A Note on Computing the Gini Inequality Measure with Weighted Data

John Creedy

WORKING PAPER 03/2015
March 2015

Working Papers in Public Finance

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

The Working Papers in Public Finance series is published by the Victoria Business School to disseminate initial research on public finance topics, from economists, accountants, finance, law and tax specialists, to a wider audience. Any opinions and views expressed in these papers are those of the author(s). They should not be attributed to Victoria University of Wellington or the sponsors of the Chair in Public Finance.

Further enquiries to:
The Administrator
Chair in Public Finance
Victoria University of Wellington
PO Box 600
Wellington 6041
New Zealand

Phone: +64-4-463-9656
Email: cpf-info@vuw.ac.nz

Papers in the series can be downloaded from the following website:
http://www.victoria.ac.nz/cpf/working-papers

A Note on Computing the Gini Inequality Measure with Weighted Data*

John Creeedy†

Abstract

This note sets out some basic results regarding calculation of the Gini measure and its standard error in the context of cross-sectional micro-datasets where sample weights are provided for aggregation from sample to population values.

1 Introduction

It is well known that there are several formulae for the Gini inequality measure.[1] It is usual to express the Gini in unweighted form, in terms of individual values. However, when using cross-sectional survey data, each observation is usually provided with a weight so that population-level values can be computed. Lerman and Yitzhaki (1989) showed how the covariance-based expression for the Gini can easily be modified to deal with sample weights. For incomes of x_i, for $i = 1, ..., n$, and incomes ranked in ascending (strictly, non-decreasing) order, the covariance expression for the Gini, G, is:

$$G = \frac{2}{\bar{x}} Cov\left(x, F\left(x\right)\right) \tag{1}$$

where $F\left(x\right)$ is the distribution function, $\bar{x}$ is the arithmetic mean of the x_i, and $Cov\left(x, F\left(x\right)\right)$ is the covariance. In samples, $F\left(x_i\right)$ is calculated as i/n.

*In preparing this note, I have benefited from discussions with Jesse Eedrah.

†Victoria University of Wellington and New Zealand Treasury.

[1] See, for example, Yitzhaki (1998).

Lerman and Yitzaki (1989) state that if each observation has a weight, w_i, with $\sum_{i=1}^{n} w_i = 1$, $F(x_i)$ is obtained, where $w_0 = 0$, as:[2]

$$\hat{F}(x_i) = \frac{w_i}{2} + \sum_{j=0}^{i-1} w_i \tag{2}$$

The estimate of the Gini coefficient is thus:

$$G = \frac{2}{\bar{x}} \sum_{i=1}^{n} w_i (x_i - \bar{x}) \left(\hat{F}(x_i) - \bar{F} \right) \tag{3}$$

where $\bar{x}$ is now the weighted mean $\bar{x} = \sum_{i=1}^{n} w_i x_i$ and $\bar{F}$ is the weighted mean of the $\hat{F}(x_i)$.[3]

The covariance form of the Gini is therefore very convenient. An alternative and widely used alternative expression, in terms of individual values, is one which has a more transparent link to the (often implicit) value judgements involved in the use of the Gini measure. The modification of this expression to deal with sample weights is set out here in Section 2. A convenient expression also holds for the Gini expression which does not use the ranks directly: this is given in Section 3. Finally, Section 4 modifies a result due to Kakwani *et al.* (1997) on the standard error of the Gini, to deal with sample weights.

2 The Gini and Value Judgements

As above, suppose individual values of x_i for $i = 1, ..., n$ are available. The Gini inequality measure, G, can be written as:

$$G = 1 + \frac{1}{n} - \frac{2}{n^2 \bar{x}} \sum_{i=1}^{n} (n + 1 - i) x_i \tag{4}$$

where $\bar{x}$ is the arithmetic mean, $\frac{1}{n} \sum_{i=1}^{n} x_i$. This is in fact a 'replication invariant' form of the Gini, because for small n the value depends on the sample size. This can be written as:

$$G = \frac{1+n}{n} - \frac{2 \sum_{i=1}^{n} (n + 1 - i)}{n^2} \left(\frac{x_R}{\bar{x}} \right) \tag{5}$$

[2] They actually write the estimate as $\hat{F}_i(x)$.

