econometrics

MDPI

Article

Estimation of Gini Index within Pre-Specified Error Bound

Bhargab Chattopadhyay [1,*] and Shyamal Krishna De [2]

[1] Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA
[2] School of Mathematical Sciences, National Institute of Science Education and Research, Jatni 752050, Odisha, India; sde@niser.ac.in
* Correspondence: bhargab@utdallas.edu; Tel.: +1-972-883-6693

Academic Editor: Kerry Patterson
Received: 19 December 2015; Accepted: 3 June 2016; Published: 24 June 2016

Abstract: Gini index is a widely used measure of economic inequality. This article develops a theory and methodology for constructing a confidence interval for Gini index with a specified confidence coefficient and a specified width without assuming any specific distribution of the data. Fixed sample size methods cannot simultaneously achieve both specified confidence coefficient and fixed width. We develop a purely sequential procedure for interval estimation of Gini index with a specified confidence coefficient and a specified margin of error. Optimality properties of the proposed method, namely first order asymptotic efficiency and asymptotic consistency properties are proved under mild moment assumptions of the distribution of the data.

Keywords: distribution-free method; fixed width confidence interval; Gini index; sample size planning; U-statistics

JEL: C130; C140; C400; C440

1. Introduction

Economic inequality arises due to inequality in the distribution of income and assets among individuals or groups within a society or region or even between countries. Economic inequality is usually measured to evaluate the effects of economic policies at the micro or macro level. Several inequality indexes that measure the economic inequality are proposed in the economics literature. Among those indexes, Gini inequality index is the most widely used measure. The most celebrated Gini index, as given in [1], is

$$G_F(X) = \frac{\Delta}{2\mu}, \quad \text{where} \quad \Delta = E\,|X_1 - X_2|\,,\ \mu = E(X) \tag{1}$$

and X_1 & X_2 are two i.i.d. copies of nonnegative random variable X with distribution function F. Gini index compares every individual's income with other individual's income. If there are n randomly selected individuals with incomes given by $X_1, \ldots, X_n$, then the estimator of the celebrated Gini index is

$$\widehat{G}_n = \frac{\widehat{\Delta}_n}{2\bar{X}_n}, \tag{2}$$

where $\bar{X}_n$ is the sample mean and $\widehat{\Delta}_n$ is the sample Gini's mean difference defined as,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \widehat{\Delta}_n = \binom{n}{2}^{-1} \sum_{1\le i_1<i_2\le n} \left|X_{i_1} - X_{i_2}\right|. \tag{3}$$

The Gini index is undefined if $\bar{X}_n = 0$. We ignore this special case.

Inference for inequality measures, including Gini index, has been an area of research interest among many economists in recent years. For the existing literature on inference problems related to inequality index, we refer to [2–7]. Even though there exist innovative methods for constructing confidence intervals for G_F (e.g., see [7]), due to large standard errors of estimated Gini index as mentioned in [8], we may not get a short confidence interval for Gini index. We know that the confidence interval varies from sample to sample and so does its width. Wider confidence intervals provide less precise information about the true value of the parameter of interest. Since it is desirable to construct shorter confidence intervals, we rather fix the length of the confidence interval, or in other words, the margin of error while achieving the confidence coefficient $(1-\alpha)$ for some specified α in $(0,1)$. This problem is known as the fixed-width confidence interval estimation problem.

No fixed sample size procedure can provide a solution to the fixed-width confidence interval estimation problem (e.g., see [9]). Thus, one must resort to sampling in stages to construct a $100(1-\alpha)\%$ confidence interval for G_F with a pre-specified width. This problem falls in the domain of sequential analysis. For details about the general theory of fixed-width confidence interval estimation, we refer to [10,11]. Sequential analysis is concerned with studies where sample sizes are not fixed in advance. Instead, the sequential estimation procedure depends on collecting observations until an a-priori specified criterion or *stopping rule* is satisfied.

We know that Gini's mean difference is U-statistic with a symmetric kernel of degree 2 and the sample mean is a U-statistic with a symmetric kernel of degree 1 (e.g., see [12]). Under distribution-free scenario, [7] used the central limit theorem for U-statistics to come up with a confidence interval for Gini index. However, this cannot be used to find out a fixed-width confidence interval for Gini index. In this article, we solve the problem of obtaining a fixed-width confidence interval for Gini index using a purely sequential procedure with a stopping rule based on several U-statistics. Apart from being unbiased estimators, U-statistics are also reverse martingales with respect to some non-increasing filtration as proven in [13]. For more literature on reverse martingales, we refer to classical textbooks on probability theory and stochastic processes such as [14,15]. We exploit the reverse martingale property of U-statistics to derive attractive asymptotic properties of our proposed estimation procedure.

In the next section, we formally state the fixed-width confidence interval estimation problem and why a fixed-sample size procedure cannot be used. In Section 3, a purely sequential procedure is proposed to construct a $100(1-\alpha)\%$ fixed-width confidence interval for unknown population Gini index and implementation and characteristics of the sequential procedure is discussed as well. Section 4 presents simulation study and validate all theoretical results related to our procedure. We conclude this article with some remarks in Section 5.

