Jackknife
1970	0.4649	0.4684	0.4685
	(0.0418)	(0.0173)	(0.0478)
1975	0.4767	0.4803	0.4802
	(0.0406)	(0.0169)	(0.0457)
1980	0.4795	0.4831	0.4827
	(0.0397)	(0.0177)	(0.0445)
1985	0.4940	0.4978	0.4974
	(0.0391)	(0.0176)	(0.0438)

Table 3. Estimates and standard errors of G

It can be seen that the jackknife does a very good job of bias correction with these data, at least according to the theory-based bias correction. On the other hand, the jackknife standard errors are similar to those produced by regression (26), and are quite different from those given by (19).

There are slight numerical discrepancies between the results given here and those given by Giles. They affect only the last two decimal places. I think that the numbers here are correct.

Table 4 gives three 95%-nominal confidence intervals for each year of data. The first is based on the standard errors from (19), along with critical values from the standard normal distribution; the second uses the jackknife standard errors with N(0,1) critical values; the third is the percentile-t bootstrap confidence interval. The asymptotic intervals with the best standard error are very similar indeed to the bootstrap intervals, and both are very much narrower than those computed with the jackknife standard errors.

Whatever confidence intervals are used, they all overlap, and so it is not possible to reject the hypothesis that the indices are the same for all four years, unless one takes into account the fact that the series are strongly correlated. By estimating

Year	Std error from (19)	Jackknife std error	bootstrap
1970	[0.4345,0.5022]	[0.3746,0.5621]	[0.4393,0.5074]
1975	[0.4470,0.5135]	[0.3906,0.5699]	[0.4477,0.5140]
1980	[0.4482,0.5179]	[0.3959,0.5702]	[0.4531,0.5219]
1985	[0.4632,0.5323]	[0.4119,0.5836]	[0.4647,0.5329]

Table 4. Confidence intervals for G

the covariance of the estimated indices for 1970 and 1985 by (25), an asymptotically standard normal statistic can be computed for the hypothesis that the indices are the same for both years. Its value is 2.462, which allows us to reject the hypothesis at conventional significance levels.

9. Extensions: Sen's poverty index and the SST index

In this section, we sketch briefly how the methods of this paper can be used to obtain an asymptotically valid standard error of Sen's poverty index; see Sen (1976). This index makes use of the Gini index of the poor in the population, that is, those whose income is less than a specified poverty line, which we here treat as exogenously given. For a poverty line z, Sen justifies the use of the following index as a reasonable measure of poverty:

$$S(z) \equiv H\big(I + (1 - I)G_p\big). \qquad (35)$$

Sen's definition is for a discrete population. His notation is as follows. H is the headcount ratio, which can generally be expressed as $F(z)$, where F is the population CDF, discrete or continuous. His I – not the same as our I defined in (8) – is given by

$$\frac{1}{q}\sum_{j=1}^{n} \mathrm{I}(y_j \le z)\Big(1 - \frac{y_j}{z}\Big) = \frac{1}{F(z)}\int_0^z \Big(1 - \frac{y_j}{z}\Big)\,\mathrm{d}F(y) = 1 - \frac{m(z)}{zF(z)}, \qquad (36)$$

where $q = nF(z)$ is the number of the poor. Here we have used the function $m(y) = \int_0^y x\,\mathrm{d}F(x)$ defined in section 3. The last two expressions in (36) can apply to either a discrete or continuous population. Sen's I is interpreted as the income-gap ratio. The Gini index of the poor, G_p, is defined by Sen as

$$G_p = 1 + \frac{1}{q} - \frac{2}{q^2\mu_p}\sum_{i=1}^{q} y_{(i)}(q + 1 - i),$$

where μ_p is the average income of the poor. The above expression can also be written as

$$G_p = \frac{2}{q^2\mu_p}\sum_{i=1}^{q} y_{(i)}(i - \tfrac{1}{2}) - 1, \qquad (37)$$

which corresponds exactly with our definition (5) for the Gini index of everyone in a discrete sample or population. In terms of the CDF F, we have

$$\mu_p = \frac{1}{F(z)} \int_0^z y \,\mathrm{d}F(y) = \frac{m(z)}{F(z)},$$

It follows that (37) can be expressed as

$$G_p = \frac{2}{F(z)m(z)} \int_0^z yF(y)\,\mathrm{d}F(y) - 1, \tag{38}$$

and so, from (35) along with (36) and (38), we find that

$$\begin{aligned} S(z) &= F(z) - \frac{2}{zF(z)} \int_0^z y\,\bigl(F(z) - F(y)\bigr)\,\mathrm{d}F(y) \\ &= \frac{2}{zF(z)} \int_0^z (z-y)\bigl(F(z) - F(y)\bigr)\,\mathrm{d}F(y) \end{aligned} \tag{39}$$

We use (39) as our definition of Sen's index for both continuous and discrete populations, in the latter case resolving the ambiguity of a left- or right-continuous CDF by splitting the difference, as in (5) and (37).

