Electronic Journal of Applied Statistical Analysis
EJASA, Electron. j. app. stat. anal. 1(2008), 33 – 41
ISSN 2070-5948, DOI 10.1285/i20705948v1n1p33
http://siba2.unile.it/ese/ejasa
http://faculty.yu.edu.jo/alnasser/EJASA.htm

ESTIMATION OF GINI-INDEX FROM CONTINUOUS DISTRIBUTION BASED ON RANKED SET SAMPLING

Mohammad M. Al- Talib[(1)] and **Amjad D. Al-Nasser**[(2)]
Department of Statistics, Yarmouk University, Science Faculty, 21163 Irbid, Jordan
[(1)] m7mdtalib@yahoo.co.uk , [(2)]amjadn@yu.edu.jo

Received 5 February 2008; accepted 23 July 2008
Available online 18 August 2008

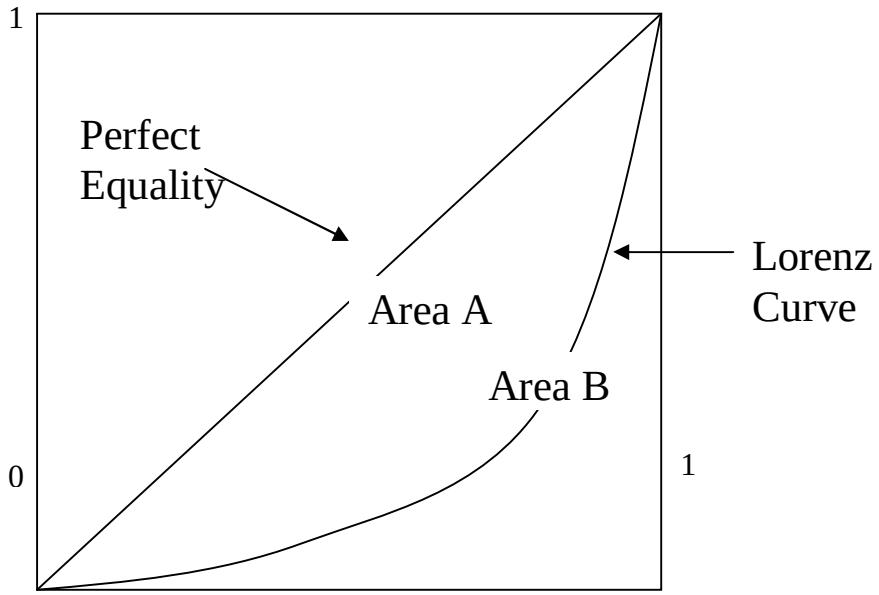
Abstract. *This paper introduces the idea of using the novel ranked set sampling scheme for estimating the Gini index from continuous distributions. A one dimensional integral estimation problem based on ranked samples was discussed. It is demonstrated by a simple Monte Carlo experiment that this approach provides an unbiased and more efficient Gini index estimators than the traditional estimators based on simple random sampling.*

Keywords. *Ranked Set Sampling, Lorenz Curve, Gini Index.*

1. Introduction

Gini index is the most common statistical index of diversity or inequality to measure the dispersion of a distribution in ecology and social sciences Allison (1978). Also, it is widely used in econometrics as a standard measure of inter-individual or inter-household inequality in income and wealth; Atkinson (1970), Sen (1973) and Anand (1983).
The Gini index is a complex inequality measure and it is strictly linked to the representation of measurement inequality through the Lorenz Curve, Figure.1. Typically, a Lorenz Curve is defined on [0, 1], continuous, increasing and concave up, and passes through (0, 0) and (1, 1). Lorenz curve is the most common device for a full description of distribution of income in a population.
To construct Lorenz curves, each measurement ranked from lowest to largest; then the cumulative distribution calculated and plotted against the cumulative proportion.

© 2008 University of Salento - SIBA http://siba2.unile.it/ese



Then, the Gini index (G) quantifies the area between the Lorenz curve and the line of perfect equality. Therefore,

$$G = \frac{Area\ A}{Area\ A + Area\ B} = 2 \times Area\ A = 1 - 2\ Area\ B \tag{1}$$

It ranges from 0, when all measurements are equal; which represents perfect equality, to $\frac{n-1}{n}$ in which all measurements but one is 0; where *n* is a sample size, for large sample a theoretical maximum is 1 which represents maximum possible degree of inequality, a comprehensive surveys for Gini index can be found in Gini (1921) and Bellu and Liberati (2006).