2. Problem Statement and Optimal Sample Size

Consider n randomly selected individuals from some population of interest with incomes denoted by $X_1, X_2, \ldots, X_n$. Suppose these are nonnegative independent and identically distributed random variables assumed to be drawn from an unknown distribution function F where the support of the distribution is $(0,\infty)$. A strongly consistent estimator of population Gini index G_F is $\widehat{G}_n$ given in (2). For pre-specified α in $(0,1)$, the goal of this paper is to develop the theory for constructing a $100(1-\alpha)\%$ fixed-width confidence interval for G_F. Formally, we would like to construct a confidence interval $J_n = (\widehat{G}_n - d, \widehat{G}_n + d)$ such that

$$P\left(\widehat{G}_n - d < G_F < \widehat{G}_n + d\right) \ge 1-\alpha, \tag{4}$$

for some prefixed margin of error $d > 0$. Using [7], we have

$$\sqrt{n}\left(\widehat{G}_n - G_F\right) \to N\left(0, \xi^2\right) \quad \text{as } n \to \infty, \tag{5}$$

where ξ^2 is the asymptotic variance given by

$$\xi^2 = \frac{\Delta^2}{4\mu^4}\sigma^2 - \frac{\Delta\tau}{\mu^3} + \frac{\Delta^2}{\mu^2} + \frac{\sigma_1^2}{\mu^2}. \tag{6}$$

Here,

$$\tau = E(X_1 |X_1 - X_2|) \text{ and } \sigma_1^2 = V\left[E\left(|X_1 - X_2| \,|X_1\right)\right].$$

Schroder and Yitzhaki [16] proposed a way to come up with the reasonable sample size related to the convergence of the distribution of $\widehat{G}_n$ to normality. In this paper, we verify via simulation study that for moderate sample sizes (see Section 4) the distribution of sample Gini index is approximately normal. Based on the asymptotic normality of $\widehat{G}_n$, we observe that the coverage probability is

$$P\left(\widehat{G}_n - d < G_F < \widehat{G}_n + d\right) \approx 2\Phi\left(\frac{d\sqrt{n}}{\xi}\right) - 1,$$

where Φ is the distribution function of standard normal random variable. In order to have $100(1-\alpha)\%$ confidence interval, sample size n must satisfy

$$2\Phi\left(\frac{d\sqrt{n}}{\xi}\right) - 1 \geq 1 - \alpha. \tag{7}$$

Solving (7) for n, we obtain $n \geq d^{-2} z_{\alpha/2}^2 \xi^2$, where $z_{\alpha/2}$ is the upper $\left(\frac{\alpha}{2}\right)^{th}$ quantile of the standard normal distribution. Provided ξ is known, the optimal (minimal) sample size required to construct a fixed-width confidence interval for Gini index with approximately $(1-\alpha)$ coverage probability is

$$C = \lceil d^{-2} z_{\alpha/2}^2 \xi^2 \rceil, \tag{8}$$

where, $\lceil w \rceil$ the lowest integer which is greater than or equal to w.

The optimal fixed sample size C is unknown since the true value of ξ is unknown in practice. If C were known, one would just draw C observations independently from the population of interest and compute $(\widehat{G}_C - d, \widehat{G}_C + d)$ which would satisfy (4) approximately. Since C is unknown, one must draw samples at least in two stages in order to achieve the desired coverage probability at least approximately. In the first stage, one must estimate C by estimating ξ, and then in the subsequent stages one should collect samples until the current sample size is more or equal to the estimated optimal sample size. In this article, we propose a purely sequential sampling procedure to estimate the optimal sample size C and ensure that the fixed-width confidence interval based on the final sample size attains the desired $(1-\alpha)$ coverage probability atleast asymptotically.

3. The Sequential Estimation Procedure

In sequential estimation procedures, the parameter estimates are updated as the data is observed. In the first step, a small sample, called the pilot sample, is observed to gather preliminary information about the parameter of interest. Then, in each successive step, one or more additional observations are collected and the estimates of the parameters are updated. After each and every step a decision is taken whether to continue or to terminate the sampling process. This decision is based on a pre-defined stopping rule.

From (8) we note that the optimal sample size needed to find a fixed-width confidence interval depends on unknown parameter ξ^2. So, let us first find a good estimator of the unknown parameter ξ^2. Following [7,17], we consider the following strongly consistent estimator of ξ^2 based on U-statistics. Let us define a U-statistic, for each $j = 1, 2, \ldots, n$,

$$\widehat{\Delta}_n^{(j)} = \binom{n-1}{2}^{-1} \sum_{T_j} |X_{i_1} - X_{i_2}|, \tag{9}$$

where $T_j = \{(i_1, i_2) : 1 \leq i_1 < i_2 \leq n \text{ and } i_1, i_2 \neq j\}$. Define $W_{jn} = n\widehat{\Delta}_n - (n-2)\widehat{\Delta}_n^{(j)}$ for $j = 1, \ldots, n$, and $\overline{W}_n = n^{-1} \sum_{j=1}^n W_{jn}$. According to [17], a strongly consistent estimator of $4\sigma_1^2$ is

$$s_{wn}^2 = (n-1)^{-1} \sum_{i=1}^n (W_{jn} - \overline{W}_n)^2.$$

Using [7],

$$\widehat{\tau}_n = \frac{2}{n(n-1)} \sum_{(n,2)} \frac{1}{2}(X_{i_1} + X_{i_2}) \, |X_{i_1} - X_{i_2}| \tag{10}$$

is an estimator of τ. Let S_n^2 be the sample variance. Thus, the estimator of ξ^2 is

$$V_n^2 = max\left(0, \frac{\widehat{\Delta}_n^2 S_n^2}{4\overline{X}_n^4} - \frac{\widehat{\Delta}_n}{\overline{X}_n^3}\widehat{\tau}_n + \frac{\widehat{\Delta}_n^2}{\overline{X}_n^2} + \frac{s_{wn}^2}{4\overline{X}_n^2}\right) \tag{11}$$

similar to [7]. Note that $\widehat{\tau}_n, S_n^2, \widehat{\Delta}_n$ are U-statistics of degree 2 (e.g., [7,18]) and the sample mean $\overline{X}_n$ is a U-statistic of degree 1. Using continuous mapping theorem [17], and Theorem 3.2.1 in [11], we observe that a strongly consistent estimator of ξ^2 is V_n^2 which will be used in our proposed sequential procedure to estimate C.

