Revista Colombiana de Estadística
July 2017, Volume 40, Issue 2, pp. 205 to 221
DOI: http://dx.doi.org/10.15446/rce.v40n2.53399

A Comparative Study of the Gini Coefficient Estimators Based on the Linearization and U-Statistics Methods

Estudio comparativo de coeficientes de estimación Gini basados en la linealización y métodos de U-statsitics

Shahryar Mirzaei[a], Gholam Reza Mohtashami Borzadaran[b], Mohammad Amini[c]

Department of Statistics, Faculty of science, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

In this paper, we consider two well-known methods for analysis of the Gini index, which are U-statistics and linearization for some income distributions. In addition, we evaluate two different methods for some properties of their proposed estimators. Also, we compare two methods with resampling techniques in approximating some properties of the Gini index. A simulation study shows that the linearization method performs 'well' compared to the Gini estimator based on U-statistics. A brief study on real data supports our findings.

Key words: Gini coefficient, Income distribution, Linearization method, Resampling techniques, U-statistics.

Resumen

En este artículo consideramos dos métodos ampliamente conocidos para en análisis del índice Gini, los cuales son U-statistics y linealización. Adicionalmente, evaluamos los dos métodos diferentes con base en las propiedades de los estimadores propuestos sobre distribuciones de la renta. También comparamos los métodos con técnicas de remuestreo aproximando algunas propiedades del índice Gini. Un estudio de simulación muestra que el método de linealización se comporta "bien" comparado con el método basado en U-statistics. Un corto estudio de datos reales confirma nuestro resultado.

Palabras clave: índice Gini, distribuciones de la renta, método de linealización, técnicas de remuestreo, U-statistics.

[a]PhD. E-mail: sh.mirzaei@stu.um.ac.ir
[b]PhD. E-mail: grmohtashami@um.ac.ir
[c]PhD. E-mail: m-amini@um.ac.ir

1. Introduction

The most common measure that economists and sociologists use is the Gini index mainly because of clear economic interpretation. The Gini concentration index has been estimated in different ways to obtain valid variance. The reliable standard error is necessary to conduct statistical inference methods, in particular to verify statistical hypothesis and construct confidence intervals. The estimator of this concentration coefficient is usually non-linear, thus it's standard error cannot be obtained easily. There are different methods of variance estimation for the Gini coefficient that can solve this problem. The Gini coefficient can be obtained from a simple ordinary least square regression based approach: see for instance Lerman & Yitzhaki (1984), Shalit (1985), Ogwang (2000), Giles (2004, 2006) and Modarres & Gastwirth (2006). Also, some authors have proposed the resampling techniques to estimate the standard error of the Gini concentration index (see Yitzhaki (1991), Mills & Zandvakili (1997), Berger (2008) and Yitzhaki & Schechtman (2013)).

Another approach to variance estimation of the Gini index is the linearization method. This way combines a range of techniques used to calculate the approximated variance of a non-linear statistic (here, the Gini index). It is based on the first-order Taylor expansion around a parameter and neglecting remaining term. References based on this approximation to variance estimation of the Gini index are such as Berger (2008), Davidson (2009), Langel & Tillé (2013) and Arcagni & Porro (2014).

The U-statistics as a unique frame for a class of statistic includes some popular concentration indices. In income inequality study, we are dealing with estimators that are U-statistic or functions of U-statistics. The theory of U-statistics states that estimators that are included in U-statistics have a desirable asymptotic behavior with nice consistency properties. Among inequality indices, the Gini index is a good applicant because its estimate can be viewed as functions of two simple U-statistics. At first Hoeffding (1948) expressed the Gini index based on function of two U-statistics and then studied its asymptotic properties. Since then, this idea was pursued by authors such as Gastwirth (1971), Wolfe & Randles (1973), Bishop, Formby & Zheng (1997), Xu (2007), Barrett & Donald (2009), Serfling (2009) and Yitzhaki & Schechtman (2013).

Since comparison of the methods to obtain a reliable estimator for the Gini index has been attention in economic and applied statistics, in the literature, according to desirable properties of the U-statistics method, we evaluate and compare this way with the linearization technique. Also, we examine some special situations where the underlying distribution follows popular income distributions.

In the next section, we discuss the concept of the Gini index which is the popular income inequality measure. The main contributions of section 3 is to present and compare some different approaches such as resampling techniques, linearization method and U-statistics to variance estimation of the Gini index. Section 4 provides simulation evidence that bears out the main conclusions of the paper and compares some inferential statistics among these methods. Also, some graphical comparisons have been done. In section 5, the results of the paper for

the real data of Austrian EU-SILC[1] data from 2006 are illustrated. Conclusions are left to the last part of the paper.