[3] The approach can easily be applied to deal with the extended Gini.

where x_R is an 'reverse-order-rank-weighted mean' of x_i, given by:

$$x_R = \frac{\sum_{i=1}^{n} (n+1-i)\, x_i}{\sum_{i=1}^{n} (n+1-i)} \tag{6}$$

That is, each value is given a weight given by its 'reverse rank' (that is, its rank when in descending order – ordered from rich to poor, rather than poor to rich). Using $\sum_{i=1}^{n} i = n(n+1)/2$, it can be seen that:

$$G = \frac{1+n}{n} - \frac{n(n+1)}{n^2}\left(\frac{x_R}{\bar{x}}\right) \tag{7}$$

For large samples this reduces to:

$$G = 1 - \frac{x_R}{\bar{x}} \tag{8}$$

The Gini is thus a member of the class, which includes the well-known Atkinson inequality measure, defined as the proportional difference between the arithmetic mean an an 'equally distributed equivalent' income, defined as the income which, if equally distributed, produces the same 'social welfare' as the actual distribution. In the present case the 'social welfare function' (summarising the value judgements of the independent judge), takes the form:

$$W = \sum_{i=1}^{n} (n+1-i)\, x_i \tag{9}$$

for which it can be shown that the equally distributed equivalent income is simply the 'reverse rank' weighted mean.

Now suppose that each x_i has an integer weight, w_i. Let $N = \sum_{i=1}^{n} w_i$, and $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i w_i$. Simply calculating a weighted mean of x, and multiplying each term in the sum in (4) by w_i, does not produce the correct value of the Gini measure. Instead, define $D_{i,j}$ as follows. For $i = 1$, and $j = 1, ..., w_1$:

$$D_{1,j} = N + 1 - j \tag{10}$$

and for $i = 2, ..., n$, and $j = 1, ..., w_i$:

$$D_{i,j} = N + 1 - \sum_{k=1}^{i-1} w_k - j \tag{11}$$

Then:

$$G = 1 + \frac{1}{N} - \frac{2}{N^2 \bar{x}} \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{w_i} D_{i,j} \right) \tag{12}$$

If the weights are non-integer, they can be converted to integer by multiplying by an appropriate constant. For example, if the weights are given to two decimal places, simply multiply all weights by 100. This can be done because G is invariant with respect to changes in the scale of the weights.

Alternatively, the reverse-order-rank weighted mean is given by:

$$x_R = \frac{\sum_{i=1}^{n} \sum_{j=1}^{w_i} \left(N + 1 - \left(\sum_{k=1}^{i-1} w_k \right) - j \right) x_i}{\sum_{i=1}^{n} \sum_{i=1}^{w_i} \left(N + 1 - \left(\sum_{k=1}^{i-1} w_k \right) - j \right)} \tag{13}$$

where it is understood that $w_0 = 0$ so that for $i = 1$, $\sum_{k=1}^{i-1} w_k = 0$. Hence it is clear that re-scaling the weights – which is equivalent to replication – has no effect on the Gini measure.

3 Weights and Non-Ordered Data

It is also possible to use an expression for the Gini measure which does not make use of ordering. For unweighted data, the standard expression involving all pairwise comparisons is:

$$G_s = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N^2 \bar{x}} \tag{14}$$

As before, suppose the weight attached to i is w_i, with $N = \sum_{i=1}^{n} w_i$. Then:

$$\begin{aligned} G_g &= \frac{\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j |x_i - x_j|}{2 \left(\sum_{i=1}^{n} w_i \right)^2 \bar{x}} \\ &= \frac{\sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j |x_i - x_j|}{2 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i} \end{aligned} \tag{15}$$

This can be rewritten as:

$$G_g = \frac{\sum_{i=1}^{n} w_i \sum_{j=1}^{i-1} w_j |x_i - x_j|}{2 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i} + \frac{\sum_{i=1}^{n} w_i \sum_{j=i+1}^{n} w_j |x_i - x_j|}{2 \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i} \tag{16}$$

The numerator of the first term is the sum of every pairwise absolute difference in x_i. It differs from the numerator of equation (15) because it does not repeat any previous comparison. These duplicate comparisons can be found in the last term of equation (16), which is equal to the first term. Thus:

$$G_g = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} w_i w_j |x_i - x_j|}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i} \tag{17}$$

4 Standard Errors of Gini

This section describes the calculation of standard errors for the Gini with and without weights.