Several plug-in estimators of the asymptotic variance parameter of Gini index are proposed in the economics and statistics literature. To find the details about several plug-in estimators of the asymptotic variance of Gini index under different sampling schemes, we refer to [4,19]. The proposed plug-in estimator in [4] is simpler than V_n^2. However, it is not known whether the estimator enjoys the almost sure convergence property which is very important for us as we need this property to prove the asymptotic optimality properties of the proposed sequential procedure. Moreover, with the high-end computing facilities available these days, V_n^2 can be computed in seconds.

Using V_n^2 as the estimator of ξ^2, we define the stopping rule N_d, for every $d > 0$, as

$$N_d = \text{ the smallest integer } n(\geq m) \text{ such that } n \geq \left(\frac{z_{\alpha/2}}{d}\right)^2 \left(V_n^2 + n^{-1}\right). \tag{12}$$

Here, m is called the initial or pilot sample size, and the term n^{-1} is known as a correction term. Note that V_n can be very close to zero with positive probability. Without the correction term, the inequality (12) may be satisfied for very small n terminating the sampling process too early. Thus the correction term n^{-1} ensures that the sampling process for estimating the optimal sample size does not stop too early. For details about the correction term, we refer to [11].

From (12), we note that, $N_d \geq \left(\frac{z_{\alpha/2}}{d}\right)^2 N_d^{-1}$, i.e., the final sample size must be at least $z_{\alpha/2}/d$. Therefore, we consider the pilot sample size to be $m = \max\{4, \lceil z_{\alpha/2}/d \rceil\}$. This technique of estimating pilot sample size can also be found in [10].

Recall that the optimal sample size required to achieve $100(1-\alpha)\%$ confidence interval for Gini index is C which is unknown in practice. The stopping variable N_d defined in (12) serves as an estimator of C. Below, we develop a purely sequential procedure to estimate the optimal sample size C.

Implementation and Characteristics

We propose the following purely sequential procedure to estimate the optimal sample size C.

Stage 1: Compute the pilot sample size $m = \max\{4, \lceil z_{\alpha/2}/d \rceil\}$ and draw a random sample of size m from the population of interest. Based on this pilot sample of size m, obtain an estimate of ξ^2 by finding V_m^2 as given in (11) and check whether $m \geq (z_{\alpha/2}/d)^2 \left(V_m^2 + m^{-1}\right)$. If $m < (z_{\alpha/2}/d)^2 \left(V_m^2 + m^{-1}\right)$ then go to the next step. Otherwise, set the final sample size $N_d = m$.

Stage 2: Draw an additional observation independent of the pilot sample and update the estimate of ξ^2 by computing V_{m+1}^2. Check if $m+1 \geq (z_{\alpha/2}/d)^2 \left(V_{m+1}^2 + (m+1)^{-1}\right)$. If $m+1 < (z_{\alpha/2}/d)^2 \left(V_{m+1}^2 + (m+1)^{-1}\right)$ then go to the next step. Otherwise, stop sampling and report the final sample size as $N_d = m+1$.

The process of collecting observations one by one is continued until there are N_d observations such that $N_d \geq (z_{\alpha/2}/d)^2 \left(V_{N_d}^2 + {N_d}^{-1}\right)$. At this stage, we stop sampling and report the final sample size as N_d.

Based on the above algorithm, the sampling process will stop at some stage. This is proved in Lemma A1 which states that if observations are collected using (12), under appropriate conditions, $P(N_d < \infty) = 1$. This is a very important property of any sequential procedure since it mathematically ensures that the sampling will be terminated eventually.

Next, we establish some desirable asymptotic properties of our proposed sequential procedure. First, we prove that the final sample size N_d required by our sampling strategy is close to the optimal sample size C at least asymptotically. We also prove the asymptotic efficiency property of sequential procedure which ensures that, on average, we collect only the minimum number of samples to achieve certain accuracy of estimation. Second, we show that the fixed-width confidence interval $\left(\widehat{G}_{N_d} - d, \widehat{G}_{N_d} + d\right)$ contains the true value of Gini index G_F nearly with probability $1-\alpha$. We formally state these results in Theorems 1 and 2.

Theorem 1. *If the parent distribution F of X is such that $E[X^4]$ exists, then the stopping rule in (12) yields the following asymptotic optimality properties:*

(i) $N_d/C \xrightarrow{a.s.} 1 \quad as \;\; d \downarrow 0.$

(ii) $P\left(\widehat{G}_{N_d} - d < G_F < \widehat{G}_{N_d} + d\right) \to 1-\alpha \quad as \;\; d \downarrow 0.$

Theorem 2. *If the parent distribution F of X is such that the support of the distribution being (t, ∞) with $t > 0$ and $E[X^4]$ exists, then the stopping rule in (12) yields*

$$E\left(N_d/C\right) \to 1 \quad as \;\; d \downarrow 0. \tag{13}$$

Theorems 1 and 2 are proved in the appendix. Part (i) of Theorem 1 implies that the ratio of final sample size of our procedure and the optimal sample size, C asymptotically converges to 1. Part (ii) of Theorem 1 implies that the coverage probability produced by the fixed-width confidence interval $\left(\widehat{G}_{N_d} - d, \widehat{G}_{N_d} + d\right)$ attains the desired level $1-\alpha$ asymptotically. Theorem 2 implies that the ratio of the average final sample size of our procedure asymptotically converges to the optimal sample size, C.

