An (almost) unbiased estimator for the S-Gini index*

Thomas Demuynck†

February 25, 2009

Abstract

This note provides an unbiased estimator for the absolute S-Gini and an almost unbiased estimator for the relative S-Gini for integer parameter values. Simulations indicate that these estimators perform considerably better then the usual estimators, especially for small sample sizes.

1 The absolute and relative S-gini indices

Assume that income is distributed according to a continuous and differentiable cumulative distribution function (cdf) $F : [0, \infty] \to [0, 1]$ with finite mean, μ, and continuous population density function (pdf) f. The absolute single-series Gini (absolute S-Gini), A^{δ}_{∞}, and the Relative single series Gini (relative S-Gini), R^{δ}_{∞}, with parameter $\delta \in \mathbb{R}_{++}$ are given by:

$$A^{\delta}_{\infty} = \mu - H^{\delta}_{\infty} \qquad \text{and} \qquad R^{\delta}_{\infty} = 1 - \frac{H^{\delta}_{\infty}}{\mu},$$
$$\text{with } H^{\delta}_{\infty} = \delta \int_0^{\infty} x \left(1 - F(x)\right)^{\delta - 1} dF(x).$$

These indices exist for all values of $\delta \geq 1$, but for values of $\delta < 1$ it is possible that H^{δ}_{∞} reaches infinity. From now on, we assume that H^{δ}_{∞} is well defined for all values of δ under consideration.

The parameter δ determines the weight attached to the income of individuals at different points in the income distribution. As δ increases, more weight is given to the bottom of the income distribution. For δ equal to one, H^{1}_{∞} is equal to the mean μ and R^{1}_{∞} and A^{1}_{∞} are both equal to zero. For δ equal to 2, the indices A^{2}_{∞} and R^{2}_{∞} reduce to the well-known absolute and relative Ginis. We refer to Donaldson and Weymark (1980), Yitzhaki (1983) and Bossert (1990) for an in depth discussion of the properties related to the S-Gini index.

*I am pleased to acknowledge the insightful comments of Dirk Van de gaer.

†University of Ghent, Sherppa, Tweekerkenstraat 2, B-9000 Gent, Belgium. E-mail: thomas.demuynck@ugent.be

The most common finite sample estimators for the S-Ginis are given by:

$$A_n^\delta = \mu_n - H_n^\delta \qquad \text{and} \qquad R_n^\delta = 1 - \frac{H_n^\delta}{\mu_n}$$
$$\text{with } H_n^\delta = \frac{\sum_{i=1}^n \left((n-i+1)^\delta - (n-i)^\delta\right)\tilde{x}_i}{n^\delta}$$

Here $\tilde{x}_i$ represents the ith smallest value in the sample (the ith order statistic) and μ_n is the sample mean, $\sum_{i=1}^n \tilde{x}_i/n$.

The estimators A_n^δ and R_n^δ are strongly consistent estimators for A_∞^δ and R_∞^δ and they are asymptotically normally distributed (Barrett and Pendakur, 1995; Zitikis and Gastwirth, 2002). Unfortunately, they are not unbiased and their bias depends on the sample size, n, the value of the parameter, δ, and the distribution, F.

The sample mean μ_n is an unbiased estimator for the population mean μ, hence, for the absolute S-Gini, A_∞^δ, we only need to construct an unbiased estimator for the term H_∞^δ. Such estimator would also provide us with an almost unbiased estimator for R_∞^δ. This last estimator is not unbiased because it is divided by the sample mean which is itself an estimator of the population mean. The next section provides an unbiased estimator of H_∞^δ and the last section provides simulation results to compare these estimators with the estimators A_n^δ and R_n^δ.