We now consider the problem of estimating $S(z)$ on the basis of a sample of size n drawn from a population characterised by the CDF F. As usual, we denote by $\hat{F}$ the empirical distribution function of the sample. The natural plug-in estimator is

$$\hat{S}(z) = \frac{2}{z\hat{F}(z)} \int_0^z (z-y)\bigl(\hat{F}(z) - \hat{F}(y)\bigr)\,\mathrm{d}\hat{F}(y). \tag{40}$$

Let $\hat{q}$ be the number of individuals in the sample whose incomes are below the poverty line z; we have $\hat{q} = n\hat{F}(z)$. Then a short calculation shows that

$$\hat{S}(z) = \frac{2}{n\hat{q}z} \sum_{i=1}^{\hat{q}} (z - y_{(i)})\bigl(\hat{q} - i + \tfrac{1}{2}\bigr). \tag{41}$$

It is of interest to observe that the estimate (41) does not coincide exactly with Sen's own definition for a discrete population, which can be written as

$$S_{\text{Sen}}(z) = \frac{2}{n(q+1)z} \sum_{i=1}^{q} (z - y_{(i)})\bigl(q - i + 1\bigr). \tag{42}$$

This point is discussed further in Appendix A.

The algorithm for computing $\hat{S}(z)$ along with its standard error from a sample of size n can be summarised as follows.

- Sort the sample in increasing order, so as to obtain the series of order statistics $y_{(i)}$.
- Determine the number $\hat{q}$ of individuals with income less than the poverty line z.
- For $i = 1, \ldots, \hat{q}$, form the series $s_i \equiv (z - y_{(i)})(\hat{q} - i + \frac{1}{2})$. Then $\hat{S}(z)$ is the sum of the s_i, $i = 1, \ldots, \hat{q}$, times $2/(n\hat{q}z)$.
- For $i = 1, \ldots, \hat{q}$, form the series $p_i = (2\hat{q} - 2i + 1)y_{(i)}/(2n) + n^{-1}\sum_{j=1}^{i} y_{(j)}$.
- Form a series $\hat{Z}_i$ with $\hat{Z}_i = 0$ for $i = \hat{q} + 1, \ldots, n$, and, for $i = 1, \ldots, \hat{q}$, $\hat{Z}_i = z\big(2\hat{q}/n - \hat{S}(z)\big)/2 - p_i$, and compute the mean $\bar{Z}$. The estimated variance of $\hat{S}(z)$ is the sum of the squares of the $\hat{Z}_i - \bar{Z}$, times $4/(z\hat{q})^2$.

The calculations that lead to this algorithm are found in Appendix B. Simulations show clearly that $\hat{S}(z)$ is downward biased. Unfortunately, estimating the bias is not as straightforward as for $\hat{G}$.

The SST index is defined by Shorrocks (1995) as

$$S_{\text{SST}}(z) = \frac{1}{n^2 z} \sum_{i=1}^{q} (2n - 2i + 1)(z - y_{(i)}). \tag{43}$$

Arguments like those leading to (39) show that this formula can be extended to deal with both continuous discrete distributions by using the definition

$$S_{\text{SST}}(z) = \frac{2}{z} \int_0^z (z - y)\big(1 - F(y)\big)\, \mathrm{d}F(y). \tag{44}$$

The plug-in estimator obtained by replacing F by $\hat{F}$ in (44) does in this case coincide exactly with (43). The algorithm for computing $\hat{S}_{\text{SST}}(z)$ using a sample of size n is much like that for $\hat{S}(z)$. The last three steps are replaced by