2. The Gini Coefficient

The most traditional member of the income inequality family is the Gini coefficient. It is widely used to measure income inequality, mainly because of its intuitive geometric interpretation. This measure can be defined in various ways (see Yitzhaki 1998 and Xu 2003). In general, the Gini index is a function $G : {R_n}^+ \to [0,1]$ that assigns to each non-negative income vector a real number between 0 and 1, which represents the society's inequality level. This measure is 0 in maximum equality and 1 in perfect inequality. The attractive definition of the Gini index is as twice the area between the equality line and the Lorenz curve in the unit box (as shown in Figure 1). The line at 45^0 represents perfect equality of incomes and the area between this line and the Lorenz curve is called concentration area. Therefore, the Gini index can be expressed as

$$G = 2\int_0^1 (p - L(p))dp, \tag{1}$$

such that $p = F(x)$ is a cumulative distribution function (cdf) of non-negative income with positive and finite expectation μ, $L(p)$- the Lorenz function given by $\frac{1}{\mu}\int_0^p F^{-1}(t)dt$, where $F^{-1}(p) = \inf\{x|F(x) \geq p : p \in [0,1]\}$.

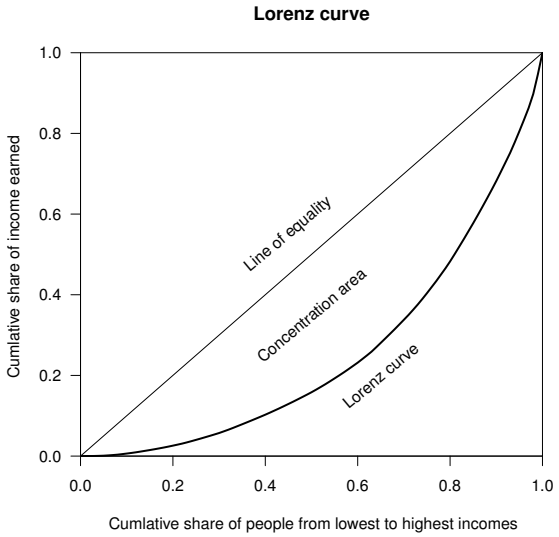

FIGURE 1: The area between the equality line and the Lorenz curve.

[1]European Union Statistics on Income and Living Conditions

Using the definition of Gini index in equation (1), as twice the area between the equality-line and the Lorenz curve, and applying a change of variable $p = F(x)$, it can be found that:

$$G = \frac{2}{\mu}\int_0^\infty xF(x)dF(x) - 1. \tag{2}$$

(for more details see Xu 2003 and Davidson 2009).

Suppose that an i.i.d sample of size n is drawn randomly from the population, and $\hat{F}$ denotes the corresponding empirical distribution function. Let $X_1, \ldots, X_n$ be a random sample and $X_{1:n} \leq \cdots \leq X_{n:n}$ be the order statistics obtained from the sample. Then, an alternative estimator of the Gini coefficient can be obtained by plug-in empirical cdf ($\hat{F}$) of income instead of its corresponding distribution function F in (2), as:

$$\hat{G} = \frac{2}{\hat{\mu}}\int_0^\infty x\hat{F}(x)d\hat{F}(x) - 1. \tag{3}$$

In this regard, the sample Gini index can be expressed as

$$\begin{aligned}\hat{G} &= \frac{1}{\hat{\mu}}\int_0^\infty xd(\hat{F}(x))^2 - 1,\\ &= \frac{2\sum_{i=1}^{n} X_{i:n}(i-\frac{1}{2})}{n\sum_{i=1}^{n} X_i} - 1.\end{aligned} \tag{4}$$

Davidson (2009) found an approximate expression for the bias of $\hat{G}$ from which he derived the bias-corrected estimator of the Gini coefficient, denoted $\tilde{G}$, which is given by:

$$\tilde{G} = \frac{n}{n-1}\hat{G}, \tag{5}$$

while the estimator (5) is still biased but it's bias is of order n^{-1}. Sometimes using this estimator is recommended because the properly bias corrected estimator is not only even easier to compute rather than the other estimators but also its bias converges to 0 faster as $n \to \infty$.

3. Variance Estimation of the Gini Index

The computation of standard error of the Gini index has been subject to numerous publications. Different approaches with complicated formula to variance estimation of the Gini index have prompted a great amount of research in statistics and economics. The main contributions of this section is to present and compare some different approaches to variance estimation of the Gini index.

- The linearization technique. The linearization combines a range of techniques used to calculate the approximated variance of a non-linear statistic. It consists of approximating a non linear or complex statistic (here, the Gini

estimator $\hat{G}$) by a sum of a set of i.i.d linearized weighted variable Z_k such that

$$\hat{G} - G \approx \sum_{k=1}^{n} w_k Z_k.$$

Next, the variance of $\hat{G}$ is simply approximated by the variance of the normalized sum of a set of i.i.d random variables. Based on linearization technique, Davidson (2009) showed that the quantity of $\sqrt{n}(\hat{G} - G)$ is approximately as

$$\sqrt{n}(\hat{G} - G) \approx \frac{2}{\sqrt{n}\mu}\Bigg(\sum_{i=1}^{n} -\frac{E(X.F(X))}{\mu} + X_i.F(X_i) \\ -E(X.I_{[X \leq X_i]}) - (2E(X.F(X)) - \mu)\Bigg),$$

which is the normalized sum of a set of i.i.d random variables (asymptotic normality is an immediate consequence) of expectation zero and the standard error that can be estimated by

$$\hat{\sigma}_{\sqrt{n}(\hat{G}-G)} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Z}_i - \bar{Z})^2}{n\hat{\mu}^2}},$$

where

$$\hat{Z}_i = -(\hat{G} + 1)X_{i:n} + \frac{2i-1}{n}X_{i:n} - \frac{2}{n}\sum_{j=1}^{i} X_{j:n},$$

and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i$ is an estimate of $E(Z_i)$.