The available information about the distributions is usually discrete; however, we are interested in this article in continuous distributions; noting that both cases can be unified. The remainder of this paper is organized as follows. In section 2, Gini index for continuous distributions is defined. Section 3 introduces the sample mean Monte Carlo method for integral estimation problems. In section 4, Gini index based on ranked set sampling. In section 5, a simple Monte Carlo experiment for estimating the Gini index is presented. The final section briefly summarizes some concluding remarks.

2. Gini Index for Continuous Distributions

Let *X* be a non negative random variable with cumulative distribution function *F(x)* = $\int_0^x f(t)dt$, which is absolutely continuous with mean μ. Then, the Lorenz curve *L(p)* for the observed value *x* is defined by:

$$L(p) = L(F(x)) = \frac{1}{\mu}\int_{a}^{x} x\, dF(x) \tag{2}$$

which represents, for example, the total income of the economy that received by the lowest *100p* of the population for all possible values of *p*. Then, the Gini index can be defined as:

$$G = 1 - 2\int_{0}^{1} L(p)dp \tag{3}$$

Clearly, the Gini index given in (3) is not easy to be evaluated when we are dealing with complicated distribution functions. Alternatively, one can use integral approximation to find an estimator for such index. Traditionally, the Monte Carlo sample mean is the appropriate method used for integral estimation based on simulated simple random samples (SRS). To improve the integral estimation, we will propose on using simulated ranked set sampling (RSS).

3. Sample Mean Monte Carlo Method for Integral Estimation

A defined integral, such as (3), which cannot be explicitly evaluated, can be obtained by a variety of numerical methods. Some of these methods were given in Rubinstein (1981). The integral given in (3) can be represented as expected value of some random variable. Let us rewrite the integral as:

$$G = 1 - 2\int_{0}^{1} \frac{L(p)}{f(p)} f(p)\, dp$$

Assuming that *f(p)* is a continuous probability density function such that *f(p)* > 0 and *0< p< 1*; then,

$$G = 1 - 2E\left[\frac{L(p)}{f(p)}\right]$$

It is very clear that, *p = F(x)* is distributed uniformly over [0,1]. Therefore, an unbiased estimator of G based on SRS is given by

$$\hat{G} = 1 - 2\frac{\sum_{i=1}^{n} g(p_i)}{n} \tag{4}$$

This is an unbiased estimator with variance

$$Var(\hat{G}) = \frac{4}{n}\left(\int_{0}^{1} g^2(p)dp - G^2\right)$$

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

4. Gini Index Based on Ranked Set Sampling

The balanced RSS (Chen et al., 2004) scheme involves of the following steps:

Step 1. Draw randomly *m* sets of SRS each of size *m* from a population,
Step 2. In each set rank the measurements with a cost free method.
Step 3.Then, from the first set the element with the smallest rank is chosen for the actual measurement. From the second set the element with the second smallest rank is chosen. The process is continued by selecting the i[th] order statistics of the i[th] random sample until the element with the largest rank from the m[th] set is chosen.

The scheme yields the following data:

$$\left\{X_{[i:m]}\ , i = 1,2,...,m\right\}$$

where $X_{[i:m]}$ is the i[th] order statistics of the i[th] random sample of size m, and it is denoted by the i[th] judgment order statistics. It can be noted that the selected elements are independent order statistics but not identically distributed.

In practice, the sample size *m* is kept small to ease the visual ranking, RSS literature suggested that m = 2, 3, 4, 5 or 6. Therefore, if a sample of larger size is needed, then the entire cycle may be repeated several times; say *r* times, to produce a RSS sample of size n = r m.

$$\left\{X_{[i:m]j}\ , i = 1,2,...,m,\ j = 1,2,...,r\right\}$$

where $X_{[i:m]j}$ is the i[th] judgment order statistics in the j[th] cycle, which is the i[th] order statistics of the i[th] random sample of size m in the j[th] cycle. It should be noted that all of $X_{[i:m]j}$'s are mutually independent.
In order to plan sample mean Monte Carlo RSS design for the problem in (3), n RSS should be selected. Then the integral estimation has the following steps:

Step 1. Generate a RSS of size $n = m \times r$ from U(0,1)

$$\left\{U_{[i:m]j}\ , i = 1,2,...,m,\ j = 1,2,...,r\right\}$$

Step 2. Compute $g(U_{(i)j})$
Step 3. Find the ranked sample-mean estimator,

$$\hat{G}_{RSS} = 1 - 2\frac{\sum_{j=1}^{m}\sum_{i=1}^{r} g(U_{(i)j})}{mr} \tag{5}$$