4. Simulation Study

In this section, we validate the asymptotic properties of our method stated in Theorems 1 and 2 through Monte Carlo study. To implement the sequential procedure, we fix $d(= 0.01, 0.02)$ and $\alpha(= 0.1, 0.05)$. Using the pilot sample size formula $m = \max\{4, \lceil z_{\frac{\alpha}{2}}/d \rceil\}$, the pilot sample size considered here is 165. Then, we implement the sequential procedure described in Section 3.1 and estimate the average sample size $(\overline{N})$, the maximum sample size $(\max(N))$, the standard error $(s(\overline{N}))$ of $\overline{N}$, the coverage probability (p), and its standard error (s_p) based on 2000 replications by drawing

random samples from gamma (shape = 2.649, rate = 0.84) distribution truncated at $t = 0.001 (G_F = 0.3308, \xi^2 = 0.0468)$, log-normal distribution (mean = 2.185, sd = 0.562) truncated at $t = 0.001 (G_F = 0.3089, \xi^2 = 0.0532)$, and Pareto (20,000, 5).

Table 1 summarizes the numerical results obtained from the simulation study. The parameters of log-normal and gamma distributions are the same as used by [20]. From the fourth column of Table 1, we find that the ratio of the average final sample size and C is close to 1. Moreover, column 6 of Table 1 illustrates that the attained coverage probability is very close to the desired level of 90%. Thus, we find that the simulation results validate all theoretical results mentioned in the previous section, and the performance of the procedure is satisfactory for the above mentioned distributions.

Table 1. Performance of the proposed sequential procedure when the data is from gamma, log-normal, and Pareto distribution.

d α	**Distribution**	$\overline{N}$ $\mathbf{s}(\overline{N})$	C	$\overline{N}/C$	p s_p
$d = 0.01$ $\alpha = 0.1$	Gamma	1283.7450 1.7561	1267	1.0132	0.9090 0.0064
$d = 0.02$ $\alpha = 0.05$	Gamma	469.1650 1.0485	450	1.0426	0.9535 0.0047
$d = 0.01$ $\alpha = 0.1$	Lognormal	1435.3640 4.091604	1440	0.9968	0.8965 0.0068
$d = 0.02$ $\alpha = 0.05$	Lognormal	509.0020 2.0538	511	0.9961	0.9430 0.0052
$d = 0.01$ $\alpha = 0.1$	Pareto	654.5364 4.2151	686	0.9541	0.9018 0.0063
$d = 0.02$ $\alpha = 0.05$	Pareto	244.3330 2.0099	244	1.0014	0.9470 0.0050

To verify whether the distribution of sample Gini index converges to normality, we do the following simulation study. We draw samples of sizes $n = 470$, 509, and 245 from gamma (shape = 2.649, rate = 0.84), log-normal (mean = 2.185, sd = 0.562), and Pareto (20,000, 5) distributions respectively and compute the sample Gini indexes. Please note that the choice of the parameters of lognormal and gamma distributions are same as in [20]. The sample size n is chosen according to the smallest $\overline{N}$ in Table 1 for different scenarios. We compute 200 replications for the sample Gini index (using set.seed(123) in "*R*" language) and observe that Shapiro-Wilk's test for normality returns p-values 0.1938, 0.5066, and 0.3984 for gamma (shape = 2.649, rate = 0.84), log-normal (mean = 2.185, sd = 0.562), and Pareto (20,000, 5) distributions respectively. Thus, we observe that for the above scenarios sample Gini index converges to normality for moderate sample sizes.

5. Discussion and Concluding Remarks

Gini index is a widely used measure of economic inequality index. In order to evaluate the economic policies adopted by a government, it is important to estimate Gini index at any specific time period. If the income data for all households in the region of interest is not available, one needs to estimate Gini index by drawing a sample of households from that region. Typically, large income surveys are associated with different sampling schemes. To review several sampling techniques, we refer to [19,21–24]. The sampling technique chosen to collect data usually depends on the socio-economic diversity and size of the country or region. For regions or smaller countries with lesser socio-economic diversity, the simple random sampling technique can be used to collect income or expenditure data to estimate Gini index. Several research articles (e.g., [4,5,7,25,26]) are devoted to drawing statistical inference on inequality indexes which are computed from household income or expenditure by means of simple random sampling from the population of interest. In this paper,

we also use simple random sampling technique to collect income or expenditure data in order to estimate Gini index accurately. Even though the sequential methodology is introduced under i.i.d. framework, which may be considered as a practical limitation of our work, sequential methodologies may be adopted to different sampling schemes (e.g., see [27]). In the following Section 5.1, we discuss the possibility of extending our work to stratified sampling.

Without assuming any specific distribution of the data, we show that the ratio of the final sample size and the optimal sample size approaches 1. We also show that the confidence interval constructed using our proposed sequential method attains the required coverage probability. Thus, based on these results, we conclude that the proposed sequential estimation strategy can efficiently construct a $100(1-\alpha)\%$ fixed-width confidence interval for Gini index. In this article, we consider that after pilot sample, one additional observation is collected in each step. If instead, a group of $r(\geq 1)$ observations are collected in each step after the pilot sample stage, the same properties will hold. The proofs will be similar to the ones in Appendix.

The theory of the sequential procedure revolves around the idea of "learn-as-you-go". In our proposed method of estimation, the final sample size is not fixed in advance, and the observations are collected until the estimated optimal sample size is obtained achieving the required $100(1-\alpha)\%$ fixed-width confidence interval for Gini index. We hope that number of sequential procedures to estimate income inequalities will be developed following this article, whereupon the idea of sequential procedure can be applied to other sampling schemes used in economic surveys, taking into account of the cost considerations as well.