- U-Statistics method. In this section the relationships between U-statistics and the Gini index are discussed. Suppose $X_1, \ldots, X_n$ be i.i.d random variables of a distribution F. Consider a parametric function θ for which there is an unbiased estimator. θ may be represented as

$$\theta = E[h(X_1, \ldots, X_m)] = \int \ldots \int h(x_1, \ldots, x_m) dF(x_1) \ldots dF(x_m),$$

where $h = h(x_1, \ldots, x_m)$ is a symmetric function of m ($m \leq n$) i.i.d random variables, called the kernel for θ.

For any kernel h, the corresponding U-statistic for estimating of θ on the basis of a random sample of size n is obtained by averaging the kernel h symmetrically over the observations:

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{m}} \sum_{c} h(X_{i_1}, \ldots, X_{i_m}),$$

where $\sum_c$ denotes summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \ldots, i_m\}$ from $\{1, \ldots, n\}$. Clearly, this estimator is an unbiased estimator of θ.

Here, we use the results from U-statistics (Hoeffding 1948) to derive the inferential statistics of Gini inequality index. The well known Gini coefficient can be expressed in terms of statistical functionals as, $G = \frac{\Delta}{2\mu}$, where $\Delta = E|X_1 - X_2| = \int\int |x_1 - x_2| F(x_1)\, dF(x_2)$ and $\mu = E(X)$, is the population mean of incomes. It is evident that consistent estimator of the population mean is the sample mean given by

$$\hat{\mu} = U_1 = \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} X_i,$$

which is the U-statistic. Also, the consistent estimator of the function Δ is

$$\hat{\Delta} = U_2 = \frac{1}{\binom{n}{2}} \sum_{i<j} |X_i - X_j|,$$

which is called as the Gini mean difference statistic is itself also a U-statistic. It can be noted that the Gini estimator can be estimated by a ratio of these two U-statistics as

$$\hat{G} = \frac{\hat{\Delta}}{2\hat{\mu}} = \frac{\frac{1}{\binom{n}{2}} \sum\limits_{i<j} |X_i - X_j|}{2\frac{1}{\binom{n}{1}} \sum\limits_{i=1}^{n} X_i}. \tag{6}$$

By using Hoeffding's theorem (1948), which concerns the joint distribution of several U-statistics, the U-statistics U_1 and U_2 are consistent estimators for two parameters $\theta_1 = \mu$ and $\theta_2 = \Delta$ and the joint asymptotic distribution of U_1 and U_2 is a bivariate normal distribution. If distribution F is continuous and has a finite variance, then, the joint distribution of two U-statistics is

$$[\sqrt{n}(U_1 - \theta_1), \sqrt{n}(U_2 - \theta_2)] \sim N(0, \Sigma),$$

as $n \to \infty$, where

$$\Sigma = \begin{bmatrix} \phi(\theta_1) & 2\phi(\theta_1, \theta_2) \\ 2\phi(\theta_1, \theta_2) & 4\phi(\theta_2) \end{bmatrix}$$

such that

$$\begin{aligned} \phi(\theta_1) &= Var(X_1), \\ \phi(\theta_2) &= Var(|X_1 - X_2|), \\ \phi(\theta_1, \theta_2) &= Cov(X_1, |X_1 - X_2|), \end{aligned}$$

with corresponding consistent estimators (Bishop et al. 1997) as follow:

$$\begin{aligned}
\widehat{\phi(\theta_1)} =& \frac{1}{n-1}(\sum_{i=1}^{n} {X_i}^2 - n{U_1}^2),\\
\widehat{\phi(\theta_2)} =& \frac{2}{n(n-1)(n-2)} \times \sum_{i<j<k} \{|X_i - X_j||X_i - X_k| + \\
& |X_j - X_i||X_j - X_k| + |X_k - X_i||X_k - X_j| - {U_2}^2,\\
\widehat{\phi(\theta_1, \theta_2)} =& \frac{1}{n(n-1)} \sum_{i<j} (X_i + X_j)|X_i - X_j| - U_1 U_2.
\end{aligned}$$

Since the Gini estimator is the ratio of the sample absolute mean difference to the sample mean, on the asymptotic multivariate normality of a vector of U-statistics together with the Delta method, the sample Gini index, $\hat{G} = \frac{\hat{\Delta}}{2\hat{\mu}}$, converges to a normal distribution with mean $\frac{\Delta}{2\mu}$ and variance:

$$\frac{1}{n}(\frac{\Delta^2}{4\hat{\mu}^4}\phi(\theta_1) - \frac{\Delta}{\mu^3}\phi(\theta_1, \theta_2) + \frac{1}{\mu^2}\phi(\theta_2)),$$

as $n \to \infty$.