2 A unbiased estimator for H_∞^δ

We denote by $\left\{ {n \atop k} \right\}$ the stirling number of the second kind with upper index n and lower index k. The number $\left\{ {n \atop k} \right\}$ represents the number of ways that a set of size n can be partitioned into k subsets. We denote by $\binom{n}{k}$ the binomial coefficient with upper index n and lower index k, i.e. the number of k element subsets of an n element set. Finally, we denote by $\langle n \rangle_k$ the falling factorial $n(n-1)\dots(n-k+1)$. The following identities[1] will be used in this section:

$$\binom{n}{k} = \binom{n}{n-k}, \qquad \text{R-1}$$

$$\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\}, \qquad \text{R-2}$$

$$x^r = \sum_{j=0}^{r} \left\{ {r \atop j} \right\} \langle x \rangle_j, \qquad \text{R-3}$$

[1] See Graham et al. (1989) for a proof of these identities.

$$\langle k \rangle_j \binom{n}{k} = \langle n \rangle_j \binom{n-j}{k-j}, \qquad \text{R-4}$$

$$(x+y)^n = \sum_{j=0}^{n} \binom{n}{j} x^j y^{n-j}. \qquad \text{R-5}$$

We focus on the case where the parameter δ takes only integer values. Assume that we have a set of observations $\{x_1, \ldots, x_n\}$ that is drawn i.i.d. from the cdf F. The ith order statistic $\tilde{x}_i$ will have pdf $f_{(i)}$ equal to:

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

The expected value of H_n^δ equals:

$$E(H_n^\delta) = \frac{1}{n^\delta} \sum_{i=1}^{n} i((n-i+1)^\delta - (n-i)^\delta) \binom{n}{i} \int_0^\infty x F(x)^{i-1}(1-F(x))^{n-i}\, dF(x)$$

In order to simplify this expression we split it up into several parts:

$$E(H_n^\delta) = \frac{1}{n^\delta} \int_0^\infty \underbrace{x \sum_{i=1}^{n} \underbrace{i(n-i+1)^\delta \binom{n}{i}}_{A_1} F(x)^{i-1}(1-F(x))^{n-i} -}_{A}$$
$$\underbrace{\sum_{i=1}^{n} \underbrace{i(n-i)^\delta \binom{n}{i}}_{B_1} F(x)^{i-1}(1-F(x))^{n-i} dF(x).}_{B} \qquad (1)$$

We have that:

$$\begin{aligned} A_1 &= i(n-i+1)^\delta \binom{n}{i} \\ &= n(n-i+1)^\delta \binom{n-1}{i-1} && \text{(R-4)} \\ &= (n-i+1)^{\delta+1} \binom{n}{n-i+1} && \text{(R-1)} \\ &= \sum_{j=0}^{\delta+1} \left\{ {\delta+1 \atop j} \right\} \langle n \rangle_j \binom{n-j}{n-i-j+1} && \text{(R-3, R-4)} \end{aligned}$$

$$\begin{aligned} B_1 &= i(n-i)^\delta \binom{n}{i} \\ &= n(n-i)^\delta \binom{n-1}{i-1} && \text{(R-4)} \\ &= \sum_{j=0}^{\delta} n \left\{ {\delta \atop j} \right\} \langle n-i \rangle_j \binom{n-1}{n-i} && \text{(R-1, R-3)} \\ &= \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n \rangle_{j+1} \binom{n-1-j}{i-1}. && \text{(R-4, R-1)} \end{aligned}$$

$$= \sum_{j=1}^{\delta+1} \left\{ \begin{matrix} \delta+1 \\ j \end{matrix} \right\} \langle n \rangle_j \binom{n-j}{i-1}. \qquad \text{(R-1)} \quad \Big\|$$

These results enable us to simplify A and B:

$$\begin{aligned} A &= \sum_{i=1}^{n} \sum_{j=1}^{\delta+1} \left\{ \begin{matrix} \delta+1 \\ j \end{matrix} \right\} \langle n \rangle_j \binom{n-j}{i-1} F(x)^{i-1}(1-F(x))^{n-i} \\ &= \sum_{j=1}^{\delta+1} \left\{ \begin{matrix} \delta+1 \\ j \end{matrix} \right\} \langle n \rangle_j \sum_{i=1}^{n} \binom{n-j}{i-1} F(x)^{i-1}(1-F(x))^{n-i} \\ &= \sum_{j=1}^{\delta+1} \left\{ \begin{matrix} \delta+1 \\ j \end{matrix} \right\} \langle n \rangle_j (1-F(x))^{j-1}. \qquad \text{R-5} \end{aligned}$$