- For $i = 1, \ldots, \hat{q}$, form the series $s_i = (z - y_{(i)})(n - i + \frac{1}{2})$. Then $\hat{S}_{\text{SST}}(z)$ is the sum of the s_i, $i = 1, \ldots, \hat{q}$, times $2/(n^2 z)$.
- For $i = 1, \ldots, \hat{q}$, form the series $p_i = (2n - 2i + 1)y_{(i)}/(2n) + n^{-1}\sum_{j=1}^{i} y_{(j)}$.
- Form a series $\hat{Z}_i$ with $\hat{Z}_i = 0$ for $i = \hat{q} + 1, \ldots, n$, and, for $i = 1, \ldots \hat{q}$, $\hat{Z}_i = z(1 - \hat{q}/n)) + n^{-1}\sum_{j=1}^{\hat{q}} y_{(j)} - p_i$, and compute the mean $\bar{Z}$. The estimated variance of $\hat{S}_{\text{SST}}(z)$ is the sum of the squares of the $\hat{Z}_i - \bar{Z}$, times $4/(zn)^2$.

Estimating the bias of $\hat{S}_{\text{SST}}(z)$ is quite feasible. Thus, for a bias-corrected estimator, we may add the step

- The bias-corrected estimator is

$$\frac{n}{n-1}\hat{S}_{\text{SST}}(z) - \frac{1}{n-1}\Big(\frac{\hat{q}}{n} - \frac{1}{nz}\sum_{j=1}^{\hat{q}} y_{(j)}\Big).$$

Again, details are found in Appendix B.

10. Conclusion

An expression for the asymptotic distribution of the plug-in estimator of the Gini index has been found that behaves at least as well as other proposed distributions. It is based on an approximation of the estimator as a sum of IID random variables. This approximation allows us to derive a reliable formula for the asymptotic variance. A somewhat more complicated argument leads to an expression for the bias of the estimator. Both bias and variance are easy to estimate in a distribution-free manner.

Similar methods can be used to estimate the variance of Sen's (1976) index of poverty, for arbitrary poverty line z. Unfortunately, it does not seem to be easy to find an expression for the bias of the estimator. This is not the case for the Sen-Shorrocks-Thon modification of the index, for which the standard error and bias are easy to estimate.

Simulations demonstrate that the asymptotic distribution derived for the Gini index is good even for quite small sample sizes, and, unless the tails of the underlying distribution are heavy, is thoroughly reliable for sample sizes greater than around 100. With heavy tails, the asymptotic distribution is a less good approximation. Use of the bootstrap, however, allows us to obtain reliable inference unless the tails are so heavy that the variance is huge or fails to exist.

Appendix A

As mentioned just before the derivation of formula (5) for $\hat{G}$, equation (4) does not define the Gini index of a discrete distribution unambiguously, since adopting right- or left-continuous forms of the CDF lead to different expressions. In this paper, we have used the average of the two different expressions, that is, the $\hat{G}$ of expression (5), for estimation purposes, although we saw that yet another expression, namely $n\hat{G}/(n-1)$, is less biased.

In much of the literature on the Gini index, it is assumed that there is a *finite* population for which the index is to be computed. There has been some discussion of just how to do so, caused by disagreement over the desirable and undesirable features of different definitions. The issues are very clearly set out in a pair of comments that appeared in the *American Sociological Review*, Jasso (1979), and the reply by Allison (1979). Gini's original idea was that the index measured the mean difference between any pair of incomes. Allison and Jasso disagreed over whether the "pair" formed by an income and itself should be counted as a pair for calculating the mean. Allison, who felt that it should, arrived at a formula numerically identical to the $\hat{G}$ of (5). Jasso, who felt that it should not, preferred the formula

$$\hat{G}_2 \equiv \frac{2}{\hat{\mu}n(n-1)} \sum_{i=1}^{n} y_{(i)} i - \frac{n+1}{n-1}, \tag{45}$$

which is readily shown to be equal to $n\hat{G}/(n-1)$, the expression we have used here as a less biased estimator. More than a few subsequent authors have shared Jasso's preference for (45), for instance Deaton (1997). However, as pointed out by Allison, (45) does not satisfy the population symmetry axiom of Sen (1973), which requires that, if a finite population is exactly replicated, the new double-size population should have the same Gini index as the old. Sen himself, in Sen (1976), as might be expected, uses the definition (5). Many economists have sided with Sen and Allison. A notable example is a paper by Donaldson and Weymark (1980), in which various generalisations of the Gini index are presented.