- Resampling techniques. The most of the formulations of the variance for the Gini index are mathematically complex. To avoid these mathematical difficulties, some authors have proposed using the resampling techniques such as bootstrap and jackknife methods. Resampling methods treat an observed sample as a finite population and random samples are generated from it to estimate population characteristics and make inferences about the sampled population.

  The bootstrap method is a class of Monte Carlo method that estimate the distribution of a population by resampling. The term bootstrap can refer to nonparametric or parametric bootstrap. Monte Carlo methods that involve sampling from a fully specified probability distribution are called parametric bootstrap. In nonparametric bootstrap, the distribution is not specified and the distribution of the finite population represented by the sample can be regarded as a pseudo population. By repeatedly generating random samples from this pseudo population, the sampling distribution of a statistic can be estimated.

  Suppose G is the parameter of interest and $\hat{G}$ is an estimator of G. Then the bootstrap estimate of distribution of $\hat{G}$ is obtained as follows:

  i) Given a sample $X_1, \ldots, X_n$ of size n and an estimate of $\hat{G}$.

  ii) Draw M bootstrap samples of size n with replacement from $X_1, \ldots, X_n$.

  iii) Calculate the estimator for each one of them and obtain M values of the estimator, denoted by $\hat{G}_1^*, \ldots, \hat{G}_M^*$.

Then, these values are used in order to estimate the variance of the original estimator. Namely, the sample variance of $\hat{G}_1^*, \ldots, \hat{G}_M^*$ is used as the bootstrap variance estimator of the variance of the original statistic. The bootstrap standard error of $\hat{G}$ can then be estimated as:

$$\hat{\sigma}_{Boot} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{G}_m^* - \overline{G^*})^2}, \tag{7}$$

where $\bar{G^*} = \frac{1}{M}\sum_{m=1}^{M}\hat{G}_m^*$.

The jackknife is another resampling method for estimating bias and standard error of an estimator when standard methods for computing bias and variance cannot be applied or are difficult to apply. Suppose that $\hat{G}$ is an estimator of the Gini coefficient (G) based on plug-in estimator in (4). If we denote by $\hat{G}^{(i)}$ the Gini estimator for the subsample of the initial sample where the i^{th} observation has been deleted, then the jackknife estimator for measuring the Gini coefficient based on the n values of $\hat{G}^{(i)}$ is defined as (Knight 1999)

$$\hat{G}_J = \hat{G} + \frac{n-1}{n}\sum_{i=1}^{n}(\hat{G} - \hat{G}^{(i)}), \tag{8}$$

The jackknife technique can also be used to variance estimation of the Gini estimator. Yitzhaki (1991) proposed the standard error of the Gini index with the jackknife method in the following form:

$$\hat{\sigma}_J = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{G}^{(i)} - \hat{G}_\bullet)^2}, \tag{9}$$

where

$$\hat{G}_\bullet = \frac{1}{n}\sum_{i=1}^{n}\hat{G}^{(i)}.$$

4. Simulation Study

In this part of the literature, we carried out a simulation study to comparison the performance of the Gini estimators. We compare the U-statistic estimator with linearization estimator (which proposed by Davidson 2009) in terms of bias and MSE. To find the bias and the MSE, 10,000 estimate of Gini index is obtained by taking the sample size 10, 20, 30, 50, 70 and 100. It is notable that the results is directly applicable to any other sample size. Also, the number of replications is a stopping criteria to access the reliable inferences.

In our study, we first generate random sample from the exponential distribution with cdf $F(x) = 1 - e^{-x},\ x > 0$. Note that the true value of the Gini index for

this distribution is 0.5. It is interesting to see that the estimator based on U-statistics has less bias and more MSE compared to the linearization estimator. So the linearization estimator perform better. The comparison results are given in Table 1.

TABLE 1: Bias and MSE of the two Gini estimate in exponential distribution.

n	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
10	−0.05067100	0.01062074	−0.00074555	0.04405025
20	−0.02503643	0.00434030	−0.00023838	0.02109458
30	−0.01671419	0.00276912	−0.00006753	0.01394532
50	−0.01023361	0.00164075	−0.00004916	0.00833662
70	−0.00707629	0.00116610	−0.00003835	0.00629282
100	−0.00487323	0.00081815	−0.00002804	0.00421893

It is evident that the two estimators are affected by negative bias, that is, they underestimate the value of the index. Also, it is possible to see that the linearization method has better performance, in particular for small samples.

Plot of the statistic distribution $\tau = \frac{\hat{G}-0.5}{\hat{\sigma}_{\hat{G}}}$ based on the linearization technique are shown in Figure 2 for $n = 10,\ 100$. It can be noted that, even for a very small sample size, the asymptotic standard normal approximation of linearization estimator in exponential distribution is high.