Possible Extension to Stratified Sampling

The sequential procedure that is proposed under i.i.d. framework may be extended to non i.i.d framework where stratified sampling is used. Suppose we divide the population into S strata. In the population, stratum s contains a mass of H_s households. The total number of households in the population is $H = \sum_{s=1}^{S} H_s$. The density of the household income or expenditure, X in the s^{th} stratum is denoted by $dF(x|s)$.

Now, a sample of n_s households (indexed by h_s) is drawn by simple random sample with replacement from every strata so that the total number of households in the sample is $n = \sum_{s=1}^{S} n_s$ and $n_s = na_s$ with $\sum_{s=1}^{S} a_s = 1$, where $a_s = H_s/H$. Let x_{sh_s} be the total income of the h_s^{th} household belonging to the s^{th} stratum. If $w_{sh_s} = \frac{H_s}{n_s}$ is the weight of h_s^{th} household in the s^{th} stratum, then following [21,22], Gini index can be estimated by the estimator

$$\widehat{G} = 1 - \frac{2}{\widehat{\mu}} \sum_{s=1}^{S} \sum_{h_s=1}^{n_s} w_{sh_s} x_{sh_s} \left(1 - \widehat{F}(x_{sh_s})\right), \tag{14}$$

where

$$\widehat{\mu} = \sum_{s=1}^{S} \sum_{h_s=1}^{n_s} w_{sh_s} x_{sh_s} \text{ and } \widehat{F}(x_{sh_s}) = \sum_{i=1}^{S} \sum_{j=1}^{n_s} w_{ij} I[x_{ij} \leq x_{hc_s}] / \sum_{i=1}^{S} \sum_{j=1}^{n_s} w_{ij}. \tag{15}$$

Now, following [21,22] we have

$$\sqrt{n}\left(\widehat{G} - G_F\right) \xrightarrow{d} N\left(0, V^*\right), \tag{16}$$

where V^* is the asymptotic variance given in [22], modified to take into account of the stratified sampling only. Now,

$$P\left[\left|\widehat{G}_n - G_F\right| \leq d\right] \to 2\Phi\left(\sqrt{\frac{n}{V^*}}d\right) - 1 \tag{17}$$

The confidence coefficient will be approximately $1-\alpha$ provided $\sqrt{n}d/\sqrt{V^*} \geq z_{\alpha/2}$. In order to have a fixed-width confidence interval, we need sample size n satisfying

$$n \geq d^{-2} z_{\alpha/2}^2 V^* \equiv C, \text{ say.} \tag{18}$$

Since C is unknown it must be estimated in the first stage and continue sampling until the sample size n is bigger than corresponding estimated value of C. Note that C is the optimum (i.e., minimum) household size to be sampled to achieve $(1-\alpha)$ confidence level provided V^* were known. The optimal number of households to be sampled in the s^{th} stratum ($s = 1, \ldots, S$) will be $C_s = Ca_s$ which is also unknown since C is unknown. Bhattacharya [22] proposed an estimator of the asymptotic variance V^* of the Gini index under complex household survey which can be used in Equation (18). Then the stopping rule developed in Equation (12) may be modified taking into account of the stratification and finite sampling scenario to find out an estimate of the optimum number of households in order to find a fixed-width confidence interval for Gini index under stratified sampling. However, we do not intend to explore this possibility in this article, and we believe that this could be a good topic of future research.

Acknowledgments: We thank the three anonymous referees and the editor whose insightful comments helped us improve the paper. We remain deeply indebted to Professor Gautam Tripathi and Professor Nitis Mukhopadhyay for their comments and suggestions.

Author Contributions: The authors contributed equally.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Lemma A1. *Under the assumption that $\xi < \infty$, for any $d > 0$, the stopping time N_d is finite, that is, $P(N_d < \infty) = 1$.*

Proof. The Lemma A1 is proved by using (12) and the fact that V_n^2 is strongly consistent estimator of ξ^2 and $N_d \to \infty$ as $d \downarrow 0$ almost surely. □

Lemma A2. *The value of sample Gini index lies between 0 and 1.*

Proof. Let $Y_1, \ldots, Y_n$ be the ordered incomes of n persons where Y_1 represents the income of the poorest person and Y_n represents the income of the richest person. Using [28], Gini index can be rewritten as

$$0 \leq \widehat{G}_n = \frac{2\sum_{i=1}^n iY_i}{n\sum_{i=1}^n Y_i} - \frac{n+1}{n} \leq \frac{2n\sum_{i=1}^n Y_i}{n\sum_{i=1}^n Y_i} - \frac{n+1}{n} = \frac{n-1}{n} \leq 1.$$

This proves the lemma. □

Appendix A.1. Proof of Theorem 1

(i) The definition of stopping rule N_d in (12) yields

$$\left(\frac{z_{\alpha/2}}{d}\right)^2 V_{N_d}^2 \leq N_d \leq mI(N_d = m) + \left(\frac{z_{\alpha/2}}{d}\right)^2 \left(V_{N_d-1}^2 + (N_d-1)^{-1}\right). \tag{A1}$$

Since $N_d \to \infty$ a.s. as $d \downarrow 0$ and $V_n \to \xi$ a.s. as $n \to \infty$, by Theorem 2.1 of [29], $V_{N_d}^2 \to \xi^2$ a.s. Hence, dividing all sides of (A1) by C and letting $d \downarrow 0$, we prove $N_d/C \to 1$ a.s. as $d \downarrow 0$.