In view of all this, it is a little strange that the definition (42) given by Sen (1976) of his poverty index does not satisfy the population symmetry axiom. On the other hand, the estimator (41) that we use does so if treated as the actual index for a discrete population. To see this, let every individual be duplicated, letting the individual whose income has rank i reappear as two individuals, each with income $y_{(i)}$ but with ranks $2i-1$ and $2i$. Then, when the population is duplicated, since q and n become $2q$ and $2n$ respectively, (41) becomes

$$\frac{2}{4nqz}\sum_{i=1}^{q}(z-y_{(i)})\big(2q-2i+\tfrac{3}{2}+2q-2i+\tfrac{1}{2}\big)=\frac{2}{nqz}\sum_{i=1}^{q}(z-y_{(i)})\big(q-i+\tfrac{1}{2}\big),$$

as for the original population. A similar calculation shows that (42) does not share this property.

What no one seems to dispute is that, for a *continuous* distribution, the appropriate definition of the Gini index is (3) or one of its many equivalents. The approach adopted in this paper is that the finite sample is drawn from an underlying continuous distribution, and our task is to estimate the population index (3) as well as possible. The plug-in estimator (4), as realised by the formula (5), takes the form of one of the possible versions of the index for a discrete distribution, the one that has the support of Sen, Allison, and others. On the other hand, the other version (45), favoured by Jasso and many applied econometricians like Deaton, is a better, because less biased, estimator of the population index.

Appendix B

We use the delta method to express (40) approximately as a sum of IID random variables. The approximation is as follows:

$$\begin{aligned}\hat{S}(z)-S(z) &= -\frac{2}{z\big(F(z)\big)^2}\int_0^z (z-y)\big(F(z)-F(y)\big)\,\mathrm{d}F(y)\\ &+\frac{2}{zF(z)}\int_0^z (z-y)\big(\hat{F}(z)-\hat{F}(y)-F(z)+F(y)\big)\,\mathrm{d}F(y)\\ &+\frac{2}{zF(z)}\int_0^z (z-y)\big(F(z)-F(y)\big)\,\mathrm{d}(\hat{F}-F)(y).\end{aligned}\qquad(46)$$

The first term on the right-hand side above is

$$-\frac{S}{nF(z)}\sum_{j=1}^{n}\big(\mathrm{I}(y_j \le z) - F(z)\big). \tag{47}$$

The third term can be written as

$$\frac{2}{nzF(z)}\sum_{j=1}^{n}\Big(\mathrm{I}(y_j \le z)(z-y_j)\big(F(z)-F(y_j)\big) - \mathrm{E}\big(\mathrm{I}(Y \le z)(z-Y)(F(z)-F(Y))\big)\Big), \tag{48}$$

and the second term as

$$\frac{2}{nzF(z)}\sum_{j=1}^{n}\int_0^z (z-y)\big(\mathrm{I}(y_j \le z) - \mathrm{I}(y_j \le y) - F(z) + F(y)\big)\,\mathrm{d}F(y)$$

Here,

$$\int_0^z (z-y)\mathrm{I}(y_j \le z)\,\mathrm{d}F(y) = \mathrm{I}(y_j \le z)\big(zF(z) - m(z)\big),$$

and

$$\begin{aligned}\int_0^z (z-y)\mathrm{I}(y_j \le y)\,\mathrm{d}F(y) &= \mathrm{I}(y_j \le z)\int_{y_j}^z (z-y)\,\mathrm{d}F(y)\\ &= \mathrm{I}(y_j \le z)\Big(z\big(F(z)-F(y_j)\big) - m(z) + m(y_j)\Big),\end{aligned}$$

Thus, on collecting terms, we see that the second term of (46) is

$$\frac{2}{nzF(z)}\sum_{j=1}^{n}\Big(\mathrm{I}(y_j \le z)\big(zF(y_j) - m(y_j)\big) - \mathrm{E}\big(\mathrm{I}(Y \le z)(z-Y)(F(z)-F(Y))\big)\Big). \tag{49}$$

The terms (48) and (49) can be combined to give

$$\begin{aligned}\frac{2}{nzF(z)}\sum_{j=1}^{n}\mathrm{I}(y_j \le z)\Big(&zF(z) - y_jF(z) + y_jF(y_j) - m(y_j)\\ &- 2\mathrm{E}\big(\mathrm{I}(Y \le z)(z-Y)(F(z)-F(Y))\big)\Big),\end{aligned}$$

and so, with (47), we find that

$$\hat{S}(z) - S(z) \approx \frac{2}{nzF(z)}\sum_{j=1}^{n}\big(\mathrm{I}(y_j \le z)Z_j - \mathrm{E}(\mathrm{I}(Y \le z)Z)\big), \tag{50}$$