$$\begin{aligned} B &= \sum_{i=1}^{n} \sum_{j=0}^{\delta} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_{j+1} \binom{n-1-j}{i-1} F(x)^{i-1}(1-F(x))^{n-i} \\ &= \sum_{j=0}^{\delta} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_{j+1} \sum_{i=1}^{n} \binom{n-1-j}{i-1} F(x)^{i-1}(1-F(x))^{n-i} \\ &= \sum_{j=0}^{\delta} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_{j+1} (1-F(x))^{j} \qquad \text{R-5} \\ &= \sum_{j=1}^{\delta+1} \left\{ \begin{matrix} \delta \\ j-1 \end{matrix} \right\} \langle n \rangle_j (1-F(x))^{j-1}. \end{aligned}$$

Substituting A and B into equation (1) gives:

$$\begin{aligned} E(H_n^\delta) &= \frac{1}{n^\delta} \int_{-\infty}^{\infty} x \sum_{j=1}^{\delta+1} \left(\left\{ \begin{matrix} \delta+1 \\ j \end{matrix} \right\} - \left\{ \begin{matrix} \delta \\ j-1 \end{matrix} \right\} \right) \langle n \rangle_j (1-F)^{j-1} \, dF(x) \\ &= \frac{1}{n^\delta} \int_{-\infty}^{\infty} x \sum_{j=1}^{\delta+1} j \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_j (1-F)^{j-1} \, dF(x) \qquad \text{R-2} \\ &= \frac{1}{n^\delta} \sum_{j=1}^{\delta} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_j H_\infty^j. \qquad (2) \end{aligned}$$

Equation (2) shows that the expected value of H_n^δ can be expressed as a weighted average of all indices H_∞^m with $m \leq \delta$. As such, the estimator H_n^δ will not be unbiased unless H_∞^m

is zero for all $m \leq \delta$. Equation 2 allows us to construct an unbiased estimator of H_∞^δ in a recursive way.

For $\delta = 1$, we have that $E(H_n^1) = H_\infty^1 = \mu$. Hence, H_n^1 is an unbiased estimator of H_∞^1. Now, assume that we have an unbiased estimator h_n^m of H_∞^m for all m in $\{1, 2, \ldots, \delta - 1\}$. Then we can construct following estimator h_n^δ of H_∞^δ:

$$h_n^\delta = \frac{1}{\langle n \rangle_\delta} \left(n^\delta \, H_n^\delta - \sum_{j=0}^{\delta-1} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_j h_n^j \right). \tag{3}$$

This estimator is unbiased:

$$\begin{aligned} E(h_n^\delta) &= E\left(\frac{1}{\langle n \rangle_\delta} \left(n^\delta \, H_n^\delta - \sum_{j=0}^{\delta-1} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_j h_n^j \right) \right) \\ &= \frac{1}{\langle n \rangle_\delta} \left(n^\delta \, E\left(H_n^\delta\right) - \sum_{j=0}^{\delta-1} \left\{ \begin{matrix} \delta \\ j \end{matrix} \right\} \langle n \rangle_j E\left(h_n^j\right) \right) \\ &= H_\infty^\delta. \end{aligned}$$

The unbiased estimator for A_∞^δ is then given by $a_n^\delta = \mu_n - h_n^\delta$ and the almost unbiased estimator for R_∞^δ is given by $r_n^\delta = 1 - h_n^\delta / \mu_n$. For the Gini index, i.e. $\delta = 2$, it can be shown that $r_n^2 = n R_n^\delta / (n-1)$. This is in agreement to the first order correction for the Gini index found in the literature (see Deaton, 1997; Deltas, 2003; Davidson, 2009).
It can be shown that h_n^δ is equal to the following expression[2]:

$$h_n^\delta = \sum_{i=1}^{n} \frac{\delta \langle n - i \rangle_{\delta-1}}{\langle n \rangle_\delta} \, \tilde{x}_i. \tag{4}$$

The multiplicators $\delta \langle n - i \rangle_{\delta-1} / \langle n \rangle_\delta$ sum to one[3] which implies that, analogue to the estimators H_n^δ, the estimators h_n^δ are a weighted average of the order statistics $\tilde{x}_i$. Also, note that the weights attached to the $\delta - 1$ highest incomes are equal to zero. This implies that the estimator h_n^δ does not use all available information. For example, the value of h_n^{10} on a sample of size 10 coincides with the smallest value in the sample.