(ii) In order to show that our procedure satisfies the asymptotic consistency property, we will derive an Anscombe-type random central limit theorem for Gini index. This requires the existence of usual central limit theorem of Gini index and uniform continuity in probability (u.c.i.p.) condition. For details about the u.c.i.p. condition, we refer to [17,30–32] *etc*.

First of all, let us define $n_1 = (1-\rho)C$ and $n_2 = (1+\rho)C$ for $0 < \rho < 1$. Now, we know from [7] that $\mathbf{Y}_n = \left(\sqrt{n}(\widehat{\Delta}_n - \Delta), \sqrt{n}(\bar{X}_n - \mu)\right)' \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 4\sigma_1^2 & 2(\tau - \mu\Delta) \\ 2(\tau - \mu\Delta) & \sigma^2 \end{pmatrix}.$$

First, let us prove that $\mathbf{Y}_{N_d} \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \Sigma)$. Define $\mathbf{D}' = (a_0\ a_1)$. Note that $\mathbf{D}'\mathbf{Y}_{N_d} = \mathbf{D}'\mathbf{Y}_C + (\mathbf{D}'\mathbf{Y}_{N_d} - \mathbf{D}'\mathbf{Y}_C)$. Thus, it is enough to show that $(\mathbf{D}'\mathbf{Y}_{N_d} - \mathbf{D}'\mathbf{Y}_C) \xrightarrow{P} 0$ as $d \downarrow 0$. We can write

$$\begin{aligned}(\mathbf{D}'\mathbf{Y}_{N_d} - \mathbf{D}'\mathbf{Y}_C) =& a_0\sqrt{N_d}(\widehat{\Delta}_N - \widehat{\Delta}_C) + a_1\sqrt{N_d}(\bar{X}_{N_d} - \bar{X}_C) \\ &+ (\sqrt{N_d/C} - 1)\mathbf{D}'\mathbf{Y}_C. \end{aligned} \tag{A2}$$

Fix some $\epsilon > 0$ and note that

$$\begin{aligned} & P\left\{|a_0\sqrt{N_d}(\widehat{\Delta}_{N_d} - \widehat{\Delta}_C) + a_1\sqrt{N_d}(\bar{X}_{N_d} - \bar{X}_C)| > \epsilon\right\} \\ & \leq P\left\{|a_0\sqrt{N_d}(\widehat{\Delta}_{N_d} - \widehat{\Delta}_C) + a_1\sqrt{N_d}(\bar{X}_{N_d} - \bar{X}_C)| > \epsilon, |N_d - C| < \rho C\right\} \\ & + P[|N_d - C| > \rho C] \\ & \leq P\left\{\max_{n_1<n<n_2} |\sqrt{n}|\widehat{\Delta}_n - \widehat{\Delta}_C| > \frac{\epsilon}{2|a_0|}\right\} + P\left\{\max_{n_1<n<n_2} |\sqrt{n}|\bar{X}_n - \bar{X}_C| > \frac{\epsilon}{2|a_1|}\right\} \\ & + P[|N_d - C| > \rho C] \end{aligned}$$

Here, $\widehat{\Delta}_n$ and $\bar{X}_n$ are both U-statistics which satisfy Anscombe's u.c.i.p. condition (for e.g., see [17]). Using u.c.i.p. condition and the fact that $N_d/C \xrightarrow{a.s.} 1$, we conclude that for given $\epsilon > 0$, there exist $\eta > 0$ and $d_0 > 0$ such that

$$P\{|a_0\sqrt{N_d}(\widehat{\Delta}_{N_d} - \widehat{\Delta}_C) + a_1\sqrt{N_d}(\bar{X}_{N_d} - \bar{X}_C)| > \epsilon\} < \eta \quad \text{for all} \ \ d \leq d_0.$$

This implies $a_0\sqrt{N_d}(\widehat{\Delta}_{N_d} - \widehat{\Delta}_C) + a_1\sqrt{N_d}(\bar{X}_{N_d} - \bar{X}_C) \xrightarrow{P} 0$ as $d \downarrow 0$. Also, note that $\left(\sqrt{N_d/C} - 1\right)\mathbf{D}'\mathbf{Y}_C \xrightarrow{P} 0$ as $d \downarrow 0$ since $N_d/C \to 1$ almost surely and $\mathbf{D}'\mathbf{Y}_C \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \Sigma)$. Thus, from (A2), we conclude $(\mathbf{D}'\mathbf{Y}_{N_d} - \mathbf{D}'\mathbf{Y}_C) \xrightarrow{P} 0$, that is, $\mathbf{Y}_{N_d} \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \Sigma)$. Now, define $G(u, v) = \frac{u}{2v}$, if $v \neq 0$. Using Taylor's expansion, we can write

$$\sqrt{N_d}(G(\widehat{\Delta}_{N_d}, \bar{X}_{N_d}) - G(\Delta, \mu)) = \sqrt{N_d}\left(\frac{\widehat{\Delta}_{N_d} - \Delta}{2\mu} - \frac{\Delta}{2\mu^2}(\overline{X}_{N_d} - \mu) + R_{N_d}\right), \tag{A3}$$

where $R_{N_d} = -2(\widehat{\Delta}_{N_d} - \Delta)(\overline{X}_{N_d} - \mu)/b^2 + 4a(\overline{X}_{N_d} - \mu)^2/b^3$, $a = \Delta + p(\widehat{\Delta}_{N_d} - \Delta)$, $b = 2\mu + p(2\overline{X}_{N_d} - 2\mu)$, and $p \in (0, 1)$. Rewriting (A3) in the vector-matrix form, we get