where

$$Z_j = zF(z) - \tfrac{1}{2}zS - y_jF(z) + y_jF(y_j) - m(y_j), \tag{51}$$

and Z is the random variable formed by replacing y_j in (51) by Y. If $y_j = y_{(i)}$, the i^{th} order statistic, for $i \le \hat{q}$ we estimate Z_j, as in (18), by

$$\begin{aligned}\hat{Z}_i &= \tfrac{1}{2}z\bigl(2\hat{q}/n - \hat{S}(z)\bigr) - (\hat{q} - i + \tfrac{1}{2})y_{(i)}/n - \tfrac{1}{n}\sum_{j=1}^{i} y_{(j)} \\ &= \tfrac{1}{2}z\bigl(2\hat{q}/n - \hat{S}(z)\bigr) - p_i,\end{aligned}$$

where $p_i = (2\hat{q} - 2i + 1)y_{(i)}/(2n) + n^{-1}\sum_{j=1}^{i} y_{(j)}$. For $i > \hat{q}$, we set $\hat{Z}_i = 0$. We can thus estimate (50) by the expression

$$\frac{2}{z\hat{q}}\sum_{i=1}^{n}\bigl(\hat{Z}_i - \mathrm{E}\bigl(\mathrm{I}(Y \le z)Z\bigr).$$

Clearly we can estimate $\mathrm{E}\bigl(\mathrm{I}(Y \le z)Z\bigr)$ by $\bar{Z}$, the mean of the $\hat{Z}_i$, $i = 1, \ldots, n$, and so can estimate the variance of $\hat{S}(z)$ by

$$\frac{4}{(z\hat{q})^2}\sum_{i=1}^{n}(\hat{Z}_i - \bar{Z})^2,$$

as claimed in the algorithm for $\hat{S}(z)$ in section 9.

For $\hat{S}_{\mathrm{SST}}(z)$, the analysis is similar, and so we can be brief. The delta method tells us that $\hat{S}_{\mathrm{SST}}(z) - S_{\mathrm{SST}}(z)$ is approximately

$$\frac{2}{z}\int_0^z (z-y)\bigl(1 - F(y)\bigr)\,\mathrm{d}(\hat{F} - F)(y) - \frac{2}{z}\int_0^z (z-y)\bigl(\hat{F}(y) - F(y)\bigr)\,\mathrm{d}F(y). \qquad (52)$$

Making the same substitutions as for $\hat{S}(z)$ leads to the result that

$$\hat{S}_{\mathrm{SST}}(z) - S_{\mathrm{SST}}(z) = \frac{2}{nz}\sum_{j=1}^{n}\bigl(\mathrm{I}(y_j \le z)Z_j - \mathrm{E}(\mathrm{I}(Y \le z)Z)\bigr),$$

with

$$Z_j = z\bigl(1 - F(z)\bigr) - y_j\bigl((1 - F(y_j)\bigr) + m(z) - m(y_j).$$

For $y_j = y_{(i)}$ with $i \le \hat{q}$, the estimate of Z_j for $i = 1, \ldots, \hat{q}$ is

$$\hat{Z}_i = z(1 - \hat{q}/n) - (n - i + \tfrac{1}{2})y_{(i)}/n + \hat{m}(z) - \hat{m}(y_j),$$

with $\hat{m}(y)$ given by (17). This leads to the variance estimator of the algorithm for $\hat{S}_{\mathrm{SST}}(z)$ in section 9.

The difference between $\hat S_{\rm SST}(z) - S_{\rm SST}(z)$ and the delta-method approximation (52) is

$$-\frac{2}{z}\int_0^z (z-y)\big(\hat F(y) - F(y)\big)\,{\rm d}(\hat F - F)(y). \tag{53}$$

Since the expectation of (52) is manifestly zero, the bias of $\hat S_{\rm SST}(z)$ is the expectation of (53). Indeed, since the integral with respect to $F(y)$ is also manifestly zero, the bias is the expectation of

$$-\frac{2}{z}\int_0^z (z-y)\big(\hat F(y) - F(y)\big)\,{\rm d}\hat F(y) = -\frac{2}{z}\sum_{i=1}^{\hat q}(z-y_{(i)})\Big(\frac{2i-1}{2n} - F(y_{(i)})\Big).$$