5. Simulation Study

In order to illustrate the performance of the suggested estimator of the Gini index based on ranked data we considered the following Lorenz Curve which is used to measure the distribution of income among households in a given country:

$$L(p) = \frac{7}{12} p^2 + \frac{5}{12} p$$

10000 random samples each of size, 6, 8, 10, 12, 18, 24, 30, 36, 60, 60, 80, 100, 120, 1500, 2000, 2500 and 3000 were generated to estimate the Gini index as given in (4) and (5), for p = 0.3, 0.5 and 1.0. Noting that, for the simulated ranked set samples we set the set size m = 3, 4, 5 and 6; and the number of repetitions equal to r = 2, 6, 20 and 500. A comparisons between the two estimators were made by computing the simulated mean, bias, mean squared error (MSE) and efficiency, where

$$MSE(*) = \frac{\sum_{i=1}^{10000} (\hat{G}_{(*)_i} - G)^2}{10000}$$, where * stands for SRS or RSS; and

$$Efficiency = \frac{MSE(SRS)}{MSE(RSS)}$$

The simulation results are given in Table.1 – Table.3.

6. Concluding Remarks

In this paper, we consider using simulated ranked set sampling for estimation the Gini index when the underlying distribution is continuous. The performance of the proposed estimator is compared with a traditional estimator based on simple random sampling based on different sample sizes. All simple simulation experiments indicated that there is an improvement, superior results with more accurate and more efficient Gini index estimator when we are dealing with ranked set sampling.

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

Table.1 Comparisons between SRS and RSS in estimation Gini index with p = 0.5

r	m	Method	Mean	Bias	MSE	Efficiency
2	3	SRS	0.19436	0.09713	0.01069	1.05666
		RSS	0.19444	0.09722	0.01011	
	4	SRS	0.19441	0.09718	0.01039	1.05490
		RSS	0.19446	0.09723	0.00985	
	5	SRS	0.19435	0.09713	0.01018	1.04767
		RSS	0.19441	0.09719	0.00971	
	6	SRS	0.19454	0.09732	0.01009	1.04721
		RSS	0.19443	0.09721	0.00964	
6	3	SRS	0.19459	0.09737	0.00990	1.02419
		RSS	0.19441	0.09719	0.00966	
	4	SRS	0.19443	0.09721	0.00976	1.01874
		RSS	0.19444	0.09722	0.00958	
	5	SRS	0.19445	0.09723	0.00970	1.01553
		RSS	0.19452	0.09730	0.00955	
	6	SRS	0.19439	0.09717	0.00965	1.01378
		RSS	0.19447	0.09725	0.00952	
20	3	SRS	0.19441	0.09718	0.00957	1.00629
		RSS	0.19440	0.09718	0.00951	
	4	SRS	0.19442	0.09720	0.00954	1.00571
		RSS	0.19443	0.09721	0.00948	
	5	SRS	0.19450	0.09728	0.00953	1.00688
		RSS	0.19442	0.09719	0.00947	
	6	SRS	0.19439	0.09717	0.00951	1.00322
		RSS	0.19446	0.09724	0.00948	
500	3	SRS	0.19445	0.09723	0.00946	1.00017
		RSS	0.19446	0.09723	0.00945	
	4	SRS	0.19445	0.09723	0.00946	1.00039
		RSS	0.19444	0.09722	0.00945	
	5	SRS	0.19445	0.09723	0.00946	1.00026
		RSS	0.19445	0.09723	0.00945	
	6	SRS	0.19444	0.09722	0.00945	1.00005
		RSS	0.19445	0.09723	0.00945	