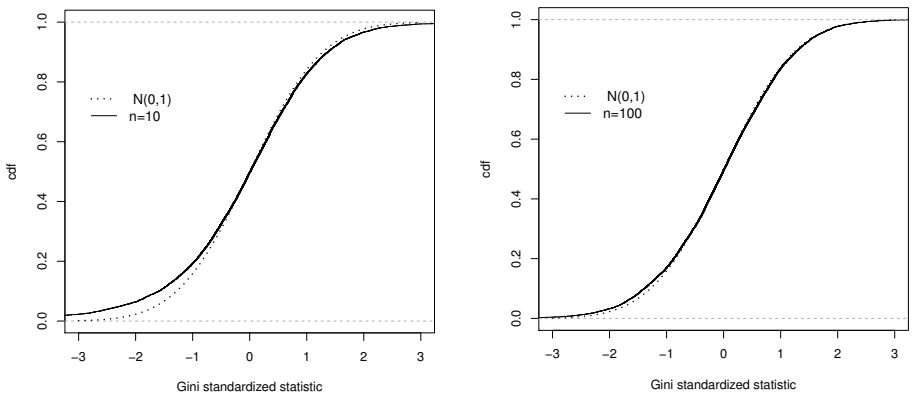

FIGURE 2: The empirical distributions of Gini linearization statistic.

The Pareto distribution is considered as the best model for income data as it capture heavy tail behavior. For our study consider Pareto distribution with cdf $F(x) = 1 - x^{-\lambda}\ ,\ x \geq 1\ ,\ \lambda > 1$. The true value of the Gini index is $\frac{1}{2\lambda-1}$. The corresponding results are given in Table 2.

For different values of λ and for $n = 10,\ 100$, the MSE and bias are given in Table 3 and 4. The bias of U-statistic is less than of linearization estimator but MSE is more.

Table 2: Bias and MSE of the two Gini estimates in Pareto distribution with $\lambda = 5$.

n	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
10	−0.01339002	0.00063490	−0.00253212	0.00244343
20	−0.00675677	0.00052846	−0.00126444	0.00149833
30	−0.00438400	0.00044062	−0.00070375	0.00106643
50	−0.00247359	0.00032622	−0.00025650	0.00048747
70	−0.00181505	0.00025644	−0.00023105	0.00042769
100	−0.00122035	0.00019486	−0.00011034	0.00031979

Table 3: Bias and MSE of the two Gini estimates in Pareto distribution with $n = 10$.

λ	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
2	−0.07303529	0.00710130	−0.04411328	0.02535909
3	−0.03079368	0.00204798	−0.01199298	0.00882469
4	−0.01874429	0.00103846	−0.00495397	0.00429905
5	−0.01339002	0.00063490	−0.00253212	0.00256978
10	−0.00550398	0.00014570	−0.00026758	0.00049986
20	−0.00253181	0.00003477	0.00003588	0.00011052
50	−0.00096728	0.00000540	0.00004758	0.00001678

Table 4: Bias and MSE of the two Gini estimates in Pareto distribution with $n = 100$.

λ	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
2	−0.01267973	0.00229515	−0.00944081	0.00846369
3	−0.00330350	0.00075055	−0.00131667	0.00268573
4	−0.00177992	0.00034570	−0.00035490	0.00115436
5	−0.00122035	0.00019486	−0.00011034	0.00066013
10	−0.00047960	0.00003781	0.00004719	0.00013214
20	−0.00021772	0.00000837	0.00003908	0.00002822
50	−0.00008265	0.00000125	0.00001854	0.00000401

In using a hypothesis test and confidence interval, it is important to have a correct method available for computing the standard error of the Gini coefficient. So, in this regard, Figure 3 compares the variance estimation of the two methods under Pareto distribution with $n = 100$. This Figure shows that how the variance of Gini index varies with parameter λ. It is evident that for values of λ greater than about 50 there is no significant difference between variance estimation of the two ways.

Finally we compare two estimators when the sample come from Loglogistic (LL) distribution as one of the simplest form of the generalized beta distribution of second kind (GB2) with cdf

$$F(x) = 1 - \frac{1}{1 + x^a}, \qquad a > 0,$$

where $a > 0$ is the shape parameter. In this distribution family, the true value of the Gini estimator is $\frac{1}{a}$. The comparison results are shown in Table 5.

For different values of a and $n = 100$, MSE and bias are also given in Table 6. It can be noted that the sample Gini index, however, is not remarkable biased for any value of a considered.

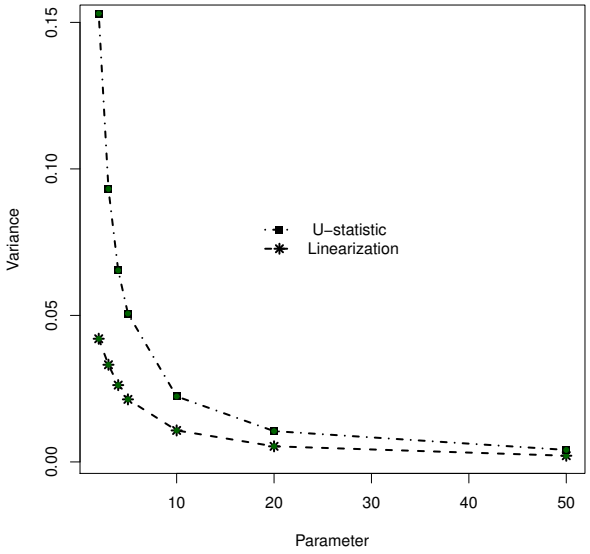

FIGURE 3: Comparison the variance estimation with different parameters.