The methods used in section 4 to find the bias of $\hat G$ can be used again to compute the expectation of this expression. The computation is slightly complicated by the fact that $\hat q$ is random. We see that ${\rm E}(\hat q) = nF(z)$, and ${\rm E}(\hat q^2) = n(n-1)\big(F(z)\big)^2 + nF(z)$. After some calculation, we find that the bias is $-n^{-1}\big(S_{\rm SST}(z) - F(z) + m(z)/z\big)$. Therefore

$${\rm E}\Big(\frac{n\hat S_{\rm SST}(z)}{n-1}\Big) = S_{\rm SST}(z) + \frac{1}{n-1}\Big(F(z) - \frac{m(z)}{z}\Big),$$

and so a suitable bias-corrected estimator is

$$\frac{n}{n-1}\hat S_{\rm SST}(z) - \frac{1}{n-1}\Big(\frac{q}{n} - \frac{\hat m(z)}{z}\Big),$$

as claimed in the algorithm.

References

Allison, P. D. (1979). “Reply to Jasso”, *American Sociological Review*, Vol. 44, pp. 870-72.

Bahadur, R. R. and L. J. Savage (1956). “The nonexistence of certain statistical procedures in nonparametric problems”, *Annals of Statistics*, Vol. 27, pp. 1115–22.

Beran, R., 1988. “Prepivoting test statistics: A bootstrap view of asymptotic refinements”, *Journal of the American Statistical Association*, Vol. 83, pp. 687–697.

Bhattacharya, D. (2007). “Inference on inequality from household survey data”, *Journal of Econometrics*, Vol. 137, pp. 674-707.

Bishop, J. A., J. P. Formby, and B. Zheng “Statistical inference and the Sen index of poverty”, *International Economic Review*, Vol. 38, pp. 381-87.

Davidson, R. and J. G. MacKinnon (2000). “Bootstrap tests: how many bootstraps?”, *Econometric Reviews* Vol. 19, pp. 55–68.

Davidson, R. and J. G. MacKinnon (2004). *Econometric Theory and Methods*, Oxford University Press, New York.

Deaton, A. S. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore, Johns Hopkins University Press for the World Bank.

Donaldson, D. and J. A. Weymark (1980). “A Single-Parameter Generalization of the Gini Indices of Inequality”, *Journal of Economic Theory*, Vol. 22, pp. 67-86.

Giles, D. E. A. (2004). “Calculating a standard error for the Gini coefficient: some further results”, *Oxford Bulletin of Economics and Statistics*, Vol. 66, pp. 425-33.

Giles, D. E. A. (2006). “A cautionary note on estimating the standard error of the Gini Index of inequality: comment”, *Oxford Bulletin of Economics and Statistics*, Vol. 68, pp. 395-96.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Jasso, G. (1979). “On Gini’s Mean Difference and Gini’s Index of Concentration”, *American Sociological Review*, Vol. 44, pp. 867-70.

Modarres, R. and J. L. Gastwirth (2006). “A cautionary note on estimating the standard error of the Gini Index of inequality”, *Oxford Bulletin of Economics and Statistics*, Vol. 68, pp. 385-90.

Ogwang, T. (2000). “A convenient method of computing the Gini index and its standard error”, *Oxford Bulletin of Economics and Statistics*, Vol. 62, pp. 123-29.

Ogwang, T. (2004). “Calculating a standard error for the Gini coefficient: some further results: reply”, *Oxford Bulletin of Economics and Statistics*, Vol. 66, pp. 435-37.

Ogwang, T. (2006). “A cautionary note on estimating the standard error of the Gini Index of inequality: comment”, *Oxford Bulletin of Economics and Statistics*, Vol. 68, pp. 391-93.

Sandström, A., J. H. Wretman, and B. Waldén (1988). “Variance estimators of the Gini coefficient: probability sampling”, *Journal of Business and Economic Statistics*, Vol. 6, pp. 113-19.

Sen, A. (1973). *On Economic Inequality*, New York, Norton

Sen A. (1976). “Poverty: An Ordinal Approach to Measurement”, *Econometrica*, Vol. 44, pp. 219-13.

Shorrocks, A. F. (1995). “Revisiting the Sen Poverty Index”, *Econometrica*, Vol. 63, pp. 1225-30.

Summers, R. and A. Heston (1995). *The Penn World Tables*, Version 5.6, NBER, Cambridge, MA. (`http://www.nber.org/pub/pwt56/`).

Xu, K. (2007). “U-Statistics and their asymptotic results for some inequality and poverty measures”, *Econometric Reviews*, Vol. 26, pp. 567-77.