$$\sqrt{N_d}(G(\widehat{\Delta}_{N_d}, \bar{X}_{N_d}) - G(\Delta, \mu)) = \mathbf{D}'\mathbf{Y}_{N_d} + \sqrt{N_d}R_{N_d}, \tag{A4}$$

where $\mathbf{D}' = \left(\frac{1}{2\mu}, \frac{-\Delta}{2\mu^2}\right)$. Note that $\sqrt{N_d}(\overline{X}_{N_d} - \mu)$ converges in distribution to a normal distribution by Anscombe's CLT and both $\left(\widehat{\Delta}_{N_d} - \Delta\right)$ and $(\overline{X}_{N_d} - \mu)$ converges to 0 almost surely. This yields $\sqrt{N_d}R_{N_d} \xrightarrow{P} 0$ as $d \downarrow 0$. Hence, $\sqrt{N_d}(\widehat{G}_{N_d} - G_F) \xrightarrow{\mathcal{L}} N(\mathbf{0}, D'\Sigma D)$ as $d \downarrow 0$. This completes the proof of Theorem 1. □

Appendix A.2. Proof of Theorem 2

In this subsection, we prove a lemma that is essential to establish Theorem 2. Note from (12) that $N_d \geq \frac{z_{\alpha/2}^2}{d^2} N_d^{-1}$, i.e., $N_d \geq \frac{z_{\alpha/2}}{d} (= m)$ with probability 1. Suppose $\boldsymbol{X}_{(n)} = (X_{(1)}, \ldots, X_{(n)})$ denotes the n dimensional vector of order statistics from the sample $X_1, \ldots, X_n$, and $\mathcal{F}_n$ is the σ-algebra generated by $(\boldsymbol{X}_{(n)}, X_{n+1}, X_{n+2}, \ldots)$. By [13], $\{\overline{X}_n, \mathcal{F}_n\}$, $\{S_n^2, \mathcal{F}_n\}$, $\{\widehat{\tau}_n, \mathcal{F}_n\}$, $\left\{\widehat{\Delta}_n, \mathcal{F}_n\right\}$, and their convex functions are all reverse submartingales. Using reverse submartingale properties, let us prove the following lemma along the lines of [33].

Lemma A3. *If $E(X_1^{2p})$ is finite for some $p > 1$, then $E\left[\sup_{n \geq m} V_n^2\right] < \infty$ for $m \geq 4$.*

Proof. To prove Lemma A3, it is enough to show that: $E\left[\sup_{n \geq m} s_{wn}^2 \overline{X}_n^{-2}\right]$, $E\left[\sup_{n \geq m} \left|\frac{\widehat{\Delta}_n}{\overline{X}_n^3} \widehat{\tau}_n\right|\right]$, $E\left[\sup_{n \geq m} \frac{\widehat{\Delta}_n^2}{\overline{X}_n^2}\right]$, and $E\left[\sup_{n \geq m} \frac{\widehat{\Delta}_n^2}{\overline{X}_n^4} S_n^2\right]$ are finite.

We note that, $0 \leq \frac{\widehat{\Delta}_n}{2\overline{X}} \leq 1$. So, it is enough to show that $E\left[\sup_{n \geq m} s_{wn}^2 \overline{X}_n^{-2}\right]$, $E\left[\sup_{n \geq m} \frac{\widehat{\tau}_n}{\overline{X}_n^2}\right]$ and $E\left[\sup_{n \geq m} \frac{S_n^2}{\overline{X}_n^2}\right]$ are finite. Following [34] (p. 338), we have $E\left[\sup_{n \geq m} s_{wn}^2\right] < \infty$ if $E[X_1^\alpha] < \infty$ for $\alpha > 2$ and $m \geq 4$. Therefore,

$$E\left(\sup_{n \geq m} s_{wn}^2 \overline{X}_n^{-2}\right) \leq t^{-2} E\left(\sup_{n \geq m} s_{wn}^2\right) < \infty. \tag{A5}$$

We note that $\widehat{\tau}_n$ and S_n^2 are U-statistics. Using Lemma 9.2.4 of [35], for $p > 1$,

$$E\left(\sup_{n \geq m} |\widehat{\tau}_n|^p\right) \leq \left(\frac{p}{1-p}\right)^p E\left(|\widehat{\tau}_m|^p\right) \text{ and } E\left(\sup_{n \geq m} \left|S_n^2\right|^p\right) \leq \left(\frac{p}{1-p}\right)^p E\left(\left|S_m^2\right|^p\right).$$

Since $E(X_1^{2p})$ if finite for some $p > 1$, $E\left(|\widehat{\tau}_m|^p\right)$ and $E\left(\left|S_m^2\right|^p\right)$ are finite which yield

$$E\left(\sup_{n \geq m} \left|\frac{\widehat{\tau}_n}{\overline{X}_n^2}\right|\right) \leq \left\{t^{-2} E\left(\sup_{n \geq m} |\widehat{\tau}_n|\right)\right\} < \infty,$$

and

$$E\left(\sup_{n \geq m} \left|\frac{S_n^2}{\overline{X}_n^2}\right|\right) \leq \left\{t^{-2} E\left(\sup_{n \geq m} \left|S_n^2\right|\right)\right\} < \infty.$$

This completes the proof of Lemma A3. □

Below, we prove Theorem 2 by using Lemma A3.
Since $N_d \geq m$ a.s., dividing (A1) by C yields

$$N_d/C - mI(N_d = m)/C \leq \frac{1}{\xi^2}\left(\sup_{d>0} V_{N_d-1}^2 + (m-1)^{-1}\right) \text{ almost surely.} \tag{A6}$$

Since $E\left(\sup_{d>0} V_{N_d-1}^2\right) < \infty$ by Lemma A3 and $N_d/C \to 1$ a.s. as $d \downarrow 0$, by the dominated convergence theorem, we conclude that $\lim_{d \downarrow 0} E(N_d/C) = 1$.