Simple manipulation of equation (4) shows that we can write h_n^δ as $\sum_{i=1}^{n} a_i \tilde{x}_i$, with

$$a_i = \begin{cases} \delta / n & \text{if } i = 1 \\ a_{i-1} \left(1 - \dfrac{\delta - 1}{n - (i-1)} \right) & \text{for } i > 1. \end{cases} \tag{5}$$

[2]See appendix A.
[3]See appendix B

For $\delta \geq 1$, as i increases, the weights attached to $\tilde{x}_i$ decrease in an increasing rate until they reach zero for $\tilde{x}_{n-\delta+2}$. The recursion (5) shows that the estimator h_n^δ is very easy to calculate. It also makes it possible to define h_n^δ for non-integer values of δ. Unfortunately, this extension has the unwanted side-effects that the weights a_i do no longer sum to unity, although is will approximate unity if n is not to small, and that the estimator is no longer unbiased.

3 Simulation

For our empirical illustration we used a lognormal distribution with parameters 9.85 and 0.6. Our population statistics A_∞^δ and R_∞^δ were calculated on the basis of a random sample of 50 million observations. We drew 200.000 independent samples of size m ($m = 10, 30, 50$). For each of these samples, we calculated the estimators A_m^δ, a_n^δ, R_m^δ and r_m^δ. Table 1 presents the averages over these 200.000 samples (standard errors are between brackets) for the values $\delta = 1.5$; 2; 5; 7.5 and 10. Simulation results for other parameter values and other distributions give similar results.

Table 1: simulation results

δ	sample size	A_n^δ	a_n^δ	A_∞^δ	R_n^δ	r_n^δ	R_∞^δ
1.5	10	4296 (1764)	4378 (1852)		0.1853 (0.0495)	0.1886 (0.0530)	
	30	4708 (1115)	4802 (1146)	4941	0.2059 (0.033)	0.2100 (0.034)	0.2177
	50	4799 (882)	4870 (899)		0.2105 (0.0262)	0.2136 (0.0268)	
2	10	6733 (2600)	7481 (2890)		0.2908 (0.0696)	0.3231 (0.0774)	
	30	7217 (1574)	7466 (1628)	7458	0.3158 (0.0431)	0.3267 (0.0446)	0.3286
	50	7307 (1227)	7455 (1252)		0.3207 (0.0344)	0.3273 (0.0347)	
5	10	11545 (3837)	12515 (4055)		0.5011 (0.0938)	0.5438 (0.0982)	
	30	12193 (2250)	12509 (2289)	12505	0.5345 (0.0535)	0.5484 (0.0541)	0.5508
	50	12319 (1751)	12508 (1768)		0.5411 (0.0415)	0.5494 (0.0417)	
7.5	10	12729 (4082)	13853 (4346)		0.5545 (0.0993)	0.6043 (0.1065)	
	30	13526 (2395)	13900 (2438)	13894	0.5936 (0.0556)	0.6101 (0.0563)	0.6123
	50	13671 (1863)	13895 (1882)		0.60113 (0.0428)	0.61098 (0.0430)	
10	10	13398 (4246)	14722 (4598)		0.5828 (0.1026)	0.6414 (0.1151)	
	30	14287 (2488)	14715 (2539)	14706	0.6268 (0.0569)	0.6457 (0.0580)	0.6480

Table 1: simulation results

δ	sample size	A_n^δ	a_n^δ	A_∞^δ	R_n^δ	r_n^δ	R_∞^δ
	50	14443 (1942)	14698 (1947)		0.6353 (0.0437)	0.6466 (0.0441)	

NOTE: These simulations were based on the lognormal distribution: $\ln X \sim N(9.85, 0.6)$. The statistics R_∞^δ and A_∞^δ were based on a random sample of 10 million observations. Each average was computed over a set of 200.000 samples. Standard errors are between brackets.