TABLE 5: Bias and MSE of the two Gini estimates in LL distribution with $a = 5$.

n	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
10	−0.02172880	0.00192282	−0.00192089	0.00343828
20	−0.01099741	0.00110183	−0.00104990	0.00197708
30	−0.00734795	0.00080539	−0.00070478	0.00137877
50	−0.00430765	0.00053468	−0.00031393	0.00068702
70	−0.00309310	0.00040239	−0.00023938	0.00056462
100	−0.00216258	0.00029543	−0.00016422	0.00041351

TABLE 6: Bias and MSE of the two Gini estimates in LL distribution with $n = 100$.

a	Bias(linearization)	MSE(linearization)	Bias(U-statistic)	MSE(U-statistic)
2	−0.01487567	0.00213197	−0.00997542	0.00254675
3	−0.00476476	0.00090349	−0.00144588	0.00165349
4	−0.00279950	0.00047809	−0.00030253	0.00085085
5	−0.00202923	0.00029428	−0.00002952	0.00041351
10	−0.00091400	0.00006922	0.00008687	0.00014300
20	−0.00045018	0.00001705	0.00005032	0.00009780
50	−0.00018015	0.00000272	0.00002005	0.00000836

Plot of the empirical distribution of statistics $\tau = \frac{\hat{G}-\frac{1}{a}}{\hat{\sigma}_{\hat{G}}}$ based on linearization technique are shown in Figure 4 for $n = 100$ and different values of a. It can be noted that for values of a greater than about 50 the distribution does not change much.

Here, we refer to some evidence about the behaviour of the bootstrap. In Table 7, coverage probabilities (C.P) and average sizes (A.S) of t-bootstrap confidence intervals (see Mills & Zandvakili 1997) are given for $n = 100$ and for nominal confidence levels %90, %95 and %99. Apart from the expected serious distortions

when $a = 2$, the coverage rate of these confidence intervals is remarkably close to nominal. It seems that, unless the tails are very heavy, the bootstrap can yield acceptably reliable inference.

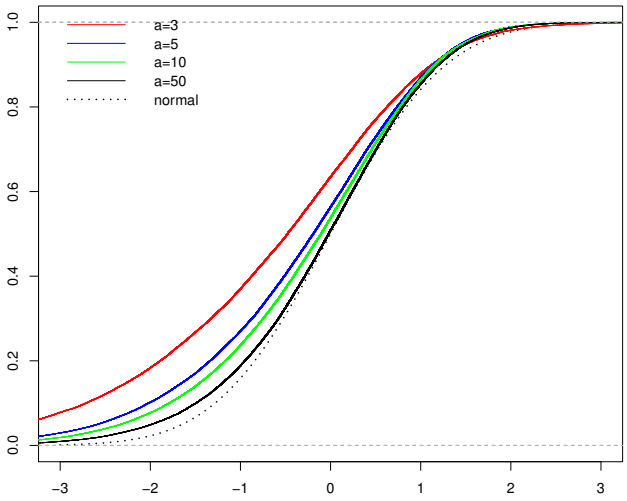

FIGURE 4: The the distribution of statistics τ for Loglogistic distribution.

TABLE 7: Comparison the C.P and A.S of Gini estimate in LL distribution.

	%90		%95		%99	
a	C.P	A.S	C.P	A.S	C.P	A.S
50	0.8798	0.0053457	0.9287	0.006378561	0.9754	0.00838285
10	0.8709	0.0271002	0.9237	0.032179858	0.9730	0.04229150
5	0.8577	0.0557920	0.9132	0.066488796	0.9680	0.08738109
3	0.8115	0.0972594	0.8803	0.116247604	0.9432	0.15277525
2	0.7225	0.1449060	0.7970	0.173864654	0.8839	0.22849689

We end this section with comparison the greatest absolute deviation of the empirical distribution of Gini standardized statistics from standard normal. Table 8 explains the divergence from normal distribution for the lineariazation, U-statistics, jackknife and bootstrap estimators under exponential distribution as a benchmark.

It can be seen that the lineariazation method does a very good job of the divergence from normal distribution. The lineariazation estimator has desirable asymptotic properties. Generally the two resampling methods provide similar results.

5. Data Analysis

In the following we will refer to the real data set with 14827 observations which is generated from real Austrian EU-SILC (European Union Statistics on Income

TABLE 8: Comparison the divergence of Gini estimates from $N(0,1)$.

n	Linearization	U-statistics	Jackknife	Bootstrap
10	0.21339112	0.28600970	0.21216048	0.21416842
20	0.16325142	0.24781780	0.16048859	0.16842558
30	0.13498691	0.23805330	0.13148433	0.12948875
50	0.10892481	0.22295260	0.09785820	0.09937744
70	0.08500432	0.22114550	0.09348386	0.07838025
100	0.07240914	0.21111600	0.06714185	0.07703274
500	0.03472761	0.09221455	0.03978515	0.03826586
1000	0.02160005	0.07214550	0.02603517	0.03083913

and Living Conditions) data from 2006[2]. We have first performed an analysis on comparison of the standard errors and %95 confidence intervals of the Gini estimators. The corresponding results are given in Table 9.