4.1 Individual Data

In the case where individual data are available and no weights are required, Kakwani *et al.* (1997) show that for $x_1 < x_2 < ... < x_n$, the standard error can be calculated as follows. Let:

$$z_i = \frac{2i - 1}{2n} \tag{18}$$

$$q_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{N} x_j} \tag{19}$$

and:

$$a_i = \frac{x_i}{\overline{x}} \left\{ 2z_i - (1 + G) \right\} + 2 - (q_i + q_{i-1}) \tag{20}$$

with $q_0 = 0$. Notice that:

$$q_i + q_{i-1} = \frac{i}{n\bar{x}} \left(\frac{1}{i} \right) \sum_{j=1}^{i} (i + 1 - j)\, x_j \tag{21}$$

And letting $x_{R,i}$ denote the 'reverse-order-rank weighted mean' of the first i values (in ascending order) of x, this becomes:

$$q_i + q_{i-1} = \left(\frac{i}{n}\right)\frac{x_{R,i}}{\bar{x}} \tag{22}$$

An estimate of the sampling variance can be obtained as:

$$V(G) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{N} a_i^2 - (1+G)^2\right] \tag{23}$$

Central limit theorems can be used to show that the sampling distribution follows the Normal distribution. For large n the term, $(1+G)^2/n$, can be neglected.

4.2 The Use of Weights

In the case where there are integer weights, w_i, the above expressions need to modified. First, for $i = 1$, and for $j = 1, ..., w_1$, with $N = \sum_{i=1}^{n} w_i$ as before:

$$z_{1,j} = \frac{2j-1}{2N} \tag{24}$$

and for $i = 2, ..., n$:

$$z_{i,j} = \frac{2\left(\sum_{k=1}^{i-1} w_k + j\right) - 1}{2N} \tag{25}$$

Letting $X = \sum_{i=1}^{n} x_i w_i$, for $i = 1$ and for $j = 1, ..., w_1$:

$$q_{1,j} = \frac{jx_1}{X} \tag{26}$$

and for $i = 2, ..., n$, for $j = 1, ..., w_i$:

$$q_{i,j} = \frac{\sum_{k=1}^{i-1} x_k w_k + jx_i}{X} \tag{27}$$

with, as before, $q_{0,j} = 0$ for all j. In the case of no weights, let $d_i = q_i + q_{i-1}$. Where weights are introduced, care is needed in defining the corresponding term, $d_{i,j}$. First, for $i = 1$, $d_{1,1} = q_{1,1}$, and for $j = 2, ..., w_1$:

$$d_{1,j} = q_{1,j} \tag{28}$$

For $j = 1$ and $i = 2, ..., n$, it is seen that:

$$d_{i,1} = q_{i,1} + q_{i-1,w_{i-1}} \tag{29}$$

and for $j = 2, ..., w_i$, and $i = 2, ..., n$:

$$d_{i,j} = q_{i,j} + q_{i,j-1} \tag{30}$$

Then writing:

$$a_{i,j} = \frac{x_i}{\overline{x}} \{2z_{i,j} - (1 + G)\} + 2 - d_{i,j} \tag{31}$$

Then:

$$V(G) = \frac{1}{N} \left[\frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{w_i} a_{i,j}^2 - (1 + G)^2 \right] \tag{32}$$

For large samples the term, $(1 + G)^2 / N$, can again be neglected. With decimal weights, these can again be converted to integers by muliplying by an appropriate amount, and then making the appropriate adjustment to the standard error.

When calculating standard errors, it is not appropriate to use the population weights used for scaling sample values to population values (since the sample size is important). Yet it is useful to ensure that the *relative* weights are maintained, so that the Gini value is the same as when weights are used. Hence the population weights can be re-scaled so that they sum to the sample size.

References

[1] Kakwani, N., Wagstaff, A. and van Doorslaer, E. (1997) Socioeconomic inequalities in health: measurement, computation and statistical inference. *Journal of Econometrics*, 77, pp. 87-103.

[2] Lerman, R.I. and Yitzhaki, S. (1989) Improving the accuracy of estimates of Gini coefficients. *Journal of Econometrics*, 42, pp. 43-47.

[3] Yitzhaki, S. (1998) 'More than a Dozen Alternative Ways of Spelling Gini'. *Research on Economic Inequality*, 8, pp. 13-30.

About the Authors

John Creedy is Professor of Public Economics and Taxation at Victoria Business School, Victoria University of Wellington, New Zealand, and a Principal Advisor at the New Zealand Treasury.
Email: john.creedy@vuw.ac.nz