We observe following regularities:

- For integer parameter values, the estimators r_n^δ and a_n^δ performs considerably better then the estimators R_n^δ and A_n^δ.
- For noninteger parameter values one can clearly see that the estimator a_n^δ is no longer unbiased although the bias decreases for larger sample sizes and larger parameter values. Furthermore, the estimators r_n^δ and a_n^δ seem to perform considerably better in comparison to the estimators A_n^δ and R_n^δ.
- The standard errors for the estimators r_n^δ and a_n^δ are slightly larger compared to the standard errors for the estimators R_n^δ and A_n^δ.

References

Barrett, G. F., Pendakur, K., 1995. The asymptotic distribution of the generalized gini indices of inequality. Canadian Journal of Economics 28, 1042–1055.

Bossert, W., 1990. An axiomatization of the single-series ginis. Journal of Economic Theory 50, 82–92.

Davidson, R., 2009. Reliable inference for the gini index. GREQAM Document de Travail nr 2007-23.

Deaton, A. S., 1997. The analysis of household surveys: a microeconometric approach to development policy. John Hopkins University Press for the World Bank, Baltimore.

Deltas, G., 2003. The small-sample bias of the gini coefficient: results and implications for empirical research. The Review of Economics and Statistics 85, 226–234.

Donaldson, D., Weymark, J. A., 1980. A single-parameter generalization of the gini indices of inequality. Journal of Economic Theory 22, 67–86.

Graham, R. L., Knuth, D. E., Patashnik, O., 1989. Concrete Mathematics. Addison-Wesley.

Yitzhaki, S., 1983. Relative deprivation and the gini coefficient. International Economic Review 93, 617–628.

Zitikis, R., Gastwirth, J., 2002. The asymptotic distribution of the s-gini index. Australian and New Zealand Journal of Statistics 44, 439–446.

A Equivalence of equation 3 and 4

The proof is by induction on δ. For $\delta = 1$ we easily establish that both equations 3 and 4 reduce to μ_n. Assume that the assertion holds for all $m < \delta$. The proof follows if we can show that:

$$n^\delta H_n^\delta = \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n \rangle_j h_n^j.$$

where h_n^j is given by equation 4.

$$\begin{aligned}
n^\delta H_n^\delta &= \sum_{i=1}^{n} (n-i+1)^\delta - (n-i)^\delta \tilde{x}_i \\
&= \sum_{i=1}^{\delta} \tilde{x}_i \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n-i+1 \rangle_j - \sum_{i=1}^{n} \tilde{x}_i \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n-i \rangle_j \qquad \text{(R-1)} \\
&= \sum_{i=1}^{n} \tilde{x}_i \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n-i \rangle_{j-1} ((n-i+1) - (n-i-j+1)) \\
&= \sum_{i=1}^{n} \tilde{x}_i \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} j \langle n-i \rangle_{j-1} \\
&= \sum_{i=1}^{n} \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n \rangle_j \frac{j \langle n-i \rangle_{j-1}}{\langle n \rangle_j} \tilde{x}_i \\
&= \sum_{j=0}^{\delta} \left\{ {\delta \atop j} \right\} \langle n \rangle_j h_n^j.
\end{aligned}$$

B h_n^δ is a weighted sum

We show that the weights $\frac{j\langle n-i\rangle_{j-1}}{\langle n\rangle_j}$ sum to one.

$$\begin{aligned}
\sum_{i=1}^{n} \frac{j\langle n-i\rangle_{j-1}}{\langle n\rangle_j} &= \sum_{i=1}^{n} \frac{j}{n} \frac{(n-j)!}{(i-1)!(n-i-j+1)!} \frac{(i-1)!(n-i)!}{(n-1)!} \\
&= \frac{j}{n} \sum_{i=1}^{n} \binom{n-j}{i-1} \Big/ \binom{n-1}{i-1} \\
&= \frac{j}{n} \sum_{k=0}^{n-1} \binom{n-j}{k} \Big/ \binom{n-1}{k} \\
&= 1.
\end{aligned}$$

The last step uses the identity: $\sum_{k=0}^{m} \binom{n}{k} \Big/ \binom{m}{k} = \frac{m+1}{m+1-n}$ (see Graham et al., 1989, problem 1, p173).