TABLE 9: Comparison the Gini estimates in real data.

Method	$\hat{G}$	S.E($\hat{G}$)	Confidence interval
Linearization	0.2628532	0.0019063	[0.262822515 , 0.26288384]
U-statistics	0.2628710	0.0019077	[0.259131908 , 0.26661009]
Jackknife	0.2628736	0.0019067	[0.262842909 , 0.26290429]

Figure 5 shows relative frequency histograms obtained on the basis of simulated 10,000 samples (simple random sampling design without replacement) of size $n = 100$ from the Austrian income data and estimated both indices in each sample. The histograms are accompanied by fitted normal density curves. It is evident that for the two Gini estimators, the consistency with the normal distribution is high. This intuition is confirmed by a small simulation study performed to estimate the skewness and excess kurtosis of the sampling distribution of both indices.

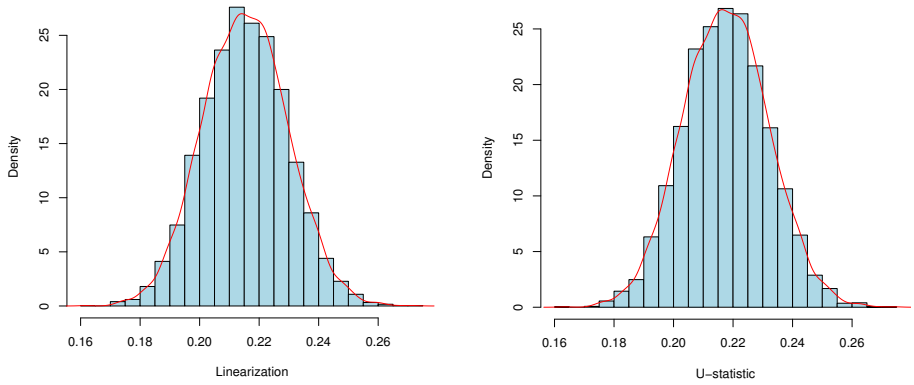

FIGURE 5: The empirical distributions of Gini linearization statistics.

The skewness and excess kurtosis for each index are then estimated on the 10,000 samples. The results, displayed in Table 10, show that the skewness and

[2]The data set is available from the laeken-package in R software environment.

excess kurtosis of the Gini estimator based on U-statistics is farther from the desired level (0 for both statistics) rather than the other estimate based on linearization technique.

TABLE 10: Comparison the skewness and kurtosis of the Gini estimators.

Method	Skewness	Kurtosis
Linearization	0.05139230	−0.02724533
U-statistics	0.08945503	−0.03652660

According to probability plots and quantile plots (Figure 6), the GB2 distribution with scale parameter equal to $b = 20933$ and shape parameters equal to $a = 5.2$, $p = 0.5$ and $q = 0.77$ fits the data well. It should be noted that the parameters considered are the maximum likelihood estimates of the GB2 distribution based on data income.

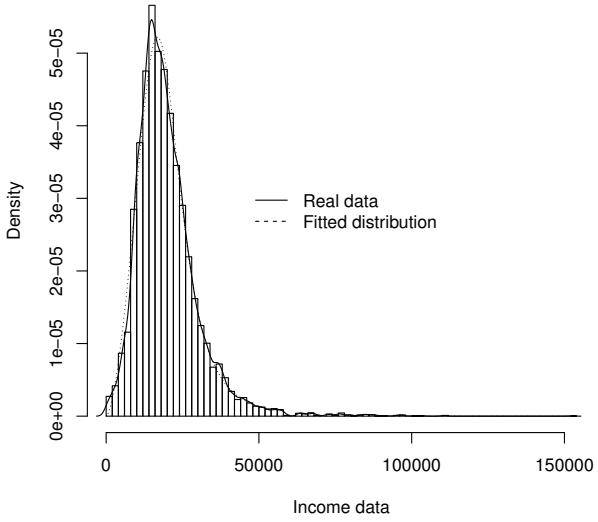

FIGURE 6: The fitted distribution to real data.

Here, we have performed an analysis on comparison the Bias and MSE of the two Gini estimates in fitted distribution to real data. For better interpretation, the results have been shown in Figure 7. This figure shows that the two estimators underestimate the value of the index. The bias of the linearization method is bigger than of the U-statistic that seem to be asymptotically unbiased. Considering the accuracy of the estimators through the value of the MSE, it is possible to see that the linearization method has better performance, in particular for small samples.