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

Table.2 Comparisons between SRS and RSS in estimation Gini index with p = 0.3

r	m	Method	Mean	Bias	MSE	Efficiency
2	3	SRS	0.17634	0.13434	0.01917	1.91670
		RSS	0.13986	0.09786	0.01000	
	4	SRS	0.18083	0.13883	0.02013	2.04137
		RSS	0.14004	0.09804	0.00986	
	5	SRS	0.18350	0.14150	0.02072	2.12170
		RSS	0.13996	0.09796	0.00976	
	6	SRS	0.18542	0.14342	0.02116	2.17311
		RSS	0.14005	0.09805	0.00973	
6	3	SRS	0.17636	0.13436	0.01842	1.89277
		RSS	0.13994	0.09794	0.00973	
	4	SRS	0.18102	0.13902	0.01961	2.02428
		RSS	0.14001	0.09801	0.00969	
	5	SRS	0.18356	0.14156	0.02027	2.10057
		RSS	0.13995	0.09795	0.00965	
	6	SRS	0.18535	0.14335	0.02074	2.15214
		RSS	0.13998	0.09798	0.00964	
20	3	SRS	0.17626	0.13425	0.01813	1.88100
		RSS	0.13998	0.09798	0.00964	
	4	SRS	0.18082	0.13882	0.01936	2.00833
		RSS	0.14005	0.09805	0.00963	
	5	SRS	0.18355	0.14156	0.02010	2.08988
		RSS	0.14000	0.09801	0.00962	
	6	SRS	0.18536	0.14336	0.02061	2.14286
		RSS	0.14001	0.09801	0.00961	
500	3	SRS	0.17630	0.13430	0.01804	1.87809
		RSS	0.14000	0.09800	0.00960	
	4	SRS	0.18082	0.13882	0.00002	2.00686
		RSS	0.14000	0.09800	0.00960	
	5	SRS	0.18354	0.14154	0.02003	2.08634
		RSS	0.13999	0.09799	0.00960	
	6	SRS	0.18536	0.14336	0.02055	2.14033
		RSS	0.13999	0.09800	0.00960	

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

Table.3 Comparisons between SRS and RSS in estimation Gini index with p = 1.0

r	m	Method	Mean	Bias	MSE	Efficiency
2	3	SRS	0.19479	0.00034	0.00124	1.10714
		RSS	0.19435	-0.00008	0.00112	
	4	SRS	0.19426	-0.00017	0.00094	1.24922
		RSS	0.19450	0.00005	0.00075	
	5	SRS	0.19438	-0.00006	0.00074	1.37719
		RSS	0.19451	0.00007	0.00054	
	6	SRS	0.19449	0.00005	0.00063	1.55606
		RSS	0.19439	-0.00005	0.00041	
6	3	SRS	0.19450	0.00006	0.00042	1.11369
		RSS	0.19442	-0.00002	0.00038	
	4	SRS	0.19458	0.00013	0.00031	1.25023
		RSS	0.19459	0.00014	0.00025	
	5	SRS	0.19433	-0.00011	0.00025	1.40257
		RSS	0.19444	-0.000004	0.00018	
	6	SRS	0.19448	0.00004	0.00021	1.58108
		RSS	0.19445	0.00001	0.00013	
20	3	SRS	0.19444	0.000004	0.00013	1.10382
		RSS	0.19447	0.00002	0.00011	
	4	SRS	0.19436	-0.00008	0.00009	1.23885
		RSS	0.19446	0.00002	0.00007	
	5	SRS	0.19448	0.00004	0.00008	1.40063
		RSS	0.19446	0.00002	0.00005	
	6	SRS	0.19449	0.00004	0.00006	1.55864
		RSS	0.19448	0.00004	0.00004	
500	3	SRS	0.19445	0.00001	0.000005	1.09834
		RSS	0.19445	0.00001	0.000004	
	4	SRS	0.19444	0.000003	0.000004	1.27188
		RSS	0.19445	0.00001	0.000003	
	5	SRS	0.19445	0.000006	0.000003	1.40112
		RSS	0.19444	-0.0000002	0.000002	
	6	SRS	0.19444	0.000002	0.000003	1.55179
		RSS	0.19444	0.0000008	0.000002	

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

References

Anand S. (1983). *Inequality and Poverty in Malaysia*, Oxford University Press, London,UK.

Atkinson, A. B. and Bourguignon, F. (2000). "Introduction: Income Distribution and Economics," Handbook of Income Distribution, New York: Elsevier.

Atkinson, Anthony B. (1970). "On the Measurement of Inequality," Journal of Economic Theory, 2, 244–263.

Bellù, L and Liberati, P. (2006). Inequality Analysis: The Gini Index. FAO, EASYPol Module 040 at: *www.fao.org/tc/easypol*.

Chen, Z., Bai, Z. D. and Sinha, B. K. (2004). *Ranked Set Sampling: Theory and Applications*. Springer.

Gini, C (1921). "Measurement of Inequality of Incomes," The Economic Journal, 31, 124–126.

Sen, Amartya K. (1973). On Economic Inequality, Oxford: Clarendon Press.

Rubinstein, R. (1981). Simulation and the Monte Carlo Method. John Wiley & sons.

© 2008 University of Salento - SIBA http://siba2.unile.it/ese