This completes the proof of Theorem 2. □

References

1. Arnold, B.C. Inequality measures for multivariate distributions. *Metron* **2005**, *63*, 317–327.
2. Andres, R.; Samuel, C. *Inference on Income Inequality and Tax Progressivity Indices: U-Statistics and Bootstrap Methods*; ECINEQ working paper 2005-9; ECINEQ: Palma, Spain, 2005.
3. Bishop, J.A.; Formby, J.P.; Zheng, B. Statistical inference and the sen index of poverty. *Int. Econ. Rev.* **1997**, *38*, 381–387.
4. Davidson, R. Reliable inference for the Gini index. *J. Econom.* **2009**, *150*, 30–40.
5. Gastwirth, J.L. The estimation of the Lorenz curve and Gini index. *Rev. Econ. Stat.* **1972**, *54*, 306–316.
6. Palmitesta, P.; Corrado, P.; Cosimo, S. Confidence interval estimation for inequality indices of the Gini family. *Comput. Econ.* **2000**, *16*, 137–147.
7. Xu, K. U-statistics and their asymptotic results for some inequality and poverty measures. *Econom. Rev.* **2007**, *26*, 567–577.
8. Maasoumi, E. Empirical analysis of inequality and welfare. In *Handbook of Applied Microeconomics*; Schmidt, S., Pesaran, H., Eds.; Blackwell Publishers Inc.: Malden, MA, USA, 1997.
9. Dantzig, G.B. On the non-existence of tests of "student's" hypothesis having power functions independent of σ. *Ann. Math. Stat.* **1940**, *11*, 186–192.
10. Mukhopadhyay, N.; de Silva, B.M. *Sequential Methods and Their Applications*; CRC Press: Boca Raton, FL, USA, 2009.
11. Sen, P.K. *Sequential Nonparametrics: Invariance Principles and Statistical Inference*; Wiley: New York, NY, USA, 1981.
12. Hoeffding, W. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **1948**, *19*, 293–325.
13. Lee, A.J. *U-Statistics: Theory and Practice*; CRC Press: New York, NY, USA, 1990.
14. Loève, M. *Probability Theory*; Van Nostrand: Princeton, NJ, USA, 1963.
15. Doob, J.L. *Stochastic Processes*; Wiley: New York, NY, USA, 1953.
16. Schröder, C.; Yitzhaki, S. *Reasonable Sample Sizes for Convergence to Normality*; No. 714, SOEP Papers on Multidisciplinary Panel Data Research. 2014. Available online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2539096 (accessed on 5 March 2016).
17. Sproule, R. A Sequential Fixed-Width Confidence Interval for the Mean of a U-Statistic. Ph.D. Thesis, University of North Carolina, Chapel Hill, NC, USA, 1969.
18. Chattopadhyay, B.; Mukhopadhyay, N. Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Seq. Anal.* **2013**, *32*, 134–157.
19. Langel, M.; Tillè, Y. Variance estimation of the Gini index: Revisiting a result several times published. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2013**, *176*, 521–540.
20. Ransom, M.R.; Cramer, J.S. Income distribution functions with disturbances. *Eur. Econ. Rev.* **1983**, *22*, 363–372.
21. Bhattacharya, D. Asymptotic inference from multi-stage samples. *J. Econom.* **2005**, *126*, 145–171.
22. Bhattacharya, D. Inference on inequality from household survey data. *J. Econom.* **2007**, *137*, 674–707.
23. Binder, D.A.; Kovacevic, M.S. Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Surv. Methodol.* **1995**, *21*, 137–146.
24. Cochran, W.G. *Sampling Techniques*; Wiley & Sons: New York, NY, USA, 1977.
25. Beach, C.M.; Davidson, R. Distribution-free statistical inference with Lorenz curves and income shares. *Rev. Econ. Stud.* **1983**, *50*, 723–735.
26. Davidson, R.; Duclos, J. Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* **2000**, *68*, 1435–1464.
27. Zacks, S. *Stage-Wise Adaptive Designs*; Wiley: New York, NY, USA, 2009.
28. Damgaard, C.; Weiner, J. Describing inequality in plant size or fecundity. *Ecology* **2000**, *81*, 1139–1142.
29. Gut, A. *Stopped random walks: Limit theorems and applications*; Springer: New York, NY, USA, 2009.
30. Anscombe, F.J. Sequential estimation. *J. R. Stat. Soc. Ser. B* **1953**, *15*, 1–29.
31. Isogai, E. Asymptotic consistency of fixed-width sequential confidence intervals for a multiple regression function. *Ann. Inst. Stat. Math.* **1986**, *38*, 69–83.

32. Mukhopadhyay, N.; Chattopadhyay, B. A tribute to Frank Anscombe and random central limit theorem from 1952. *Seq. Anal.* **2012**, *31*, 265–277.
33. De, S.K.; Chattopadhyay, B. Minimum Risk Point Estimation of Gini Index. Available online: http://arxiv.org/abs/1503.08148 (accessed on 27 March 2015).
34. Sen, P.K.; Ghosh, M. Sequential point estimation of estimable parameters based on U-statistics. *Sankhyā Indian J. Stat. Ser. A* **1981**, *43*, 331–344.
35. Ghosh, M.; Mukhopadhyay, N.; Sen, P.K. *Sequential Estimation*; Wiley: New York, NY, USA, 1997.