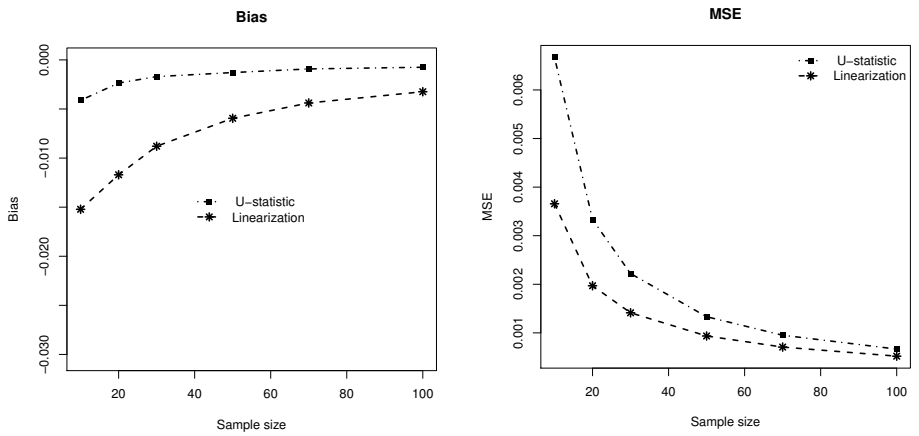

FIGURE 7: Bias and MSE of the two Gini estimates in the fitted distribution to real data.

6. Conclusion

In this paper, we consider two well-known methods for analysis of the Gini index, which are U-statistics and linearization for some inequality distributions. In these distributions, in addition, we evaluate two different methods for some properties of their estimators. Also via some figures, we compare the two methods with jackknife technique in approximating variance and convergence rate of the Gini estimator. Overall, in this note, the results are all favor of linearization method compared to U-statistic technique. Also, a brief study on real data income supports our findings.

[Received: October 2015 — Accepted: January 2017]

References

Arcagni, A. & Porro, F. (2014), 'The graphical reprensentation of inequality', *Revista Colombiana de Estadística* **37**, 419–436.

Barrett, G. F. & Donald, S. G. (2009), 'Statistical inference with generalized Gini indices of inequality, poverty, and welfare', *Journal of Business & Economic Statistics* **27**, 1–17.

Berger, Y. G. (2008), 'A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient', *Journal of Statist* **24**(1), 541–555.

Bishop, J. A., Formby, J. P. & Zheng, B. (1997), 'Statistical inference and the Sen index of poverty', *International Economic Review* **150**, 381–387.

Davidson, R. (2009), 'Reliable inference for the Gini index', *Journal of econometrics* **150**(1), 30–40.

Gastwirth, J. L. (1971), 'A general definition of the Lorenz curve', *Econometrica* **39**, 1037–1039.

Giles, D. E. (2004), 'Calculating a standard error for the Gini coefficient: some further results', *Oxford Bulletin of Economics and Statistics* **66**(1), 425–433.

Giles, D. E. (2006), 'A cautionary note on estimating the standard error of the Gini index of inequality: comment', *Oxford Bulletin of Economics and Statistics* **68**(1), 395–396.

Hoeffding, W. (1948), 'A class of statistics with asymptotically normal distribution', *The Annals of Mathematical Statistics* **19**(1), 293–325.

Knight, K. (1999), *Mathematical Statistics*, John Wiley & Sons, New York.

Langel, M. & Tillé, Y. (2013), 'Variance estimation of the Gini index: revisiting a result several times published', *Journal of the Royal Statistical Society-Series A* **176**, 521–540.

Lerman, R. I. & Yitzhaki, S. (1984), 'A note on the calculation and interpretation of the Gini index', *Economics Letters de Estadistica* **15**, 363–368.

Mills, J. A. & Zandvakili, S. (1997), 'Statistical inference via bootstrapping for measures of inequality', *Journal of Applied econometrics* **12**, 133–150.

Modarres, R. & Gastwirth, J. L. (2006), 'A cautionary note on estimating the standard error of the Gini index of inequality', *Oxford Bulletin of Economics and Statistics* **68**(1), 391–393.

Ogwang, T. (2000), 'A convenient method of computing the Gini index and its standard error', *Oxford Bulletin of Economics and Statistics* **47**, 123–129.

Serfling, R. J. (2009), *Approximation theorems of mathematical statistics*, John Wiley & Sons, New York.

Shalit, H. (1985), 'Calculating the Gini index of inequality for individual data', *Oxford Bulletin of Economics and Statistics* **47**, 185–189.

Wolfe, D. & Randles, R. (1973), *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York.

Xu, K. (2003), 'How has the literature on Gini's index evolved in the past 80 years?', *Dalhousie University, Economics Working Paper* .

Xu, K. (2007), 'U-statistics and their asymptotic results for some inequality and poverty measures', *Econometric Reviews* **26**, 567–577.

Yitzhaki, S. (1991), 'Calculating jackknife variance estimators for parameters of the Gini method', *Journal of Business & Economic Statistics* **9**, 235–239.

Yitzhaki, S. (1998), 'More than a dozen alternative ways of spelling Gini', *Research on economic inequality* **8**, 13–30.

Yitzhaki, S. & Schechtman, E. (2013), *The Gini Methodology: A primer on a Statistical Methodology*, Springer, New York.